# Food Recognition: Can Deep Learning or Bag-of-Words Match Humans?

Pedro Furtado[a]
*CISUC, Universidade de Coimbra, Polo II, Coimbra, Portugal*
*pnf@dei.uc.pt*

Keywords: Deep Learning, Bag-of-Words, Food Recognition.

Abstract: Automated smartphone-based food recognition is a useful basis for applications targeted at dietary assessment. Dish recognition is a necessary step in that process. One of the possible approaches to use is deep learning-based recognition, another one is bag-of-words based classification. Deep learning has increasingly become the preferred approach to use in either this or other image classification tasks. Additionally, if humans are better recognizing the dish, the automated approach is useless (it will be less error-prone for the user to identify the dish instead of capturing the photo). We compare the alternatives of Deep Learning (DL), Bag-of-words (BoW) and Humans (H). The best deep learner beats humans when on few food categories, but looses if it has to learn many more food categories, which is expected in real contexts. We describe the approaches, analyze the results, draw conclusions and design further work to evaluate further and improve the approaches.

## 1 INTRODUCTION

Food recognition from plates is a well-known problem in computer vision. It has important applications in dietary assessment for healthy lifestyles. One important step for food recognition is dish recognition, the capacity to detect which dish is presented in front of the "machine's eye". Using a smartphone and an adequate piece of software, it should be possible to identify the dish automatically. That capacity would relieve the user from having to identify the food manually, instead he would only point the smartphone at the plate in front of him and the dish would be classified automatically. This vision would be impractical previously, because machine learning classifiers were too inaccurate, but the last few years have seen exciting developments in image classification and object recognition based in convolution neural networks (CNNs). Naturally, these have been considered as an option for food recognition in the type of context that we described above. However, CNNs have to reach comparable or better accuracy than humans in order to be accepted as a possible substitute to manual specification of the plate contents by the person herself. We would not trust an automated food recognizer that would fare worse than humans. In that case we would instead ask the human. The relevant question to be answered is whether CNNs are better or worse than the more traditional machine learning approaches in the task of dish recognition, and whether they or the traditional approaches are worse or better than humans in that task. To evaluate this, in what concerns more traditional approaches, we needed to setup one of the most popular and accurate alternatives, Bag-of-Words (BoW) (Petraitis et al., 2017). Regarding deep learning, we needed a setup with state-of-the-art CNN architectures, and in what concerns humans, we needed to evaluate the capacity of a set of subjects to recognize dishes that were previously unknown to them.

In this paper we describe the architectures of the BoW pipeline and of the CNNs, and we review the design of the survey we created for testing human ability in food recognition. Then we describe and analyze the results we obtained, drawing conclusions and future work based on those results. The paper is structured as follows: section 2 reviews related work. Section 3 discusses methodology, describing how we have built a set of alternative approaches, including bag-of-words and deep learning architectures, the human survey, implementation and training details, plus experimental setup. Section 4 describes the experimental results and section 5 concludes the paper and previews future work.

## 2 RELATED WORK

Dish recognition is a reasonably challenging task for machine learning approaches. Most difficulties arise

[a] https://orcid.org/0000-0001-6054-637X

from food being a plastic material, in the sense that shape, colour, texture and mixtures change significantly and present huge variability. Examples of works using "traditional" machine learning classification approaches to detect food include (Yang et al., 2010), where statistics on spatial relationships between a few ingredients were explored, or (Matsuda et al., 2012), which detected candidate regions, then applied bag-of-features (BoF) on SIFT and CSIFT with spatial pyramid (SPBoF), histogram of oriented gradient (HoG), and Gabor texture features (56% accuracy with 10 types of foods).

