PIC2DISH: A Customized Cooking Assistant System

Yongsheng An
Tsinghua University
316991611@qq.com

Yu Cao
National University of Singapore
caoyu.cy@u.nus.edu

Jingjing Chen
City University of Hong Kong
jingjchen9-c@my.city.edu.hk

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

Jia Jia
Tsinghua University
jjia@mail.tsinghua.edu.cn

Huanbo Luan
Tsinghua University
luanhuanbo@gmail.com

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

ABSTRACT

The art of cooking is always fascinating. Nevertheless, reproducing a delicious dish that one has never encountered before is not easy. Even if the name of dish is known and the corresponding recipe could be retrieved, the right ingredients for cooking the dish may not be available due to factors such as geography region or season. Furthermore, knowing how to cut, cook and control timing may be challenging for one whose has no cooking experience. In this paper, an all-around cooking assistant mobile app, named Pic2Dish, is developed to help users who would like to cook a dish but neither know the name of dish nor has cooking skill. Basically, by inputting a picture of the dish and the list of ingredients at hand, Pic2Dish automatically recognizes the dish name and recommends a customized recipe together with video clips to guide user on how to cook the dish. Importantly, the recommended recipe is modified from a retrieved recipe that best matches the given dish, with missing ingredients being replaced with the available ingredients that match dish context and taste. The whole process involves the recognition of dishes with convolutional neural network, classification of key and non-key ingredients, and context analysis of ingredient relationship and their cooking/cutting methods. The user studies, which recruit real users to cook dishes by using Pic2Dish, shows the usefulness of the app.

KEYWORDS

Cooking recipe; Cooking instruction; Food recognition

1 INTRODUCTION

In social networks, there are plenty of delicious food pictures shared by users. When browsing those food pictures, people may always ask "how to cook them". However, answering this question is not easy. The solution of this problem is far more than returning a simple recipe, it should also consider the practical usage scenario, such as ingredient shortage and less experienced user. For example, the right ingredients for cooking the dish might not be available due to factors such as geography region or season. Moreover, the cooking skills of the user largely affects the final results even when the ingredients are ready. Therefore, this paper focuses on the problem of assisting users to cook the dish according to a picture with ingredients at hand and provides an all-around cooking support. The key technologies for addressing this problem include food recognition, recipe retrieval, replaceable ingredient mining and instruction video generalization.

Existing work includes automatic cooking instruction video generation [3], smart video cooking [2] and interactive kitchen counter [4]. These work mainly focus on generating the cooking videos to instruct users, and none of them considers the situation when some ingredients are not available. Another work is intelligent menu planning [5], which recommends users recipes based on what ingredients users have. Such kind of recipe recommendation is useful when users have no idea on what kinds of dish they want to cook. Different with [5], our work considers the situation when users want to reproduce a particular dish with ingredients at hand.

In this paper, we design a cooking assistant system - Pic2Dish which contains three key features, food recognition, recipe optimization and instructive video recommendation. By inputting a dish picture and the list of ingredients at hand, our system will generate the most appropriate recipe with existing ingredient for cooking the dish in the picture. Considering users without cooking experience, our system will also provide instructive cooking videos during the whole cooking process. To evaluate the proposed system, we conduct several groups of user study.

Figure 1 shows the overview of Pic2Dish system. It includes three key modules: food recognition, customized recipe generazation and instructive video recommendation. For food recognition module, we use deep convolutional neural network that trained on a large Chinese food dataset to predict the name of dish and obtain the recipe. For customized recipe generalization module, based on the current ingredient list, we consider both co-occurrence context relations with key ingredients and cutting/cooking actions of each ingredient to find the alternative ingredients. For instructive video recommendation module, it matches the guidance video clips with

MM'17, October 23–27, 2017, Mountain View, CA, USA.
 ISBN 978-1-4503-4906-2/17/10…$15.00
DOI: https://doi.org/10.1145/3123266.3126490

