Food Image Captioning with Verb-Noun Pairs Empowered by Joint Correlation

Anonymous ACCV 2016 submission

Paper ID ***

Abstract. Studies of image captioning explosively emerge in recent two years. Though many elegant approaches have been proposed for general-purposed image captioning, considering domain knowledge or specific description structure in a targeted domain still remains undiscovered. In this paper, we concentrate on food image captioning where a food image is better described by not only what food it is but also how it was cooked. We propose neural networks to jointly consider multiple factors, i.e., food recognition, ingredient recognition, and cooking method recognition, and verify that recognition performance can be improved by taking multiple factors into account. With these three factors, food image captions composed of verb-noun pairs (usually cooking method followed by ingredients) can be generated. We demonstrate effectiveness of the proposed methods from various viewpoints, and believe this would be a better way to describe food images in contrast to general-purposed image captioning.

1 Introduction

Image captioning has recently attracted much attention because of its extensive potentials in bridging the semantic gap between visual features and high-level semantics. Thanks to the rapid advancement of deep visual representation by convolutional neural network and the language generation by recurrent neural network, performance of image captioning scales up to a large factor in just recent two years [1][2]. These emerging techniques have also been extended to achieve video captioning [3].

The elegant models mentioned above, however, largely put efforts on general-purposed image captioning, which may be widely adopted in general image classification or retrieval, but may not well catch the uniqueness of images in a specific domain. In this work, we concentrate on food image captioning for three reasons. First, tremendous amounts of food images are daily shared on social media platforms. These images not only show what a user eats, but also present the user's life style. With food image captioning, we would more deeply describe a user's living experience. Second, food image captions facilitate many valuable applications, such as health management, recipe recommendation, and restaurant recommendation. Third, unlike general image captioning, an appropriate food image caption would show not only the food name (food recognition in the literature), but also the way it was cooked. For example, description like "roasted

beef with soft-boiled eggs" is richer than "beef and egg" when a user tries to order a meal with a menu showing food images. The verb-noun pair showing the cooking method and the ingredient makes food image captioning distinct from general-purposed image captioning.

In order to generate food image captions consisting of verb-noun pairs (VNPs), we propose neural frameworks that take a food image as the input and infer the confidence distributions of cooking methods as well as ingredients. With the correlation learnt from recipes, this framework finally outputs the food image caption that well describes what this food is and how it was cooked.

Contributions of this work are summarized as follows.

- We propose a learning framework that can separately work for food recognition, ingredient detection, and cooking method detection, as well as transfer information from one modality to another modality in order to improve performance of a targeted task.
- Based on the proposed multi-task learning framework, we propose food image descriptions as a set of verb-noun pairs, generally a cooking method followed by an ingredient. Correlations between cooking methods and ingredients with given visual information are learnt based on recipe information.
- To facilitate the proposed food image captioning, we collect a food image dataset associated with well-organized recipe information. In contrast to previous datasets where recipe data are usually just for food recognition, we analyze cooking steps and associated ingredients, and summarize a recipe as a set of verb-noun pairs to facilitate construction of the proposed system.

The rest of this paper is organized as follows. Section 2 provides literature survey on food image analysis and image captioning. Section 3 describes the framework generally adopted to food recognition, cooking method recognition, and ingredient recognition. Details of the learning framework transferring information from one modality to another modality will be provided. With these recognition results, the proposed food image captioning in the representation of VNPs is described in Section 4. Section 5 describes experimental results in several aspects, followed by the concluding remarks given in Section 6.

2 Related Works

Food image analysis emerges recently not only because there is a large number of food images shared on the social platforms, but also it is the foundation to achieve health management that has attracted much attention due to the popularity of wearable devices. Most existing works focus on food recognition. In [4], a food image is segmented into regions, where each region is viewed as an ingredient, and correlations between regions are considered as a feature to do food recognition. This approach is limited to food where different ingredients can be clearly separated. Bossard et al. [5] adopted random forests to find discriminative components in food images to enhance performance of food recognition. They proposed one of the largest food image dataset consisting of 101 classes.

