Technology Reports
Image Recognition Deep Learning AI
Special Articles on AI Technologies Contributing to Industry and Society

A Retail Shelving Analysis Solution Using Image Recognition —Recognizes Shelving Allocation and Quantifies Inventory by Analyzing Photos of Shelved Merchandise—

Service Innovation Department Hayato Akatsuka† Issei Nakamura
Kansai Branch, Corporate Sales Dept. Sungmyeong Koh

In consumer retail, it is important to understand the condition of product display (planograms, retail space management) when analyzing factors affecting sales. Conventionally, obtaining planogram data has been done manually, but it has been costly and involved a heavy workload for employees. NTT DOCOMO has developed retail shelving image recognition technology and used it to implement creation of planogram data automatically, from photographs of retail shelving taken with a smartphone or other camera. This greatly reduces the amount of time required for this work. A solution using this technology was awarded the 19th Automatic Identification Systems Prize, Award of Excellence, from the Japan Automatic Identification Systems Association in 2017, and has been provided by NTT DOCOMO to partner enterprises since April 2018.

1. Introduction

In consumer retail, it is important to know the condition of product displays (planograms*1) when analyzing factors affecting sales. Currently, consumer goods manufacturers create planogram data manually, writing documents from notes taken by sales representatives*2. With the recent growth in large-scale retailers, sales representatives need to visit an increasing number of stores, and with the continually shrinking workforce, a shortage of workers can be expected in the near future. To solve

† Currently Innovation Management Department

*1 Planogram (tana-wari): Refers to the layout of how products are displayed on retail shelving.

this problem, it is critical to reduce operating costs and improve work efficiency. We have estimated that there are tens of thousands of sales representatives performing such tasks in Japan, and we can see huge demand in the market for automating creation of planogram data.

We have proposed a method using image recognition technology to automatically extract planogram data from photographs of retail shelving. There were two main issues with using existing image recognition technologies in real retail environments.

- The first is that object detection is difficult for products when they are packed tightly onto compact displays[*3]. When shelving space is limited, as it often is in Japan, products like shampoo refill pouches are placed in compact displays and possibly only the part of product face is shown or the shape of the product is distorted.
- The second is that it may be difficult to uniquely identify products because they are not always facing the front. In a real store, a product could be facing in any direction on the shelf.

In earlier solutions, the photographing process was manually adjusted in order to make image recognition easier. Product overlap was eliminated and packages were faced to the front as much as possible. However, such adjustment was time consuming and not practical.

For these reasons, NTT DOCOMO has developed a retail shelving image recognition engine that resolves issues with compact displays and photographing angle, using deep learning technology to analyze photographs of retail shelves automatically and create planogram data. Our image recognition engine was recognized by the Japan Automatic Identification Systems Association for "Innovative reproduction of retail shelving using image recognition" and "Solution feasibility," receiving the 19th Automatic Identification Systems Prize, Award of Excellence in 2017 [1]. It has been provided to Cyber Links Co., Ltd. since April 2018 as its first user [2]. This article describes technical details of our image recognition engine together with practical use cases.

2. Image Recognition Overview

Our image recognition engine is composed of two technologies.

The first is a deep-learning based object detection technology that detects whether shelves and products are present in a photograph and determines the spatial position of each product region. The outlines shown in **Figure 1** are the results of object detection. NTT DOCOMO's object recognition technology uses deep learning trained with images of retail shelves in various states, so products can be detected accurately, even if they are packed into shelves in small spaces.

The second is a specific-object recognition technology that uses local feature values[*4] and is able to identify the product in each object region on the shelf. Product region partial images detected as described above are compared with a large number of product images stored in an image database. Images taken from various angles are stored in a large image database beforehand, which enables

*2 Sales representative: An employee that travels to supermarkets, department stores and other retail outlets to promote sales of a company's products. Specific duties could include making sales recommendations and understanding display conditions of the company's products in the stores, studying products from other companies, and getting feedback and any requests from the stores.