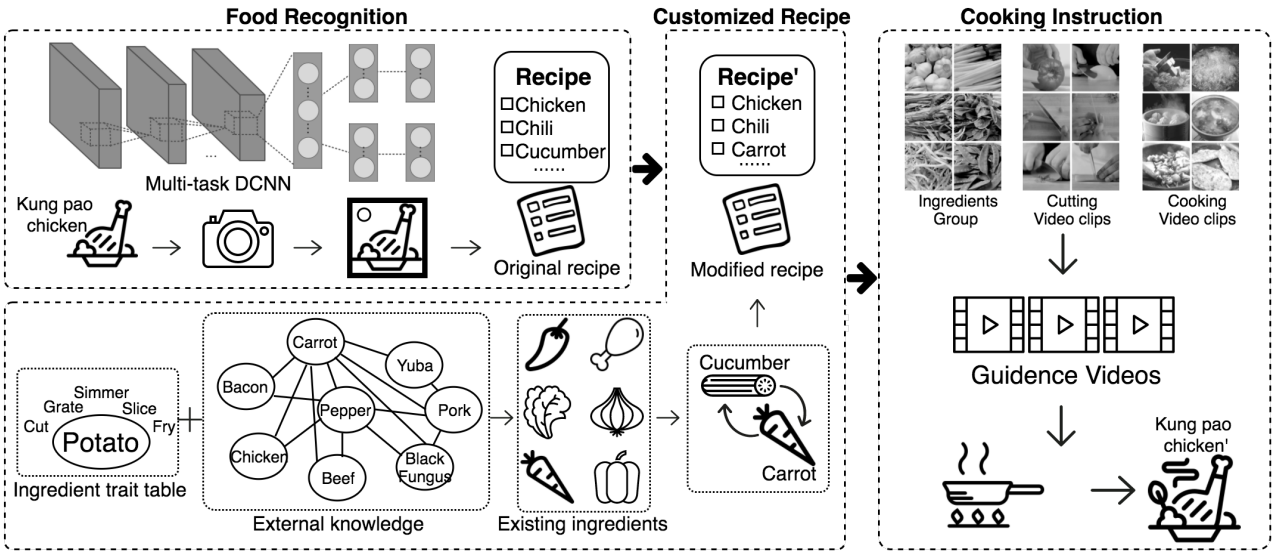

Figure 1: Overview of PIC2DISH

description in customized recipe to generate a series of specific video clips, guiding the users.

2 PIC2DISH SYSTEM

This section gives a detailed introduction on the three modules of the proposed Pic2Dish system: food recognition, customized recipe generazation and instructive video recommendation.

2.1 Deep convolutional network for food recognition

We adopt the multi-task architecture which presented in [1]. The model is modified from VGG 16-layers network [7], with the last two layers being split into two branches, respectively, for single-label food categorization and multi-label ingredient recognition. Figure 2 shows the architecture. The model is trained on VIREO-FOOD172 [1] dataset, and it is able to recognize 172 Chinese dishes and 353 ingredients. When users input dish images, the model will return the name of dish as well as its ingredients. Basically, by matching the dish name against recipe database, we are able to find different versions of recipes, and the ingredient compositions of those recipes are quite diverse. Figure 3 shows three versions of "Kung pao chicken" on "Go Cooking"[1] website with different ingredient composition. For these three recipes, there is even no overlap in ingredients except "chicken" and "peanut". Hence, apart from matching the name of recipes, it is also necessary to match the ingredient in order to find the most appropriate recipe. Denote $\mathbb{U}$ as the a set of retrieved recipes by name matching. We do re-ranking of recipes in $\mathbb{U}$ by ingredient matching in the same way with [1]. Denote Q as the probability distribution of ingredients. Every element in Q corresponds to an ingredient and its value indicates the probability output by DCNN. On the other hand, the ingredients extracted from a recipe are represented as a binary

[1]http://www.xiachufang.com

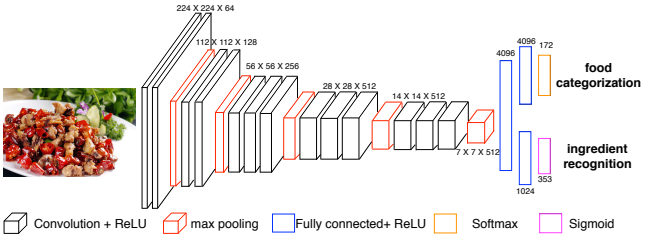

Figure 2: Multi-task DCNN for food categorization and ingredient recognition

vector O. The matching score, s, between them is defined as

$$s = \sum_{c \in O \cap c \in Q} x_c \tag{1}$$

Note that the score is not normalized in order not to bias recipes with a small number of ingredients. As a result, Eqn-1 tends to give a higher score for the recipes with excessive number of ingredients. To prevent such cases, the matching between Q and O is performed only for the top-k predicted ingredients with higher probability scores. The value of k is empirically set to 10 as there are few recipes with more than 10 ingredients in our dataset.