To better represent food images, Kagaya et al. [6] extracted convolutional neural network features based on a pre-trained deep model, and verified that performance better than conventional hand-crafted features can be obtained.

External information can be utilized to facilitate food recognition. Xu et al. [7] considered the GPS information of a given food image, and search for possible restaurants to reduce the range of possible recognition candidates, so that better recognition performance can be obtained. Wang et al. [8] showed that by jointly considering recipe information and visual descriptors, much accurate food recognition can be achieved. However, this approach is limited to the case where users simultaneously have the recipe and the food image.

In addition to food recognition, Myers et al. [9] proposed a system to estimate calories of a given food image. Given a food image, they first estimate what ingredients are included, estimate their volumes, and finally estimate calories. Few works have been proposed to specially focus on food image captioning. Hessel et al. [10] studied whether image modeling or language modeling acts as bottleneck of image captioning performance. They found that, by adding more image training data or changing deep architecture to get more complex image representation, only slight or even no performance improvement can be obtained for image captioning. They collected the Yummly food image dataset to support this research, where for any given food image the recipe title is generated as the image caption. In contrast to recipe title, we argue that further considering cooking methods and ingredients can more appropriately describe food images.

3 Food Recognition, Cooking Recognition, and Ingredient Recognition

In contrast to general-purposed image captioning that employs general language models, to generate VNP-based food image captions we focus on factors that are highly related to food, and devise a relatively simpler model with the consideration of multiple factors to generate verb-noun pairs. In this section, we especially discuss three important factors as pre-processes for food image captioning: food recognition, cooking method recognition, and ingredient recognition. We advocate that recognition of one factor, e.g., food recognition, can be benefited from the results of two other factors, i.e., cooking method and ingredient. In the following, we would take food recognition as the main example to show how other factors are considered, though the same idea can be employed to enhance cooking method recognition (by considering results of food recognition and ingredient recognition) or ingredient recognition (by considering results of food recognition and cooking method recognition).

3.1 Databases

We first describe the two databases we use for this study. The first is the UPMC Food-101 dataset [8] that covers 101 food categories and includes totally 90,840 images. Images were retrieved by Google Image search, with queries from the

101 labels taken from the ETHZ Food-101 dataset [5]. In order to study recipe recognition, raw HTML pages that embed these images were also collected. The number of images that have corresponding HTML text is 86,574.

The UPMC Food-101 dataset is quite challenging because images were collected from uncontrolled sources, and the top-returned images from the web search engine may include noises. In addition, the collected HTML pages may consist of content not highly related to the embedded food images. To facilitate more precise food image captioning, we need a dataset consisting of food images associated with clean recipe information that clearly describes what ingredients are used and how they are cooked. For this purpose, we crawl ten types of images and recipes from Recipe.com[1], including *beef, bread, burger, cake, casseroles, chicken, chili, cookies, fruit, and grilling*. We collected 9,363 images in total, with each image associated with clean recipe. Figure 1 shows a food called *Beef Wellington* and the corresponding recipe consisting of ingredients and cooking directions. This example also shows the shortage of food name for us to realize what this food really is. From the food name, we only know it is a kind of beef meal, but have no idea about how it was cooked and hardly imagine how it tastes. Associating cooking methods and ingredients (verb-noun pair) is therefore important in food image captioning.

Fig. 1. A sample food image (*Beef Wellington*) and its corresponding ingredients and cooking directions.

