CH-016

# Estimation of Product Amount on Store Shelves Using Image Change Classification Based on Background Subtraction and CNN

Kyota Higa† and Kota Iwamoto†

## 1. Introduction

The reduction of sales opportunity loss is effective for improving business profits in retail stores such as supermarkets and convenience stores. The sales opportunity loss happens when products which customers want to buy are not available on store shelves. To reduce such loss, we need to accurately track changes of product amount on shelves and replenish products when stockout occurs. One method to track the changes of product amount is to attach integrated circuit tags (IC tags) on each product and detect their movement using sensor sheets. However, a large amount of time and cost is needed for attaching IC tags on all products in a retail store.

To apply an image processing technology to a video of shelves captured from a fixed camera can be low-cost solution for tracking changes of product amount on shelves. A desired requirement for such technology is to update display condition (presence or absence) of products which changes every moment. In order to satisfy such requirement, it is necessary to accurately detect "product taken (decrease)" and "product replenished (increase)", i.e. actual change in product amount, so that the current display condition can be accurately updated. Although background subtraction [1][2][3] can be applied to detect the changes in the image as foregrounds, it cannot distinguish between "taken" and "replenished" since it detects the regions which do not match to its background model. Therefore, the background subtraction alone is not sufficient for accurately estimating the changes of product amount on shelves.

In order to solve this problem, it is necessary to classify the change regions in the image. We propose a method to estimate product amount on shelves by classifying the change regions detected by background subtraction using a convolutional neural network (CNN) [4].

## 2. Proposed Method

### 2.1 Overview

Figure 1 shows the block diagram of the proposed method. First, the proposed method detects change regions in the image by background subtraction followed by moving object removal. Then, the detected image change regions are classified into four classes representing the accrual change in product amount such as "taken (decrease)" or "replenished (increase)" by using a CNN. Finally, the display condition (presence or absence) of products is accurately updated using classification results, and then product amount on shelves is computed.

† NEC Data Science Research Laboratories

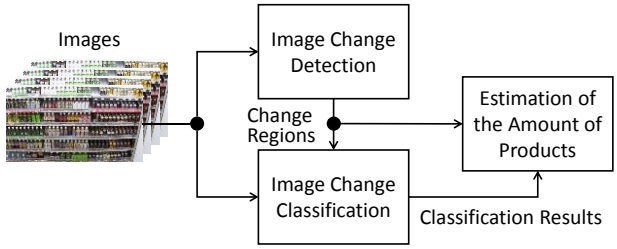


Fig. 1 Block diagram of the proposed method.

### 2.2 Image Change Detection

Our method detects foregrounds based on statistical information of pixels in images in the same way as the conventional background subtraction [1]. However, the background subtraction detects not only the change regions of shelves but also false positives such as moving objects (i.e. customers). We remove false positives by determining correspondence of the foregrounds between consecutive images. The foregrounds with a moving distance larger than $T_d$ pixels are regarded as moving objects and are removed, while the static foregrounds detected for more than $T_s$ seconds are determined as the change regions of shelves.

The Hungarian method [5], which is a combinatorial optimization algorithm which solves the assignment problem, is used for determining the correspondence of the foregrounds in the consecutive images. The correspondence of the foregrounds are determined by minimizing the summation of assignment cost among $N$ foregrounds detected in a current image and $M$ foregrounds detected in a previous image. The cost matrix of the Hungarian method used in our method is as follows.

$$C_{i,j} = \overbrace{\begin{bmatrix} c_{1,1} & \cdots & c_{1,N} \\ \vdots & \ddots & \vdots \\ c_{M,1} & \cdots & c_{M,N} \\ c_{def} & & c_{max} \\ & c_{max} & \ddots \\ & & c_{def} \end{bmatrix}}^{N} \begin{matrix} \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} M \\ \left.\vphantom{\begin{matrix} a \\ a \\ a \end{matrix}}\right\} N \end{matrix} \quad (1)$$

Here, the parameters $c_{def}$ and $c_{max}$ are a default value and a value sufficiently larger than the default value, respectively.

The assignment cost $c_{i,j}$ consists of similarity of color histogram in the foregrounds, area ratio of the foregrounds, and aspect ratio of bounding rectangles of the foregrounds.

