Development of A Food Categories and Calories Estimation Full Stack System Based on Multi-CNNs Structures

Kanjanapan Sukvichai[1]* Warayut Muknumporn[2]

[1*,2]*Faculty of Engineering, Kasetsart University, Bangkok, Thailand*

*Corresponding Author. fengkpsc@ku.ac.th, warayut.m@ku.th

Received: 5 December 2019; Revised: 27 April 2020; Accepted: 10 June 2020
Published online: 25 June 2020

Abstract

Humans require different food amounts and nutrition depended on age, gender and health. Amount of food intake can create health problems especially for infants, elderly or diabetics. Tradition nutrition booklet is not suitable for most people since it is hard to understand. Thai-foods are hard to extract the nutrition and most of the Thai dishes are not in the book since it focused on Western dishes. This research focused on development on a full stack AI system that categorizes Thai-food dishes, classifies and localizes the ingredients in each dish and estimate nutrition and calories. Multi-Convolutional Neural Networks (CNNs) are used to achieve these categorize, classify and localize tasks. The designed system is separated into AI backend and Mobile application frontend based on OpenCV in an Android smartphone. MobileNet is used as a food categorizer while You-Only-Look-Once (YOLO) network works as the ingredient's classifier and localizer. Then, ingredients in the pictures are cropped and passed through traditional image processing algorithm with predetermined parameters to calculate and transformed pixel into real-dimension area referenced by Thai coins. Pixel area of non-uniform shape ingredients are segmented and the nutrition and calories can be estimated via a standard reference lookup table. Full stack system is developed in this research based on RESTful protocol with JSON format that used to communicate between a smartphone and AI server. The designed CNNs and full stack system are trained, tested, verified and deployed then the food image captured from a smartphone application can be used to estimated nutrition and calories. Finally, useful information is display on a smartphone screen.

Keywords: Convolutional Neural Networks, YOLO, image processing, full stack, RESTful.

I. INTRODUCTION

Human health is directly related to the food that they consume. The phrase "You Are What You Ate" reflexes the importance of food to human lives. In this present day, Thai people suffer from overweight and diseases caused by food consumption such as arthritis, diabetes, and heart disease. It easily happens to humans who ate too much food or bad ingredients. This situation will lead to economic problems because the demand for medical care will increase. Therefore, the food nutrition information is always important in order to make human has a good health and prevent diseases especially for elderly who need to focus on the calories and the diversity of nutrients in each meal. According to Wardle, Parmenter and Waller [1], the relationship between nutrition knowledge and food intake had been investigated from 750 men and 750 women. The result of this research shows that people who have more knowledge in nutrition, they tend to have 25 times better-eating behavior on a healthy diet than others who have low knowledge. The easy way to educate people about nutrition knowledge is by using the foods they ate every day as examples. This approach is also can be applied to Thai people especially for the young generation since they are more concerned about the healthy diet than the previous generation [2]. The normal way to estimate nutrition in each food dish is by looking from a nutrition guild booklet or asking the nutritionists. Nutrition guild booklet is the simplest way to estimate nutrition but it is not easy to read, understand and calculate and not all Thai food dishes were included in the booklet. Nutritionists can provide information about food nutrition and correct amount intake for every age but patients cannot give precise information about their meals to nutritionists thus the estimation is not correct. Moreover, most people believe that going to talk with nutritionists is a waste of time. Thanks to the development of the image processing techniques and Convolutional Neural Networks (CNNs), the food categories and ingredients nutrition and calories can be determined.

The aim of this research is to develop an easy way for people to get information about nutrition and calories in Thai fast food dishes. The food categorization and nutrition estimation system are developed to extract food information from Thai fast food dish images such as Krapaw rice, Thai omelet rice and Kuey-teow or Thai noodle soup. In this research, a full stack AI system is developed and separated into two parts which are an AI server worked as a backend and application (App) in a mobile phone worked as a frontend. These parts exchange their information using RESTFul protocol. First, food image is taken by a mobile phone App based on OpenCV library and sent that image to AI server. Food picture is received by AI server and fed into a MobileNet Convolutional Neural Networks (CNNs) to categorize the food dishes without extract any nutrition. Then, the food image is fed to the second CNNs to extract the ingredients classes and locations by using the specific YOLO CNNs network corresponding to the food categories obtained from the MobileNet. Finally, useful food information is calculated by using image processing approaches. CNNs are easily to be implemented to the variety of hardware because of its flexibility. It can also achieve good performance for pattern recognition and localization application. New food categories networks can simply be added into the existing networks without alter the whole networks structure because they are independent. Finally, the AI server sends result and estimated nutrition and calories to display on the mobile phone screen.

