ALTERNATOR
Misliti znanost.

The State of the Art of Automated Food Recognition

26. 4. 2021

Number: 25/2021

Author:

- Simon Mezgec

Photo: Arne Hodalič

With the development of new recognition technologies, the automatic recognition of food in photographic images has become an increasingly popular computer vision problem. Applications that emerge from the study of this problem include the automatic recording of food intake which could be used as a tool for improving the quality of nutritional habits. Traditionally, methods such as food journals and nutrition questionnaires were employed to record food intake. However, such methods require the manual recording of the type and amount of food consumed as well as the method of preparation and other information. Moreover, these methods often fail to guarantee an exact record of consumed foods as people tend to make manual errors, such as misreporting quantities, and often lose motivation to diligently record their intake. The automatic recognition of food images reduces the skill and time required to record consumed foods and removes the subjective human element from this task. Computer algorithms can now automatically recognize food in images, log the nutritional data of the depicted food, and analyse the nutritional quality. These procedures could be utilized with smart device applications that enable both the recording and analysis of food consumed. This article will explain how the automatic recognition of food images works, the beginnings of these technologies, and the current state of food recognition.

Beginnings (and subsequent difficulties) in the recognition of food images

The recognition of food images is a problem that belongs in the category of computer vision. The discipline of computer vision began in the 1960s and has recently become popular due to the promising results that have been achieved in the field. The goal of computer vision is high-level understanding of images and videos that computer algorithms acquire by processing input images or videos with various methods such as object detection, object recognition, and image segmentation, and the transmission of intelligible data about the image or series of images to users. In object recognition, the algorithm classifies objects in images into a number of output categories. In the case of food recognition, these categories represent individual types of food and beverages. Classification takes place after the algorithm attributes *features* to an individual object, and this makes it possible to reliably distinguish the individual object from other objects. In the past, this process posed the greatest problem to computer vision technology because the features of objects (colour, identifying the edges of the object, size, etc.) must first be defined by accounting for all possible visual variations of objects, and then these features must be interpreted in order to place the individual objects into the appropriate category.

The problem of food image recognition is one of the more complicated challenges in computer vision because of the visual properties of food. The following is a list of factors that increase the complexity of the task:

- The same food may appear in many forms in different images, which is a result of both the preparation of the food and the way the photograph was taken.
- Different foods or dishes may appear similar in images, making it difficult to effectively distinguish them from each other.
- Much of the visual information about the raw ingredients in food is obscured in the preparation of a finished dish.
- Beverages typically offer very little visual information – generally only colour, quantity, and the way in which they are served.
- There is an enormous number of different foods and beverages.

Other areas of computer vision, such as iris recognition, car detection, the tracking of athletes across a field, or quality control of a given manufacturing process, do not have the same difficulties or, in any case, have fewer of them. Therefore, the question is how can an algorithm, independent of human intervention, recognize what kinds of foods are depicted in a given image?

Researchers in this field initially developed the idea of the *manual definition of features* of various foods that the algorithm should identify in the image. For example, if they wish to identify a tomato, they explicitly define the shape, colour, and texture of a tomato, as well as other elements of the image and its background. This data is input into the code of the algorithm, which then proceeds to employ these definitions to search for corresponding objects in images. This process is then repeated for each kind of food. However, as described above, the visual appearance of the same kind of food may radically differ from one image to the next, especially if photographs are taken with smartphones under less than ideal conditions. Something round in one image may appear oblong in another, red may look like orange, and a smooth surface in one image may appear rough in another. Due to this visual diversity, the process of manually defining features was successful when the approach was used with a series of images selected specifically for the purposes of the research, but did not achieve sufficient accuracy with images from the real world because researchers could not anticipate the diversity of visual variations that a single kind of food assumed in different images as there are too many of them, and consequently the method of manually defining features achieved a classification accuracy of well under 40 % (https://ieeexplore.ieee.org/document/5539907) , and this when applied to images that feature only one food item (as in the photograph below). Unfortunately, these results are not sufficient for practical use.

An example of a photograph featuring one food item. This type of photograph is much less frequent than images depicting several food items. As regards this specific photograph, automatic recognition technology would provide a simple textual label "French fries".

Neural networks to the rescue

The best results in the recognition of food images were achieved with approaches that employ *artificial neural networks*. As the name suggests, artificial neural networks imitate the activities of neurons in the human brain. Artificial neural networks are made of layers of neurons, through which input data travels: in our case, images of food. Each layer of the neural network learns to interpret the features of images with various levels of complexity. The initial layers interpret basic features, such as shapes and object edges, while the deepest layers interpret more complex features, such as the type of food. It is important to note that all of these activities take place automatically. Neural networks determine which features are necessary to identify objects and "learn" them on their own, which means that researchers do not need to manually define the features of objects (as is the case with the previously described approach), but only need to define the output categories (types of food). Neural networks learn in such a way that the network processes a relevant dataset of images, modifying the connections between neurons during each step of the training process to minimize the so-called loss function. This function evaluates the accuracy of the neural network's descriptions of different types of foods, with the evaluation taking place on a dataset of images that differs from the training dataset. For the purposes of neural network training, images are typically divided into three kinds of datasets: a training dataset, which is employed to learn the features, a validation dataset, which validates the accuracy of recognition at each step of the neural network training process, and a test dataset, which is used for a one-time evaluation of the accuracy of recognition at the conclusion of the training process.