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

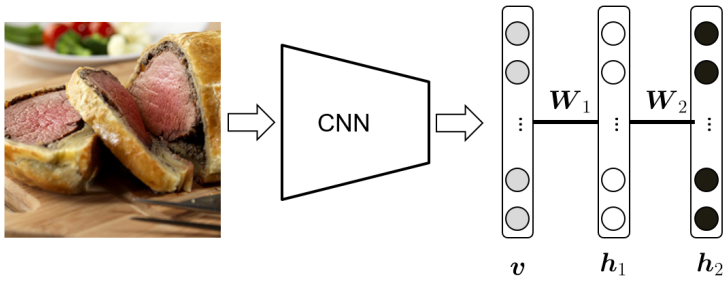

Fig. 2. The baseline deep framework for food recognition.

3.2 Baseline Model

Food Recognition. We first describe how a basic learning model is developed to achieve food recognition. Figure 2 shows the framework of a baseline model for food recognition. Given a food image, we extract its convolutional neural network (CNN) features based on the MatConvNet toolbox [11] with the vgg-f pretrained model [12]. Results of the seventh layer, i.e., the last fully-connect layer, are taken as the image representation, which is 4,096-dimensional and is shown as $\boldsymbol{v}$ in Figure 2. Based on this image representation, a three-layer neural network is constructed to do recognition. The input layer consisting of 4,096 nodes is for input entries, the hidden layer consisting of 40,960 nodes is for mid representation, and the output layer consisting of 101 nodes is a softmax layer and outputs the probabilities of the given image belonging to 101 food classes defined in the UPMC dataset. The relationship between input, mid representation, and output can be represented as follows.

$$\boldsymbol{h}_1 = \boldsymbol{W}_1\boldsymbol{v} + \boldsymbol{b}_1, \tag{1}$$

$$\boldsymbol{h}_2 = \boldsymbol{W}_2\boldsymbol{h}_1 + \boldsymbol{b}_2, \tag{2}$$

where $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are weighting matrices, and $\boldsymbol{b}_1$ and $\boldsymbol{b}_2$ are bias vectors.

To train the neural network, from each food class we randomly select 500 images, constituting a training set consisting of 50,500 images. The remaining images are used for testing. For each image, we have the image representation $\boldsymbol{v}$ and its associated food class vector $\boldsymbol{y}$, which is a one-hot representation showing which class the image belongs to. With the set of training tuples $\{(\boldsymbol{v}, \boldsymbol{y})\}$, the objective of this neural network is to find the best weighting matrices and bias vectors such that $\sum -\boldsymbol{y}^T\boldsymbol{h}_2$ is minimized. That is, we want to maximize the summation of inner products between the estimated probability vector $\boldsymbol{h}_2$ and the ground truth vector $\boldsymbol{y}$. In the experiment, we perform training and testing based on the random-split scheme for five times, and report the average recognition result.

Ingredient Recognition. The same framework as shown in Figure 2 is also adopted to achieve ingredient recognition. In this task, we build the ingredient recognition model based on our dataset, because it has clean recipe information. Given training tuples $\{(\boldsymbol{v}, \boldsymbol{y})\}$, the neural network is trained, and the softmax layer outputs a vector showing the probabilities of ingredients embedded in the given test image.

The representation of ingredients is worth mentioning specially. Based on the ingredient information of our dataset, we manually filter out stop words and commonly-used units like spoon and jar. Finally 314 different ingredients are retained in total. Unlike the one-hot representation in food recognition, the ground truth vector $\boldsymbol{y}$ is a 314-dimensional binary vector where multiple entries would be unity because a food often contain multiple ingredients.

Cooking Method Recognition. Based on the cooking direction information of our dataset, we conclude ten common actions: *cook, bake, toast, roast, season, grill, broil, heat, simmer, and stew*. A food may be prepared with several actions, like *roast the beef, and cook the mushrooms*. The ground truth vector $\boldsymbol{y}$ is thus a 10-dimensional binary vector where multiple entries would be unity. The same framework as shown in Figure 2 is also adopted for cooking method recognition. Note that the output is a vector showing the probabilities of cooking methods to prepare the given food image.

