This paper was presented in a forum without peer-review.

DEIM2021 C13-3

Stage-Aware Recognition Method for Foodstuffs Changing in Appearance in Different Cooking Stages on Chinese Recipe

Yixin ZHANG†, Yoko YAMAKATA††, and Keishi TAJIMA†

† Graduate School of Informatics, Kyoto University
36-1 Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501 Japan
†† Graduate School of Information Science and Technology, The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8654, Japan
E-mail: †zhangyx@dl.soc.i.kyoto-u.ac.jp, ††yamakata@mi.u-tokyo.ac.jp, †††tajima@i.kyoto-u.ac.jp

Abstract In this research, we aim to recognize foodstuffs in recipe instructional images, which change their appearance as they are processed during cooking. Good recognition of instructional images could enhance the illustrative relationship between images and text, therefore improve the quality of recipes and help make understanding of recipes less difficult. It is also necessary for machines to automatically understand the recipe data. The contributions of this research are as follows. (1) We construct a dataset of Chinese recipes consisting of 12,548 recipes with 136,209 steps and 136,209 instructional images in total. In this recipe dataset, each procedural step is illustrated by an instructional image of the action in this procedure. (2) We propose a method for recognizing foodstuff with changing appearance in the recipe instructional images. The general object recognition methods, which assume that the appearance of an object is constant, does not perform well for such food. The appearance of food changes not randomly, but depending on the progress of the cooking process with direction. The same ingredients always show similar tendency of change. Therefore, we propose a stage-aware recognition method to train separated datasets according to the progress of the procedure in order to improve the accuracy of image recognition for foodstuffs. This methods contains two parts: Food Subset-Based Recognition Model Selector and Food Curriculum Learning in Stages. The method proposed in this paper improves the accuracy recognition compared with the baseline model.

Key words Recipe Data, cooking, multimedia, dataset, food recognition

1 Introduction

In recent years, many user-submitted recipe sites, such as Allrecipes(注1) in North America and UK, Cookpad(注2) in Japan, and Haodou(注3) in China, have become popular. Nowadays millions of recipe data posted by users are shared on those recipe sites. Each recipe on these sites consists of the title of recipe, outline, material and ingredient lists, cooking procedures, and tips.

The cooking procedure part in a recipe data is usually structured well into a list of procedural steps. In addition, most of recipe data are multi-modal, and includes both text and images. Figure 1 shows an example of such recipe data on a recipe site Haodou. On this site, a cooking procedure is clearly organized into a list of procedural steps, and each procedural step is associated with an instructional image. This kind of recipe data contain rich and valuable multimedia information. As a result, research on extracting and analyzing useful information in recipe data has become an important research issue in both natural language processing and computer vision fields [1, 2].

The method of analyzing multimedia data is useful in numerous recipe retrieval applications or scenarios and provides important implications in natural language processing and image recognition in general. Recently, some studies such as [2–5] investigated the automatic understanding of recipe data. They focus on analysis of recipe images with captions, whole dish recognition, and recipe text semantic analysis. However, those studies barely takes the problem of food which changes its appearance into account. This problem has several challenges and provides more research opportunities:

- Object recognition in images is a useful method to recognize food or other kind of objects. It is useful for those appearance-stable objects such as tool entities. However, the state and shapes of food is always changing during cooking

(注1)：https://www.allrecipes.com/
(注2)：http://cookpad.com/
(注3)：http://www.haodou.com/recipe/

Figure 1 Example of recipe data posted on Haodou

procedures. Existing image recognition method barely take the problem of object changing in appearance into account. Therefore it is a good opportunity to focus on solving the appearance-unstable object recognition problem.

- When users upload recipes onto the recipe websites, partly due to users' word habits, text descriptions of procedural steps often have typos and infrequent expressions in them, and even omit some important information such as tool or food entities. When we use smart devices (e.g., smart speakers) for automatic reading of recipes and cooking support, it is difficult for a machine to understand such incomplete descriptions. However, those food entities omitted in text descriptions are sometimes shown in the attached instructional images. It is a great challenge to enrich those incomplete textual information using text-image pair information.

