Enhancement of Food Images via Machine Learning

Ali Eren Gokcelioglu[1]
aegokce16@earlham.edu
Department of Computer Science
Earlham College
Richmond, Indiana

ABSTRACT

When ordering food, people rely on visuals. This creates the need for food based businesses to have aesthetically pleasing images of their food. However, there are no tools to automate the process of taking professional quality pictures of food. For many small food based businesses, this can be a roadblock to competing with large chain restaurants that have professional photographers. We propose a novel way to automate this process. We use Convolutional Neural Networks and photographical heuristics to develop a ranking algorithm that grades the aesthetic quality of an image of food. Next, we devise algorithms to process the images. Using the ranking algorithm, we pick the best looking ten images to be processed again. This process is repeated several times. At the end, the best looking ten processed images are returned to the user.

KEYWORDS

Machine Learning, Convolutional Neural Network, Image Processing, Food Aesthetic, Photography Heuristics
.

1 INTRODUCTION

While making decisions about food, people rely on visual cues as much as textual cues[13]. With increased commercialization of food ordering, recommendation, and delivery systems, there is a large amount of commercial demand in the food sector for high quality images of food. As a result, many people want to take delicious looking pictures of food [6]. However, many of them do not know how to do this. Although widespread accessibility of digital cameras and smartphones has given everybody the ability to take pictures, this does not guarantee that the images will look good. There are several aspects to taking a good picture, such as composition, lighting, color, and focus.[10]. The problem is that most people do not how to properly apply these techniques[6]. This creates a demand for automating the process of taking aesthetically pleasing images of food.

While there is research in grading the aesthetics of images of food, and research in automatically processing generic images to look better, there is currently no published work on using machine learning to automate the task of improving the aesthetics of images of food. To close this gap in research, we suggest an automated system to improve the quality of images of food. Our system will take a video displaying the food from a variety of angels as an input. It will then apply a variety of image processing techniques to the image. With the help of a machine learning algorithm that grades how good an image of food looks, the system will return ten images of the dish that are the best looking. This will allow businesses to generate professional looking images of their food without the assistance of professional photographers. This will especially be helpful to small businesses, who want to compete with larger chain companies with access to the resources to hire professional photographers.

In the following section, we will examine related work in the field. This section will be divided into three subsections according to the researches focus: detecting images of food, labeling aesthetic assessment of food images, and processing images to look more aesthetic.

In the section after that, we will demonstrate the design and overview of our framework. We will start by a general overview, and then continue by explaining each module in detail. These modules are: positional module, food detection, and food Enhancement. After that, we put forward the experiment design.

Following, we will describe the budget and the timeline for the proposed project.

2 RELATED WORK

In this section we will discuss notable research related to the goal at hand. As mentioned, there is no published work in enhancing images of food specifically. The first subsection of this section introduces research in detecting the location of food in images. The following subsection summarizes notable work in grading the aesthetic quality of images of food. The final subsection presents published work on enhancing generic images.

2.1 Detecting Images of Food

Detecting types of objects is currently one of the most focused areas in computer vision. In this subsection, we will introduce research on detecting regions of food in images. Solutions using CNNs and SVMs have both been attempted, although the mainstream consensus seems that CNNs yield better results. The success rate rapidly drops when images of multiple dishes are introduced.

Yang et. al. argued that commonly used techniques focusing on local or global features are not useful in detecting images of food due to their varying shapes[14]. They instead suggested an SVM system where each pixel is grouped into a category of ingredients such as chicken, bread or butter. To increase accuracy, rather than assigning a category to each pixel, they associated each pixel with a probability vector containing probable ingredients. They used Semantic Texton Forest to characterize pixels. For a dataset, they used the PFID[1]. Their work achieved 80% accuracy in recognizing the type of food, and 98% in detecting if there is food in the image. However, the PFID only contains images of fast food that