3.3 Enhanced Models Considering Correlation

To enhance performance of food image analysis, we consider that food name, ingredients, and cooking methods are correlated, and by jointly taking two other results (e.g., ingredient recognition and cooking method recognition) into account, performance of the targeted domain (e.g., food recognition) can be improved. In this section, we propose two frameworks to join multiple recognition results in order to improve one targeted domain. We still take food recognition as the main targeted domain for instance.

Joint Model 1: Similar Mid Representations. The first idea to boost food recognition is that the mid representation learnt for three recognition domains should be similar. The three output layers for three different recognition tasks respectively yield good results from three perspectives if the mid representation well describe the given food image. Figure 3 shows the framework of the first joint model. To make the figure simple and clear, we just illustrate the case where the mid representation learnt for ingredient recognition is considered to improve food recognition. Basically, the same framework can be adopted to further consider the mid representation learnt for cooking method recognition.

With the learnt mid representation $\boldsymbol{h}_{1i}$ for ingredient recognition, the cost function to learn the food recognition model is defined as

$$\sum(-\boldsymbol{y}^T \boldsymbol{h}_{2f} + |\boldsymbol{h}_{1f} - \boldsymbol{h}_{1i}|), \tag{3}$$

where $\boldsymbol{y}$ is the ground truth vector, $\boldsymbol{h}_{2f}$ is the output probability vector of food names, and $\boldsymbol{h}_{1f}$ is the mid representation for food recognition. By considering

the Euclidean distance $|\boldsymbol{h}_{1f} - \boldsymbol{h}_{1i}|$ between $\boldsymbol{h}_{1f}$ and $\boldsymbol{h}_{1i}$ in the objective function, we take mid representation similarity into account in learning weighting matrices and bias vectors in the food recognition model.

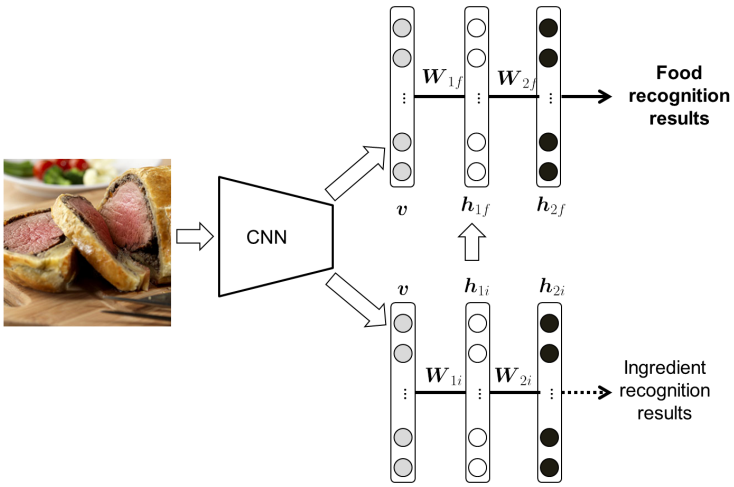

Fig. 3. Illustration of the first joint model, with the idea of similar mid representations.

Joint Model 2: Early Fusion. The second idea to improve food recognition is simply concatenating the probability vector of ingredient recognition with the CNN feature of the given image to form an *enhanced* image descriptor. This idea is similar to early fusion widely used in integrating multimodal features, and Figure 4 illustrates this idea. Note again that, to make the figure simple and clear, only using ingredient to enhance food recognition is illustrated here. Based on such enhanced image descriptor, named $\boldsymbol{u} = (\boldsymbol{v}, \boldsymbol{h}_{2i})$, the objective function to be maximized in training the neural network is the same as that in the baseline model.

Note that the proposed joint models can be adopted to improve ingredient recognition based on results of food recognition and cooking method recognition, or improve cooking method recognition based on results of food recognition and ingredient recognition. In the evaluation section, we will verify and compare recognition performance of the baseline model and two joint models.