Problem Statement: A good recognition of instructional images could enhance the illustrative relationship between images and text. There are two problems we need to solve in this paper: one is that the appearance of the same food differs when the stage differs; and the other is that the further forward the cooking process goes, the more diverse the appearance of food becomes. Existing image recognition method barely take the problem of object changing in appearance into account. Therefore, we are going to study the problem of recognizing appearance-changing foodstuff in recipe instructional images using the feature of instructional images in different stages.

The main contributions of this paper can be summarized as follows: First, since the recipe dataset with text-image procedures is rare in any language, we construct our own text-image instructional recipe dataset in Chinese, which provides a wealth of data and language options for future research. Second, since the shapes and states of food in different cooking stages could be very different, we propose a novel stage-aware image recognition method to identify food in different stages in order to improve the recognition accuracy.

The rest of this paper is organized as follows. Related work is reviewed in Section 2. Section 3 contains the detail information of our dataset. Section 4 explains the details of our proposed methods. After that, experiments and evaluation results are shown in Section 5. Section 6 summarizes this paper and discuss the future work.

2 Related Work

Generally speaking, information supplementation in multi-modal recipe descriptions using features of procedural images is a practical research topic. It is related to the following research lines.

2.1 Recipe Text Processing

There has been research on recipe text processing. Recipe text processing has some differences from general text processing, which makes it difficult to apply the existing text processing method easily to recipe data [6]. A specific method for the recipe domain, which could even be applied for the multilingual environment, is desired.

The analysis of a set of words attached to images is also a research issue nowadays, such as tag importance estimation from a tag set attached with an image [7] and inferring the semantic relationship between them [8]. They focus on the data from image posting sites like Flickr, where the images are attached with tags by the posting user. In this paper, we use the text description along with the attached instructional images and try to infer the relationship between images and text.

2.2 Recipe Image Recognition

Currently, research on the recognition of images in cooking recipes mainly focuses on the whole dish's appearance without explicit analysis of ingredient composition [2]. Ingredient and material estimation only from a completed food image is a task far harder than food categorization. Our method intends to recognize the food which is in changing state during cooking procedures which takes the feature of foodstuffs' appearance changing into account.

2.3 Multi-modal Correlation Learning

Multi-modal Correlation Learning between images and

texts is a hot research issue in computer vision and natural language processing in recent years, such as CCA [9], which study the general cross-modal correlation between images and text information and Bi-linear model [10]. However, multi-modal learning about procedural text-image data in recipes has more unique characteristics compared with general problems. For example, food is changing gradually with procedures. This means that the study of recipes requires a more specific analysis and the effectiveness of multi-modal correlation learning in recipe data can be improved.

It is difficult for general methods to work well on recipe data, which means that the study of recipes requires more specific analysis so that the characteristics of recipes can be better grasped and the accuracy of multi-modal correlation learning in recipe data can be improved.

In summary, recipe text processing and image feature analysis are important issues, and in this paper, we specifically focus on recognizing foodstuffs in recipe instructional images, which changes its appearance as they are processed during cooking, which needs specific consideration and has not been studied in the existing research.

3 Dataset

In this section, we discuss about the recipe dataset we construct.

3.1 Dataset Structure

Our dataset contains 12,548 structured cooking recipes with 136,209 associated images (which is also the number of instructions). The resolution of each image is around 700×500. Figure 2 shows the dataset statistics. We are also continuously adding data into the dataset to reach 1 million cooking recipes.

To analyze procedural text of Chinese recipes, we constructed a Chinese recipe corpus of 50 recipes, and manually annotated recipe named entities (r-NEs) according to guidelines previously defined for Japanese and English [11] and adopt the BERT-NE(注4) (a state-of-the-art named entity recognizer which is constructed based on the BERT neural network architecture [12]) with the model trained by the annotation corpus.

Based on the result obtained from r-NE recognizer [13], in our dataset, we had obtained 5,685 distinct food entities in which 24 food entities appear more than 1,000 times, 298 food entities more than 100 times but less than 1000 times, and 5,355 food entities only 50 times or less, which accounts for 94.20% in total.

In this work, we focus on images of 20 food classes selected from the 24 food entities which appear more than 1,000 times. The other four classes, "water", "flour", "salt" and "milk", were removed because it is hard to recognize their shapes and states. The 20 food classes are as follows:

> Potato, ginger, onion, pork, shrimp, chicken, corn, carrot, eggplant, shallot, tofu, spinach, sauce, chili, bread, dough, fish, egg, cucumber and soybean.