*3 Compact displays: A way of completely filling shelves in retails spaces with limited space, often resulting in products becoming somewhat compressed.

*4 Local feature values: Extracted from data, values (numbers) that characterize the data. In this article, "feature values" refers specifically to image feature values, which are characteristic points (corners) extracted from the image and the surrounding distribution of brightness.

Figure 1 Retail shelving image recognition process

our specific-object recognition technology to recognize products accurately, even if the image is not taken from the front.

The processes used by these technologies are shown in Fig. 1.

2.1 Object Detection Technology

1) Algorithm Details

The object detection technology developed by NTT DOCOMO is an application of deep learning and is able to estimate the position (coordinates of upper left and lower right of a rectangle) and a category for each object in an image. To do so, the object detection engine must be trained beforehand in a machine learning*5 process for the category of objects to be detected. For this process, we prepared hundreds to thousands of images with annotation data*6. The annotation data consisted of position data, indicating where the object is depicted in the image (upper-left (x_{min}, y_{min}) and lower-right (x_{max}, y_{max}) coordinates of rectangle) and the object category (e.g.: fabric softener, laundry detergent, drinking water, shelf, etc.) as shown in **Figure 2**.

The machine learning process creates a trained model, which is loaded into the object detection engine, which then performs its inference*7 process to detect objects in input images.

The inference process consists of the following four phases (**Figure 3**).

(1) Extract feature values

Input images are converted to feature values through several convolution*8 layers and pooling*9 layers.

(2) Estimate candidate regions

Using the feature values extracted above, multiple regions are cut out at various aspect ratios and scales, and a fully-connected layer*10 is applied to each to predict whether it is object or background.

*5 Machine learning: Technology that enables computers to acquire knowledge, decision criteria, behavior, etc. from data, in ways similar to how humans acquire these things from perception and experience.

*6 Annotation data: In this article, refers to metadata indicating what is in an image.

*7 Inference: In this article, refers to use of a previously trained model to predict what is depicted in an image.

*8 Convolution: A process of scanning an input such as image or feature value horizontally and vertically, multiplying by a vector of certain size and outputting the value. Extracts patterns similar to the vector used.

Annotation

Object	Type	Position data			
		x_{min}	y_{min}	x_{max}	y_{max}
1	Fabric softener	640	30	700	460
2	Laundry soap	…	…	…	…
3	Laundry soap	…	…	…	…
…	…	…	…	…	…