More recent works apply deep learning to achieve higher accuracies. Deep learning-based image recognition makes use of Deep Convolution Neural Networks (CNNs). These systems first raised to prominence in the year of 2012, when Alex Krizhevsky, Ilya Sutskever and Geoffrey Hinton introduced Alexnet (Alom, 2018) as a new approach to recognize objects in images, in the context of the ImageNet Large Scale Visual Recognition Challenge (ILSVR) (ILSVRC, 2018). Its results beat those of then state-of-the-art approaches in object recognition, and they became frequent in any kind of task (e.g. gender recognition from facial images (Arora and Bhatia, 2018)). Two of the main features of CNNs is that they add convolution layers that process images to extract, filter and abstract features automatically, and they extend back-propagation based learning to those convolution layers to automatically extract information from images. Posterior CNN architectures perfected Alexnet to increase accuracy even further in image recognition tasks. Resnet (He et al., 2016) added residual blocks to handle vanishing gradients in first layers on back-propagation runs resulting from increasing depth. Inception (Szegedy, 2015) introduced sparse connections between layers, and it also uses different filter sizes to capture features with varied degrees of detail.

Authors using CNNs on the food recognition task obtained much better accuracies than previous approaches in general. For instance, (Kawano and Yanai, 2014) achieved 72.26% top 1 accuracy and 92% top 5 accuracy for 100 class food dataset. These results improve over previous machine-learning non-deep learning approaches, however it makes sense to ask whether humans would do better than the 72.3% accuracy reported in that work. In another work, (Pouladzadeh and Shirmohammadi, 2017) proposes a mobile multi-food recognition system using deep learning as well. Their approach requires the user to make a bounding circle around the food elements. It was applied experimentally to 7000 food images of 30 categories of the FooD dataset, showing an average recall rate of 90.98%, precision rate of 93.05%, and accuracy of 94.11% compared to 50.8% to 88% accuracy of other existing food recognition systems. The accuracy of this approach is impressing, however there are three details worth noting: (1) it requires the user to draw a bounding circle, therefore it is not directly comparable with approaches requiring no user interaction besides taking the photo; the dataset contains many easy to recognize food items (e.g. bananas); and it has only 30 categories, we will see that the number of categories has a huge impact on accuracy.

In (Yanai and Kawano, 2015) the authors examined the effectiveness of deep convolutional neural network (DCNN) for food photo recognition task, seeking the best combination of DCNN-related techniques such as pre-training with the large-scale ImageNet data, fine-tuning and activation features extracted from the pre-trained DCNN. From the experiments they concluded the fine-tuned DCNN which was pre-trained with 2000 categories in the ImageNet, including 1000 food-related categories, was the best method, achieving 78.77% as the top-1 accuracy for UEC-FOOD100 and 67.57% for UEC-FOOD256 (Foodcam, 2018), both of which were the best results so far.

When analyzing the results of (Yanai and Kawano, 2015), it becomes clear that the top accuracy is still far from 100%, (e.g. UEC-FOOD256 was 67.57%). Our own previous work on this issue involved setting up the survey and preliminary work on comparison (Caldeira et al., 2019). We want to compare BoW with CNN, and both with humans. Is BoW and is CNN sufficiently better than humans to replace them in the task of food recognition?

# 3 METHODOLOGY

Existing food recognition techniques fall under the categories of classification-based and deep learning-based approaches. A third category is human-based identification, whereby a human is asked to recognize each food. This third alternative is necessary for comparison reasons, since we would not trust an automated food recognizer that would fare worse than humans and would instead ask the human. In the following we discuss how the practical systems were created, so that we can compare them and reach conclusions. These approaches were all setup and ran in Matlab 2018a.

## 3.1 Bag-of-Words (BoW)

Bag of Word or Bag of Features (Csurka et al., 2004) is a machine learning approach for image category

classification by creating a bag of visual words. The process generates a histogram of visual word occurrences that represent an image. These histograms are used to train an image category classifier. BoW creates a visual vocabulary, or bag of features, by extracting feature descriptors from representative images (train dataset) of each category. Figure 1 illustrates the steps of BoW. The first step (a) involves feature extraction. In our case we created a custom feature extractor that would extract texture (GLCM, binary patterns), colour histograms, SURF and geometry features of regions obtained by thresholding the images into 6 levels. In BoW, the features extracted from each image are represented as a vector of features (descriptors). Given all vectors of features from all the training images, the k-means clustering algorithm is applied with k clusters (the vocabulary size) to automatically obtain k feature vector representative centroids, or visual words, step (b). The algorithm iteratively groups all descriptors into k mutually exclusive clusters. The resulting clusters are compact and separated by similar characteristics.