The similarity cost $c_{i,j}^{color}$ is computed as,

$$c_{i,j}^{color} = 1.0 - \frac{\sum_k \left(H_t^j(k) - \bar{H}_t^j\right)\left(H_{t-1}^i(k) - \bar{H}_{t-1}^i\right)}{\sqrt{\sum_k \left(H_t^j(k) - \bar{H}_t^j\right)^2 \sum_k \left(H_{t-1}^i(k) - \bar{H}_{t-1}^i\right)^2}}, \quad (2)$$

$$\bar{H}_m^n = \frac{1}{K}\sum_k H_m^n(k), \quad (3)$$

where, $H_t^j$ and $H_{t-1}^i$ are color histograms of $j$th and $i$th foregrounds in the current and previous images, respectively. $K$ is the number of bins. Note that the total area of a histogram has been normalized to 1.0 and $K$ is empirically set to 32.

The area ratio cost $c_{i,j}^{area}$ is computed as,

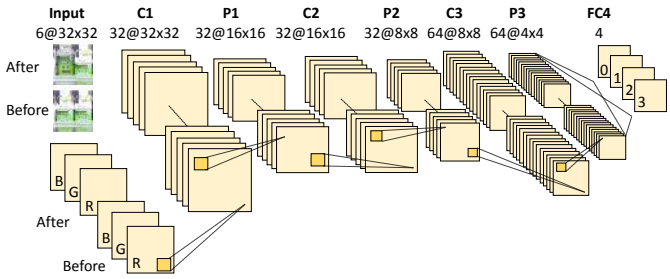


Fig. 2 Neural network architecture.

Table 1 Classes of the change regions of shelves.

| Class | Description |
|---|---|
| 0 | A product was taken. Product amount decrease. |
| 1 | A product was replenished. Product amount increase. |
| 2 | A position or direction of product changed slightly since a customer touched it. Product amount does not change. |
| 3 | It was a false positive due to illumination changes. Product amount does not change. |

$$c_{i,j}^{area} = 1.0 - \begin{cases} S_t^j / S_{t-1}^i & if\ S_t^j < S_{t-1}^i \\ S_{t-1}^i / S_t^j & else \end{cases}, \quad (4)$$

where, $S_t^j$ and $S_{t-1}^i$ are area of $j$th and $i$th foregrounds in the current and previous images, respectively.

The aspect ratio cost $c_{i,j}^{aspect}$ is computed as,

$$c_{i,j}^{aspect} = 1.0 - \begin{cases} A_t^j / A_{t-1}^i & if\ A_t^j < A_{t-1}^i \\ A_{t-1}^i / A_t^j & else \end{cases}, \quad (5)$$

where, $A_t^j$ and $A_{t-1}^i$ are aspect ratio of the bounding rectangle for $j$th and $i$th foregrounds in the current and previous images, respectively.

Finally, the assignment cost $c_{i,j}$ is computed as,

$$c_{i,j} = w_1 \times c_{i,j}^{color} + w_2 \times c_{i,j}^{area} + w_3 \times c_{i,j}^{aspect}, \quad (6)$$

where, $w$ means a weight.

## 2.3 Image Change Classification

The detected change regions of shelves are classified into four classes using a CNN. The CNN used in our method is constructed based on the CIFAR-10 network. Our network architecture is illustrated in Fig. 2. It consists of three convolutional layers, three pooling layers, and a fully-connected layer. Input data of our network is a six-channel image constructed from images before and after the change. Table 1 shows output of our network. The class 0 and 1 show “product taken” and “product replenished”, respectively. Other classes show that product amount of shelves does not change. We use training samples collected form videos captured in real stores.

In our method, the images before and after the change are used as input data of our network to classify the change regions of shelves. There has been a work [6] on classifying a difference image by a deep neural network for image change detection of synthetic aperture radar images. However, the accrual change in product amount such as “taken” and “replenished” cannot be distinguished by the difference image since the difference of those


Fig. 3 Examples of difference images for the change regions of shelves.


Fig. 4 Examples of images before and after the change for each class.

change regions are very similar as shown in Fig. 3. Therefore, we use the images to which extra operations are not applied.

The images before and after the change are bounding rectangles of the change regions. Examples of them are illustrated in Fig. 4. The left side of each image is the image before the change, while the right side of each image is the image after the change. Note that the images before and after the change are normalized to 32x32 pixels before constructing the six-channel image.

