Mixed Dish Recognition through Multi-Label Learning

Yunan Wang
Jilin University
Changchun, Jilin, China
yunan17@mails.jlu.edu.cn

Jing-jing Chen
National University of Singapore
Singapore
chenjingjing.tju@gmail.com

Chong-Wah Ngo
City University of Hong Kong
Hong Kong, China
cscwngo@gapps.cityu.edu.hk

Tat-Seng Chua
National University of Singapore
Singapore
dcscts@nus.edu.sg

Wanli Zuo
Jilin University
Changchun, Jilin, China
zuowl@jlu.edu.cn

Zhaoyan Ming
National University of Singapore
Singapore
dcsming@nus.edu.sg

ABSTRACT

Mix dish recognition, whose goal is to identify each of the dish type presented on one plate, is generally regarded as a difficult problem. The major challenge of this problem is that different dishes presented in one plate may overlap with each other and there may be no clear boundaries among them. Therefore, labeling the bounding box of each dish type is difficult and not necessarily leading to good results. This paper studies the problem from the perspective of multi-label learning. Specially, we propose to perform dish recognition on region level with multiple granularities. For experimental purpose, we collect two mix dish datasets: mixed economic rice and economic beehoon. The experimental results on these two datasets demonstrate the effectiveness of the proposed region-level multi-label learning methods.

CCS CONCEPTS

• **Computer systems organization** → **Embedded systems**; *Redundancy*; Robotics; • **Networks** → Network reliability.

KEYWORDS

Mix dish recognition, Multi-label recogniition, region-wise, multi-scale

ACM Reference Format:
Yunan Wang, Jing-jing Chen, Chong-Wah Ngo, Tat-Seng Chua, Wanli Zuo, and Zhaoyan Ming. 2019. Mixed Dish Recognition through Multi-Label Learning. In *11th Workshop on Multimedia for Cooking and Eating Activities (CEA'19), June 10, 2019, Ottawa, ON, Canada.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3326458.3326929

1 INTRODUCTION

Automatic dietary recording service is becoming increasingly important today with the attractive concept of "health lifestyle". With automatic dietary assessment tools, people can easily track their food log and get the nutrition analysis. Therefore, food recognition, which is the key technology for such kind of automatic dietary assessment tool, has become a hot research topic in recent years. Recent efforts on food recognition mostly devoted to recognize food images that only contain one type of dish [7] [9], yet problem of multiple food recognition, especially the situation where different types of dish presented in one plate, has been less studied.

While various technologies and attempts are booming, one thing that cannot be ignored is the recognition of mix dish. It's not practical to study just individual dishes and ignore set meals, and in many cases, such as in canteens and food courts, several dishes are placed on one same plate especially for Chinese food stalls. This kind of mix dish is very popular and widely distributed, and asking users to take several separate photos of each dish in the plate and uploading them one by one will definitely kill their thin patience soon. Study of mix dish can help us get nutrition or other information for set meals without repetitive heavy labor. One photo of all dishes is enough for analysis. Till now, many scholars have studied food recognition. With the development of various smart applications, simple usage of deep learning methods such as CNN for identifying single-dish photos has been difficult to meet actual needs. In this sense, multi-dish classification reduces the burden on users. Faster R-CNN, proposed by Ross Girshick et al. [16] in 2014 which detects objects as well as their bounding boxes, can be used as a very effective classifier. In addition, segmentation algorithms which can get accurate pixel-wise prediction area of each object is also helpful for food-ness problem such as calorie estimation, considering that calorie of a dish has a significant proportional relationship with its amount. Unfortunately, the above methods has either limited effect or high cost on the mix dish problem where all dishes are on the same plate.

In our case, the shape of each dish is irregular. Overlap among dishes is quite common, and borders of dishes are not clear. Therefore, square bounding boxes may not accurately frame the position of each dish during process of both annotation and prediction. For segmentation, overlap and boundary mixing may be a problem, and it has higher labeling cost and more complicated algorithm. We want to solve the problem with a simple and effective solution, reducing unnecessary manual labeling and excessive running time. Therefore, we study this problem from the perspective of multi-label classification.

Advantages of treating mix dish problem as multi-label learning problem are obvious: it requires neither bounding box annotation nor pixel-wise annotation, and aims at only categories, rather than

CEA'19, June 10, 2019, Ottawa, ON, Canada

ACM ISBN 978-1-4503-6779-0/19/06…$15.00
https://doi.org/10.1145/3326458.3326929

the location of each dish, which will greatly reduce the burden of annotation work and simplify the design of the network. Conversely, problems that can be brought about are equally obvious: it is already very difficult to correctly identify each one of the dishes in a food image, not to mention that we only use image-level annotation. Without help of detailed labeling information, the supervision information available during training is greatly weakened, and mixed margins of dishes will make recognition more difficult. How to get acceptable and competitive classification results for such a troublesome recognition problem under weak supervision, this is an important issue that we need to consider.