There are 35,401 images in these 20 classes in total for this experiment.

3.2 Image Classification For Food Recognition

Since food appearances are changing during the cooking procedure, images located in different cooking stages of the same food may be very different. The difficulty of food recognition is different for different positions. For example, the potato in the beginning stage is often in its initial round shape, while in the intermediate stage, it may be cut into slices or blocks, and in the finishing stage, it may be plated with other foodstuffs together, which is noisier which makes it harder to recognize. Therefore we classify them into 3 subsets according to their relative position in the whole recipe (beginning, intermediate, and finishing stage) as shown in Figure 3.

Our method is to calculate the 1-Dimension feature for each step. The 1-Dimension feature means the relative position of steps in the whole recipe. For example, the 1-Dimension feature is 0.267 if the step number is 4 and there are 15 steps in total in this recipe. Then we classify the images into 3 subsets (beginning stage, intermediate stage and finishing stage).

3.3 Data Features

One of the key concept this research is based on is that, the food's appearance is changing as they are cut or mixed during cooking. This kind of change is not random, but in accordance with the direction of the cooking process. In other words, food is always in their raw state in the very beginning and their state and shapes start changing during the cooking procedure.

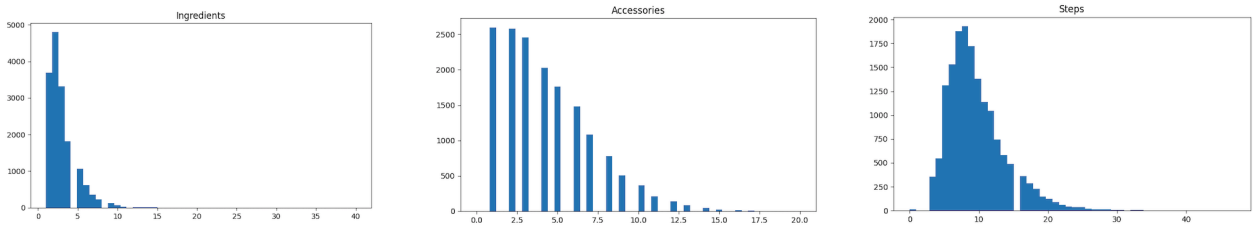
As shown in Figure 4, potatoes in Subset 1 are in their original shape. In Subset 2, they are always cut into slices and put in the wok, while in Subset 3, they are in the dish or along with other food.

3.4 Subsets Division According to the Procedural Stages

We select 20 food classes with high frequencies of occurrence as shown in Table 2 and divide them into three subsets according to their relative positions in recipes, taking into account the balance of the number of images in each subset as shown in Table 1.

In Table 2, from a vertical perspective, the number of images in each subset of each class is relatively average distribu-

(注4)：https://github.com/kyzhouhzau/BERT-NER



Num. of Ingredients Num. of Accessories Num. of Instructions

Figure 2 Dataset Statistics

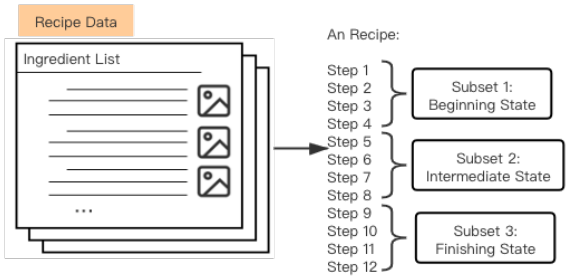


Figure 3 Subsets Division

Table 1 Division of Subsets

	1-D score	Images	Rate
Subset 1	(0, 0.3)	14,135	0.399
Subset 2	[0.3, 0.6]	12,863	0.363
Subset 3	(0.6, 1)	8,403	0.237
Total	1	35401	1

tion. From a horizontal perspective of this table, the image volume of seasonings such as ginger and shallot classes are higher in each subset.

To evaluate the classification method we proposed, we use the PCA method to visualize the image distribution of the same food in different subsets.

We first vectorize images by using a convolutional neural network VGG16 [16], which is widely used for image recognition, trained on ImageNet data. We use the output of two fully-connected layers in VGG16, which is a 4096-dimensional vector. Then we projects the original data into 2 dimensions by using PCA. Figure 5 shows some examples. It is not difficult to find that the images in Subset 1 tend to be more clustered and show stronger similarity, while the images in Subset 3 are more scattered, in other words, they are noisier than images in Subset 1.