II. CONVOLUTIONAL NEURAL NETWORKS

Convolutional Neural Networks (CNNs) is a popular tool for modern image recognition application since it is easy to be used by assigning an input image data set with output set to CNNs network and let it learns until the error is small enough. This approach achieves by using deep-learning techniques. The problem of patterns, objects or faces recognition can be solved by using CNNs. There are success network developed in the past few years such as GoogLeNet [3], Inception V3[4] and AlexNet[5]. MobileNets, the CNNs that select in this research as the food categorize network, is one of the most popular networks developed by Google that used in many researches because this network can be implemented into mobile applications [3] easily without consumes a large amount of calculation power and resources. Unlike other networks, the image classification using MobileNets is fast and the network weights variables are smaller than others which is suitable for limited resources system such as a server without any Graphic Processing Unit (GPU). MobileNets have a fast classification performance and have a small network size, but with a trade-off on its advantages, the accuracy of this network is less than other networks. The main difference between the MobileNet architecture and other traditional CNNs is instead of a single 3x3 convolutional layer followed by the batch norm and ReLU activation function, MobileNets split the convolution into a 3x3 depth-wise convolution and a 1x1 point-wise convolution yield a better calculation speed. MobileNets is run based on TensorFlow [6]. TensorFlow is an open-source library released by Google to build and design deep learning models. Although MobileNets doesn't provide the highest accuracy but it is good enough to categorize the food dishes in order to be used to select suitable correspondence ingredients extractor networks in the second layer CNNs in this multi-CNNs structure. There are six food categories focused on this research which are Krapaw rice, Mooping (Thai grilled pork on a stick), Thai stewed pork rice, Chicken rice, Roast duck noodle and Kuey-teow because these foods are the most popular fast food dishes consumed by Thai people. Example of Thai foods are shown in Figure 1.

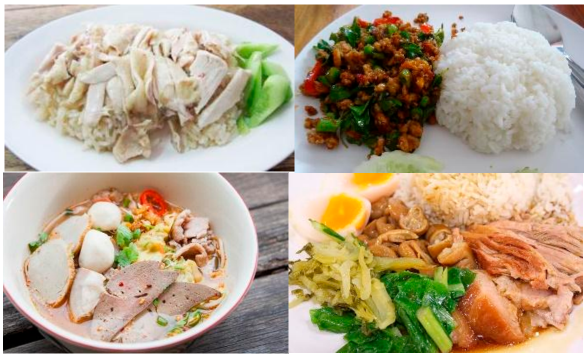


Figure 1 Popular Thai food dishes

Second Convolutional Neural Networks used in the second layer of multi-CNNs structure is the You Only Look Once or YOLO [7] networks. This network is used as the ingredients extractor and localizer which is used to extract ingredients specifically to the food category obtained from the food categorizer network in the first layer with ingredient localization. YOLO network is very similar to region proposal classification networks or RCNNs which perform detection on various region proposals and also performing prediction multiple times for various regions in an image in order to perform location and classification on an image. YOLO uses a single CNN network for both classification and localizing the object using bounding boxes therefore YOLO is faster than other RCNNs or Fast-RCNNs. In this research, YOLO v3 tiny network is used because it is light weight, fast calculation and easy to be taught in a short period of time. YOLO v3 tiny network architecture [8] is a combination of 24 layers consisted of convolution, max pooling, full-connected, up-sample and decision layers. This network is considered as a shallow network because it has only around 3 million parameters. YOLO is originally implemented in C language using Darknet platform. Thus, in this research, YOLO will be wrapped by Python language using CDLL library in order to make it easier to be combined with the server program which based on python language. There are many YOLO networks used in this research because ingredients in each type of Thai food dish are differences and the network must distinguish those ingredients separately. Moreover, YOLO networks are required for each category since many ingredients in Thai food dishes share the same properties such as shape and size but different in nutrition and calories for example a hard-boiled egg in Kuey-teow, chicken rice, and stewed pork rice. Thus, one YOLO network must be used for one food category. Thanks to this approach, the network can be shrunk into a small network for one dish and the designed structure is easy to add new food dish categories into the network without altering existing structure since all YOLO networks are independent. For example, the YOLO network for Kuey-teow will focus on determine the location and size of the porkball and shrimpball as shown in Figure 2 while YOLO network for stewed pork rice will focus on stewed pork and hard-boiled egg.

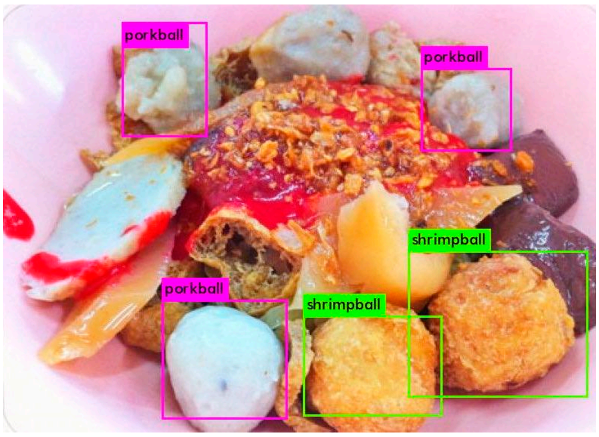


Figure 2 Pork and shrimp ball from Kuey-teow network

The overall multi-CNNs structure in this research is displayed in Figure 3. The focused ingredients for each dish are displayed in Table 1. The total classes of ingredients for each Thai food dish are differences after food categories and ingredients are recognized, localized and classified by using CNNs via MobileNet CNNs and YOLO CNNs, then, the ingredients are cropped and passed through series of image processing algorithms in order to determine areas of the interested ingredients for estimating calories and nutrition. For example, a stewed pork rice bowl has two main ingredients which are stewed pork and vegetable. The shapeless ingredient can be segmented from the food dish image by using 4 image processing steps. First, the food image is blurred using a box blur filter. Then, the food image's color space is transformed from RGB into HSV color space.