To this end, we propose a multi-label learning framework that performs dish recognition at region level under different scales. Region-level recognition enables to detect dish locations for better recognition performance while multi-scale recognition enables to handle the variations in dish size. Besides, to transfer the knowledge learned from single dish image, we initialize the weights of convolutional layers with the DCNN model trained on a large single food image dataset. Compared to detection and segmentation schemes, our approach has lower cost, less processing time and potentially better results, and compared to other multi-label classification methods, our method achieves much higher accuracy. The contribution of this work can be summarized as follows:

- We study the problem of mix dish recognition from the perspective of multi-label learning and propose a framework that recognize the dish at region level with multiple granularities.
- We collect two challenging mix dish datasets and provide them with image-level labels. To the best of our knowledge, they are the only datasets for mix dish.
- We verify the proposed framework on two datasets.

2 RELATED WORK

Food recognition has attracted lots of research interest in recent few years. Existing efforts include deep-based recognition [7] [23] [22] that leverage different deep models for food recognition, context based recognition by GPS and restaurant menus [3] [17] [2], personalized food recognition by history data [19], multiple-food recognition [1] [14] [24], multi-modal fusion [18] and real-time recognition [20] [26]. This section mainly reviews previous works on multiple food recognition as well as multi-label learning in food domain.

Compared with single food recognition, multiple food recognition receives fewer research attentions. Early works on multiple food recognition mainly follow a two-step pip-line [24], which performs plate detection with circle detector or deformable part models (DPM) [15] followed by feature fusion based food recognition method on the detected plate regions. Recent works are mostly based on deep models [5], such as YOLO [27] or faster-RCNN [28] for dish detection and recognition [13] [14]. For example, in [13], Ege el al. proposed to leverage faster-RCNN to obtain the candidate bounding box of the dish, then apply multi-task learning on the candidates for simultaneously dish recognition and calories estimation. Later, Ege el al. [14] proposed a framework that leverage YOLO for simultaneously bounding box detection, food recognition and calorie estimation. To further boost the performance of multiple dish recognition, Aguilar et al. [1] proposed to combine semantic segmentation model with YOLO for dish detection and recognition. In this work, semantic segmentation is applied to segment food and non-food regions, to refine the dish detection results from YOLO. Since both semantic segmentation and YOLO model are trained in fully supervised fashion, this work requires both bounding-box-level and pixel-level labels. Different to all the works mentioned above, Shimoda et al. [29] proposed to generate foodness proposals with a fully convolutional neural network for multiple dish recognition. Compared to Faster-RCNN based or YOLO based food detection methods, this work does not require any bounding box labels. Nevertheless, the aforementioned approaches were proposed under the assumption that different dishes are in different containers and each plate only contains one type of dish, which is different from the situation we consider in this work.

There are also a few works that aim to recognize multiple dish items which are presented on one plate. These works are mostly based on semantic segmentation, whose goal is to assign the dish label to each pixel [25] [12]. For example, Myers et al. [25] proposed to use the CNN model and conditional random field (CRF) to predict pixel-level labels for multiple dish segmentation. In [12], Dehais et al. proposed a CNN-based food border map to guide the region growing for food segmentation. Both methods have shown quite promising semantic segmentation results on western food, where each food item mostly contains one single ingredient. Since we are dealing with a more challenging situation where each food item may composite of multiple ingredients, these methods are not directly applicable to our problem.

Multi-label learning has also been studied in food domain in recent years [4] [7] [8]. Nevertheless, most of these works are focused on ingredient recognition. For example, Marc Bolanos et al. [4] propose a deep multi-ingredients recognition method which uses Inception-V3 and ResNet-50 as basic deep architectures. For both deep models, the last layer is modified to apply multi-label classification over N possible outputs to predict the list of ingredients in a food image. In [7], Chen et al. also study ingredients recognition problem through multi-label learning by proposing a deep multi-task learning model to simultaneously recognize food categories as well as their ingredients. In addition, conditional random field (CRF) are utilized to incorporate the co-occurrence context information to refine the ingredient recognition performance. In [8], a multi-task learning model is proposed to recognize ingredient, cooking and cutting attributes of a food picture. They divided the *Pool5* feature correspond to the last convolution layer into $m \times m$ grids and applied region-level dependency pooling on these grids. Meanwhile, instead of fixed resolution of grids, multi-scale recognition is used to handle the change in scale. Compared to the previous two papers, this paper uses a very efficient region-wise method which significantly improves the results. Different to the aforementioned works that focus on ingredient recognition, this paper studies the problem of mix dish recognition with multi-label learning.

2.1 Methodology

Figure 1 presents an overview of the proposed framework. Given an image I, a pyramid of multi-resolution images is generated and input to a deep convolutional network. The corresponding feature