Food entities in the images become increasingly difficult to be identified as the step in which they are located increases. Therefore, a method which could take this properties into account may could improve the recognition accuracy. Our method is aim to focus on this problem firstly by classifying them into three subsets (the beginning stage, the intermediate stage, and the finishing stage).

4 Proposed Method

We have briefly explained our methods to recognize foodstuff in images attached to the recipe procedural text whose appearance changes as they are processed during cooking in the first section. In this section, we will explain some details of our proposed methods.

Stage-Aware Recipe Image Recognition For Food: We propose a more specific approach – stage-aware recipe image recognition that considers the different difficulty of food identification problems for different stages to obtain a higher degree of accuracy.

We adopted the image recognition method of ResNet50 [14] to identify food in different stages in cooking procedures. The general object recognition methods in the literature, which assume that the properties of an object are constant, perform well for parsing the information in the picture for many general tasks. For entities like tools whose shapes and states do not change during the cooking process, the general recognition method is enough to obtain a good result. However, as mentioned above, the shape and states of food are constantly changing as being cooked, from being easily recognizable to difficultly recognizable. Therefore, it may be challenging to achieve a high degree of accuracy in food recognition with the model trained with images extracted from all stages.

Our method basically contains two methods.

4.1 Food Subset-Based Recognition Model Selector

It is to divide images into three subsets according to their relative positions. Since the features of the same food in different stages may be very different, we could train models for images in different subsets as training or test dataset in order to figure out the models with higher accuracy. Then we could apply the model with highest accuracy for images in each subset.

4.2 Food Curriculum Learning in Stages

It is based on the idea of Curriculum Learning [15]. Humans and animals learn much better if examples are not presented randomly but are organized in a meaningful order

Beginning Stage Subset | Intermediate Stage Subset | Finishing Stage Subset

Potato

Beginning Stage Subset | Intermediate Stage Subset | Finishing Stage Subset

Ginger

Beginning Stage Subset | Intermediate Stage Subset | Finishing Stage Subset

Shrimp

Figure 4 Subsets Example of Some Classes

Table 2 Division of 20 Classes of Food

	potato	ginger	onion	pork	shrimp	chicken	corn	carrot	eggplant	shallot
Subset1	499	1887	330	461	730	170	806	856	169	1980
Subset2	409	2108	320	231	600	155	507	653	120	2147
Subset3	413	638	175	72	440	108	246	406	108	1725
Total	1321	4633	825	764	1770	433	1559	1915	397	5852
	tofu	spinach	sauce	chili	bread	dough	fish	egg	cucumber	soybean
Subset1	479	221	34	199	618	1574	1081	1690	222	129
Subset2	342	109	114	184	280	2240	900	1143	187	104
Subset3	283	96	220	204	454	1003	824	698	194	96
Total	1104	426	378	587	1352	4817	2805	3531	603	329

that progressively illustrates more and more complex concepts. The basic idea is to start learning the easier sub-tasks and then gradually increasing the difficulty level.

Food will change its state and shape from raw in the beginning stage to being mixed up or being cut into slices in the intermediate or finishing stage, which means from being easily recognizable to difficultly recognizable.

We can know the related positions of sentences from which food entities were extracted, and their attached instructional images. To improve the recognition accuracy, we start by