The images in the training dataset are then encoded into histograms of visual words using the created vocabulary (c), representing occurrences of the code-words. This way the characteristics of any image will be captured as a k-length histogram of occurrences of the k codewords. An artificial neural network classifier is then trained to classify images as one of the categories based on the histograms (d). After the BoW is created, given any new image to classify, the approach detects and extracts features from the image and constructs the codewords occurrence histogram for the image. The trained classifier is then used to classify the image as one of the categories.

Our BoW implementation in Matlab used the "bagOfFeatures" BoF object. The BoF object allows developers to write custom feature extractors. We created a customized feature extractor that, besides SURF, would compute texture (GLCM, binary patterns), colour histograms and geometry features (Matlab regionprop function) over regions obtained by Othsu's thresholding of the images into 6 levels.

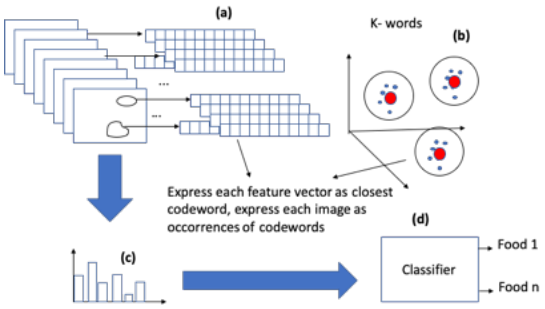


Figure 1: Steps of Bag-of-Word.

## 3.2 Deep Learning Architectures

Our objective setting up deep learning architectures for the food recognition task was to apply state-of-the-art standard DCNNs, while at the same time comparing their accuracy on the job. We have chosen GoogLeNet (Szegedy, 2015), Inception-v3 (Szegedy and Vanhoucke, 2016) and Resnet101 (He et al., 2016). The Resnet architecture is an important milestone in the history of deep learning, since it demonstrated that extremely deep networks can be trained using standard SGD through the use of residual modules. Before Resnet, deeper networks suffered from vanishing gradients problem during backpropagation learning, which severely limited the number of layers in practice. To improve on this problem, Resnet residual modules are (local) micro-architectural blocks that add the identity locally (a layer feeds into the next layer but also directly into the layers about 2–3 hops away). This enables very deep networks to learn classifications appropriately, minimizing the vanishing gradients problem. Figure 2 shows the Resnet architecture (we show the smaller 50 stages version for illustration). At each stage in Figure 2 we can see a set of layers (left), and the adding of the output of the stage to prior layer (center).

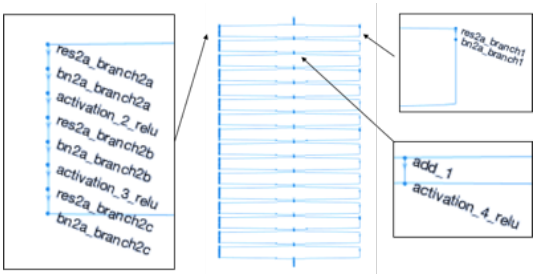


Figure 2: Resnet architecture.

The Inception architecture introduced another micro architecture pattern, the inception module. The inception module is a multi level feature extractor by computing 1x1, 3x3, and 5x5 convolutions within the same module of the network. The output of these filters are then stacked along the channel dimension and before being fed into the next layer in the network. Googlenet was the initial incarnation of this architecture, improvements were then named InceptionVx. Figure 3 illustrates the graph of InceptionV3 architecture.