Table 1 Food Ingredients for each Thai dish

Food Dish	Ingredients	
	details	classes
Krapaw rice	Basil Pork, Fried Egg	2
Mooping	Pork	1
Roast Duck noodle soup	Wonton, Grilled Duck, Crispy pork	3
Stewed pork rice	Thai strews pork, Boiled Egg	2
Chicken rice	Boiled chicken, Fried chicken, Boiled Pork Blood, Boiled Pork Liver	4
Kuey-teow	Shrimp ball, Pork ball, Fish ball, Boiled pork blood, Fried tofu, Wonton fish, Chinese roll fish	7

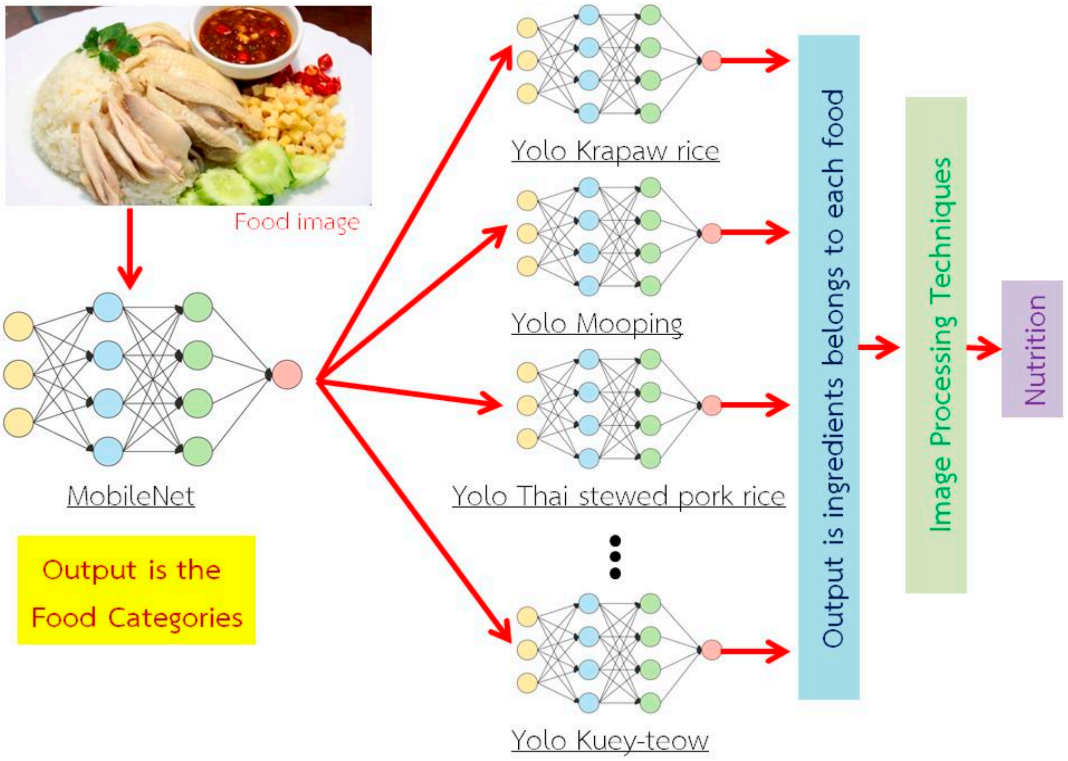

Food image
MobileNet
Output is the Food Categories
Yolo Krapaw rice
Yolo Mooping
Yolo Thai stewed pork rice
Yolo Kuey-teow
Output is ingredients belongs to each food
Image Processing Techniques
Nutrition

Figure 3 Overall multi-CNNs structure

Finally, dilation and erosion techniques are used to filter noise and get a better segmented result. In the stewed pork rice case, the vegetable and stewed pork can be segmented out from the food image. The predetermined color segmentation parameters are specifically defined for particular ingredients from specific Thai dish which are selected correctly based on corresponding YOLO networks for each Thai dish category. The output from the segmented process is shown in Figure 4. In order to transform the pixel area into the real dimension area, the ground truth must be provided. In this research, Thai coins are selected as the standard ground truth because it has a fixed size and it is easy to find and detect. Moreover, most people carry coins around while they order food. YOLO network detects 1, 5 and 10 baths coins. The coins detection result is shown in Figure 5.

