The Data-Driven Deep Learning to Localization Product in Convenience Store

Muslikhin¹,², Y Indrihapsari³,⁴, A A Baiti¹,² and M S Wang²
¹ Department of Electronics Engineering Education, Universitas Negeri Yogyakarta, Yogyakarta City 55281, Indonesia
² Department of Electrical Engineering, Southern Taiwan University of Science Technology, Tainan City 710, Taiwan (R.O.C)
³ Department of Information Technology, Universitas Negeri Yogyakarta, Yogyakarta City 55281, Indonesia
⁴ Department of Informatics Management, National Taiwan University of Science Technology, Taipei City 106, Taiwan (R.O.C)

E-mail: ¹muslikhin@uny.ac.id

Abstract. The penetration of Industry 4.0 in the convenience store is increasingly visible. The combination of Artificial Intelligence (AI), the Internet of Things (IoT), and robotic were always under evaluation, including applied AI in the whole system that is demanded robust when integrated with online shop. Meanwhile, the problem of speed detection and precision is still a challenge in AI. In this study data-driven deep learning was developed to improve the speed and maximize localization of consumer products. After the selected data-driven received from the online shop platform then triggers deep learning and activates a modified YOLOv2 in parallel. Thus, each detector per product is needed. Our system performance is proven by experiments to meet expectations in evaluating speed, precision, and performance in recognition and localization.

Keywords: convenience store; data-driven; deep learning; industry 4.0.; localization; modified YOLOv2; product recognizing.

1. Introduction
Certainty, Industry 4.0 enters the sector closest to society, namely the convenience store cannot be predicted. However, big wholesale like Amazon and Walmart have initiated it [1]. Amazon and Walmart use an integrated, marker-based and positioning system. Separately, traditional grocery stores are still semi-structured even though they are equipped with a barcode or QR code, item name, and price. If industry 4.0 is involved soon, then the study of AI, robots, IoT, and big data as a welcome need to be accelerated [2].

Since the Japanese Cabinet introduced Society 5.0, the role of AI has been the backbone of the system [2]-[3]. With the emergence of AI in the last decade, Krizhevsky et al. [4] have introduced AlexNet to the CNN concept. Furthermore, Girshick et al. [5] proposed the Faster R-CNN, which had better speed and accuracy than AlexNet, CNN, R-CNN, and Fast R-CNN. Another quite capable method is offered by Redmon et al. [6] using the YOLO (You Only Look Once) method, the latter being YOLOv3. As for YOLO v2, its detection speed is still superior to Faster R-CNN.
YOLO continues to be refined and applied in the robotics, automotive, medical, and military fields. Mao et al. [7] using a modified YOLOv3 technique to lessen the Floating Point Operation (FLOP), which can improve achievement 2.5 times faster. YOLOv3 also involves K-means to classify clusters in specific targets and the accuracy over 90% [8]. However, YOLOv3’s speed requires a fast GPU and a computer. Not only the hardware requirements but also some detection approaches focus on the value of confidence. In specific applications, the capture process is essential once the target is known [9]-[10].

References [11] - [13] their studies are focused on AI, robots, and data mining for online shops. Yamamoto et al. [11] work analysing consumer behavior of various attributes stare, listen and smell while shopping on an online site. Meanwhile, [13] tends to data mining which is presented in a preference matrix so that within a certain time it can be seen which products are selling well. Both [11] and [24] use an online shop platform, IoT, and applied data mining. Even so, it has not been explained how the recognition process on the side of the offline shop is, and we suspect that the item recognition system has not been integrated with the vision robot. Furthermore, Li et al. [14] have engaged online shop platforms with robots. However, the process of taking off the shop shelves still involves humans; here, there is a chance that an error may occur before it is passed on to the packaging robot.

We work between Li and Yamamoto to ensure the validity of the recognition of consumer groceries by the vision robot. In this study, we improve from previous work. The object localization, depth estimation, and eye-in-hand R-CNN construction with a mono camera were adopted by our prior work [8]. This method [15] recognizes and sorts the targets then determines the centroid and the target depth (Z). Meanwhile, from the disparity of right and left images, the other two coordinates (XY) are obtained. In additional work [16], we also used a stereo camera for the selected selection using CNN with an eye-to-hand structure. However, on-target applications are relatively easy to recognize with almost uniform shapes and colors even in cluttered and cluttered environments. Our challenge now lies in the heterogeneous and data-driven environment that comes from an online app or website store.