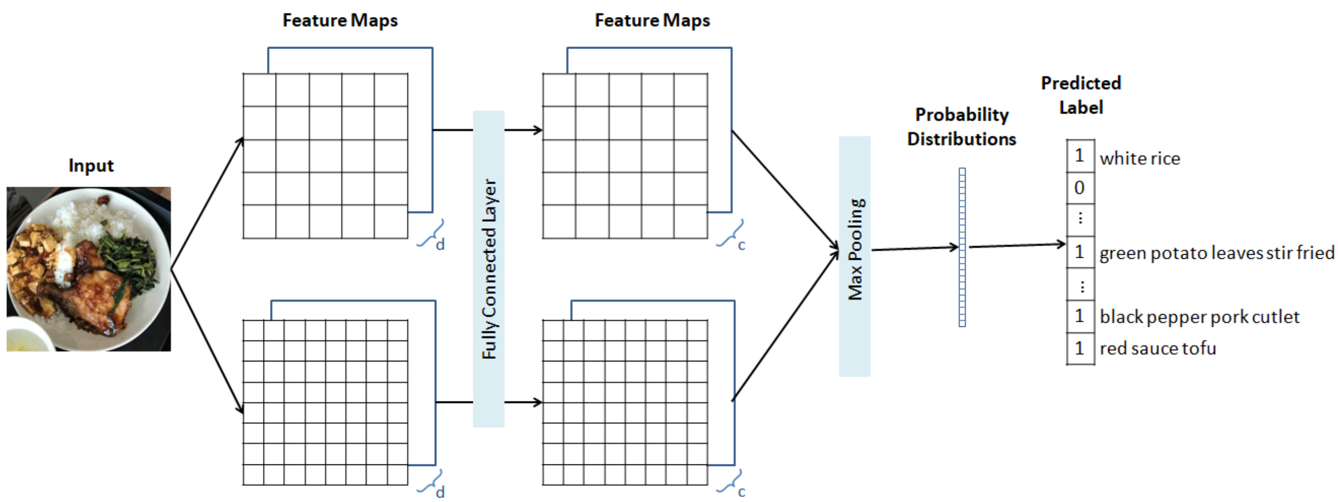


Figure 1: Framework overview. d is the dimension of embedding feature, and c is the number of classes. Given an image, multi-scale region-wise classification is performed first, then max pooling is done across regions and scales to get the global probability distribution.

maps are transformed into embedded features for the prediction of dishes. Max-pooling is done across different regions and scales to get the global probability distribution.

2.2 Region-wise classification

Our DCNN architecture uses the Inception-V4 [30] network. We obtain the feature map from 3× Inception-C layers, which retain the spatial information of the original image. The obtained feature map is divided into $n \times n$ grids, where each grid is presented by a vector of 1536 dimensions. The value of n varies depending on the image size. For an image of size 299×299, $n = 8$ and each grid corresponds to a receptive field of 37×37 resolution. The classification is performed on each region with the assumption that each grid contains part of one dish among all categories. In this way, we make sufficient use of the regional information of a food image. For each grid, a shared fully connected layer is applied. Denote the feature vector for i^{th} grid as f_i, we have

$$v_i = tanh(W_I f_i + b_I) \tag{1}$$

where $v_i \in \mathbb{R}^c$ is a vector of c dimensions corresponding to the i^{th} region $f_i \in \mathbb{R}^d$. c is the number of dish categories. Therefore, v_i can also be considered as the response vector of dish categories. $W_I \in \mathbb{R}^{c \times d}$ is the learnt transformation matrix and $b_I \in \mathbb{R}^c$ is the bias term. Then these response vectors will be max pooled to get the global response vector V.

$$V = max\left\{v_i|_{i=1}^{n^2}\right\} \tag{2}$$

where $V \in \mathbb{R}^c$ is a vector whose dimensions are the same as number of dish categories.

2.3 Multi-scale classification

When dealing with food pictures, one thing to note is that the angle and distance of the shot, as well as the size of the area occupied by each dish in the picture, are not fixed. Therefore, choosing fixed-size grids as basic units of recognition may affect the results to some extent. Based on this consideration, we introduce the multi-scale food recognition method. It provides us with multi resolution of the image, combining different granularity of features to predominantly enhance model efficiency.

Our model is easy to extend to multi-scale recognition. Instead of single-scale, pyramid images in multiple resolutions are put into the network, resulting in regions with different amounts and receptive fields. An image of size 299×299 obtains 8×8 grids after feature embedding layer, while an image of size 598×598 obtains 17×17 grids. All these regions experience the same operation as above, the only difference is that the vectors obtained by different scales will be put together for max pooling. The process is as follows:

$$v_{s,i} = tanh(W_I f_{s,i} + b_I) \tag{3}$$

Where $v_{s,i} \in \mathbb{R}^c$ is a vector of c dimensions corresponding to the i^{th} small region $f_{s,i} \in \mathbb{R}^d$ with scale s, $W_I \in \mathbb{R}^{c \times d}$ is the learned transformation matrix and $b_I \in \mathbb{R}^c$ is the bias term.

$$V = max\left\{max\left\{v_{s,i}|_{i=1}^{n^2}\right\}|_{s=1}^{S}\right\} \tag{4}$$

where $V \in \mathbb{R}^c$ is a vector whose dimensions are the same as number of categories, n^2 is the number of divided grids, s is a certain scale, and S is the number of scales.

2.4 Loss Function

Binary Cross Entropy Loss. As mix dish recognition is a multi-label classification problem, we hence use binary cross entropy as the loss function. As the value of V is in the range of $[-1, 1]$, hence sigmoid is applied to transform the response into category probabilities.

$$P = \frac{1}{1 + e^{-V}}, \tag{5}$$

Table 1: Statistics of food datasets. Ecominc rice and Economic beehoon are the mixed dish datasets collected by this work. (* UEC food-100 contains both single and multiple dish images, and the number of multiple image is 1,027.)