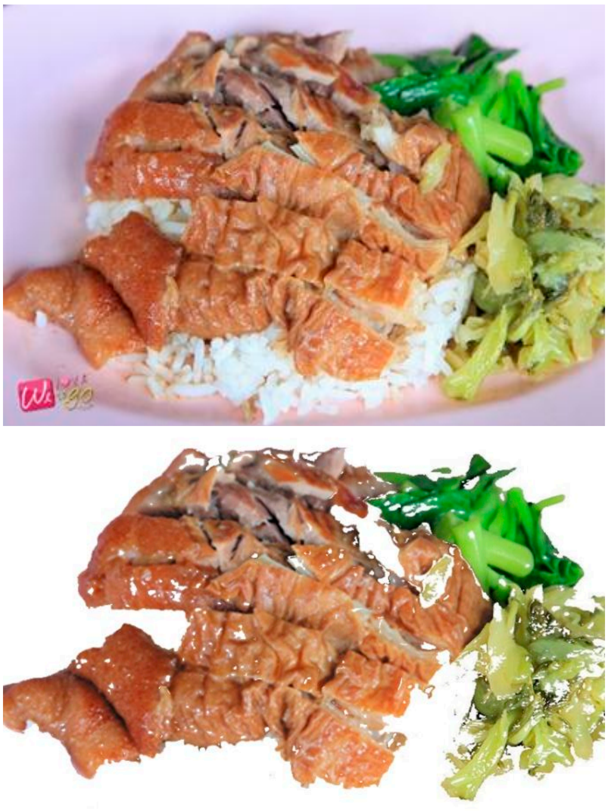


Figure 4 Stewed pork rice dish (top) and segmented ingredients (bottom)

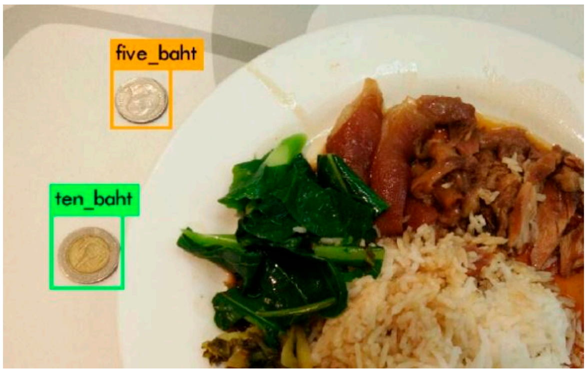


Figure 5 Coin detection using YOLO CNNs network

Thai coins are found and categorized but the area of the coin is in the unit of pixel and it is required to transform into the area in matric unit. Normally, coins in the picture are not perpendicular to the camera then they have eclipse shapes. Therefore, it will be simply transformed into a circular shape by using projection relation as shown in Figure 6. The radius of the circular coin can be estimated by determining the half of the longest line between the left and right points of the eclipse coin image as shown in Figure 7. Finally, the area in pixel square unit can simply calculate by $A = \pi r^2$ then compared with real coin dimension in meter square unit to obtain ration value for translate pixel area to matric area unit.

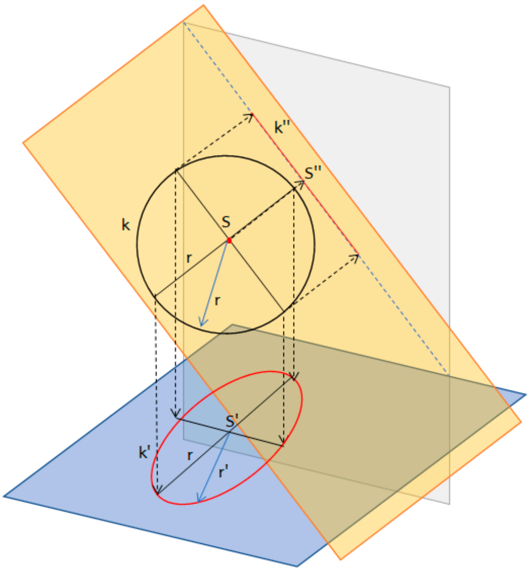


Figure 6 Relationship between circular and eclipse coin shape

III. FULL STACK SYSTEM

This topic explains a designed full stack system for the food categories and calories estimation system based on RESTful protocol. The designed system is separated into frontend and backend parts. App in mobile phone works as the frontend that has a duty to capture food image with Thai coin via smartphone camera and sent it via the internet to AI server and display the result from AI server for users. Backend is an AI server that receives food images from frontend and process image using Multi-CNNs structures and image processing in order to obtain ingredients and calories of the interested food image. The small phone that is used in this research is Samsung A9 with the Android operating system. The App is developed by using Android studio with OpenCV library.