Figure 2 Example of image and annotation needed for training

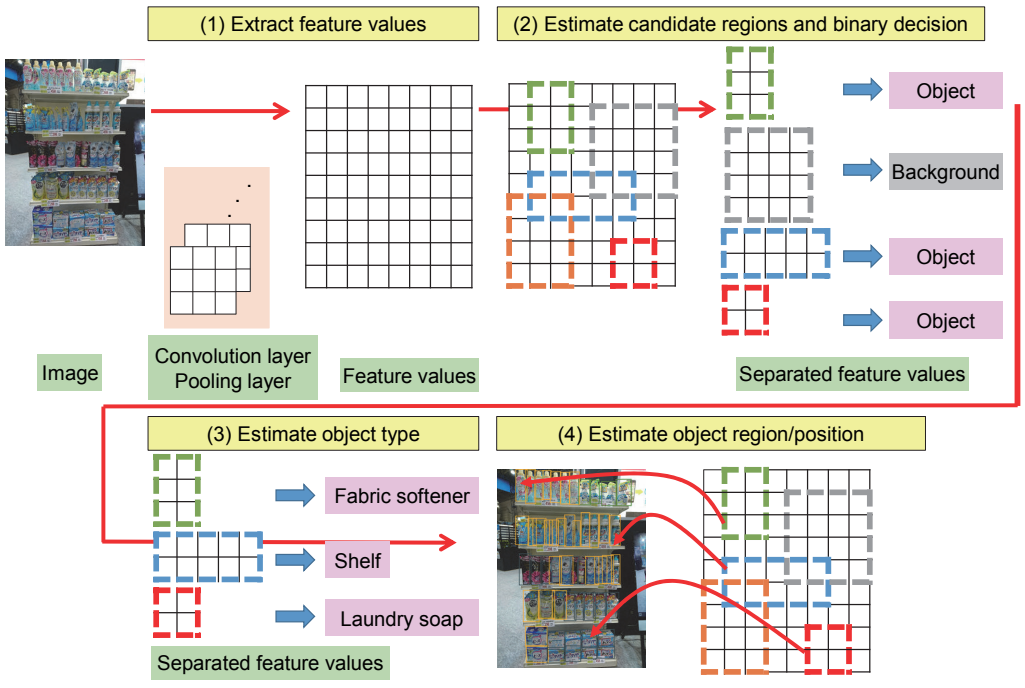


Figure 3 Detection algorithm process

*9 Pooling: A process of scanning against a feature value horizontally and vertically, and outputting the maximum or average value within a fixed size (2 × 2, 3 × 3, etc.). Reduces dimension of the feature value and increases robustness of inference.

*10 Fully-connected layer: Weighting in which all feature values are multiplied by a weighting and added to output a single value. Feature values extracted using convolution are converted to product categories by multiplying them by these weightings.

(3) Estimate object category

For regions judged to be "object" in (2), multiple fully-connected layers with different weights than in (2) are applied to infer the object category. This yields multiple candidate categories and corresponding probabilities, and the region is assigned the category with the highest probability.

(4) Estimate location of object region within the image

The position of the object within the image is estimated using the object category probability found in (3), the position in the image, and the feature values.

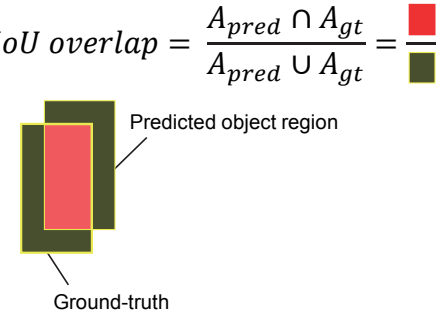
The above process is able to estimate the position and the category of multiple objects depicted in a single image.

Note that the method is able to detect shelves as well as products. This enables it to also estimate factors such as on which shelf a product is placed, and how high products are stacked.

2) Recognition Accuracy

We performed the machine learning process using several thousand images of retail shelves with many products on them, and then evaluated the accuracy of object detection by applying the object detection engine to 100 images taken in real stores and not used for training. Products were in two categories: laundry products such as laundry detergent and fabric softener, and beverages such as drinking water in PET bottles and cans of beer. Recall (No. of products correctly detected/total no. of products) and Precision (No. of correctly detected products/No. of detected products) were used as indices of accuracy. To evaluate accuracy, the corresponding ground-truth region (correct region) for each detected region must first be selected. To do so, an index called IoU overlap is used, which expresses the amount of overlap between two regions as a value from 0.0 to 1.0, as shown in **Figure 4**. The higher the value of IoU overlap, the greater the amount of overlap between the two regions. Detection is judged to be

$$IoU\ overlap = \frac{A_{pred} \cap A_{gt}}{A_{pred} \cup A_{gt}} =$$



A_{gt} : *Area of ground-truth region*
A_{pred} : *Area of predicted object region*

Figure 4 IoU overlap

correct only if the IoU overlap is 0.6 or greater, and the predicted object category is the same as the object category of the ground truth. However, if the probability of the object category computed in (3) was below 0.6, the confidence level is considered low, and such cases are excluded.