4 Food Image Captioning

Given a test image I, the cooking method recognition model mentioned above would output a probability vector $\boldsymbol{c} = (p_1^c, p_2^c, ..., p_M^c)$ showing the probability of I being cooked by each method. Similarly, the ingredient recognition method would output a probability vector $\boldsymbol{g} = (p_1^g, p_2^g, ..., p_N^g)$ showing the probability of

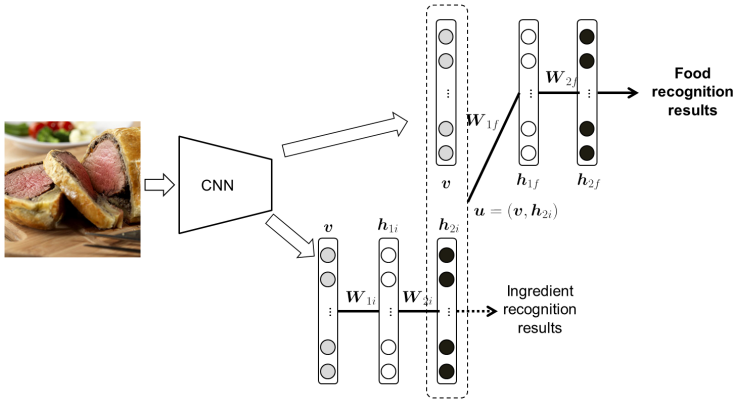

Fig. 4. Illustration of the second joint model, with the idea of early fusion.

I containing each ingredient. Conceptually, the most probably verb-noun pair can be found by finding (c^*, g^*) such that

$$(c^*, g^*) = \arg \max_{\substack{i=1,\ldots,M \\ j=1,\ldots,N}} p_i^c \times p_j^g. \tag{4}$$

To further consider the correlation between cooking methods and ingredients, we process the collected recipes by pairing each detected cooking term, e.g., roast, with its closest succeeding ingredient, e.g., beef. We thus found a large number of verb-noun pairs, like *roast beef* and *cook tomato*, from the recipe set. The frequency of each VNP is then normalized to be the prior probability of an action taken to cook an ingredient. The prior probability of the cooking method i used to cook the ingredient j is denoted as p_{ij}^r. With this consideration, the most probably VNP is then determined by finding (c^*, g^*) such that

$$(c^*, g^*) = \arg \max_{\substack{i=1,\ldots,M \\ j=1,\ldots,N}} p_{ij}^r \times p_i^c \times p_j^g. \tag{5}$$

In the evaluation, we in fact find the five VNPs with five largest probabilities to be image captions.

5 Experimental Results

5.1 Performance of Enhanced Models

We evaluate performance of the proposed enhanced models based on the UPMC dataset. To more finely evaluate the influence of different factors, we evaluate the two joint models enhanced by different combinations of factors. Table 1 shows performance of food recognition obtained based on the baseline model,

and the two joint models with various settings. As can be seen, both joint models enhanced by results of ingredient outperform the baseline model, but that enhanced by results of cooking methods don't. This result looks reasonable because with ingredient information we may be able to more accurately recognise what food it is. The cooking method, however, may give little information for food recognition, not to say that the result of cooking method recognition may be wrong. The second observation is that jointly considering more information does not necessarily yield better performance. In Table 1, the best performance is obtained by the early fusion method (Joint Model 2) that fuses the result of ingredient recognition with the visual descriptor. For the UPMC dataset that includes 101 food classes, we obtain 37.53% recognition accuracy, which significantly outperforms 33.91% accuracy reported in [8].

Table 2 shows recognition accuracy of ingredients based on three models, with different settings. Here we again see superior performance of joint models, while two joint models obtain similar performance. Results of food recognition provide more benefits in recognising ingredients. Jointly considering more information doesn't yield better performance. Similar trends can also be seen in recognising cooking methods, as shown in Table 3. From Table 3, we further observe that, when recognising cooking methods, the Joint Model 2 consistently outperforms the baseline model and the Joint Model 1.