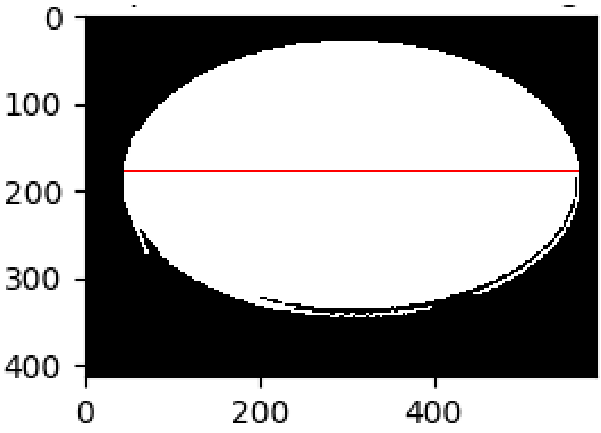


Figure 7 Diameter of eclipse coin shape

AI server consists of an i-5 CPU gen 4, 8 GBytes of DDR4 ram, 1 Tbytes of HDD and GT1030 with 2 GBytes of DDR5 ram graphic card. Server runs on Ubuntu 16.04 LTS environment with cuDNN 7.5 and CUDA 9.0 libraries. Flask is used to generate RESTful server and written in Python language. Mobile Application (App) in smartphone communicates with an AI server using JavaScript Object Notation or JSON which is a lightweight data-interchange format. Food image is capture and coded into raw format in order to be packed into JSON image package since JSON is not support the RGB image format. POST method is used to transfer JSON image package to the AI server. AI server gets the JSON image package and unpack into the original RGB image format. The food image is now ready to be processed by CNNs and image processing algorithms. Thai coins are localized, categorized, separated and selected in order to transform into real dimension by calculation. Then, calories and nutrition can be estimated using a lookup table from nutritionists. Finally, the result is sent back to App in smartphone. The system flow is shown in Figure 8.