CS388 Methods for Research and Dissemination, , Earlham College

are similar in shape and ingredients, and their methodology might yield less successful results in a dataset with larger variety. Kagaya et. al. found that SVM trained on color features confuse different types of food with similar color features[5]. For a dataset consisting of more diverse cuisine, they found that CNNs outperform SVMs significantly, with a 93% success rate vs. 60% respectively. The previous works assumed that only one type of food is present in the image. For images with more than one type of food, Matsuda et. al. generated an algorithm where they fuse outputs of several region detectors to draw candidate regions of food[11]. To do this, they used four different candidate region detection methods: whole image, the deformable part model method, a circle detector, and the JSEG region segmentation. Next, they apply a feature-fusion-based food recognition method for bounding boxes of the candidate regions. Their work had a 55.8% detection rate for multiple foods.This is not a very high rate compared to similar research. They do not mention how they construct their dataset. Matsuda et. al. have a recognition rate of only roughly half, and the Yang et. al. can only detect one type of food[11][14]. To detect multiple types of food in the same image with high accuracy, Kawano and Yanai created a system where the user encircles the food themselves, significantly decreasing computational cost and implementation difficulty[8]. They argued that this is a good way to detect food on smartphone apps, which have lower computational power and may decrease data cost. While they do not publish the success rate of humans at detecting food, it is presumably high. The obvious downside is that this method requires human input.

In this proposal, we want to use the methods developed by Kagaya et. al. to achieve what Matsuda et. al. tried to achieve[5][11]. Unlike Kawano and Yanai, we want to do this without human input[8]. To do this, we will train a CNN to detect if there is food in the image. Algorithms based on this approach are publicly available, and we will use one of those. To locate the location of the food in the image, we will use an algorithm that crops the image, and then passes it to the AI to detect if there is food in cropped image. We will do this repeatedly until we find all the parts of the image with no food. The rest of the image is the part with the food. The coordinates of these parts will be returned. Although Yang. et. al. has the highest detection rates, will not use their methods as there is no evidence their methods would work on images with a wider variety of food in them[14].

2.2 Labeling Aesthetic Assessment of Food Images

In this section we will present the literature on labeling how aesthetically pleasing an image of food looks. There are two papers of significance that we will go over. Both of them compare CNNs and SVMs, and find that CNNs outperform SVMs. However, both of the proposed methods only return whether an image looks good or not, and not how good an image looks.

Kekai et. al. used four different vision learning algorithms to compare how well they did on assessing the aesthetic quality of images of food, using the Gourmet Photography Dataset[13]. To quantify the machines success, they calculated how consistent the machine is with human experts. They compared the results to curations by human experts. They used three different SVM algorithms. The most sucesfull method is a GDP supervised CNN. CNNs outperformed the best SVG method (VGG-food + SVM) by 71.16% vs. 60.06% in positive assessment, but underperformed in negative assessment by 75.16% vs. 75.66%. It is important to note that human experts performed 72.10% on positive images, and 81.02% on negative images. To improve the performance of their CNNs, they pre-trained their networks on ImageNet. Zhang and Chen on the other hand, found different results[15]. They found that an SVM trained on nine local features (layout, GLCM, color moment, CCV, color histogram, HOG, SIFT, SUF, ORB) achieved a success rate of 99.63%, compared to a 96.67% random forest and 95.80% for their best CNN (ResNet). However, this research used only 1067 images, selected randomly from the Yelp dataset, which is a much smaller sample size. They used random image processing techniques on the images they have to enlarge their dataset. This makes the research less reliable then Kekai et. al[13].

We will use the GDP trained AlexNet based solution devised by Kekai et. al. We have gained access to their improved dataset, which contains twice as many curated images as the dataset they used for their paper. To further improve the performance of their research, we will use the Yelp dataset consisting only of images of food used by Zhang and Chen to pre-train our neural network, rather than generic ImageNet pictures. We will also use data augmentation methods, as devised by Kawano[7]. Furthermore, we will use the confidence interval of the CNN to grade how good an image is, rather than just whether it is good or not. This will be used to rank images in a specific order, rather than just group them as good looking or bad looking like Kekai et. al and Zhang and Chen did.

2.3 Processing Images to Look more Aestethic

There are three major works on cropping and retargeting generic images to look better. In this section we will sum up those works. Cropping and retargeting images to look better seems to work reasonably well, even when compared to image manipulations by human professionals. This can be done both with and without the help of machine learning.