Dataset	#image	#class	Multiple food	Mix dish	#dishes/image
PFID [9]	4,545	61	×	×	-
Chinese Food Dataset [10]	5,000	50	×	×	-
VIREO Food 172 [7]	100,241	172	×	×	-
UEC Food-256 [21]	31,397	256	×	×	-
Food-101 [6]	101,000	101	×	×	-
UNIMIB2016 [11]	1,027	73	✓	×	-
School Lunch Image Dataset [13]	3,940	21	✓	×	-
UEC Food-100 * [24]	9,060	100	✓	×	-
Economic Rice	9,254	164	✓	✓	4.07 ± 0.59
Economic Beehoon	2,851	54	✓	✓	3.71 ± 1.71

where $P \in \mathbb{R}^c$ is the learned possibility distribution whose dimensions are the same as number of categories. Then binary cross entropy loss is calculated as follows:

$$L = -\sum_{i=1}^{c}(g_i log(p_i) + (1 - g_i)log(1 - p_i)), \qquad (6)$$

where p_i is the predicted probability for i^{th} dish while $g_i \in \{0, 1\}$ is the ground-truth label. During the training process, the error will propagate through the whole network, and weights of the network will be updated to optimize the recognition performance.

Negative Sampling As each food image only contains a small number or dishes out of the available c dish categories, the ground-truth vector G is very sparse. So we adopt negative sampling during the training process. Denote $R \in \mathbb{R}^c$ as the randomly generated binary vector. The binary mask vector $M \in \mathbb{R}^c$ is obtained as follows:

$$M = R \mid g, \qquad (7)$$

where | is the or operation, used to make sure that all the positive samples are selected for loss calculation, and g is the binary ground-truth label vector. With negative sampling, the loss function can be rewrite as follows:

$$L_{NS} = \frac{-\sum_{i=1}^{c}(g_i log(M_i p_i) + (1 - g_i)log(1 - M_i p_i))}{\sum_{i=1}^{c} M_i} \qquad (8)$$

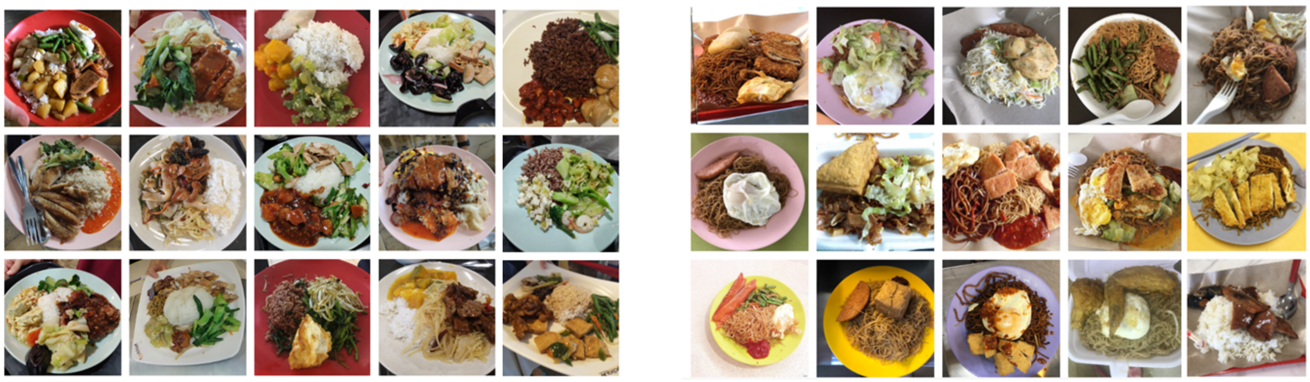
where L_{NS} is the binary cross entropy with negative sampling, p_i is the predicted probability for i^{th} dish, $M_i \in \{0, 1\}$ is the binary mask and $g_i \in \{0, 1\}$ is the ground-truth label.

3 DATASET

There are several public food datasets, including PFID [9], Chinese Food Dataset [10], VIREO Food-172 [7], UEC Food-100 [24], UEC Food-256 [21], Food-101 [6], UNIMIB2016 [11] and School lunch image dataset [13]. Table 1 summarizes the statistics of the above mentioned datasets. Basically, most of these food datasets are collected for single dish recognition. Exceptions include UEC Food-100, School lunch image dataset and UNIMIB2016. Nevertheless, all there three datasets are relatively small dataset, ranging from 1,027 to 3,940 multiple-item food images that covers less than 100 dish categories. Besides, in these three datasets, different dishes are presented in different plates which is the simplest situation for multiple food recognition. Different to these datasets, we collect two mix dish datasets: Economic Rice and Econonmic Beehoon, which contains 9,254 and 2,851 food images respectively. Both datasets contain 4 dishes per image on average. We followed the principles listed below at the time of shooting: (1) All the dishes on the plate should appear in the photo so that each dish can be seen and recognized. (2) The camera-to-plate distance should not be too far, and the plate should occupy at least two-thirds area of the entire photo. (3) Angle changes should be made instead of shooting all the plates from directly above.