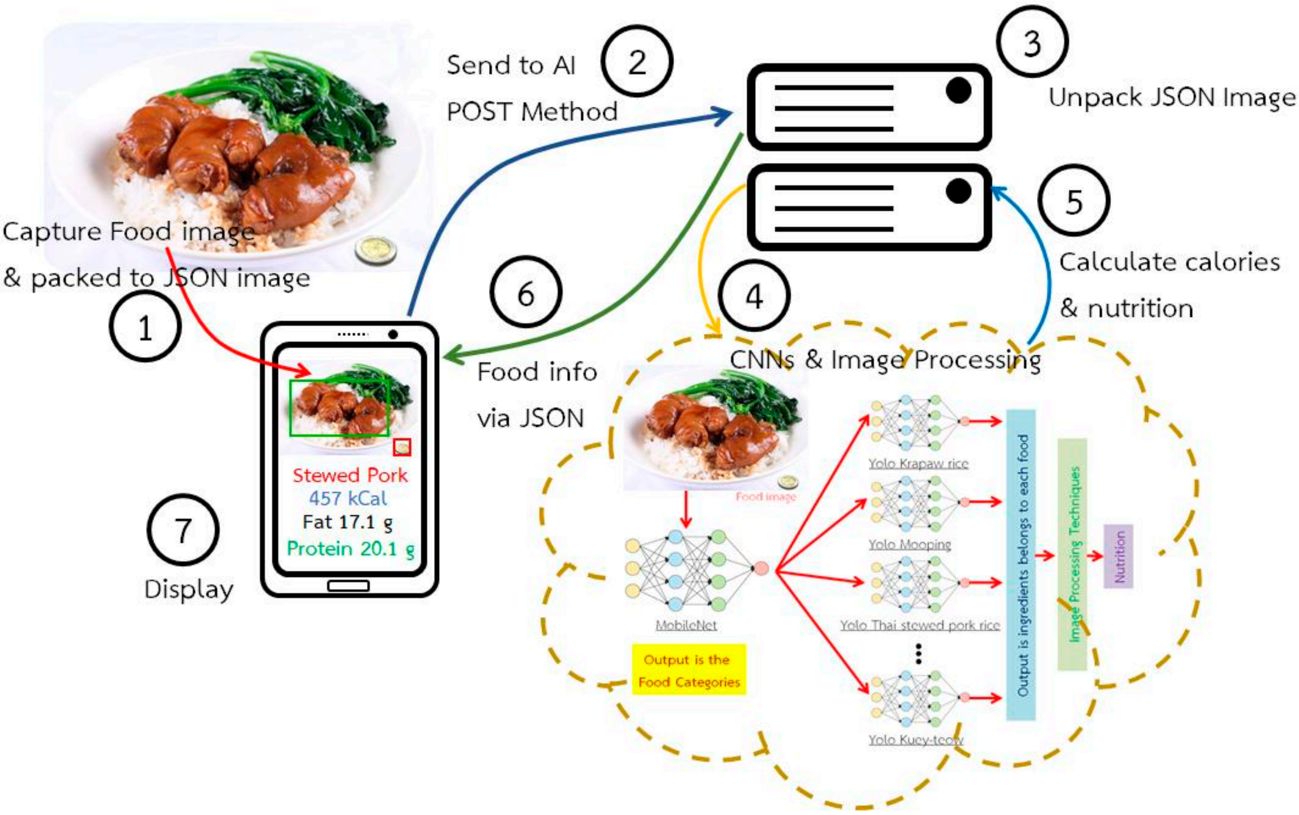


Figure 8 Flow of the designed full stack AI system

Figure 9 Example of prepared training data set

IV. RESULTS AND DISCUSSION

The food pictures were prepared for both food categorization and ingredients localization in the experiments. Food images are gathered from the internet and also taken by researchers. Five hundred pictures of each Thai fast food dish were prepared as a training data set for MobileNet network without labeling as shown in Figure 9. MobileNet learns on these food pictures and adapted its weights in every layer by using a back-propagation algorithm on the desired loss function. Then the trained network is validated by test images. The error of MobileNet happens when the network predicts the wrong food category.

Food images data set are trained for 10,000 iterations for MobileNet CNNs in AI server with the help of GPU from Nvidia graphic card. MobileNet network requires 5 hours to train on the dataset for food categorization. Each food dish is tested by 100 test images and the result is shown in Table 2. The average output accuracy is around 92.83% which calculated from the ratio between the number of wrong and total test images.

Table 2 Result from MobileNet for each Thai dish

Thai Dish	Test Images	Correct	Wrong
Krapaw rice	100	96	4
Mooping	100	99	1
Roast Duck noodle soup	100	92	8
Stewed pork rice	100	89	11
Chicken rice	100	85	15
Kuey-teow	100	96	4

Next, six YOLO networks are trained separately according to corresponding food categories. In this step, interested ingredients of each food pictures are labeled and converted the labeled bounding boxes into YOLO boxes format using YOLO labeling software as shown in Figure 10. Then, each YOLO is trained separately. The training as done on the server with GPU speed up. Each YOLO CNNs requires 4 days to learn from the dataset before the loss function value down to 0.2 and the multiple ingredients can be identified and localized with good accuracy. Roast Duck noodle soup and Kuey-teow dish ingredient classification and localization results are shown in Figure 11 and Figure 12 respectively. Each food YOLO trained network is validated and tested by using 100 test images and the result shown in Table 3.

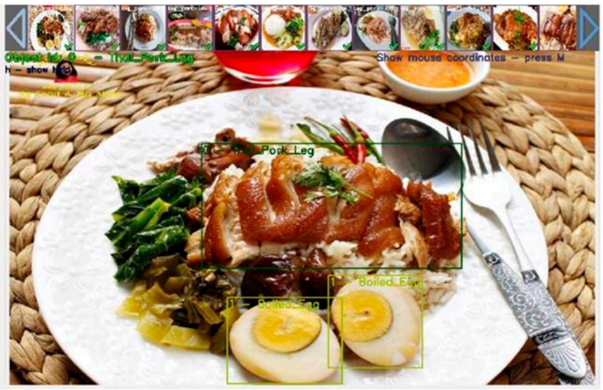


Figure 10 YOLO labeling software

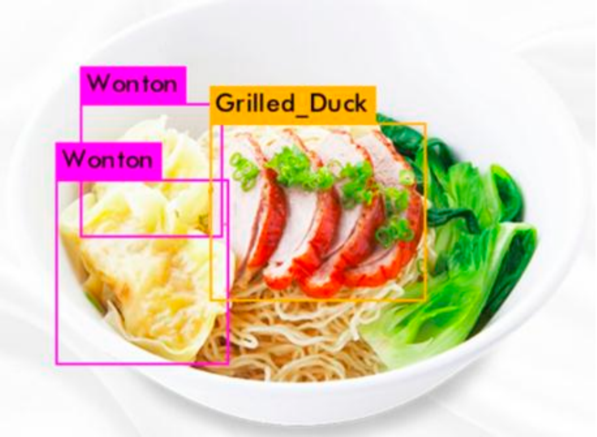


Figure 11 Result from a Roast Duck noodle soup YOLO

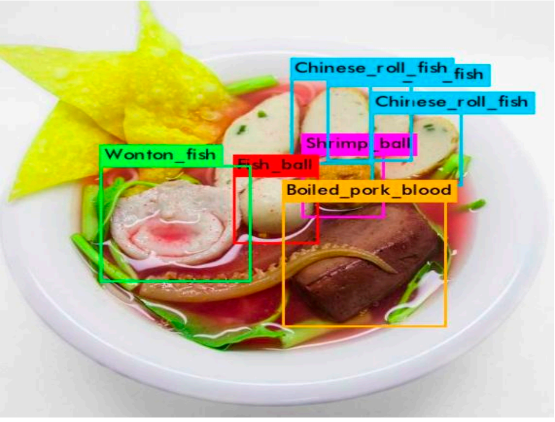


Figure 12 Result from a Kuey-teow YOLO

There are three errors in this experiment which are the wrong food categories (MobileNet), detected with wrong ingredient (YOLO) and missing ingredients (YOLO). Therefore, accuracy in Table 3 reflects overall accuracy since it combined both error from the designed multi-CNNS structures (MobileNet and YOLOs). The average accuracy of the multi-CNNs structures is 72%. The wrong and missing ingredient error is shown in Figure 13.