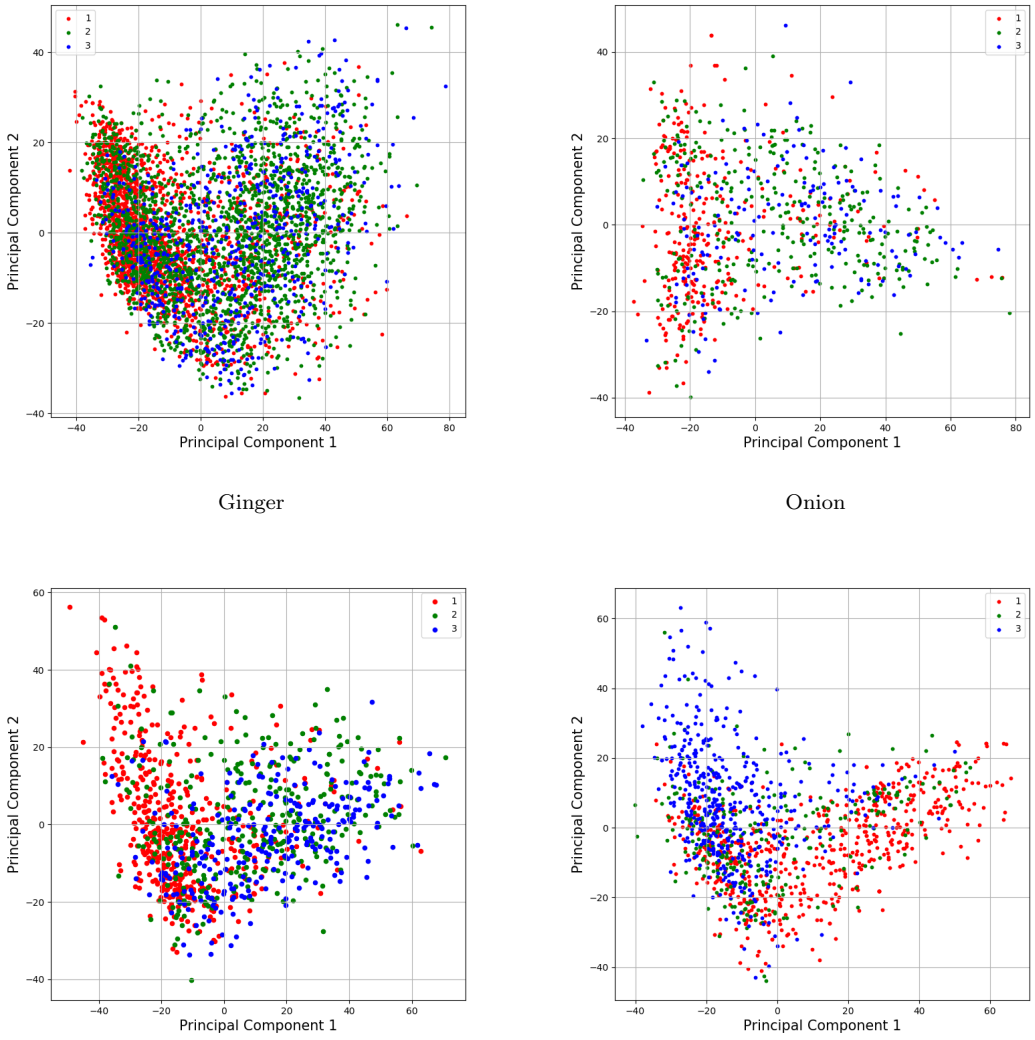


Figure 5 Subsets Distribution of Some Classes

training the model on a subset of the simple data that contains all the classes. Here, it is assumed that the simple data subset (Subset 1) contains more clean images with accurate labels.

Then, we add more and more complex data (data in Subset 2 and 3) during the training process to improve the recognition ability of our model.

The workflow of our two proposed methods is shown in Figure 6.

5 Experiment

In this section, we explain the content of the experiment and then discuss the experiment results to evaluate the proposed method.

The details and evaluation of the data we used in the experiment has been discussed in Section 3. 4.

Our proposed method – **Stage-Aware Image Recognition** mainly contains two parts, as mentioned in Section 4. One is to train the model using images from different subsets as training dataset. The other is based on the idea of Curriculum Learning. Our method takes the different difficulty of food identification problems for different stages into account in order to obtain a higher degree of accuracy.

We conduct experiments to evaluate the performance of the proposed method. We compare various training plans using the ResNet50. The data amount used in each experiment is the same and the ratio of training data to test data is 7:3 for all the experiments. As shown in Figure 6, the proposed models are designed as follows:

- Model 1(1): To train the model directly using the whole training dataset (Our Baseline Model).
- Model 1(2): Food Subset-Based Recognition Model Selector. To train the model directly using each subset as training and test data.