## 2.4 Estimation of Product Amount on Store Shelves

Our method updates status of the display condition (presence or absence) of products using the classification results, and then computes product amount on shelves for the predefined monitoring areas.

An example of the display condition of products and how to update it are illustrated in Fig. 5. Here, the display condition of products is represented as a binary image. The white and the black regions of the display condition show presence and absence of products, respectively. If the classification results are “taken (class 0)”, the display condition corresponding to the change regions are set to black. If the classification results are “replenished (class 1)”, the display condition corresponding to the change regions are set to white. The display condition is not changed in the case of other classification results.

Figure 6 shows the monitoring areas on shelves. Each white area in the right image is the monitoring area and corresponds to each

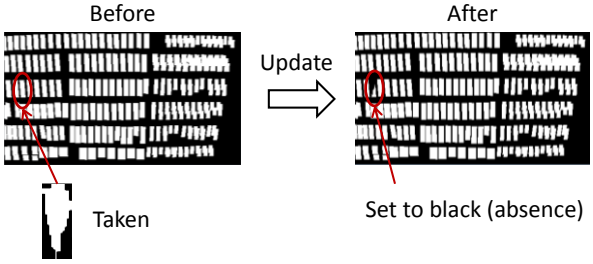


Fig. 5 Example of the display condition of products and how to update it.


Fig. 6 Monitoring areas on shelves. The left is the image captured from a fixed camera. The right is the monitoring areas with shelf numbers.

shelf in the left image. For each shelf, product amount is computed as,

$$A(n) = S_d(n)/S_m(n), \quad (7)$$

where, $A(n)$, $S_d(n)$, and $S_m(n)$ are product amount on $n$th shelf, the product area of $n$th shelf, and $n$th monitoring area, respectively.

## 3. Evaluation

### 3.1 Experimental Conditions

We evaluate estimation accuracy of our method using two videos of store shelves captured from a fixed camera attached on the ceiling in a real store. Duration, resolution, and framerate of the videos are approximately 100 and 200 minutes, 1920x1080 pixels, and 1 fps, respectively. The videos are resized to 480x270 pixels. A sample frame of the videos and the monitoring areas used in the experiment are shown in Fig. 6. The thresholds for moving object detection $T_d$ and change region detection $T_s$ are 10 pixels and 30 seconds, respectively. The default parameter $c_{def}$ in the cost matrix of the Hungarian method is set to 1.0. All weights for the assignment cost $c_{i,j}$ in equation (6) are also set to 1.0. The number of training samples for each class in the image change classification is shown in Table 2. Training samples are extracted from other videos captured in the real store with same conditions. The initial display condition of products is set manually. For every minute, we compute product amount on shelves between the range of 0.0 and 1.0 using equation (7), and decide if the difference between the estimated product amount $A(n)$ and ground truth is within tolerance (acceptable error). If so, the estimated product amount is determined as "correct". Else, the estimated product amount is determined as "incorrect". Then, the estimation accuracy for each shelf is computed as,

$$Accuracy = C_n/(C_n + I_n), \quad (8)$$

where, $C_n$ and $I_n$ are the number of "correct" and "incorrect", respectively.

### 3.2 Results

Figure 7 shows the overall estimation accuracy of product amount on shelves at various tolerance value. The vertical and horizontal axes are the accuracy and the tolerance, respectively. The blue line shows the accuracy of the proposed method, while the red one shows that of the proposed method without the image change classification (i.e. the image change detection only). All the detected change regions including such as "taken" and "replenished" are regarded as "taken" in the case of the method without the image change classification. The proposed method achieves the estimation accuracy of 89.2% when the tolerance is 0.10 which is equivalent to an error of approximately ±1 product or less since there are approximately 10 products on each shelf as

Table 2 Number of training samples for each class.

| Class | Training Samples | Class | Training Samples |
|---|---|---|---|
| 0 | 7144 | 2 | 900 |
| 1 | 7144 | 3 | 1652 |

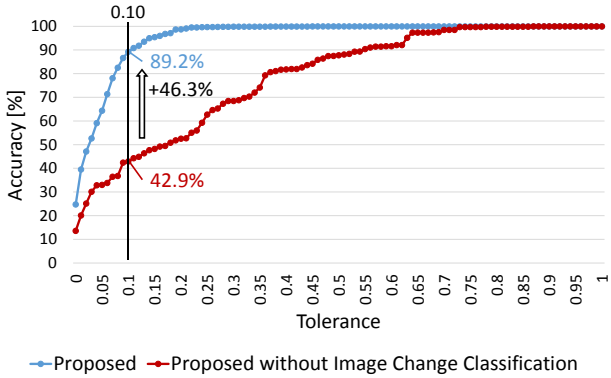


Fig. 7 Overall estimation accuracy of product amount on shelves at various tolerance value.