Liu et. al. suggested that making images look better is possible without the use of any machine learning, by just cropping and retargeting the image[10]. They developed an aesthetic grading approach, using well-known photography heuristics. By using these heuristics, they picked salient region in the image, and pick the pixels that score highest with their approach. They then cropped and resized the best looking part of the image. They quantified this approaches success by asking human test subjects to grade the images: 93% of humans agreed that the cropped image looked better than the original, and 83% said that the machine-cropped image is similar in quality to an image cropped by a human professional. The main limitation with this work is that cropping certain regions could fundamentally changed the meaning being conveyed by a picture, and the research does not control for that.

To deal with this limitation, Guo et. al. have developed an algorithm that uses the same heuristics and techniques, but penalizes changes in the structure of the image to maintain the image as faithful to the original as possible[4]. They do this by developing a measure of structural similarity called SSIM. Their algorithm carves out seams, and inserts the same number of seams in the image to maintain structural similarity. The algorithm checks for high SSIM during the insertions to pick the least image altering insertion. However, they did not test their algorithm against human experts, nor have they used human testers to measure whether the image has actually been made to look more aesthetic. A major limitation of this work is that seam carving breaks images with complex structures, making it obvious that the image has been altered.

These two methods rely on heuristics used by humans, but there is an intuitive component to taking good pictures as well. Chen et. al. used machine learning to mimic this intuition[2]. To do this, they used a CNN with no handcrafted features. They used a dataset consisting only of professionally taken images, and devise a ranking algorithm where two different parts of the image are cropped, and aesthetically more pleasing one is rewarded. They used AlexNet and AVA for deciding which cropped image should be chosen[12][9]. They compare their dataset to various heuristic based cropping algorithms, and find that their method slightly outperforms handcrafted heuristic based cropping methods[2].

Our proposal combines the heuristics aspect of Liu et. al.[10] and Guo et. al.[4] with the AI implementation of Chen et. al.[2], and applies it to images of food specifically. Instead of an image, our algorithm will take a video as an input, and we will use positional heuristics on each frame to decide which food has a better position. The best looking frames will be passed to a Convolutional Neural Network similar to Guo et. al., however ours will be trained on curated images of food, rather than generic images.

3 DESIGN AND OVERVIEW

3.1 Overview/Framework

The proposed project takes a video of food as an input. The positional module picks the ten frames with the best positioning of the food. This module will use a CNN to detect the location of the food, and photographic heuristics to decide if the location of the food in respect to the image is aesthetically pleasing. These ten frames are passed on to the food enhancement module. In the food enhancement module, each image then is transformed via several image processing techniques that use color and position altering to improve the quality of the image. A certain number of transformed images are sent to the ranking module, which will be a CNN trained on the curated dataset to judge whether an image off food is aesthetically good or bad. The ranking module will return the best looking images back to the food enhancement module, and each image will be transformed again. This will be done for a certain number of times. The number of transformation will need to be determined by experiments, considering run time constraints. At the end, the ranking module will return the ten best images to the user. The framework of our project is shown in figure 1.