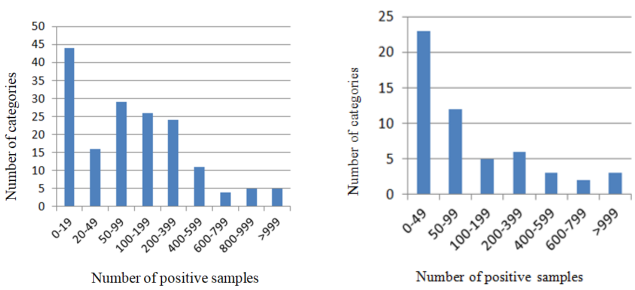
Economic rice. Economic rice is one of the most common food type in southeast Asia and most popular lunch/dinner choice for general public in Singapore since it is cheap and can be served quickly. To this end, we collect a economic rice dataset where food images are captured by cell phones from 6 different canteens. At the same time, we also collect the list of dish names among the canteens to instruct the labeling process and hire 7 students for labeling work. In total, 9,254 food images in 164 dish categories are collected, and for each image, only the dish names are labeled. Figure 2(a) shows several examples of Economic rice. As can be seen, this dataset is quit challenging as different dishes may mix with each other and there may be no clear boundaries between two dishes. Besides, even the same dish cooked by different canteen may have large visual variance because of different cooking/cutting methods, which brings certain challenges to the recognition. Figure 3(a) further shows the distribution of positive examples in dish categories. On average, there are 225 positive samples per dish category.

Economic beehoon. To demonstrate the effectiveness of our method more convincingly, we further collect the economic beehoon dataset. Economic beehoon is a popular food type for breakfast in Singapore and is also a type of mix dish that usually combines several dish categories on one plate, however, the number of dish categories are much less compared to economic rice. Figure 2(b) shows several examples of the dataset. Basically, the visual appearance of dish is quite standard in beehoon dataset and there is less visual variance among dishes in the same categories. Therefore, the recognition of economic beehoon is less challenging compared to economic rice. In total, we collect 2,851 images in 54 dish categories of economic beehoon from different hawker centers, and



(a) Economic Rice (b) Economic beehoon

Figure 2: Sample images of the collected datasets.



(a) Economic Rice (b) Economic beehoon

Figure 3: Sample distribution.

dish names of each images are labeled similar to economic rice dataset. Figure 3(b) shows the distribution of positive examples in dish categories. On average, there are 191 positive samples per dish category.

4 EXPERIMENT

4.1 Experimental Setting

Our Inception-V4 network is pre-trained on a large single dish dataset, which covers 264,048 Singapore food images from 751 categories. The top-1 recognition accuracy on this dataset is 77%. Through pre-training, we hope to transfer the knowledge learned from single dish for mix-dish recognition. On both mixed dish datasets, 70% of images are picked for training, 10% for validation and the remaining 20% for testing. For training, RMSprop [31] is chosen as the optimizer with learning rate set to 0.01, and batch size is set to 32. The model is trained around 20 epochs. For multi-scale recognition, we used two-level of pyramid images, respectively at resolutions of 598×598 and 299×299, also, due to memory limitations, batch size is set to 8 here. Considering number of categories and average dishes per image, the final prediction result is obtained by P in formula (5) retaining the probability values of top-4 with length of 164 and 54 respectively. A multi-hot predicted label is obtained by setting reserved values to '1' and other values to '0'. Due to the sparsity of ground truth, the sampling rate of negative sampling is set to 0.1, that is, 10% of negative samples are randomly selected for training. Note that we set 4 as the number of predicted labels because average number of dishes per image in Economic Rice dataset is very close to 4 with a small variance. As for Economic Beehoon dataset, the calculated mean number of dishes is 4, and our experiment also proved that the selection of 4 is better than 3 or 5 considering overall effect.

4.2 Recognition performance

We first study the effects of negative sampling as well as pre-training. Table 2(a) and Table 2(b) list the performances of economic rice and economic beehoon recognition, respectively. Basically, mix dish recognition is a challenging task as the recall, precision and F1 score are very low on both datasets. From the results, we have following observations. First, pre-training on single dish dataset improves the performance of mix dish recognition, which demonstrates that the knowledge learned from single dish is also useful for mix dish recognition. In terms of F1, pre-training improves 9% on economic rice and 5% on economic beehoon. Second, negative sampling is also effective in improving the mix dish recognition performance. With negative sampling, the recognition performances has gained more than 13% of improvement on economic rice and more than 6% of improvement on economic beehoon. Therefore, we adopt both negative sampling and pre-training in the following experiments as both of them have been demostrated to be effective in improving the mixed dish recognition performance.

Next, we evaluate the performances of region-wise mix dish recognition as well as multi-scale recognition. The difference between image-level method and region-wise method is illustrated in Figure 4. As can been seen, region-wise recognition performs recognition on each region of the image while image-wise recognition pools the feature maps as a vector and perform recognition on the pooled vector. Table 3 summarizes the performances. Basically, region-wise recognition performs much better than image-level recognition. It improves the recognition performance around 18% on economic rice dataset and 9% on economic beehoon dataset in

Table 2: Comparison between the model with (*) and without pre-training on single dish dataset, as well as with and without NS (Negative sampling).