When evaluating under these conditions, detection accuracy for beverages had recall of 91.2% and precision of 92.1%. Laundry product detection accuracy had recall of 92.0% and Precision of 99.7%. As shown in **Figure 5**, product packages on real store shelves were not all facing the front, products in pouch-type packaging were easily deformed, and packages of the same product overlapped and could be partially hidden, but the technology was still able to detect them accurately.

3) Processing Speed

We measured processing speed for detection using a GPU and using a CPU. Processing completed in approximately 0.3 s when performing inference with a single NVIDIA Tesla*11 M40 GPU. On the other hand, performing all computations on the CPU (Intel® Xeon®*12 CPU E5-2630L v3 @ 1.80 GHz) required approximately 7.0 s, showing that speed increases greater than ten-times can be achieved using a GPU. System requirements differ according to the application scenario, but this suggests that with the current system, use of a GPU will be essential for real-time processing. Note that the processing speeds described here depend on the size of the input image files, so shorter processing times may be possible using low resolution images.

2.2 Specific-object Recognition Technology

1) Algorithm Details

The specific object recognition technology compares input images with images pre-registered in

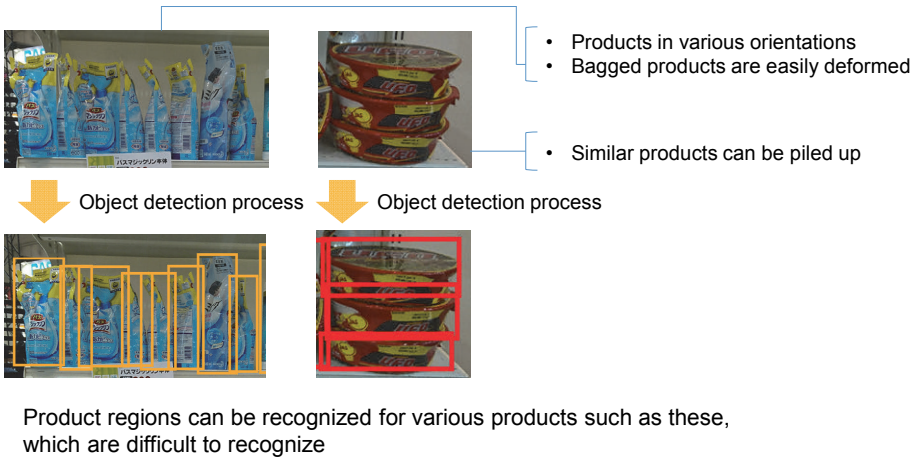


Figure 5 Features of products on retail shelves

*11 NVIDIA Tesla: A trademark or registered trademark of NVIDIA Corporation in the USA and other countries.

*12 Intel® Xeon®: A trademark of Intel Corporation or a subsidiary in the USA and other countries.

an image database, and finds a registered image that is similar to the input image (**Figure 6**).

When recognizing the retail shelf image in Fig. 1, the partial images in the regions found by object detection are input to specific-object recognition engine. The engine is able to recognize products displayed at various angles by comparing with pre-registered images of the products taken from various angles in the database. For details of NTT DOCOMO's specific object recognition algorithm, see reference [3].

2) System Requirements

NTT DOCOMO's specific-object recognition technology is able to process images in real time, even with several million images registered in the image database, but to do so, the database must be loaded in memory. The amount of memory required depends on the number of images and the data-size allocated to each image, but there is an upper limit to the amount of physical memory that a single server can have, so there is a limitation on the number of images a server can handle. As such, a large-scale image database can be built by scaling out*13 with multiple servers. Specific-object recognition can operate at high speed on a CPU, recognizing an image in a few hundred milliseconds, depending on engine settings and the number of registered images.

3) Recognition Accuracy

We evaluated recognition accuracy by preparing retail shelves in a test environment reproducing scenes of products displayed in a real store. Types of products we used included noodles (cup noodles, bagged noodles), detergent (bottles, refill pouches), and beverages (cans, PET bottles). In results of evaluating several hundred product images, the top-ranked candidate was correct 95.96% of the time, demonstrating very high accuracy.