The number of layers of the tested Resnet, inception and googlenet architectures was 101, 48 and 22 respectively. The CNN training setup for each of the architectures was simple: Imagenet pre-trained networks of each were adapted to train and classify either all 256 or 16 categories of the food dataset UEC-

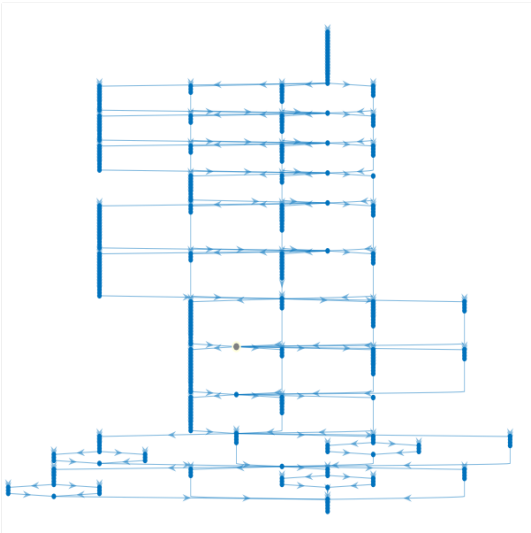

Figure 3: Sketch of InceptionV3 architecture.

FOOD256 (Foodcam, 2018) followed by training for a large number of epochs to ensure convergence (see experimental setup section). We used transfer learning. To do that the fully connected layers of the pretrained networks were replaced by ones that would classify 256 or 16 types of food. Softmax activation outputs probabilities of each class. The training epochs were configured as 300 (with the option to stop manually if, upon visual inspection, convergence into final stable accuracy is observed). The learning rate was set initially to 0.05, validating every 4 iterations, and we verified that accuracy would stabilize/converge in every run.

The trained CNNs adapted their coefficients to classify food classes, the resulting models were then used to classify food as one of those types. The CNN procedure for classifying new images consists of reading the new image of food to be classified, extracting features automatically through convolutions layers that apply learned weights. This procedure ends with the output which is a set of probabilities that the food submitted could be of each one of the types learnt before. For evaluation of accuracy, an independent test dataset is used, with images that are submitted to the system for classification, but of which we already know the correct classes.

### 3.3 Humans Survey (HVS)

The survey on humans was based in human subjects that were trained to classify Asian food dishes they did not know about, both sexes and ages between 18 and 55 years old (respondents were European and only those not knowing at least 80% of the dishes were accepted) (Caldeira et al., 2019). The dataset also included two universal food items such as pizza. The challenge then is to train the subject to be able to pick the correct name of food dishes that are presented to him, and the accuracy is measured as the percentage of right choices. Next we review some of the details of the survey.

Human training was designed based on a sequence of screens shown to teach to recognize the food types of the dataset. There were 32 screens designed to teach 16 different categories of food, a number that was deemed sufficiently small to allow humans to keep attention and simultaneously not too small to be too easy. Humans were asked to answer a quiz as the learning process. The person was expected to figure out what a certain class of food is like for the screen that he is evaluating. An example screen is illustrated in Figure 4.

Figure 4: Example of training screen.

The person needs to decide which of the 8 images is not a certain food class, and the correct name of the dish is presented as the title of the image. After thinking about the answer, the subject clicks to see the answer. Given the correct answer, the respondent improves his knowledge about the specific food class presented in that screen. In Figure 4 all dishes except one are a specific category identified by the title of the image. The one marked with an x belongs to another type of food. Given the universe of 256 types of food of the UEC-FOOD256 dataset that we used, 16 types of food were chosen for the experiment with humans, based on equally spaced percentiles of accuracy of the DCNN classifying the same data. This way both the human subjects and the machine were put before dishes that were “easy”, halfway and difficult to identify by the best-performing CNN. The classification session for humans is based on 32 screens (two of each food class) that the person needs to identify as one of the types available (16 types of food).

Figure 5 illustrates the classification quiz. In the figure we can see two example screens with images of dishes and image identification numbers, and a sheet of paper where the respondent is expected to write each identification number in the row of the correct

food name. The respondent is shown a screen, fills the image identification number on the correct row of the questionnaire and then follows to the next screen.


Figure 5: Example of testing screen.

### 3.4 Additional setup

The machine used for the experiments was configured as follows: PC, windows, processor was an Intel i5 at 3.4 GHz, RAM 16 GB, SSD 1TB; NVIDEA GForce GTX 1070 GPU, having 1920 cores, GDDR5 8 GB, memo speed ofg 8 Gbps). As an example, the training time for Resnet101 was more than 5 days (7978 mins) for 256 food categories and almost 5 hours (277 mins) for 16 categories.