(a) Economic Rice

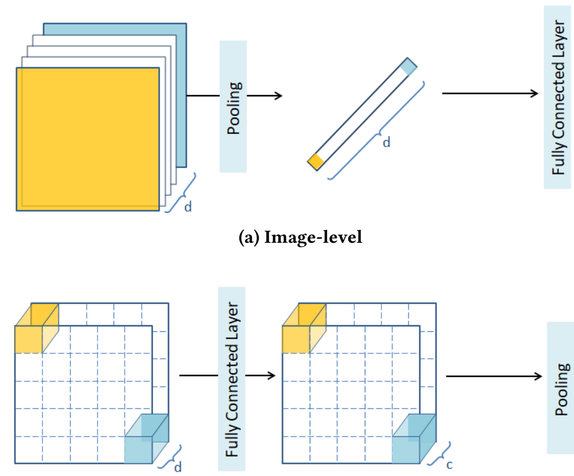
	Recall	Precision	F1
Inception-V4	0.254	0.293	0.271
Inception-V4*	0.362	0.364	0.360
Inception-V4 + NS	0.447	0.447	0.434
Inception-V4* + NS	**0.504**	**0.502**	**0.498**

(b) Economic Beehoon

	Recall	Precision	F1
Inception-V4	0.456	0.507	0.459
Inception-V4*	0.566	0.502	0.508
Inception-V4 + NS	0.602	0.532	0.541
Inception-V4* + NS	**0.645**	**0.561**	**0.571**

Table 3: Mix dish recognition comparison: image-level versus region-wise; single-scale versus multi-scale.

	Economic rice			Economic beehoon		
	Recall	Prec.	F1	Recall	Prec.	F1
Image-level	0.504	0.502	0.498	0.645	0.561	0.571
Region-wise	0.681	0.680	0.675	0.748	0.646	0.662
Multi-scale	**0.719**	**0.721**	**0.714**	**0.776**	**0.685**	**0.697**



(a) Image-level

(b) Region-wise

Figure 4: Comparison of image-level and region-level. d is the dimension of embedding feature, and c is the number of classes. For image-level method, classification is performed on the global image feature vector that obtained by average pooling operations on feature maps, while region-wise method performs dish classification on the feature vector of each region and max pools the results across regions to get the global probability distribution.

terms of F1 measure. By considering multi-scale recognition, the F1 score can be as high as 0.71 and 0.70 on economic rice and economic beehoon datasets, respectively. The results demonstrate that region-wise multi-scale recognition is effective in improving the mix dish recognition performances. In addition, from the results, the improvement gained from region-wise multi-scale recognition on economic beehoon dataset is much less than that on economic rice dataset. This is probably due to the fact that dishes in economic beehoon images are not as mixed as dishes in economic rice and there are still clear boundaries among them.

To get deep insights on how region-wise multi-scale model improves the mix dish recognition performance, we visualize the top-4 predictions for three examples, which is shown in Figure 5. As shown in the figure, the image-level recognition model predicts "white rice" or "beehoon" with the highest probability for all three examples, because these two types of dish are most common in economic rice and beehoon dataset, which is a manifestation of data imbalance that means certain label(s) are extremely frequent among all labels and may have an impact on prediction results. As we can see, in the third example, the model wrongly predicts "kway teow" as "beehoon". This basically indicates that despite the assistance of pre-trained model and negative sampling, the image-level recognition model is easy to be affected by the unbalanced data and has a certain tendency to randomly guess the predictions.

Results of region-wise and multi-scale are much better. For the third example, both of them correctly predict "kway teow" with the highest probability, and the latter even does not take "beehoon" in the top-4 predictions. Region-wise method divides the feature map into multiple grids and uses the same classifier on each small region which intuitively contains a small piece of a dish, thus, it can capture more information than directly processing the whole feature map. And multi-scale approach further enhances the model with generation of pyramid images instead of single-scale ones. This multi-granular approach helps handling images with different angles and camera-to-dish distances more flexibly. As can be observed from the third example, with finer resolution regions, the multi-scale recognition model is able to predict "luncheon meat" successfully, while the single-scale region-wise model ignores "luncheon meat" as it occupies a small region in the image. Besides, considering finer resolutions for recognition can better capture the textual information of dish, hence helps to reduce the confusion between the dishes with similar appearances. In the first and second example, with finer resolution regions, the multi-scale recognition model is able to successfully predict "braised chicken", while the single-scale region-wise model confuses it as "stir fired eggplant" or "beancurd skin strips".

