Understanding Food Images to Recommend Utensils During Meals

F. Ragusa, A. Furnari(✉), and G.M. Farinella

Department of Mathematics and Computer Science, Image Processing Laboratory, University of Catania, Catania, Italy
francescoragusa@outlook.com, {furnari,gfarinella}@dmi.unict.it

Abstract. Understanding food images can be useful to enable different technologies aimed at improving the quality of life of the society. We focus on the problem of analyzing food images to recognize the utensils to be used to consume the meal depicted in the image. The proposed investigation has both a practical and a theoretical relevance, since (1) it can contribute to the design of intelligent systems able to assist people with mental disabilities and (2) it allows to assess if high level concepts related to food (e.g., how to eat food) can be inferred from visual analysis. We augment the FD1200 dataset with labels related to utensils and perform experiments considering AlexNet features coupled with a multi-class SVM classifier. Results show that, even such a simple classification pipeline can achieve promising results.

Keywords: Image understanding · Food analysis
Assistive technologies

1 Introduction and Motivations

Analysis and understanding of food images is a challenging Computer Vision task which has gathered much interest of the research community due to its potential impact on the quality of life of modern society [1]. In this context, the main problems considered by the community are related to the discrimination of food images vs other images [11,15,16], the detection/localization of food in images [17,23], the recognition and classification of the food depicted in an image [19–21], the segmentation of food images to distinguish the different parts and ingredients [18,22,25], the estimation of the volume and nutrients contained in a food plate detected in an image [24,26,27]. A big issue in this application domain is the availability of public datasets, as well as the lack of common procedures for testing and evaluation of the different tasks. Despite some food datasets exist [1,28], their size and variability is still limited to properly feed modern supervised learning approaches currently employed to solve different computer vision tasks [29].

In recent years, considering the advancement in the fields of Computer Vision and Machine Learning, the research community is making a great effort in

S. Battiato et al. (Eds.): ICIAP 2017 International Workshops, LNCS 10590, pp. 419–425, 2017.
https://doi.org/10.1007/978-3-319-70742-6_40

designing and investigating intelligent systems able to help people in their daily activities [2]. Different studies have been proposed to design robotic personal assistants [4,5], advanced wearable vision systems to help people to augment their memory [6–8], as well as to monitor daily activities in order and improve quality of life [9,10]. The main motivation behind these studies is to help society by exploiting the advancements of computer and engineering science. In this regard, this paper builds on the following question: can we train a computer vision system to recognize the eating utensils to be used during a meal in order to help patients with dementia diseases in reminding how to eat food?

Eating is an important aspect of life. It is important to satisfy hunger, to stimulate our senses, to share moment with others, but, most importantly, to acquire the needed nutrients to live and be in good health. While the recognition of which utensils to use when eating a meal might seem to be straightforward, it is not so simple for people affected by dementia disease, such as Alzheimer. When the disease start to become severe, patients can experience problems in using utensils during a meal because of loose memory and other mental disabilities. The ability to correctly recognize which utensils to use during a meal is one of the aspect that is analyzed to monitor the functional abilities of dementia patients [3]. Patients often do not remember how to use eating utensils, and in late stages of the disease, food which can be eaten with fingers is usually adopted by caregivers to help patients.

Figure 1 illustrates the investigated problem: given a food image, the computer vision engine should be able to predict which utensils are to be used to consume the meal. We would like to note that the proposed investigation is of interest both from an application and a theoretical standpoint. On one hand, the proposed system can be used for practical purposes. For instance, it could be exploited in a wearable device or in a robotic personal assistant to provide suggestions to patients during meals, a task usually performed by the caregivers in real life. On the other hand, we find interesting to investigate up to what extent visual features can be used to infer higher level concepts such as utensils to be used for meal consumption.

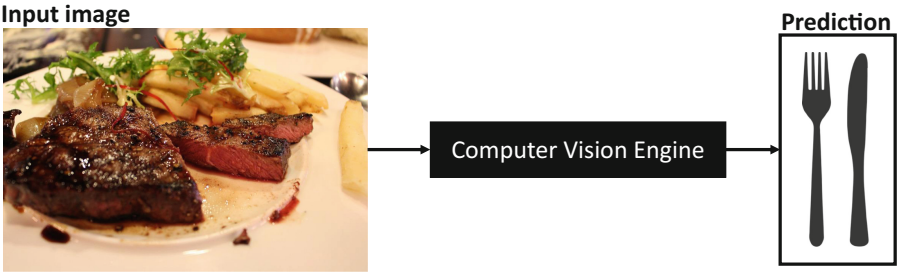

Fig. 1. The investigated problem.