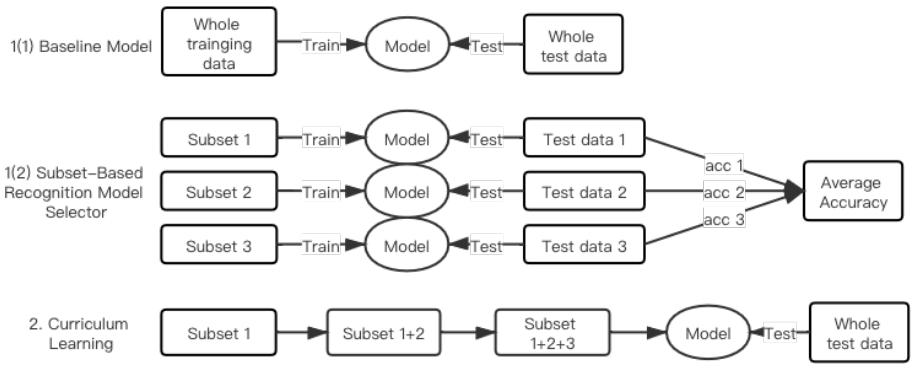


Figure 6 Workflow of the Proposed Method

Table 3 Accuracy of Model 1

Training\Test	Subset1	Subset2	Subset3	All
Subset1	**61.49%**	52.05%	38.22%	47.29%
Subset2	51.90%	**56.65%**	44.24%	43.42%
Subset3	39.49%	46.11%	**50.23%**	44.16%
All	45.36%	47.28%	42.73%	46.41%

- Model 2: Food Curriculum Learning in Stages. To train the model using the proposed method, with a 3-subset curriculum.

(Our Proposed Model metioned in Section 4)

The structure of our experiment is shown in Figure 6.

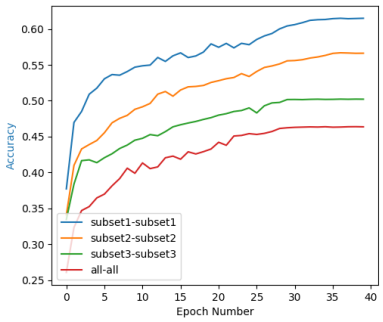
5.1 Food Subset-Based Recognition Model Selector

For Model 1(1) and 1(2), we use the data from the whole dataset and the three subsets as both training data and test data (the size of data for training and test is the same). The accuracy matrix of each training plan is shown in Table 3. We could find that the models which take training data and test data from the same subset could achieve higher accuracy. The accuracy and loss graphs are shown in Figure 7. Then we calculate the average accuracy from the three models with higher accuracy (bolded ones). The average accuracy is 56.12%.

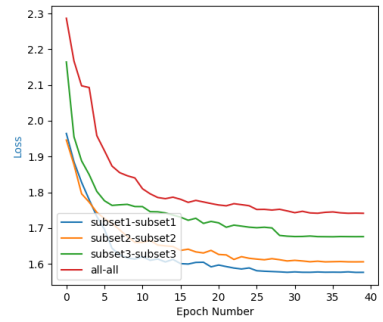
5.2 Food Curriculum Learning in Stages

For Model 2, we define the concept of easy tasks and hard tasks in curriculum learning first. Since images in Subset 1 are the easiest one to be recognized and images in Subset 3 are the hardest recognizable one, we develop a subset-level balance approach: we set the batch size as 256 and epoch number as 40. The training samples are selected in each min-batch as follows: *Stage*1: (256, 0, 0), *Stage*2: (128, 128, 0), *Stage*3: (128, 64, 64) and *Stage*4: (64, 64, 128), respectively. We compare the accuracy of Model 1(1), Model 1(2) and Model 2 as shown in Table 4.

Food appearance is changing during the cooking stages. The further back the cooking process goes, the more diverse the appearance of food becomes. We figure out a way to take this special feature into account in order to improve recognition accuracy. From the comparison result, we could find that both the proposed method (Model 1(2) and Model



(a) Accuracy



(b) Loss

Figure 7 Accuracy and Loss in Model 1

Table 4 Comparison of the Models

Plan		**Accuracy**
Model 1(1)		46.24%
Model 1(2)		56.12%
Training&Test on Subset1 only		61.49%
Curriculum Learning (Model 2)	Test: Subset1	57.34%
	Test: Subset2	56.08%
	Test: Subset3	54.12%
	Test: All	55.20%

2) achieve higher accuracy than the baseline model (Model 1(1)). For images in Subset 1, we could use the model trained using data only from Subset 1 in order to achieve higher accuracy. The food in Subset 1 is usually easy to be recognized. For images in Subset 1 or 3, we could apply the curriculum learning method to achieve a higher accuracy of recognition. In the curriculum learning method, since we gradually add images from Subset 1 to 3 to retain as many features of food as possible, we achieve a better result for the noisier data, such as data in Subset 1 or 3.