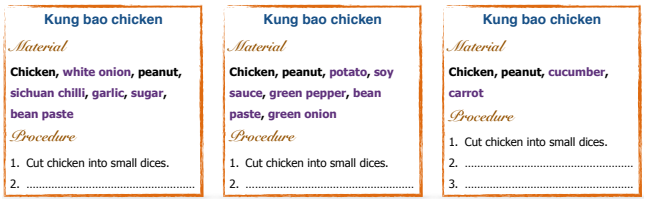

Figure 3: Different versions of kung bao chicken, except "chicken" and "peanut", all the other ingredients are different.

2.2 Customized recipe generation

After obtaining the recipe of dish picture, users may still cannot cook the dish because of lacking some ingredients. To address this problem, the customized recipe generation module will replace the missing ingredients with existing ingredients. We propose a method for finding the replaceable ingredients considering both co-occurrence context relations with key ingredients as well as characteristics like cutting and cooking actions of ingredients. Our method contains two parts: i. off-line part which relates cooking and cutting actions with each ingredient and learns a graph that encodes the context relations among ingredients from a large recipe corpus; ii. on-line customized recipe generation which includes key and non-key ingredients detection, extracting cutting and cooking actions of ingredients, finding substitute ingredients and generating customized recipes.

2.2.1 Off-line ingredient characteristic table and context graph construction. Similar to [6], we construct an ingredient characteristic table from a recipe database which contains 47,882 recipes crawled from "Go cooking website". For each ingredient, we manually labeled its cutting and cooking actions in the recipes to construct the ingredient characteristic table. In the constructed table, it includes 1,276 ingredients, 9 cutting methods and 37 cooking methods.

We mine the statistics from a large corpus composed of more than 60,000 Chinese cooking recipes. The major advantage of doing so is to learn a graph modeling ingredient relationships. We extract ingredients from recipes and construct a graph modeling their co-occurrences based on conditional random field (CRF). Denote $\mathcal{N} = \{c_1, ..., c_I\}$ as the set of available ingredients and I as its set cardinality. The graph G is composed of the elements of $\mathcal{N}$ as vertices and their pairwise relationships, denoted as $\phi(\cdot)$, as edges. Further let l_i as an indication function that signals the presence or absence of an ingredient c_i. The joint probability of ingredients given the graph is

$$p(l_1, ...l_I) = \frac{1}{Z(\phi)} exp(\sum_{i,j \in \mathcal{N}} l_i l_j \phi(i,j)) \tag{2}$$

where $Z(\cdot)$ is a partitioning function. To learn the graph, we employ Monte Carlo integration to approximate $Z(\cdot)$ and the gradient descent to estimate $\phi(\cdot)$ to optimize the data likelihood.

2.2.2 On-line customize recipe generation. Note that for a given dish image, by matching the dish name, we are able to get a set of recipes $\mathbb{U}$ whose cardinality is denoted as $|\mathbb{U}|$, and then through ingredient matching, we can obtain the most appropriate recipe r. In order to detect the key and non-key ingredients in r, we calculate the frequency of each ingredient appears in $\mathbb{U}$, for i^{th} ingredient, its frequency F_i is obtained by

$$F_i = \frac{Number\ of\ recipes\ contains\ i^{th}\ ingredient}{|\mathbb{U}|} \tag{3}$$

If the frequency of an ingredient is large than 0.9, then the ingredient is considered as key ingredient, otherwise, it is non key ingredient. Take Figure 3 for example, "chicken" and "peanut" appear in every recipe, so they are considered as key ingredient of dish "kung bao chicken", while "cucumber" appears in 1 recipe, therefore it is considered as non-key ingredients. In this way, we can obtain the key and non-key ingredient in recipe r.