The approach we propose is modifying the front of the YOLOv2. In the detector section, a parallel number of product data items are sold. In the modified YOLOv2 the training process will be very tiring. From the online shop side, the delay time for REST API replies as data-driven becomes a problem for the system. For this reason, this work is focused on the modification of detectors in deep learning combined with vision robots to make them accurate according to what consumers buy. In this paper, we ascertain the accuracy and speed of recognition of the store shelf environment.

In this paper, we discuss parallel detectors in Section 2. Section 3 introduces data-driven online shopping until the system sends the last transaction to an offline store and the localization process uses the SURF Disparity Map in Section 4. Section 5 next details the empirical results. Finally, we conclude the work and give in Section 6 ideas for potential future studies.

2. Parallel Detector YOLOv2

2.1. YOLOv2

In practice, YOLO (You Only Look Once) ran much faster than Faster RCNN due to its simpler architecture. Unlike CNN, RCNN, Fast RCNN, and Faster RCNN are trained to perform constraint box classification and regression simultaneously [16]. YOLO architecture is inspired by GoogleNet, see Figure 1.
When the output feature width and height are between 8 and 16 times smaller than the input image, the YOLOv2 feature extraction layer is most effective. This quantity of down-sampling is a trade-off between spatial resolution and performance function efficiency. Note that it requires an analytical evaluation to select the optimal feature extraction layer. This is means that consistency is needed when labelling. YOLOv2 uses anchor boxes to detect object classes in an image, shown in Figure 2. YOLOv2 uses anchor boxes, shown in Figure 2, to recognize object classes in an image. For each anchor box, YOLOv2 predicts these three attributes; 1) intersection over union (IoU) predicts each anchor box's objectivity ranking, 2) anchor box offsets as an improvement to the anchor box's location, 3) class likelihood-predicts the class mark assigned to each anchor box.

The basis of the YOLOv2 architecture is the reorganization layer (Reorg Layer). For mixing low-level and high-level features, the Reorg Layer (created using the yolov2ReorgLayer object) and the depth concatenation layer (created using the depthConcatenationLayer object) will be used. By incorporating low-level image information and enhancing detection precision for smaller objects, these layers enhance detection. Usually, within the feature extraction network, the reorganization layer is connected to a layer whose output feature map is larger than the output of the feature extraction layer.

2.2. Selected Parallel Detector
In general, the YOLOv2 detector can detect multiclass. However, the weakness that arises is that the true negative (TN), false positive (FP), and false negative (FN) values still occur. To minimize this, we modified the YOLOv2 detector part. The consequence of this modification appears that many detectors
are in accordance with the number of products available in online or offline stores. The detector on the YOLOv2det system will determine which part of the detector responds to the data-driven \( A_{det}[i] \)

\[
YOLOv2_{det}[i] = A_{det}[1] \| B_{det}[2] \| C_{det}[2] \| \ldots N_{det}[n]
\]

When data-driven is received by the system through the data parsing process, the product name attributes A, B, C… N will be automatically selected. This selection reduces computation time significantly.

3. The Data-Driven in AI Store

The online shop's minimal features are added, as we concentrate on building AIoT for selecting algorithms. These features include the display of offers, purchases, checks and updates on stocks. As stated previously, prospective customers must be registered for shopping and served by the system [1],[4]. This online shop uses the Laravel PHP framework with a MySQLi database and has adopted the Transport Layer Security (TLS) protocol for its standard security.

The data-driven principles are like event-driven programming, are structured as pattern match and subsequent processing, and are typically implemented via the main loop. The condition or model of action is similar to aspect-oriented programming, in which some actions are performed after approaching a joint point (condition). In this study, the action to be taken in the form of a vision robot will recognize products that have been purchased by consumers on the online platform, see Figure 3.

![Figure 3. REST API request-response in store powered by deep learning](image)

Online shop data (app/website) is reviewed by robots on a regular basis. The last data of the transaction is often compared with the previous data in order to ensure that the data is accurate. The connection between the online and offline stores via the REST API. When a client makes a REST API request with the URL endpoint format, method, and parameter, the response from the web server is in the form of raw data with many categories. Starting from the order number, user identity, order status, ordered items, to the detailed address sent to the system to carry out the localization process.

4. Purchased Product Localization using SURF Disparity Map

Purchased products are essential data for system performance. Items that have been purchased must be accurately identifiable. Apparently not only that, accuracy alone is not enough for the robot to grasp the target. Therefore a localization process is needed to ensure each XYZ coordinate.