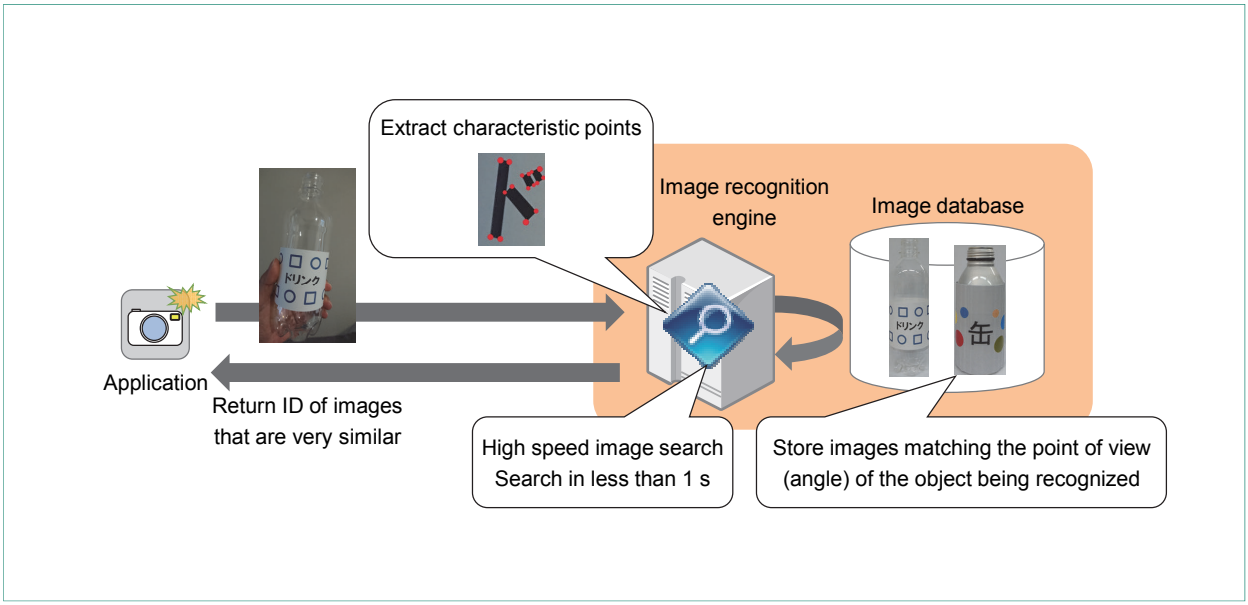


Figure 6 Specific object recognition

*13 Scale out: Adding and assigning new resources to increase processing capacity when service requests increase and there is insufficient processing capacity on the network.

On the other hand, recognition failed in cases where the size or color of the product was wrong, or the image was poor due to camera shake. Products of different size or color, or where just one character is different, are so similar that a person could also easily make a mistake in distinguishing them, so they are generally difficult to distinguish with image recognition as well. For camera-shake, the necessary information cannot be extracted from the image, and this degrades recognition accuracy. Besides these cases, recognition accuracy tended to drop when the image resolution was low, and when the product region in the image was small, so that very little image information could be extracted. To increase recognition accuracy, camera shake must be minimized, and retail shelves must be photographed with as high resolution as possible.

3. Creating Planogram Data

The image recognition engine can be used to identify products in photographs of retail shelving, to analyze the state of product displays, and to output the results as planogram data. This planogram data can be loaded into planogram simulation software to visualize and analyze planograms (**Figure 7**). Typical planogram software used in Japan includes Tana POWER®*14 [4] from Cyber Links Co., Ltd. and StoreManager®*15 [5] from Nippon