The above experimental results and examples demonstrate the effectiveness of region-wise and multi-scale methods. In addition, we also found that although all images have only image-level annotations, the response map $v_{s,i}$ (in Equation 3) obtained by the model can form rough bounding areas of the dishes. As shown in Figure 6, even we don't provide any location information of the dish during the training process, the areas of "beehoon", "fish cake" and "spring

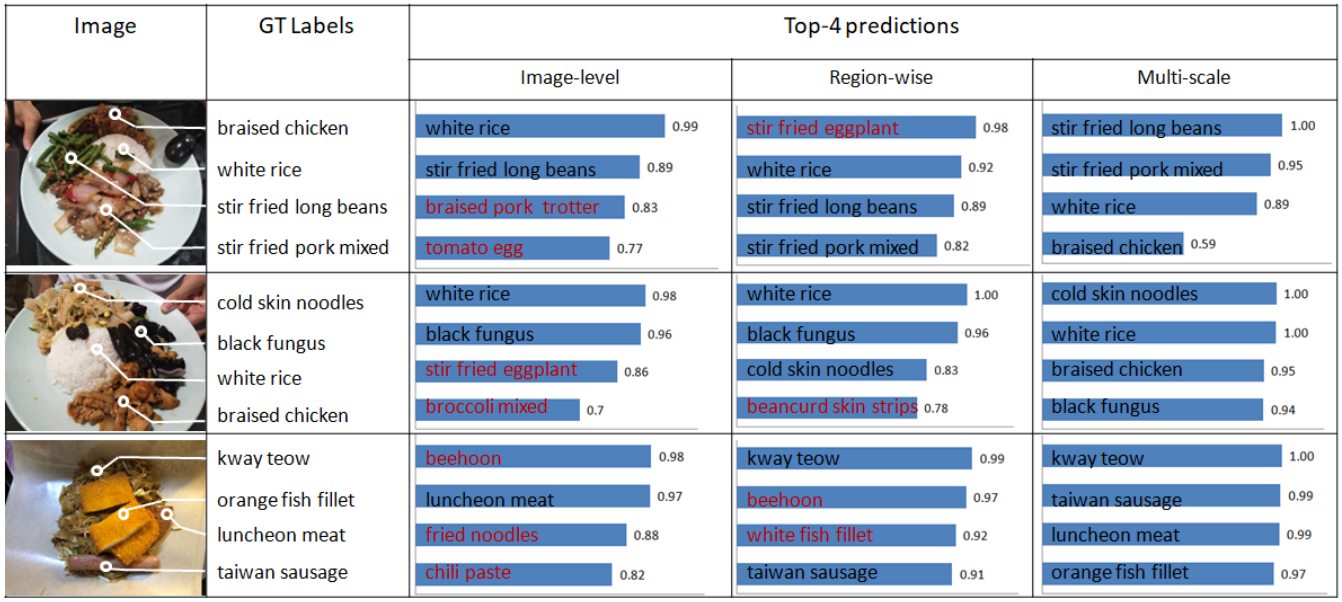


Figure 5: Examples of mix dish prediction. False positives are marked in read.

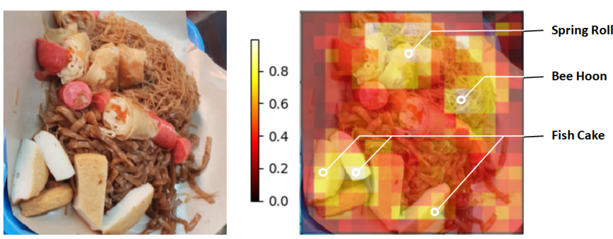


Figure 6: Detection visualization

roll" can still be roughly detected on the response map. This is owing to the advantages of performing classification at regional level. For a given dish category, the regions with higher response will be highlighted with region-wise recognition, which helps to localize the dish. Generally speaking, the better the result of localization is, the higher the prediction accuracy can be achieved.

5 CONCLUSION

In this paper, we studied mix dish recognition problem from the perspective of multi-label learning, and performed dish recognition on region level with multiple scales. Accompanied with Negative Sampling and targeted pre-trained model, we use several simple yet efficient methods to improve performance of the classification model and got competitive results with image-level annotation. For the difficult mix dish problem, our approach eliminates the heavy labor of manual labeling and significantly increased all indicators comparing to plain multi-label classification. We collected two real data sets and experimented on them, yielding convincing results. The experimental results show that the proposed method is very effective.

Future Work The effectiveness of region-wise has been well proven, but the division of grids is still a manual work. If the process of region-wise and choice of multi-scale resolutions can be done automatically according to the characteristics such as camera-to-dish distance of the image, then better results may be achieved. In the future, we plan to explore adaptive pooling to further reduce manual setup work and improve classification accuracy.

ACKNOWLEDGMENTS

This research is part of NExT++ project, supported by National Research Foundation, Prime Minister's Office, Singapore under its IRC@SG Funding Initiative. This work was also partly supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 11203517), and a grant from Nature Science Foundation of Jilin Province, China (No. 20180101330JC), Scientific and Technological Development Program of Jilin Province, China (20190302029GX).