4.1. Stereo Camera-Like

The configuration scheme of the stereo camera like is based on stereo vision, shown in Figure 4. Mono camera set in parallel line \( x \) is fixed in the work space. Two optical center cameras have a baseline of \( b \), and they have the same \( f \) focal length. The projection is \( p_1(x_1, y_1) \) and \( p_2(x_2, y_2) \) in the image plan 1.
and image plan 2, respectively as given reference point \( P(X_p, Y_p, Z_p) \). Then we have the image coordinates of \( P \) in two image planes by perspective projection, and we can see in (2) to simplify the calculation [8].

\[
\frac{X}{x_1} = \frac{Z}{f}; \quad \frac{Y}{y_1} = \frac{Z}{f}; \quad \frac{b - X}{x_2} = \frac{Z}{f}
\]

(2)

In Figure 4, we assume two cameras have the criteria for defining the object. His images are and is the parallax on the two cameras, and the Y-axis is perpendicular to the page. We are able to get the equation according to the theory of identical triangles (2). Figure 4 we can also see that \( b \) can be written in equation (3), also \( Z \) is the point depth \( P \) obtained by equation (4).

\[ b = \frac{Z}{f} x_1 + \frac{Z}{f} x_2 \]

(Error! No text of specified style in document.)

\[ Z = \frac{b^* f}{x_1 + x_2} \]  

(4)

In pair images of 1 and 2 as written in equation (5), \( d \) is the disparity in difference \( x \) coordinate.

\[ d = x_1 + x_2 \]

(5)

After \( Z \) has been obtained, we can know the \( P, X, \) and \( Y \) coordinates using equations, respectively.

\[ Z = \frac{b^* f}{d}; \quad X = \frac{Z^* x_1}{f}; \quad Y = \frac{Z^* y_1}{f} \]

(6)

Where \( x_1 \) and \( x_2 \) are the pixel locations on the 2D image, even the actual 3D image positions are \( X, Y, \) and \( Z \).

4.2. Ensuring Location with SURF
The box filter is used in the conventional SURF algorithm to approximate the convolution of the 2D image and the second order derivative of Gaussian, which can simplify the calculation and enhance the efficiency [17]. The SURF detector is using Hessian matrix because of its good performance in time of
computation and accuracy. The Hessian matrix $\mathcal{H}(x, \sigma)$ in $x$ at scale $\sigma$ is defined as follows, provided the point $x = (x, y)$ in an image $I$.

$$\mathcal{H}(x, \sigma) = \begin{bmatrix} L_{xx}(x, \sigma) & L_{xy}(x, \sigma) \\ L_{xy}(x, \sigma) & L_{yy}(x, \sigma) \end{bmatrix}$$

(7)

where the Gaussian second order derivative $\frac{\partial^2}{\partial x^2} g(\sigma)$ in the image $I$ within $x$ point, and similarly for $L_{xy}(x, \sigma)$ and $L_{yy}(x, \sigma)$ then were convolute $L_{xx}(x, \sigma)$.

**Figure 5.** The 9-9 box in the y-direction (crop-discretized) of Gaussian second-order partial derivatives, the regions of grey are equal to zero.

$$\text{det(}\mathcal{H}_{\text{approx}}\text{)} = D_{xx} \cdot D_{yy} - (0.9 \cdot D_{xy})^2$$

(8)

where $D_{xx}$, $D_{yy}$, $D_{xy}$ are approximation. With this detector, the image to be detected must be converted into a gray image first.

### 4.3. Verifying SURF by Disparity Map

The $D(x, y)$ is a disparity map represents that displacement in terms of similar pixels between the left and right images [18]. But locating corresponding pixels is difficult in a real application. Some variables for instance; camera noise, repeated-texture, and untextured-homogeneous can cause trouble in the nonocclusion pixels. The estimation of the disparity is achieved by matching block for all pixels, and the sum of the validity of disparity shall be as follows.

$$D_{L \rightarrow R}(x, y) = \arg \min_{\epsilon \in [0, D_{\text{max}}]} \epsilon_{L \rightarrow R}^d(x, y)$$

(9)

$$\epsilon_{R \rightarrow L}^d(x, y) = \frac{\sum_{(u,v)} \sum_{W} |f_r(x-u,y-v) - f_l(x-u+d,y-v)|}{\sum_{(u,v)} \sum_{W} |f_r(x-u,y-v) + f_l(x-u+d,y-v)|}$$

(10)

The pair images (left image and right image) disparities are received from (9) and (10), where $\epsilon_{R \rightarrow L}^d(x, y)$ is as the normalized block matching error. This $\epsilon_{R \rightarrow L}^d(x, y)$ with a horizontal disparity $d$, a block matching window $W$, and the disparity of maximum value $D_{\text{max}}$ within the permitted limit. To check the observed disparity between $f_r$ is the right frame to $f_l$ is the left frame, while $u$ and $v$ are the number of pixels in xy-image plane of camera, respectively.