Table 1: Ingredient groups

Groups	Examples
meat	beef, pork
seafood	fish, shrimp
leaf vegetables	lettuce, cabbage
gourd vegetable	bitter gourd, pumpkin
solanaceous vegetables	eggplant, tomato
legume vegetable	dutch beans, green bean
fungus	pleurotus, shiitake
nut fruits	peanut, cashew nuts

Denote $\mathbb{M}$ as a set of non-key ingredients which need to find the corresponding substitute ingredients. For each ingredient in $\mathbb{M}$, we extract their associated cutting and cooking actions in the recipe. Through searching the ingredient characteristic table, for each ingredient in $\mathbb{M}$, we are able to find a set of candidate ingredients $\mathbb{C}$ with same cutting and cooking actions. Denote K as a set of key ingredients in the retrieved recipe, for each missing ingredient, our goal is to find the most suitable substitute ingredient that has large co-occurrence relations with key ingredients. Recall $\phi(\cdot)$ is the learnt pairwise relations among ingredients. For each missing ingredient, the corresponding substitute ingredient $\hat{x}$ can be obtained by:

$$\hat{x} = \arg\min_{x \in \mathbb{C}} \sum_{k \in \mathbb{K}} \phi(x,k) \tag{4}$$

Then the new recipe can be generalized by replacing the missing ingredient with obtained substitute ingredients.

2.3 Instructive video recommendation

Most of recipes provide text and static image to show the cooking instruction, which is difficult to follow for less experienced user. As discussed in [2], video guidance is able to provide more detailed and precise instructions for audience. However, with the increased number of cooking videos on the web, finding the exactly matched cooking videos remains technically challenging. An simple yet effective solution is providing the users with video clips showing the cutting and cooking techniques for each ingredient in the recipe. However, collecting those video clips for each ingredient is not easy. Since we have 1,276 ingredient, 9 cutting methods and 37 cooking methods, the number of instruction video clips that need to collect could be close to 0.4 million. Based on the observation that ingredients with the same type (for example, leaf vegetables) share the same cutting or cooking techniques, for example, the procedure of cutting cucumber into slices is the same with eggplant, we divide ingredient into 8 groups based on the characteristic of the ingredients, which is shown in Table 1. For each cutting/cooking techniques, only one video clip need to be collected for each group. With these video clips, it could generate the cooking instruction step by step automatically from preparation stage to cooking stage and cover the majority of cooking instructions of Chinese recipes.

3 EXPERIMENT RESULTS

We evaluate the performance of multi-task model for food categorization and ingredient recognition performance on VIREO-FOOD172 dataset. Table 2 shows the performances. The top-1 food categorization performance can be as high as 82%. Next, we

	Categorization		Ingredient recognition	
	Top-1 (%)	Top-5 (%)	*Micro-F1 (%)*	*Macro-F1 (%)*
Multi-task	82.06	95.88	67.17	47.18

Table 2: Recognition performances of multi-task DCNN model

show two example of customized recipe generation results in Figure 4. The first example shows that our method replaces "peanut" with "cashew nuts" for "kung pao chicken", because user only has "cashew nuts". "Cashew nuts" has the same characteristics with "peanut" in the constructed ingredient characteristic table, and it co-occurs with "chicken" in some recipes. Therefore, our system replaces "peanut" with "cashew nuts". The second example shows the situation when "bean sprout" is not available and user still wants to cook "Sichuan boiled beef". Our system generates a new recipe which replaces "bean sprout" with "emoki mushrooms" for user.