Table 3 Overall result from Multi-CNNS structures

Thai Dish	Test Images	Correct	Wrong
Krapaw rice	100	81	19
Mooping	100	80	20
Roast Duck noodle soup	100	85	15
Stewed pork rice	100	89	11
Chicken rice	100	81	19
Kuey-teow	100	88	12

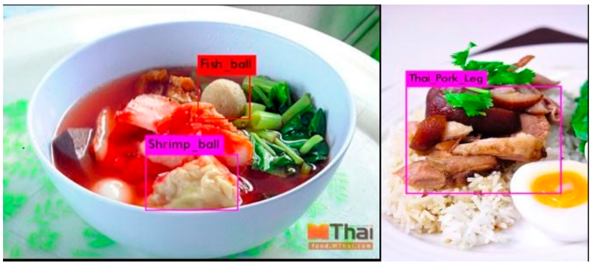


Figure 13 Wrong output from YOLO networks

After ingredients are identified and localized, each ingredient is cropped and processed through a series of image processing techniques with filter and color thresholding predetermined parameters. Finally, the non-uniform shape ingredients are extracted from the rest and the calories and nutrition are estimated. The example of segmented shapeless ingredients is shown in Figure 14. Standard Thai coin is used as the real-world reference standard object size. A coin is identified and located along with the food dishes by using coin

YOLO. The area of the coin can easily be calculated by comparing counted coin pixels to the real coin dimension and create the pixel area to square meter area ratio. Then, this ratio is used to calculate the areas of the ingredients. After ingredient areas are estimated, then, the calories can be estimated. There are several kinds of nutrition and calorie lookup booklets in Thailand and most of them give different nutrition and calorie values for the same Thai food dish. Therefore, in this research, Nutritive Values of Thai Foods form the Department of Health Ministry of Public Health is selected as the main referenced lookup table [9].

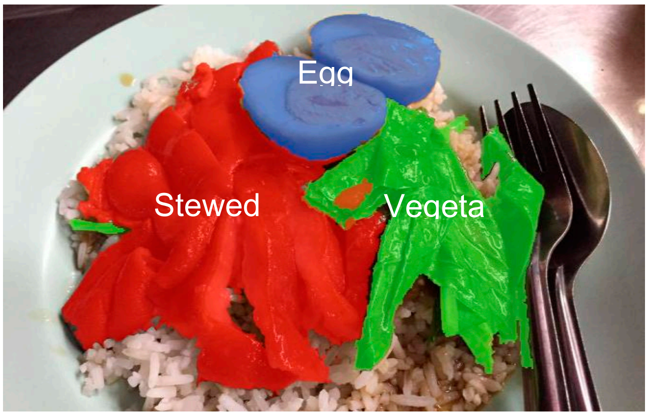
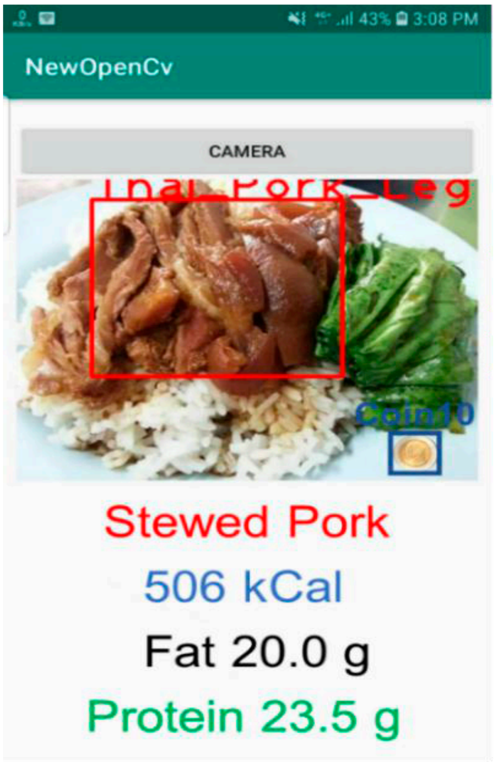