To benchmark the problem, we consider the UNICT-FD1200 dataset [1]. To perform the experiments, each image of the dataset has been labeled according to five different classes related to the utensils to be used for meal consumption: *Chopsticks*, *Fork*, *Fork and Knife*, *Hands* and *Spoon*. We investigate an approach based on the combination of features extracted using the AlexNet CNN architecture proposed in [12] and a Support Vector Machine to perform classification [14]. This simple pipeline has obtained a classification accuracy of 86.27%.

The paper is organized as following. Section 2 summarizes the representation and classification components adopted to address the considered problem. Section 3 details the experimental settings and discusses the results. Conclusions are given in Sect. 4.

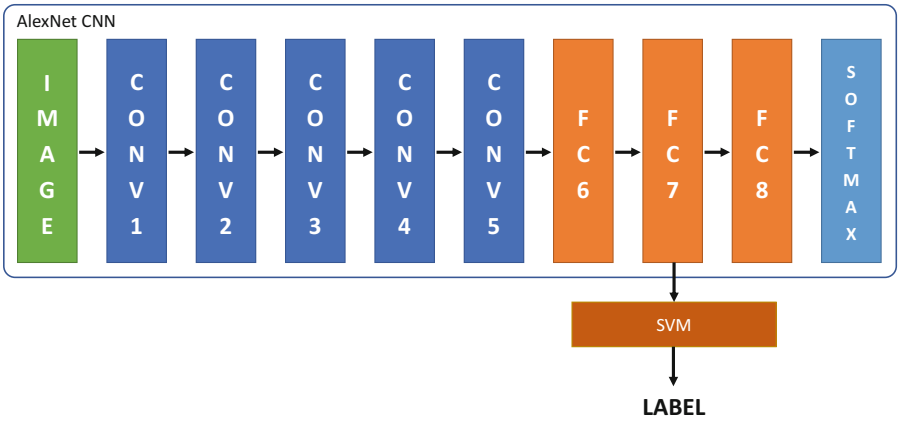

Fig. 2. The considered approach based on features extracted using the AlexNet CNN architecture and an SVM classifier.

2 Food Image Representation and Classification

In our experiments, we considered an image representation based on deep features. In particular, we use the AlexNet deep learning architecture proposed in [12]. The model has been pre-trained to categorize images from ImageNet into 1000 different object classes. The AlexNet architecture has 8 layers, plus a Softmax module at the end of the network. In our experiments, we used the activations of the fully connected seventh layer (FC7) as features. We choose to extract features form the FC7 layer since such activations are believed to have a high semantic relevance but are more general than the 1000 features of the FC8 layer which are to be considered as class-related scores. Classification is performed using a multiclass SVM classifier [14] with an RBF (Radial Basis Function) kernel. See Fig. 2 for a diagram of the approach.

The SVM has been trained on a balanced set of images with equal amounts of images for each of the considered class, i.e., *Chopsticks, Fork, Fork and Knife, Hands and Spoon*. Hyper-parameters (e.g., cost C and γ in the RBF kernel) are optimized using cross-validation.

3 Experimental Settings and Results

We consider the UNICT-FD1200 [1] for our experiments. The dataset contains 4754 images of 1200 distinct dishes of food plates characterized by different nationalities (e.g., English, Japanese, Indian, Italian, Thai, etc.). Each dish has been acquired with a smartphone several times to introduce geometric and photometric variability in the dataset (such as Flash vs. No Flash, different rotations, multiple scale, different points of view). To carry out the proposed investigation, each image of the dataset has been manually labeled considering the following classes: *Chopsticks*, *Fork*, *Fork and Knife*, *Hands* and *Spoon*. Examples of images belonging to the UNICT-FD1200 dataset are shown in Fig. 3, whereas the number of images belonging to each class is reported in Table 1.