Through the results mentioned above, we verify effectiveness of joint models, and can make two remarks. (1) The early fusion method (Joint Model 2) works consistently better than the baseline model and the mid representation approach (Joint Model 1) in food recognition and cooking method recognition. (2) Joining appropriate information is the key to get performance gain.

Table 1. Accuracy of food recognition performance based on the baseline model, the two joint models enhanced by ingredient, cooking, or both.

	with ingredient	with cooking method	with both
Joint Model 1	0.3648	0.3257	0.3212
Joint Model 2	**0.3753**	0.2550	0.2575
Baseline	0.3435		
[8] (visual only)	0.3391		

Table 2. Accuracy of ingredient recognition performance based on the baseline model, the two joint models enhanced by food, cooking, or both.

	with food	with cooking method	with both
Joint Model 1	**0.5759**	0.5751	0.5682
Joint Model 2	**0.5748**	0.5426	0.5737
Baseline	0.5343		

Table 3. Accuracy of cooking method recognition performance based on the baseline model, the two joint models enhanced by food, cooking, or both.

	with food	with ingredient	with both
Joint Model 1	0.5400	0.5403	0.5406
Joint Model 2	0.5616	**0.5621**	**0.5622**
Baseline	0.5461		

5.2 Performance of Food Image Captioning

Effectiveness of VNPs Unlike general-purposed image captioning tasks that have caption ground truth provided by the MSCOCO dataset [13], currently there is no well-developed food image captioning dataset. We thus evaluate the proposed food image captioning based on subjective tests. Ten subjects were invited to join the experiment, where each person was randomly given fifteen to twenty food images with the automatically generated captions. Two types of captions were generated for food images randomly selected in the test set of the UPMC database: (1) only ingredient recognition results were shown to the subjects; and (2) the generated verb-noun pairs were shown to the subjects. These two types of captions were randomly juxtaposed, and the subjects were asked to select which one was better to annotate the given food image.

Overall, the second type of caption (VNP) is viewed better in 86 of 155 food images, while the rest 69 of 155 food images are viewed to better be annotated by the first type of caption. The VNP-based captioning is not significantly better because both results of ingredient recognition and cooking method recognition are still not good enough (see Table 2 and Table 3). The effectiveness of VNPs might be largely elevated if these two recognition results can be improved.

General-Purposed Image Captioning vs. VNPs Here we would like to design a subjective experiment to compare results of general-purposed image captioning with food-specific VNPs. Given a food image selected in the test set of the UPMC database, or our dataset, we generate VNPs as well as a general-purposed image caption by Microsoft CaptionBot service[2]. Two types of captions were presented to subjects, who were then asked to measure which caption is better to describe the given food image. Table 4 shows the number of images viewed to have better caption results in two different datasets. Among the 58 test images in the UPMC dataset, our VNP-based captions are viewed to be better description in 39 images. Among the 100 test images in our dataset, the VNP-based captions are viewed to be better description in 84 images. Results for images from our dataset are much better than that from the UPMC dataset. The main reason is that the prior probability described in eqn. (5) can be more accurately estimated in our dataset, because our dataset have clearer recipe

[2] https://www.captionbot.ai/

information while the UPMC dataset's recipe information is from noisy HTML pages.

Table 4. Performance of food image captioning, in the representation of the number of images that are viewed to have better caption results.

Dataset	CaptionBot	Our method
UPMC (totally 58 images)	19	39
Our dataset (totally 100 images)	16	84

Fig. 5 shows a food image representing *Burger*, and Table 5 shows recognition results and image captions in the representation of VNPs as well as the image caption generated by CaptionBot. As can be seen, we correctly recognize the given image is Burger. The top five recognized ingredients are egg, milk, cheese, corn, and lemon, which are satisfactory results. The top three recognized cooking methods are heat, bake, and cook, which can be imagined to make a burger. By finding the most likely VNPs, the top five VNPs are bake egg, bake onion, bake tomato, bake beef, and bake cheese. All of them are appropriate VNPs because they are all necessary processes in making a burger. By showing food recognition result followed by VNPs, we can generate a description that provides information richer than a general-purposed image caption, as shown in the last row of Table 5.