$$D_{R \rightarrow L}(x, y) = \arg \min_{\epsilon \in [-D_{\text{max}}, 0]} \epsilon_{R \rightarrow L}^d(x, y)$$

(11)

The minimum matching error (MME) calculates then whereby to bind the values of both images at $(x,y)$ in the left image plane and the corresponding point $(x+d, y)$ on the right image plane and Eq. (12) is composed of MME.

$$\text{MME}(x, y) = \epsilon_{L \rightarrow R}^d(x, y)|_{d=D_{R \rightarrow L}(x,y)}$$

(12)
Thus, the layer heterogeneity clustering algorithm was used to refine out unclear layers. Verifying the disparity map by map-based clustering of inequalities is a method in which variations in each variety of pair images have different pixel positions to address \( Z_c \) to \( Z_r \) for the same product.

5. Experiment Work

5.1. Experimental Settings

In this paper, we divide into two parts. Firstly, the online shop has been develop using Laravel platform and MySQLi database and hosted in https://indoaltantis.com for the website version. In an app version also, we converted it for Android and iOS. The last is an offline shop: the experiment is conducted using MATLAB 2020a with C920 webcam mounting in the end-effector and connected to MELVA RV-3SD 6 DOF manipulator. Dataset for YOLOv2 detectors are around 500 images. We set our system as minimum as possible using Windows 10 Pro 64-bits operating system, Intel Core i5 CPU @ 3.0 GHz (6 CPUs), 16 GB of RAM, and with internal GPU. Figure 6 presents the overall schematic of our framework.

Figure 6. The overall of our system, the data-driven is input from online shop app triggers a modified YOLOv2 algorithm.

5.2. Experimental Results

In this experiment we carried out a retrieval experiment 87 times as in Table 1. Metrics for the detection precision, yield, and F1 score were evaluated as described in (13). To test the detection method, the amount of True Positive (TP) and False Positive (FP) is involved in a total of 500 recorded images [19]. We are set the threshold of confidence value of 0.86 and assigned to calculate precision, recall, and performance (F1 score).

\[
\begin{align*}
\text{precision} &= \frac{TP}{TP + FP} \\
\text{recall} &= \frac{TP}{TP + FN} \\
F_1 &= \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\end{align*}
\]  

The results are shown in Table 1, where products purchased by consumers can be detected with a perfect level of detection precision. The use of a parallel detector makes it easier to identify products in offline stores.
In this experiment, we conducted trials on online shop with a limited number of products according to the reach of the manipulator robot. There are three column shelf and three rows of shelf, each filled with a different product and some in one shelf containing more than one type of product. Our products are limited to the following items: ABC Ketchup (1), British Milk Tea (2), Gau Vermicelli (3), Cup Noodle (4) Apple Yogurt (5) Soto Noodle (6) Lageo Wafers (7) Master, Sardines (8) Tai Lemon Tea (9) or can be seen in Figure 7e.

The product localization process starts from taking an image by the camera, this image is labeled as the right image. Then the camera moves on the robot Yr-axis or shop shelf Xt-axis as far as the baseline b of 10 mm and takes a second picture or left image. After the two images are obtained, the right and left image composite process is carried out. Furthermore, the disparity map will be obtained as verification, shown in Fig. 7a. The result in Fig. 7c with XYZ localization is obtained from process of Fig. 7d where SURF determines the accuracy of the right and left image disparity.

We did the test with the test as numerous as 87 times taken by the robot. The results are shown in Table 1. The fastest recognition time for Soto Noodle products (6) with an average of 0.054 s, while localization with an average error of only 1,050 mm. More detail related to the whole system can be watching in this link https://youtu.be/xiVZ6fd8Zlg.

| Parameters | Classes (Products) |
|------------|--------------------|
| Confidence level |
| Precision |
| Recall |
| Performance (F1) |
| Time (s) |
| Loc.(mm) |
| Error (µ) |
| Error (σ) |

Table 1. The results of product localization testing use data-driven deep learning

![Figure 7. a) Red-cyan composite view of the stereo pair image rectified, b) disparity map from (a), c) output of YOLOv2 with confidence level and XYZ, d) merger of left and right image, e) offline store with robotic manipulator.](https://youtu.be/xiVZ6fd8Zlg)
Technically, the localization evaluation solution using SURF, disparity map, and stereo vision-like is quite robust. The identification results are shown for each product sequence as confirm in Table 2. In contrast to the detection results of the purchased product, the depth was found to be significant at small differences, only at 2.5 mm in maximum and the overall mean detection time \( \mu = 0.067 \) second for each target.