To perform evaluation, the dataset has been randomly divided into three balanced non-overlapping subsets. The three different splits allow to obtain three independent training set/test set pairs. Once FC7 features are extracted for all

Fig. 3. Example of images belonging to the UNICT-FD1200 dataset. Each row correspond to a specific class: (1) Chopsticks, (2) Fork, (3) Fork and Knife, (4) Hands, (5) Spoon

Table 1. Per-class number of images in the UNICT-FD1200 dataset.

Class	Number of images
Chopsticks	248
Fork	1468
Fork and Knife	2048
Hands	699
Spoon	291

images in the dataset, the SVM classifier is trained and tested considering the three different split. Accuracy values over the three runs are hence averaged to assess overall performance. The proposed method is implemented using the Caffe library [13] to extract FC7 features form a pre-trained AlexNet model [12] and LibSVM [14] to implement the multi-class classifier.

Table 2 summarizes the results and reports the performances of the classifier in the different runs. In Table 3 the confusion matrix with respect to the five considered classes is reported. The approach obtains good results for the *Fork and Knife* class, probably because in the dataset there are more images than in the other classes. The method has difficulties in recognizing images belonging to the *Chopsticks* class (e.g., noodle plates), which are confused in those in which a fork utensil is used during meals.

Table 2. Accuracy of our classification model.

Run 1	Run 2	Run 3	Average
87.24	85.06	86.50	86.27

Table 3. Confusion Matrix. Rows report real classes, while columns report predicted ones.

Classes	Chopsticks	Fork	Fork and Knife	Hands	Spoon
Chopsticks	**60.87%**	6.52%	28.26%	2.17%	2.17%
Fork	0%	**86.47%**	11.88%	0.82%	0.82
Fork and Knife	0.27%	17%	**92.88%**	2.19%	0%
Hands	0.84%	2.52%	14.28%	**82.35%**	0%
Spoon	1.78%	7.14%	17.85%	1.78%	**72.43%**

4 Conclusions

We have considered the problem of recognizing utensils to be used during meal consumption. The investigation is both of practical interest (e.g., to design systems to assist people with mental disabilities) and theoretical interest (i.e., to

assess whether higher level concept related to how to eat food can be obtained from visual features). To address the problem, we augment the FD1200 dataset introducing labels related to utensils to be used to consume the food detected in the images. Experiments show that even a simple pipeline based on AlexNet features and an SVM classifiers can be leveraged to perform classification despite it can be considered only as a baseline approach to be improved.

References

1. Farinella, G.M., Allegra, D., Moltisanti, M., Stanco, F., Battiato, S.: Retrieval and classification of food images. Comput. Biol. Med. **77**, 2339 (2016)
2. Leo, M., Medioni, G., Trivedi, M., Kanade, T., Farinella, G.M.: Computer vision for assistive technologies. Comput. Vis. Image Underst. **154**, 1–15 (2017)
3. Razani, J., Wong, J.T., Dafaeeboini, N., et al.: Predicting everyday functional abilities of dementia patients with the mini mental state exam. J. Geriatr. Psychiatry Neurol. **22**(1), 62–70 (2009)
4. Vincze, M., Bajones, M., Suchi, M., Wolf, D., Weiss, A., Fischinger, D., da la Puente, P.: Learning and detecting objects with a mobile robot to assist older adults in their homes. In: Hua, G., Jégou, H. (eds.) ECCV 2016. LNCS, vol. 9914, pp. 316–330. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-48881-3_22
5. Yamazaki, K., Ueda, R., Nozawa, S., Kojima, M., Okada, K., Matsumoto, K., Ishikawa, M., Shimoyama, I., Inaba, M.: Home-assistant robot for an aging society. Proc. IEEE, Centennial Year, Special Issue, Quality Life Technol. **100**(8), 2429–2441 (2012)
6. Kanade, T., Hebert, M.: First-person vision. Proc. IEEE **100**(8), 2442–2453 (2012)
7. Damen, D., Leelasawassuk, T., Mayol-Cuevas, W.: You-Do, I-Learn: egocentric unsupervised discovery of objects and their modes of interaction towards video-based guidance. Comput. Vis. Image Underst. (CVIU) **149**, 98–112 (2016)
8. Soran, B., Farhadi, A., Shapiro, L.: Generating notifications for missing actions: don't forget to turn the lights off! In: IEEE International Conference on Computer Vision (2015)
9. Furnari, A., Farinella, G.M., Battiato, S.: Recognizing personal locations from egocentric videos. IEEE Trans. Human-Mach. Syst. **47**, 6–18 (2017)
10. Ortis, A., Farinella, G.M., Damico, D., Addesso, L., Torrisi, G., Battiato, S.: Organizing egocentric videos for daily living monitoring. In: Lifelogging Tools and Applications in Conjunction with ACM Multimedia, Amsterdam (2016)
11. Ragusa, F., Tomaselli, V., Furnari, A., Battiato, S., Farinella, G.M.: Food vs non-food classification. In: 2nd International Workshop on Multimedia Assisted Dietary Management (MADiMa) in Conjunction with ACM Multimedia, Amsterdam (2016)
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classication with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (2012)
13. Berkeley Vision and Learning Center (BVLC). Cae, http://caffe.berkeleyvision.org/
14. Chih-Chung, C., Chih-Jen, L.: Libsvm: a library for support vector machines (2001)
15. Singla, A., Yuan, L., Ebrahimi, T.: Food/non-food image classification and food categorization using pre-trained GoogLeNet model. In: Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management (2016)