REFERENCES

[1] Eduardo Aguilar, Beatriz Remeseiro, Marc Bolaños, and Petia Radeva. 2018. Grab, Pay, and Eat: Semantic Food Detection for Smart Restaurants. *IEEE Transactions on Multimedia* 20, 12 (2018), 3266–3275.
[2] Oscar Beijbom, Neel Joshi, Dan Morris, Scott Saponas, and Siddharth Khullar. 2015. Menu-match: Restaurant-specific food logging from images. In *2015 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 844–851.
[3] Vinay Bettadapura, Edison Thomaz, Aman Parnami, Gregory D Abowd, and Irfan Essa. 2015. Leveraging context to support automated food recognition in restaurants. In *2015 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 580–587.
[4] Marc Bolaños, Aina Ferrà, and Petia Radeva. 2017. Food ingredients recognition through multi-label learning. In *International Conference on Image Analysis and Processing*. Springer, 394–402.
[5] Marc Bolanos and Petia Radeva. 2016. Simultaneous food localization and recognition. In *2016 23rd International Conference on Pattern Recognition (ICPR)*. IEEE, 3140–3145.
[6] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In *European Conference on*

Computer Vision. Springer, 446–461.

[7] Jingjing Chen and Chong-Wah Ngo. 2016. Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the 24th ACM international conference on Multimedia*. ACM, 32–41.

[8] Jing-jing Chen, Chong-Wah Ngo, and Tat-Seng Chua. 2017. Cross-modal recipe retrieval with rich food attributes. In *Proceedings of the 25th ACM international conference on Multimedia*. ACM, 1771–1779.

[9] Mei Chen, Kapil Dhingra, Wen Wu, Lei Yang, Rahul Sukthankar, and Jie Yang. 2009. PFID: Pittsburgh fast-food image dataset. In *2009 16th IEEE International Conference on Image Processing (ICIP)*. IEEE, 289–292.

[10] Mei-Yun Chen, Yung-Hsiang Yang, Chia-Ju Ho, Shih-Han Wang, Shane-Ming Liu, Eugene Chang, Che-Hua Yeh, and Ming Ouhyoung. 2012. Automatic chinese food identification and quantity estimation. In *SIGGRAPH Asia 2012 Technical Briefs*. ACM, 29.

[11] Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. 2017. Food recognition: a new dataset, experiments, and results. *IEEE journal of biomedical and health informatics* 21, 3 (2017), 588–598.

[12] Joachim Dehais, Marios Anthimopoulos, and Stavroula Mougiakakou. 2016. Food image segmentation for dietary assessment. In *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*. ACM, 23–28.

[13] Takumi Ege and Keiji Yanai. 2017. Estimating food calories for multiple-dish food photos. In *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*. IEEE, 646–651.

[14] Takumi Ege and Keiji Yanai. 2018. Multi-task learning of dish detection and calorie estimation. In *Proceedings of the Joint Workshop on Multimedia for Cooking and Eating Activities and Multimedia Assisted Dietary Management*. ACM, 53–58.

[15] Pedro F Felzenszwalb, Ross B Girshick, and David McAllester. 2010. Cascade object detection with deformable part models. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2241–2248.

[16] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 580–587.

[17] Luis Herranz, Shuqiang Jiang, and Ruihan Xu. 2017. Modeling restaurant context for food recognition. *IEEE Transactions on Multimedia* 19, 2 (2017), 430–440.

[18] Hajime Hoashi, Taichi Joutou, and Keiji Yanai. 2010. Image recognition of 85 food categories by feature fusion. In *2010 IEEE International Symposium on Multimedia*. IEEE, 296–301.

[19] Shota Horiguchi, Sosuke Amano, Makoto Ogawa, and Kiyoharu Aizawa. 2018. Personalized classifier for food image recognition. *IEEE Transactions on Multimedia* 20, 10 (2018), 2836–2848.

[20] Yoshiyuki Kawano and Keiji Yanai. 2013. Real-time mobile food recognition system. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 1–7.

[21] Yoshiyuki Kawano and Keiji Yanai. 2014. Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In *European Conference on Computer Vision*. Springer, 3–17.

[22] Yoshiyuki Kawano and Keiji Yanai. 2014. Food image recognition with deep convolutional features. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, 589–593.

[23] Niki Martinel, Gian Luca Foresti, and Christian Micheloni. 2018. Wide-slice residual networks for food recognition. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 567–576.

[24] Yuji Matsuda, Hajime Hoashi, and Keiji Yanai. 2012. Recognition of multiple-food images by detecting candidate regions. In *2012 IEEE International Conference on Multimedia and Expo*. IEEE, 25–30.

[25] Austin Meyers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin P Murphy. 2015. Im2Calories: towards an automated mobile vision food diary. In *Proceedings of the IEEE International Conference on Computer Vision*. 1233–1241.

[26] Zhao-Yan Ming, Jingjing Chen, Yu Cao, Ciarán Forde, Chong-Wah Ngo, and Tat Seng Chua. 2018. Food Photo Recognition for Dietary Tracking: System and Experiment. In *International Conference on Multimedia Modeling*. Springer, 129–141.

[27] Joseph Redmon and Ali Farhadi. 2017. YOLO9000: better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7263–7271.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*. 91–99.

[29] Wataru Shimoda and Keiji Yanai. 2016. Foodness proposal for multiple food detection by training of single food images. In *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*. ACM, 13–21.

[30] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[31] Tijmen Tieleman and Geoffrey Hinton. 2012. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning* 4, 2 (2012), 26–31.