Table 3 Estimation accuracy for each shelf when tolerance is 0.10. The shelf numbers are shown in Fig. 6.

| No. | Accuracy | No. | Accuracy | No. | Accuracy |
|---|---|---|---|---|---|
| 1 | 92.1% | 7 | 100% | 13 | 83.8% |
| 2 | 78.4% | 8 | 90.3% | 14 | 100% |
| 3 | 98.9% | 9 | 97.1% | 15 | 58.3% |
| 4 | 84.5% | 10 | 64.4% | 16 | 93.2% |
| 5 | 81.7% | 11 | 98.9% | 17 | 98.6% |
| 6 | 99.6% | 12 | 99.3% | 18 | 87.1% |

shown in Fig. 6. The result of the proposed method is also 46.3% higher than that of the method without the image change classification.

Table 3 shows the estimation accuracy for each shelf when the tolerance is 0.10. The shelf numbers are shown in Fig. 6. As shown in Table 3, our method achieves high accuracy in more than half of shelves.

Figure 8-10 show the changes of product amount on shelves in the case of high and low estimation accuracy, respectively. The results of shelf #3, #11, and #17 are shown in Fig. 8 and Fig. 9 as the case of high estimation accuracy. The results of shelf #10 and #15 are shown in Fig. 10 as the case of low estimation accuracy. The vertical and horizontal axes are product amount on shelves and elapsed time of the videos, respectively. Dotted lines show ground truth, while solid ones show the estimated product amount on shelves. As shown in Fig. 8 and Fig. 9, our method accurately tracks the changes of product amount on shelves. In the result of shelf #10 shown in Fig. 10 (purple lines), our method fails to classify the change regions when the elapsed time is approximately 90 minutes; that is, product amount of ground truth (the purple dotted line) increases greatly, while that of our method (the purple solid line) increases slightly. Consequently, the average of the estimation accuracy for the shelf #10 is low since the difference between our results and ground truth is larger than 0.10 (the tolerance value of estimation accuracy shown in Table 3) between 90 and 180 minutes in Fig. 10. However, tendencies of

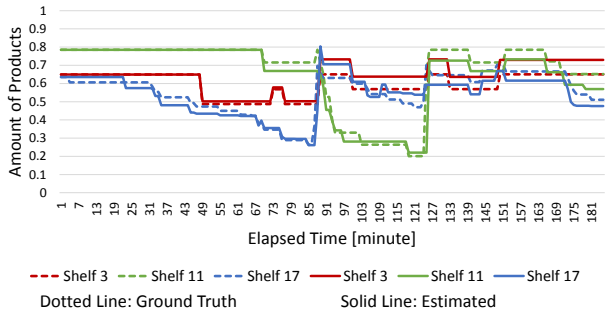


Fig. 8 Changes of product amount on shelves in the case of high estimation accuracy in the video 1.

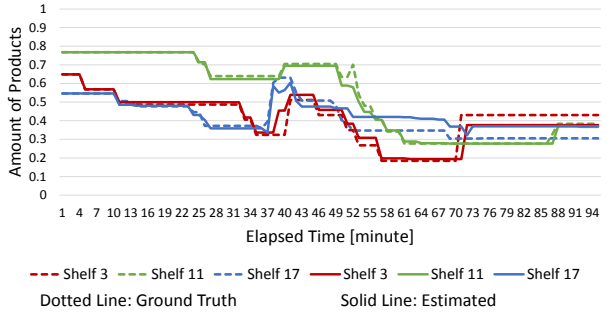


Fig. 9 Changes of product amount on shelves in the case of high estimation accuracy in the video 2.

the changes between our results and ground truth are almost same even in the case of low estimation accuracy.

Figure 11 shows the estimation results of display condition and that our method accurately tracks the changes of shelves.