Artificial neural networks have existed since 1958. Until recently, their size was the primary reason for their inefficiency at recognizing objects in photographic images. The training process of a neural network is computationally extremely complex, and the consequence of this complexity was the development of neural networks with a relatively small number of layers, which were not capable of learning a sufficient number of features in an image. In the last decade, however, due to the increasingly powerful computational capacities of computers, in particular graphics processing units (GPUs), and more efficient approaches in general, deeper neural networks – or more broadly *deep learning* – have been developed. These networks are composed of a much greater number of layers, in some cases, over a thousand layers, which can mean over ten billion connections between neurons. As a result, these new networks are capable of learning an enormous number of features and therefore more effectively differentiating between various types of food. Nevertheless, this technology is considered a form of narrow artificial intelligence, meaning that it is limited to a narrowly-defined task – in this case, effectively analysing images of food – but cannot solve other tasks. Today, two kinds of deep neural networks are particularly popular: first, *recurrent neural networks* that contain a temporal component and are appropriate for problems such as natural language processing and speech recognition; second, *convolutional neural networks* (CNNs) that are used for image recognition and contain convolutional layers that search for local features in images. CNNs are composite functions that are implemented in such a way that they contain a great number of operations (the multiplication of matrices) in each neuron layer. With the help of faster GPUs, CNNs learn much more quickly and effectively, and concurrently compute an extremely large number of operations on an enormous number of processor cores. With deep CNNs, researchers have been able to achieve accuracy of over 90 % (https://ieeexplore.ieee.org/document/8354172) in the recognition of food images.

The recognition of multiple food items per image

The results mentioned above apply to the recognition of one food or beverage item per image. With images that contain two or more food or beverage items, such as the one displayed below, the use of single-output CNNs is no longer adequate. Fortunately, deep learning can be applied to more advanced forms of recognition, the task of which is to identify each and *every pixel in an image*. When we know which pixels correspond to the object of beef broth in an image, for example, there is no limit to the number of different food and beverage items that can be identified in an image. This kind of recognition supplies the neural network with additional information so that the neural network can recognize where a specific food item appears in an image and thus can more easily determine what kind of food is depicted.

An example of an image containing multiple food items. Left: the original photograph. Right: the result of recognition on the pixel level (the recognition of each food item in the photograph).

If deep neural networks are computationally extremely demanding, this is even more the case for deep neural networks that identify images on the pixel level. Indeed, the latter require a longer training time by one order of magnitude. The use of this approach has thus only become popular in the recent past (https://pubmed.ncbi.nlm.nih.gov/29623869/) with the further improvement of computational capacities. This technology is now considered the most promising approach because it is not limited by the number of food items in an image or by the way the image is captured, and enables automatic calculation of the quantity or mass of different food items in an image because it is known what portion of the image is taken up by each food item. An international Internet-based tournament, the Food Recognition Challenge (https://www.aicrowd.com/challenges/food-recognition-challenge) , was recently held in order to evaluate various technologies for food image recognition on the pixel level. The most successful algorithms attained results with an accuracy of approximately 60 % (https://www.aicrowd.com/challenges/food-recognition-challenge) . This lower level of accuracy, compared with the classification accuracy of solutions limited to only one food item per image, is the result of the complexity of the problem – namely, it is much more difficult for neural networks to successfully identify each pixel than for them to identify the image as a whole. Nonetheless, these results are encouraging and will certainly improve in the coming years. It is not outside the realm of possibility that at some point in the future deep learning will attain an accuracy of recognition that *exceeds human recognition*, a development that has already occurred in other areas of computer vision, such as the recognition of geographical location in an image.

Data dependency

Deep neural networks have a significant disadvantage. The accuracy of their recognition ability is dependent to a great degree on the number and quality of images in the network's training dataset. If we offer a neural network only a few images, it will, in our example, only learn to recognize a few basic features of foods. A truly staggering number of images is required for a sufficient level of learning – tens of thousands, hundreds of thousands, or even millions of images. In fact, deep neural networks are so complex that, in a number of areas of computer vision, no upper limit for the optimal number of images used in the training process has been discovered: *the more images, the better*. It is interesting to note that, in the development of deep learning solutions, the human element is still crucial because the images used for training are in large part gathered and annotated manually by humans, and not by the neural networks themselves. Not only is the total sum of images important, but it is also crucial that the dataset of images has an adequate level of diversity to account for every type of food and beverage. Moreover, an appropriate dataset should include the greatest possible level of visual variation that an individual object of recognition can assume in a given image. For example, if the collection of images is composed only of those that depict food prepared and served in a particular way and taken in natural sunlight, the algorithm will only effectively recognize images created under the same conditions. If the same food was photographed under different conditions, the algorithm would most likely fail to effectively recognize the image. At issue here is the problem of *overfitting*, a broad and significant challenge to the development of solutions using deep neural networks. In the case of a too small or too specific dataset of images, overfitting can be avoided by an early termination of training – this way, the neural network learns fewer features that are found only in the training dataset of images.

The future of food recognition technology

Despite the fact that current results in food recognition technology are encouraging, there is still much room for improvement. Deep learning represents the most accurate approach but still lags behind human recognition. The reason for this lag is the absence of robustly developed solutions. Food in images appears in an enormously varied amount of forms, a fact that makes solutions in this field heavily dependent on the size and diversity of the image datasets used for training. Today the datasets do not yet satisfy the appropriate conditions, but researchers believe that the expansion and diversification of these datasets will lead to improvements in the accuracy of food image recognition. In other words, future improvements in accuracy do not require innovation or the development of new methods but rather the acquisition of a greater amount of higher quality input data.

Translated by Erica J. Debeljak (https://www.alternator.science/sl/avtorji/erica-johnson-debeljak/)