16. Farinella, G.M., Allegra, D., Stanco, F., Battiato, S.: On the exploitation of one class classification to distinguish food vs non-food images. In: Murino, V., Puppo, E., Sona, D., Cristani, M., Sansone, C. (eds.) ICIAP 2015. LNCS, vol. 9281, pp. 375–383. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23222-5_46
17. Kagaya, H., Aizawa, K., Ogawa, M.: Food detection and recognition using convolutional neural network. In: ACM International Conference on Multimedia, pp. 1085–1088 (2014)
18. Shimoda, W., Yanai, K.: Foodness proposal for multiple food detection by training of single food images. In: International Workshop on Multimedia Assisted Dietary Management, pp. 13–21 (2016)
19. Merler, M., Wu, H., Uceda-Sosa, R., Nguyen, Q.B., Smith, J.R.: Snap, eat, RepEat: a food recognition engine for dietary logging. In: Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management, pp. 31–40 (2016)
20. Farinella, G.M., Moltisanti, M., Battiato, S.: Classifying food images represented as bag of Textons. In: IEEE International Conference on Image Processing (ICIP), Paris, pp. 5212–5216 (2014)
21. Martinel, N., Piciarelli, C., Micheloni, C.: A supervised extreme learning committee for food recognition journal article. Comput. Vis. Image Underst. **148**, 67–86 (2016)
22. Dehais, J., Anthimopoulos, M., Mougiakakou, S.: Food image segmentation for dietary assessment. In: International Workshop on Multimedia Assisted Dietary Management, pp. 23–28 (2016)
23. Shimoda, W., Yanai, K.: CNN-based food image segmentation without pixel-wise annotation. In: Murino, V., Puppo, E., Sona, D., Cristani, M., Sansone, C. (eds.) ICIAP 2015. LNCS, vol. 9281, pp. 449–457. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23222-5_55
24. Ciocca, G., Napoletano, P., Schettini, R.: Food recognition and leftover estimation for daily diet monitoring. In: Murino, V., Puppo, E., Sona, D., Cristani, M., Sansone, C. (eds.) ICIAP 2015. LNCS, vol. 9281, pp. 334–341. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-23222-5_41
25. Ciocca, G., Napoletano, P., Schettini, R.: Food recognition: a new dataset, experiments and results. IEEE J. Biomed. Health Inf. **21**(3), 588–598 (2017)
26. Myers, A., Johnston, N., Rathod, V., Korattikara, A., Gorban, A., Silberman, N., Guadarrama, S., Papandreou, G., Huang, J., Murphy, K.: Im2Calories: towards an automated mobile vision food diary. In: IEEE International Conference on Computer Vision, pp. 1233–1241 (2015)
27. Beijbom, O., Joshi, N., Morris, D., Saponas, S., Khullar, S.: Menu-match: restaurant-specific food logging from images. In: IEEE Winter Conference on Applications of Computer Vision, pp. 844–851 (2015)
28. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 446–461. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10599-4_29
29. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**(7553), 436–444 (2015)