Fig. 5. A food image representing *Burger*.

6 Conclusion

We have presented a food image captioning approach that jointly considers results of food recognition, ingredient recognition, and cooking method recognition, and generates verb-noun pairs that not only shows what food it is but also

Table 5. Recognition and image caption (VNP) results corresponding to Fig. 5.

Types	Results
Food recognition	Burger
Ingredient recognition (top 5)	egg, milk, cheese, corn, lemon
Cooking method recognition (top 3)	heat, bake, cook
Image captioning (food name followed by top 5 VNPs)	Burger: bake with egg, bake with onion, bake with tomato, bake beef, bake with cheese
Microsoft CaptionBot	I think it's a sandwich on a plate.

show how it was cooked. We propose two schemes to embed correlation between different recognition results in a learning framework, and verify that recognition performance can be improved with the help of other recognition results. We generate VNP-based food image captions by maximizing the probability of the combination of cooking methods and ingredients. Based on subjective tests, the proposed VNPs are verified to be more effective in describing food images, as compared to general-purposed image captioning.

In the future, there is still much room to improve various recognition results. How to more tightly integrate different recognition results or intermediate representation in a learning framework is also an important issue. Moreover, we are interested in building an augmented dataset that consists of multi-scale patches extracted from food images. Many works [14] have shown that deep features can be extracted from multiple patches and then combined to achieve better recognition or classification performance.

References

1. Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (2015) 3128–3137
2. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: A neural image caption generator. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (2015) 3156–3164
3. Yu, H., Wang, J., Huang, Z., Yang, Y., Xu, W.: Video paragraph captioning using hierarchical recurrent neural networks. Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (2016)
4. Yang, S., Chen, M., Pomerleau, D., Sukthankar, R.: Food recognition using statistics of pairwise local features. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (2010) 2249–2256
5. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. Proceedings of European Conference on Computer Vision (2014)
6. Kagaya, H., Aizawa, K., Ogawa, M.: Food detection and recognition using convolutional neural network. Proceedings of ACM International Conference on Multimedia (2014) 1085–1088
7. Xu, R., Herranz, L., Jiang, S., Wang, S., Song, X., Jain, R.: Geolocalized modeling for dish recognition. IEEE Transactions on Multimedia **17** (2015) 1187–1199

8. Wang, X., Kumar, D., Thome, N., Cord, M., Precioso, F.: Recipe recognition with large multimodal food dataset. Proceedings of IEEE International Conference on Multimedia and Expo Workshops (2015)
9. Myers, A., Johnston, N., Rathod, V., Korattikara, A., Gorban, A.N., Silberman, N., Guadarrama, S., Papandreou, G., Huang, J., Murphy, K.: Im2calories: Towards an automated mobile vision food diary. Proceedings of IEEE International Conference on Computer Vision (2015) 1233–1241
10. Hessel, J., Savva, N., Wilber, M.J.: Image representations and new domains in neural image captioning. Proceedings of Workshop on Vision and Language Integration (2015)
11. Vedaldi, A., Lenc, K.: Matconvnet: Convolutional neural networks for matlab. Proceedings of ACM International Conference on Multimedia (2015) 689–692
12. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: Delving deep into convolutional nets. Proceedings of British Machine Vision Conference (2014)
13. Lin, T.Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C.L., Dollar, P.: Microsoft coco: Common objects in context. Proceedings of European Conference on Computer Vision (2014)
14. Lu, X., Lin, Z., Shen, X., Mech, R., Wang, J.Z.: Deep multi-patch aggregation network for image style, aesthetics, and quality estimation. Proceedings of IEEE International Conference on Computer Vision (2015) 990–998