6. Conclusion
This system takes a pair of product images on the shop shelves using a mono camera with a baseline. The experiment was completed by using MATLAB. Localizing and recognizing time can be triggered with data-driven deep learning which is YOLOv2 parallel detectors are applied in a convenience store. The SURF with disparity map is built to obtain XYZ position and location of the purchased product also double-checked by disparity map. The localization and recognition time show that performs well. In terms of performance, the system depicts acceptable errors for the gripper within a value of 1.06 mm~2.5 mm. In future work, we discretion use a selective detector on YOLOv2. We assume the process will computationally accelerate at least twice its current speed.

References
[1] Mantha A et al 2020 A large-scale deep architecture for personalized grocery basket recommendations *IEEE International Conference on Acoustics, Speech and Signal Processing* (Barcelona) vol – (Barcelona: Spain) p 3807
[2] Matsuda K et al 2019 Technologies of production with society 5.0 2009 6th Int. Conf. on Behavioral, Economic and Socio-Cultural Computing (Beijing) vol – (Beijing: China) p 1
[3] Martynov V V et al 2019 Information technology as the basis for transformation into a digital society and industry 5.0,” in 2019 Int. Conf. Quality Management, Transport and Information Security, Information Technologies (Sochi) vol – (Sochi: Russia) p 539
[4] Shibata M et al 2017 Toward an efficient search method to capture the future mot curriculum based on the society 5.0 *Portland Int. Conf. of Management of Eng. and Tech. (Portland)* vol– (Portland: USA) p 1
[5] Chang W et al 2020 ScalpEye: A deep learning-based scalp hair inspection and diagnosis system for scalp health *IEEE Access* 8 134826
[6] Krizhevsky A et al 2012 Imagenet classification with deep convolutional neural networks *Proc. Adv. Neural Inf. Process. Syst.* p 10971105.
[7] Girshick R et al 2015 Fast r-cnn *IEEE Int. Conf. Computer Vision* p 14401448.
[8] Redmon R et al 2016 You only look once: unified, real-time object detection *IEEE Conf. Comput. Visual Pattern Recognition* p 779788
[9] Mao Q et al 2019 Fast and efficient non-contact ball detector for picking robots *IEEE Access* 7 175487
[10] Muslikhin et al 2020 Object localization and depth estimation for eye-in-hand manipulator using mono camera *IEEE Access* 8 121765
[11] Ge Y et al 2019 Fruit localization and environment perception for strawberry harvesting robots *IEEE Access* 7 147642
[12] Nguyen P V et al 2019 Grasping interface with wet adhesion and patterned morphology: case of thin shell *IEEE Robotics and Automation Letters* 4 792
[13] Yamamoto Y et al 2016 Enhanced iot-aware online shopping system 12th International Conference on Signal-Image Technology & Internet-Based Systems (Naples) vol 14 (Naples: Italy) p 31
[14] Rashid Z et al 2015 Bringing online shopping experience to offline retail through augmented reality and RFID 5th International Conference on the Internet of Things (Seoul) vol – (Seoul: Korea) p 45
[15] Gaggero M et al 2019 When time matters: predictive mission planning in cyber-physical scenarios *IEEE Access* 7 11246
[16] Li L et al 2016 Design patterns and extensibility of rest api for networking applications *IEEE Transactions on Network and Service Management* 13 154
[17] Song C et al 2019 Inferring 3d shapes of unknown rigid objects in clutter through inverse physics reasoning *IEEE Robotics and Automation Letters* 4 201
[15] Du Y C et al 2020 Stereo vision-based object recognition and manipulation by regions with convolutional neural network Electronics 9 210

[16] Yang H et al 2019 Tender tea shoots recognition and positioning for picking robot using improved yolo-v3 model IEEE Access 7 180998

[17] Li Y et al A novel surf based on a unified model of appearance and motion-variation IEEE Access 6 31065

[18] Park J H 2003 Fast view interpolation of stereo images using image gradient and disparity triangulation Signal Process.: Image Communication 18 401

[19] Bargoti S et al 2017 Image segmentation for fruit detection and yield estimation in apple orchards Journal of Field Robot 34 10391060

Acknowledgments
The authors thank the anonymous reviewers for their valuable comments. The research was financially supported by Department of Electronics Engineering Education under Universitas Negeri Yogyakarta and Ministry of Research, Technology, and Higher Education (Indonesia). Authors also thank to Southern Taiwan University of Science and Technology (Taiwan) for setting of robotic manipulators.