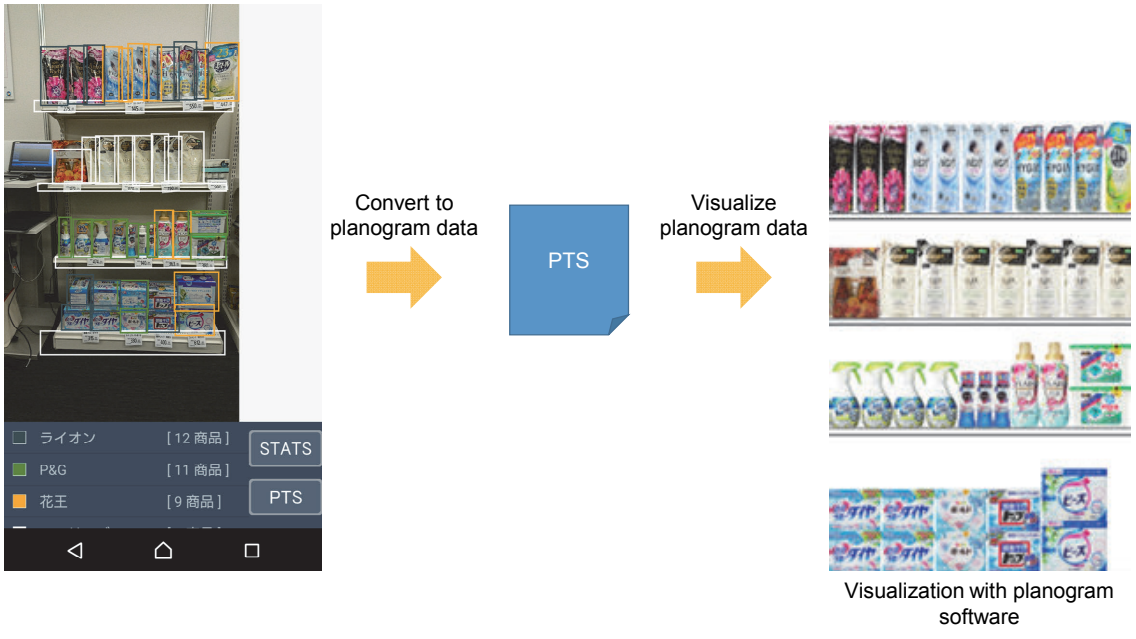


Figure 7 Visualization of planogram data

*14 Tana POWER®: A registered trademark of Cyber Links Co., Ltd.
*15 StoreManager®: A registered trademark of Nippon Sogo Systems, Inc.

Sogo Systems, Inc.

In Japan, a common format used for representing planogram data is Planogram Transfer Specification (PTS) [6]. PTS includes information needed for visualizing products, such as the number of faces*16, stacked height, and which shelf the product is on for each product. The object detection technology described above is able to detect shelves as well as products, so it is able to count number of faces and stacked height accurately for each product on each shelf, and to create planogram data conforming to PTS specifications. This enables the original product displays to be reproduced accurately without needing to reconfigure the shelves when loading data into planogram software and visualizing the planogram.

4. Use Cases

As mentioned earlier, this image recognition engine can be used to create planogram data for sales representatives. We estimate that by creating planogram data automatically from photographs, work time is reduced by a factor of ten compared to creating it by hand. In addition to reducing labor costs, it can also mitigate heavy labor. Specifically, current methods for creating planogram data involve using a barcode reader, which places a heavy burden on the back and legs due to movement up and down from the top to bottom shelves. Using the image recognition engine, this work can be completed with just two or three photographs from the front, simplifying and reducing the workload.

Other scenarios could include use by sales representatives for other consumer products, or for photographs used by retail staff to create work reports. Currently, shops use photographs of retail shelving to report completion of work preparing product displays, but these are limited to a simple visual check of the photographs. Automatically creating planogram data from retail shelving photographs could help head office or other remote offices with management work, inspecting displays at remote stores, checking for errors, and understanding any new display techniques being used at the stores.