Figure 14 Stewed pork rice dish segmentation

After CNNs and image processing parameters are learned and calibrated then they are transferred into the AI server in order to perform full system experiments. In this step, Application (App) is deployed into Samsung Android smartphone and AI server is setup. The smartphone and AI server are connected by the same router with the same subnet mark. Smartphone is used to capture a stewed pork rice along with 10 bath coin. This food picture is packed into JSON image format and submitted to the server via RESTful POST method. Then, AI server unpacks JSON image format to regular image format and passes to the MobileNet network to categorize food dish. Corresponding YOLO network is used to find and localize interested food ingredients. Non-uniform shape ingredients are processed via image processing with predetermined parameters for each ingredient in specific Thai dish. The coin image region is cropped from the food image and transform into a circular shape. The relationship between coins pixel area and the real coins dimension area is found and used to calculate the area of the non-uniform shape ingredients. Next, the nutrition and calories are estimated via a standard lookup table and all information is packed into JSON format and sent back to App in smartphone. Finally, App displays all information as shown in Figure 15 on the smartphone screen. Calories of the stewed pork rice dishes can be estimated to 506 kcal compared to the standard calories of 456 kcal from the Nutritive Values of Thai Foods booklet. From the experiment, the error happens because of the 2D to 3D estimation technique using Thai coins and also the lookup table. There is the limit of the referenced lookup table since it is no information about the shape, size, weight, and ration of each ingredient shown in the food nutrition table in the Department of Health Ministry of Public Health booklet. Therefore, in order to overcome this limitation, nutrition consultation is required for better estimation. The error of 3D estimation is also can be solved by using multiple view images or videos with advanced 3D point-cloud estimation techniques such as the optical flow algorithm or AI-based approaches. These issues will be the main focus in further research.



Stewed Pork

506 kCal

Fat 20.0 g

Protein 23.5 g

Figure 15 Output from full stack system displayed on the mobile screen.

V. CONCLUSION

Food is one of the most important things for human life but traditional nutrition estimation via comparison booklet has limitations and hard to understand for normal people. Nutrition and calories in Thai food are important especially for the elderly or diabetic persons. This research proposed the easy way for people to obtain useful information from intake food pictures via the designed full stack system to determine nutrition and calories by using Convolutional Neural Networks (CNNs) and image processing techniques. In this research, full stack system is constructed by the AI server and mobile application (App) in a smartphone. RESTful protocol and JSON format are used for communication and data transfer between the App and server. AI server consists of multi-CNNs structure and image processing layers. Mobile Net network is selected as a food categorization network and YOLO v3 tiny network is selected as the ingredient's classifier and localizer. Thai fast food dishes images are used to train the networks in server environment and used in full-stack system. Image processing parameters are predetermined specifically for individual Thai food categories. The designed networks can categorize food and identify and localize ingredients with accuracy of 72%. Networks computation time is not fast but good enough thanks to GPU from Nvidia graphic card. The application is designed and deployed into the Android smartphone. Application captures food image, packs and sends JSON image package to the server via POST method. The estimated nutrition and calories then send back from the AI server and displayed on the smartphone screen. There are errors that happen in this research caused by the technical complexity and limitation of referenced nutrition lookup table. The best way to profit users such as elderly and/or diabetes from this research is to display a range of nutrition instead of a number. For example, a stewed pork dish output will show 356kCal – 643kCal, fat about 18g – 20 g and 20 g – 30 g of protein on the screen then users can self-calculate the maximum intake per day. Other problems will be considered in future research. The main focus of the next step is an exploration in the Mask R-CNNs [10], YOLACT [11], and custom 3D AI-based regression neural network in order to improve the accuracy of the estimated nutrition and calories.

REFERENCES

[1] J. Wardle, K. Parmenter, and J. Waller, "Nutrition knowledge and food intake," *Appetite*, vol. 34, no. 3, pp. 269–275, Jun. 2000, doi: 10.1006/appe.1999.0311.

[2] T. Waratornpaibul, "Consumption behavior: consumerism food and health-conscious food," *Panyapiwat Journal*, vol. 5 no. 2, pp. 255-264, Oct. 2015.

[3] C. Szegedy *et al.*, "Going deeper with convolutions," in *2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, Oct. 2015, pp. 1-9.

[4] X. Xia, C. Xu, and B. Nan, "Inception-v3 for flower classification," in *2017 2nd International Conf. on Image, Vision and Computing (ICIVC)*, Chengdu, Jun. 2017, pp. 783-787.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. 25th International Conf. on Neural Information Processing Systems (NIPS'12)*, Tahoe, NV, USA, 2012, pp. 1097-1105.

[6] M. Abadi, *et al*., "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX conference on Operating Systems Design and Implementation (OSDI'16)*, Savannah, GA, USA, Nov. 2016, pp. 265–283.

[7] R. Joseph, D. Santosh, G. Ross, and F. Ali., "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 779-788.

[8] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," Apr. 2018. [Online]. Available: arXiv:1804.02767.

[9] S. Boonvisut, *Nutritive values of Thai foods*, Bangkok, Thailand: Nutrition Division, the Department of Health Ministry of Public Health (in Thai), 2001.

[10] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *2017 IEEE International Conf. on Computer Vision (ICCV),* Venice, Italy, Oct. 2017, pp. 2980-2988.

[11] D. Bolya, C. Zhou, F. Xiao, and Y. J. Lee, "YOLACT: Real-time instance segmentation," Oct. 2019. [Online]. Available: arXiv:1904.02689.