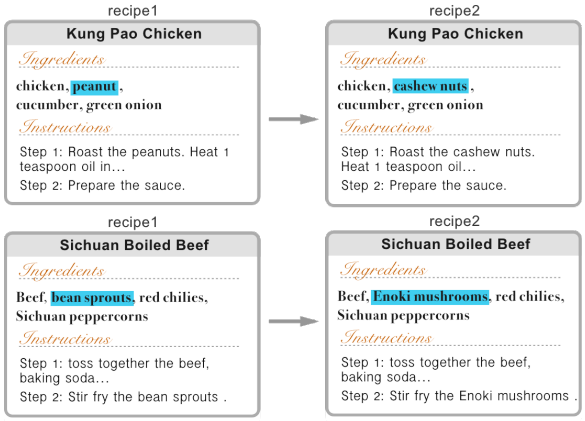

Figure 4: Examples of customized recipe generation

4 USER STUDY

4.1 The method of user study

To evaluate the performance of the proposed Pic2Dish system, we conduct user study. In total, 20 subjectives are invited to participate in user study. Among them, 10 subjectives have cooking experience and the remaining 10 have few cooking experience. The goal of the user study is twofolds: 1). to evaluate whether the generated customized recipe is reasonable or not. 2). to evaluate whether our cooking guidance is sufficient or not. For evaluating the customized recipe generation module, 10 experienced subjectives are given different dish images and required to cook the dish using our Pic2Dish app. They are required to cook the dish under two different conditions: with and without sufficient ingredients. When there are not sufficient ingredients provided, our Pic2Dish app will generate the customized cooking recipe by replacing the missing ingredients

	Customized recipes		Guidance
	With sufficient ingredients	Without sufficient ingredients	
Satisfactory Level Avg.	4.08	3.82	3.93

Table 3: Results of user study

with existing ingredients. After they finish cooking the dish, they are asked to rate the satisfactory degree of the app. For evaluating our cooking guidance module, 10 experienced subjectives are required to chose a dish they had tried before and cook the dish without Pic2Dish application assistant while the remaining 10 less experienced subjectives are asked to cook the same dish with our Pic2Dish system. Sufficient correct ingredients are provided during the cook process. When finish cooking, 10 less experienced are required to rate the satisfactory level of our app based on comparing the difference between the dish they cook and the dish cooked by experienced subjectives in terms of appearances and tastes.

4.2 User study results

We define 5 levels to evaluate the satisfactory degree: from one to five, each score corresponds to very undesirable, mildly undesirable, moderately desirable, above avg. desirable and great, respectively. The evaluation results are shown in Table 3. For customized recipes evaluation, the averages of satisfactory level are 4.08 and 3.82. The difference between with sufficient ingredients group and without sufficient ingredients group is 5.2%. For cooking guidance evaluation, the satisfactory level is 3.93. Therefore, Pic2Dish system can generate reasonable recipes and make an efficient cooking guidance.

5 FUTURE WORK

In this work, Pic2Dish has fulfilled the basic cooking support task for users. Basically, there is a considerable gap between practical application and experiment in the laboratory since we only collect the limited data of ingredients and the classification is not elaborate. Hence, there is a huge work to deal with in the future. Besides, with the development of augmented reality technology and the improvement of its hardware device, it provides a totally new interactive experience between human and machine. Once if we can embed a human cooking action recognition system into augmented reality device and combine what we implemented in this paper, it will create a real time supervised and guiding assistant system for cooking. This is our target in the future.

6 ACKNOWLEDGMENT

This work is part of the NExT++ project, supported by the National Research Foundation, Prime Minister's Office, Singapore under its IRC@SG Funding Initiative. It is also supported by National Natural Science Foundation of China (61370023).

REFERENCES

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

[2] Keisuke Doman, Cheng Ying Kuai, Tomokazu Takahashi, Ichiro Ide, and Hiroshi Murase. 2012. Smart VideoCooKing: a multimedia cooking recipe browsing

application on portable devices. In *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 1267–1268.

[3] Yasuhiro Hayashi, Keisuke Doman, Ichiro Ide, Daisuke Deguchi, and Hiroshi Murase. 2013. Automatic authoring of a domestic cooking video based on the description of cooking instructions. In *Proceedings of the 5th international workshop on Multimedia for cooking & eating activities*. ACM, 21–26.

[4] Wendy Ju, Rebecca Hurwitz, Tilke Judd, and Bonny Lee. 2001. CounterActive: an interactive cookbook for the kitchen counter. In *CHI'01 extended abstracts on Human factors in computing systems*. ACM, 269–270.

[5] Fang-Fei Kuo, Cheng-Te Li, Man-Kwan Shan, and Suh-Yin Lee. 2012. Intelligent menu planning: Recommending set of recipes by ingredients. In *Proceedings of the ACM multimedia 2012 workshop on Multimedia for cooking and eating activities*. ACM, 1–6.

[6] Yuka Shidochi, Tomokazu Takahashi, Ichiro Ide, and Hiroshi Murase. 2009. Finding replaceable materials in cooking recipe texts considering characteristic cooking actions. In *Proceedings of the ACM multimedia 2009 workshop on Multimedia for cooking and eating activities*. ACM, 9–14.

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