5. Conclusion

This article has described two elemental technologies of a retail shelving image recognition engine: one that detects products on shelves and another that recognizes specific products. It also described a capability to create planogram data automatically, by identifying products from images of retail shelving, and detecting their position and shelf within the display using these technologies.

Further work on this image recognition engine includes testing the technology in real shops and improving accuracy. In particular, further study is needed on specific-object recognition technology that can robustly handle products that are difficult, as described above: differing in size, color, or text. Some stores also have price rails*17 attached to shelves, which hide parts of products and obscure information, degrading recognition accuracy. In the future, we intend to collaborate with more partners, testing this technology and identifying cases that make recognition difficult such as these, and continuing to improve our recognition algorithms.

*16 Number of faces: Distribution term indicating the number of items that can be seen displayed when facing the retail shelf.

*17 Price rail: A rail on each shelf in a product display with price tags attached.

Besides product display image recognition, NTT DOCOMO is also providing image recognition solutions to government and partner enterprises in a wide range of fields such as sports video analysis [7], AR service applications [8], business optimization applications by digitizing name cards [9], and detecting pine wilt in coastal forest reserves using drones [10]. Although there are restrictions, some of the image recognition functionality developed by NTT DOCOMO is published by docomo Developer support [11], as Application Program Interfaces (API)*18 for use in developing applications and services. These can be accessed by simply joining docomo Developer support and applying for access. NTT DOCOMO will continue development on image recognition technologies to provide value to partner enterprises in various fields, including retail shelving image recognition.

REFERENCES

[1] Japan Automatic Identification Systems Association: "Japan Automatic Identification Systems Prize | JAISA Japan Automatic Identification Systems Association." http://www.jaisa.jp/award.php

[2] NTT DOCOMO Press Release: "DOCOMO Launches AI Engine for Fast, Accurate Shelf Analysis —Recognizes shelf allocation by analyzing photos of shelved merchandise—," Mar. 2018. https://www.nttdocomo.co.jp/english/info/media_center/pr/2018/0316_00.html

[3] H. Akatsuka et al.: "High-speed, Large-scale Image Recognition API," NTT DOCOMO Technical Journal, Vol.17, No.1, pp.10–17, Jul. 2015.

[4] Cyber Links Co., Ltd: "Basic Operation | Tana POWER | Cyber Links Co., Ltd. Tana POWER/Mise Power." https://www.tanapower.com/tpower/basic_operation.htm

[5] Nippon Sogo Systems, Inc.: "StoreManager | Nippon Sogo Systems Inc." https://tanawari.jp

[6] Japan Planogram Research Association: "Introducing Planogram Transfer Specifications (PTS): Japan Planogram Research Association." https://www.planet-van.co.jp/planogram/pts/index.html

[7] Soccer.com, NTT DOCOMO: "AI Technology "Sports Video Sensing" Development, Multi-angle Automatic Video System linked trials begin —Providing a new sports experience enabling review of one's own plays easily on video—," Feb. 2018. https://www.nttdocomo.co.jp/binary/pdf/info/news_release/topics/topics_180201_00.pdf

[8] NTT DOCOMO: "Cybernet Systems provides image recognition system to cybARnet," Oct. 2015. https://www.nttdocomo.co.jp/binary/pdf/corporate/technology/rd/topics/2015/topics_151510.pdf

[9] NTT DOCOMO Press Release: "Sansan Inc. adopts NTT DOCOMO image recognition system," May 2016. https://www.nttdocomo.co.jp/info/news_release/notice/2016/05/25_00.html

[10] NTT Press Release: "Collaboration agreement reached on drone test project in Niigata City," Sep. 2016. https://www.nttdocomo.co.jp/info/news_release/2016/09/21_00.html

[11] NTT DOCOMO: "Image Recognition | docomo Developer support | NTT DOCOMO." https://dev.smt.docomo.ne.jp/?p=docs.api.page&api_docs_id=102

*18 API: An interface that enables software functions to be used by another program.