The dataset used – UEC-FOOD256 (Foodcam, 2018) – is a publicly available food dataset consisting of around 31000 food images organized into 256 classes of food. These are physically organized into different folders named with the dish name. All food categories have more than 120 representative samples (images), although some categories have many more samples than others.

We designed different setups of the dataset to allow comparison with the results of the human survey. Note that the human survey was based on a restricted set of 16 categories and only 32 images of each. The corresponding setups were: Food16 = a restricted version of the dataset with the same 16 categories of food as used in the human survey, but keeping all images of each of those categories (120 or more of each category). Training CNNs and BoW with only these 16 categories allow direct comparison with the results of the human survey; Food256 = 256 foods, full dataset. CNNs and BoW were trained with this full version of the dataset, which allows comparison among them and also with the results of human survey. Note that, although the survey trains and tests only 16 food categories for practical reasons, adult respondents to the human survey know of thousands of food items and dishes, therefore it makes sense to compare with CNNs or BoW having to learn from many food categories as well;

The labels assigned to the techniques tested and compared are: Resnet101 (R101), InceptionV3 (Iv3), Googlenet (G), BoW500 (bag of features with a vocabulary of 500), BoW1000 (bag of features with a vocabulary of 1000), and Humans (H, the results of the survey). For the automated techniques, 5-fold cross-validation was used.

## 4 RESULTS

In this subsection we show the experimental results, which are analyzed and discussed in detail later in the next subsection. Table 1 shows the accuracy each technique achieved.

Table 1: Comparison of accuracy (technique and data setup).

| Approach | Food16 | Food256 |
|---|---|---|
| H | 83% | - |
| R101 | 93% | 73% |
| Iv3 | 92% | 67% |
| G | 89% | 56% |
| BoW500 | 59% | 49% |
| BoW1000 | 63% | 53% |

Tables 2 shows a few per-class precision values, as classified by humans (H 16), InceptionV3 (Iv3 256) and Resnet (Res 256).

Table 2: Some food classification scores).

| Approach | Pizza | Fried fish | Sashimi | Sweet S pork |
|---|---|---|---|---|
| H 16 | 96% | 59% | 96% | 67% |
| Iv3 256 | 10% | 75% | 36% | 40% |
| Res 256 | 40% | 58% | 45% | 50% |
| | Natto | Stir fried beef | Steam dumpling | chinese soup |
| H 16 | 96% | 69% | 83% | 96% |
| Iv3 256 | 90% | 92% | 33% | 67% |
| Res 256 | 100% | 92% | 50% | 25% |
| | Laksa | Mie goreng | Nasi campur | Curry puff |
| H 16 | 44% | 63% | 85% | 92% |
| Iv3 256 | 92% | 80% | 91% | 69% |
| Res 256 | 83% | 93% | 91% | 85% |

### 4.1 Analysis of Comparison Results

The results in Table 1 show that the accuracy of Bag-of-Words (BoW) was not competitive with that of either CNNs or humans. Increasing the number of codewords did improve the results slightly, but performance of BoW was still much lower than CNNs or humans. Table 1 also shows that Humans achieved 83% accuracy while, for 16 food types, the best CNN achieved 89% to 93%, depending on the CNN architecture used. The conclusion is that, given a small number of food categories to learn, CNNs are very accurate and can surpass humans.

A more realistic application of CNNs for food recognition requires learning thousands of food categories. Table 1 also shows that CNN accuracy with 256 categories of food (Food256) decreases significantly, in spite of having the full dataset to train with. The best performing CNN, which was R101, achieved 73% (down from 93% on 16 categories), Iv3 and G achieved 67% and 56% respectively. As the number of food categories to learn increases, accuracy of all tested CNN architectures decreases significantly. Practical food recognition systems should be able to learn thousands of food categories instead.

Table 2 shows that the most difficult categories for CNN and for humans are different. The conclusion is that humans and CNNs behave very differently classifying food types.