6 Conclusion

In this paper, we propose a method of stage-aware recognition for foodstuff that changes its appearance as cooking progress. We first construct a Chinese recipe dataset and analyze recipe structure and entities in recipes. Currently, there are few such recipe datasets that have a one-to-one correspondence between instructional image and text, and we provide a useful one. Besides, our dataset could provide one more choice of language in the recipe dataset. Because there is currently not enough data for our experiments, our dataset is still being added and expected to be made public. Then we proposed two methods of food recognition in images in order to improve the accuracy considering the feature of appearance changing. The recognition of foodstuffs is a difficult problem because they are changing in appearance during being cooked. Since food entities are the main entities in recipes, solving this problem is relevant in recipe research and could help understand and enrich the content of the recipe. Our food subset-based recognition model selector aims to separate images in the beginning, intermediate, and finishing stage (in other words, clean and noisy labels) recognition problem to improve the recognition accuracy, at least for the easily recognizable images. The curriculum learning method improves the recognition accuracy by starting learning the easier sub-tasks and then gradually increasing the difficulty level. It improves recognition accuracy in general, including images in the finishing stage.

In future work, since we now define food images in the beginning stage as the clearest one and images in the finishing stage as the noisiest one (based on the thought of easy recognizable to hard recognizable), we could also define this definition in reverse and compare the results. We focus on using the relative position feature of instructional images, therefore, we could consider the method of adding the relative position feature as an additional feature of image feature vectors when training the model.

Acknowledgements

This work was supported by JST CREST Grant Number JPMJCR16E3, JSPS KAKENHI Grant Number JP18H03245, and JP18K11425, Japan.

References

[1] Wang, Xin, et al. "Recipe recognition with large multimodal food dataset." 2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, 2015.

[2] Chen, Jingjing, and Chong-Wah Ngo. "Deep-based ingredient recognition for cooking recipe retrieval." Proceedings of the 24th ACM international conference on Multimedia. ACM, 2016.

[3] Kawano, Yoshiyuki, and Keiji Yanai. "Foodcam: A real-time food recognition system on a smartphone." Multimedia Tools and Applications 74.14 (2015): 5263-5287.

[4] Sasada, Tetsuro, et al. "Named entity recognizer trainable from partially annotated data." Conference of the Pacific Association for Computational Linguistics. Springer, Singapore, 2015.

[5] Mikolov, Tomas, et al. "Efficient estimation of word representations in vector space." arXiv preprint arXiv:1301.3781 (2013).

[6] Mori, Shinsuke, et al. "Flow Graph Corpus from Recipe Texts." LREC. 2014.

[7] Li, Shangwen, et al. "Measuring and predicting tag importance for image retrieval." IEEE transactions on pattern analysis and machine intelligence 39.12 (2017): 2423-2436.

[8] Katsurai, Marie, Takahiro Ogawa, and Miki Haseyama. "A cross-modal approach for extracting semantic relationships between concepts using tagged images." IEEE Transactions on Multimedia 16.4 (2014): 1059-1074.

[9] Hotelling, Harold. "Relations between two sets of variates." Breakthroughs in statistics. Springer, New York, NY, 1992. 162-190.

[10] Tenenbaum, Joshua B., and William T. Freeman. "Separating style and content with bilinear models." Neural computation 12.6 (2000): 1247-1283.

[11] Yamakata, Yoko, John Carroll, and Shinsuke Mori. "A comparison of cooking recipe named entities between Japanese and English." Proceedings of the 9th Workshop on Multimedia for Cooking and Eating Activities in conjunction with The 2017 International Joint Conference on Artificial Intelligence. ACM, 2017.

[12] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[13] ZHANG, Yixin, Yoko YAMAKATA, and Keishi TAJIMA. "Complementation of Food and Tool Information in Multi-Modal Recipe Procedural Descriptions." Data Engineering and Information Management (DEIM) 2020.

[14] Ren, Shaoqing, et al. "Faster r-cnn: Towards real-time object detection with region proposal networks." Advances in neural information processing systems. 2015.

[15] Bengio, Yoshua, et al. "Curriculum learning." Proceedings of the 26th annual international conference on machine learning. 2009.

[16] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).