From these results it can be concluded that our method accurately estimates product amount on shelves by classifying the detected change regions into the four classes shown in Table 1.

## 4. Conclusion

This paper proposed a method to estimate product amount on store shelves from a video captured from a fixed camera attached on the ceiling. First, the proposed method detects change regions in the image using background subtraction followed by moving object removal. Then, the detected image change regions are

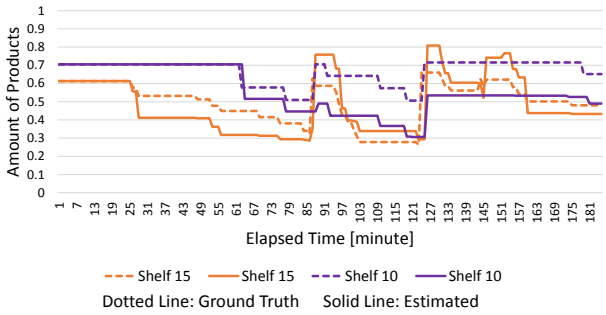


Fig. 10 Changes of product amount on shelves in the case of low estimation accuracy in the video 1.

classified into four classes representing the accrual change in product amount such as “product taken (decrease)” or “product replenished (increase)” by using a convolutional neural network. Finally, the display condition (presence or absence) of products is accurately updated using classification results, and then product amount on shelves is computed. Experimental results using two videos captured in a real store show that the proposed method achieves estimation accuracy of 89.2% when tolerance is approximately one product. With high accuracy, store clerks can replenish products when stockout occurs, enabling the reduction of sales opportunity loss in retail stores.

## References

[1] Z. Zivkovic, “Improved Adaptive Gaussian Mixture Model for Background Subtraction”, In Proceedings of the 17th International Conference on Pattern Recognition, Vol. 2, pp. 28-31, 2004.

[2] P. KaewTraKulPong, *et al.*, “An Improved Adaptive Background Mixture Model for Real-time Tracking with Shadow Detection”, In Proceedings of the 2nd European Workshop on Advanced Video Based Surveillance System, 2001.

[3] A. B. Godbehere, *et al.*, “Visual Tracking of Human Visitors under Variable-Lighting Conditions for a Responsive Audio Art Installation”, American Control Conference, pp. 4305-4312, 2012.

[4] Y. LeCun, *et al.*, “Backpropagation Applied to Handwritten Zip Code Recognition”, Neural Computation, Vol. 1, Issue 4, pp.541-551, 1989.

[5] J. Munkres, “Algorithms for the Assignment and Transportation Problems”, Journal of the Society for Industrial and Applied Mathematics, Vol. 5, No. 1, pp. 32-38, 1957.

[6] J. Zhao, *et al*, “Deep Learning to Classify Difference Image for Image Change Detection”, International Joint Conference on Neural Networks, pp. 411-417, 2014.

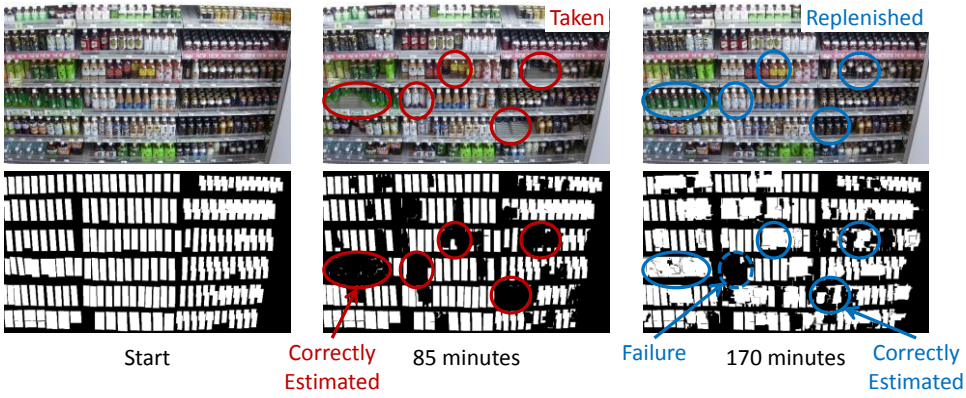


Fig. 11 Estimation results of display condition. First row is captured image. Second one is estimated display condition.