### 4.2 Conclusions from Experiments

Analysis of the results confirm that CNNs are very accurate, and much more accurate than BoW. But the results also show that when CNNs had to learn 256 food categories they were less accurate than humans. As a consequence, in realistic contexts with many food categories and prior knowledge, humans are expected to still be more accurate than CNNs. There is however a need to investigate the comparison to humans in more detail. Note also that other factors, such as illumination, perspective, occlusion and others, will further influence automated recognition capacity negatively, and other human capabilities and contextual information may aid humans improving their guesses in real environments.

## 5 CONCLUSIONS AND FUTURE

The promise of smartphone-based capturing of the dish to be eaten for dietary assessment makes it important to evaluate feasibility. This work analysed bag-of-words (BoW) and deep learning-based solutions for food recognition (CNNs), comparing them to humans as well. The approaches were compared experimentally and we analyzed the results. This allowed us to conclude that CNNs beat BoW significantly. But we also concluded that CNNs accuracy decreases when they have to learn more food categories. Our current and future work related to this issue is focused on the need to analyse this issue in more detail, evaluating how deep learning compares with humans, the deficiencies, and how to improve the approaches.

Future work on practical food recognition systems can ask the user which of top-3 possibilities is the right one, since the top-3 accuracy should be much higher, based also on results by other authors. Another line of work concerns feeding different kinds of contextual information to the CNN classifier stage, both during training and use, to improve automated classification accuracy.

Finally, it is very important to experiment with difficulty inducing-factors, such as bad illumination and shadows, perspective, occlusion and others, which will further influence recognition capacity negatively.

## REFERENCES


Alom, Z. M. (2018). The history began from alexnet: A comprehensive survey on deep learning approaches. In *in ArXiv Preprint ArXiv:1803.01164,.*

Arora, S. and Bhatia, M. (2018). A robust approach for gender recognition using deep learning. In *2018 9th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–6. IEEE.

Caldeira, M., Martins, P., and Cecílio, J. (2019). Comparison study on convolution neural networks (cnns) vs human visual system (hvs). In . *BDAS 2019: 111-125. Beyond Databases, Architectures and Structures. Paving the Road to Smart Data Processing and Analysis - 15th International Conference, BDAS 2019,*, pages 978–3. Ustro, Poland,.

Csurka, G., Dance, C., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague.

Foodcam, U. (2018). Uec food dataset. *[URL accessed in 10/2018] [http://foodcam.mobi/dataset256.html].*

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *In Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

ILSVRC (2018). Large scale visual recognition challenge. *[URL Accessed 10/2018] http://www.imagenet.org/challenges/LSVRC/.*

Kawano, Y. and Yanai, K. (2014). Food image recognition with deep convolutional features. In *In Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (pp. 589593). ACM. https://doi.org/10.1145/2638728.2641339*.

Matsuda, Y., Hoashi, H., and Yanai, K. (2012). Recognition of multiple food images by detecting candidate regions. In *In Multimedia and Expo (ICME), 2012 IEEE International Conference on*, pages 2530–10.

Petraitis, T., Maskeliūnas, R., Damaševičius, R., Połap, D., Woźniak, M., and Gabryel, M. (2017). Environment recognition based on images using bag-of-words. In *Proceedings of the 9th International Joint Conference on Computational Intelligence - IJCCI,*, pages 166–176. INSTICC, SciTePress.

Pouladzadeh, P. and Shirmohammadi, S. (2017). Mobile multi-food recognition using deep learning.". In *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM) 13.3s*, page 3610.

Szegedy, C. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition.*

Szegedy, C. and Vanhoucke, Vincent Sergey Ioffe, J. S. Z. W. (2016). Rethinking the inception architecture for computer vision. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR),*, pages 28182826–10.

Yanai, K. and Kawano, Y. (2015). Food image recognition using deep convolutional network with pre-training and fine-tuning.". In *IEEE International Conference on Multimedia & Expo Workshops (ICMEW).* IEEE,.

Yang, S., Chen, M., Pomerleau, D., and Sukthankar, R. (2010). Food recognition using statistics of pairwise local features. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2249–2256. IEEE.