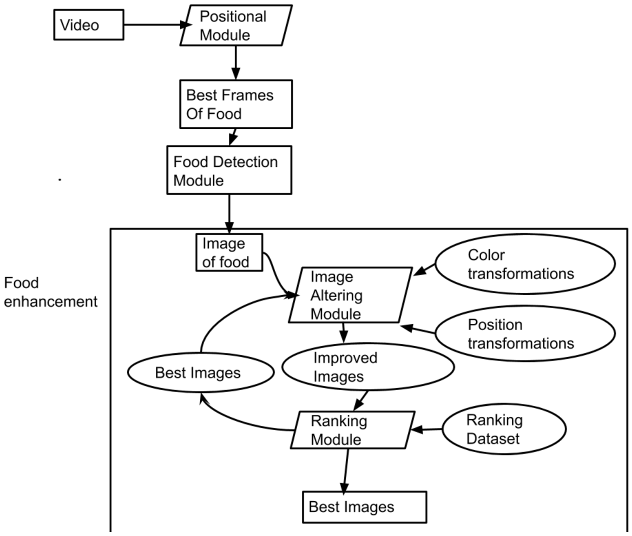

Figure 1: Architecture of Software

3.2 Positional module

This module takes a video as an input, and return the frames where the image is positioned in the most aesthetically pleasing manner. The user takes the video by moving the camera 360 degrees around the dish, three times at different angles and different distances. This video is showing the dish from as many angles and distances as possible. The angles and distances are too difficult to communicate to an amateur photographer, and are left to the judgement of the user. The positional module is a separate module than the image enhancing module because re-positioning the dish in a meaningful way, while preserving the nature of the image, is a problem that is not within the reach of current technology. The module works by detecting salient regions[10], and applying photography heuristics[6], to detect which images project the dish in a flattering way. Due to performance concerns, to split the video into frames, we will use OpenCV, implemented in C++ and interfaced with Python using the Ctypes module. To determine the position of food an image, we will use a publicly available food detection AI implemented in Keras, and to determine whether the position is aesthetically pleasing we will use OpenCV in Python and hard-coded positional heuristics. If the frames do not fit the positional heuristics, the module will crop and re-target the images to fit the heuristics better, hence look more pleasing. The 10 most flattering frames will be passed to the food detection module.

3.3 Food Detection

The positional module uses AI only to detect the region with food, which is not very accurate, but computationally cheap. Because the food and the background is processed differently, in order to obtain the optimal results, we need to find which pixels are food and which pixels are not food with high accuracy. Once the positional module has narrowed down the input video to ten images, we can start looking for the specific pixels with food. The location of these pixels will be located by using a a CNN trained on detecting food[5]. The module crops the image in different ways, showing different parts of the food, and passes them to the CNN. The cropping will be done using OpenCV. If a cropped image contains a high number of pixels that are food, it will be cropped again passed back to the AI. Cropped sections with high food occurrence are rewarded, while cropped sections with low food occurrence are punished. This way, we detect all pixels with food in them, which narrows down the region which contains food to specific pixels. The coordinates of the food, together with the image itself will be passed to the food enhancement module, as a list of tuples, each tuple containing an image and a position.

3.4 Food Enhancement

This module is the core module. It takes an image of food, with the coordinates of the dish in the food as an input, and return a certain number of images that are processed to aesthetically pleasing. To better explain this, this module is further divided into sub-modules.

3.4.1 Image Altering Module. This module applies a variety of color and positional transformation techniques to alter the image. These techniques are scaling, cropping and re-targeting, resizing, color transformation, color enhancement, and brightness/shade altering[3]. One thousand improved images are passed on to the ranking algorithm.

3.4.2 Ranking Module. The ranking module will be a CNN trained on the Gourmet dataset[13]. It will be an AlexNet implemented in Tensorflow. The algorithm returns a label for an image (positive or negative) and a confidence for the assessment. This module receives one thousand images, and will randomly sort them into 10 groups. From each group we will pick the best image, in parallel, as follows:

(1) The ranking module will take two images, and compare them using the aforementioned CNN to decide which one looks better. The better looking one will be passed onto the next round, to be compared to another image. There are three possibilities:
   - If only has a positive assessment, then we obviously pick that one.
   - If two images have a negative assessment then we will pick the one with the smaller confidence value.
   - On the other hand, if both have a positive assessment, the one with the higher confidence value will be picked.

(2) The winning images will be passed back to the image altering module, which will be transformed and passed back to the ranking algorithm again. This process will be repeated a certain number of times, and then the final images will be returned to the user. The number of repetitions will be determined by experimenting once we have a working prototype. More than one image will be returned so the user has some choice in picking an image.

3.5 Experiment Design

To verify the success of our project, we will test the ranking algorithm and the image altering module separately.

3.5.1 Ranking Module. To test the ranking module, we split our dataset into three parts. 70% of the dataset are used for training, 15% for validation, and 15% for testing. Since the images are already labeled, we can test whether the ranking module picks aesthetic images over non-aesthetic images. Without a dataset that consisting of images labeled by quantified aesthetic assessment, it is impossible to quantify how much each image has been improved. Compiling such a dataset is outside the scope, and budget, of this proposal.

3.5.2 Image Altering Module. To test the image altering module, we pick negatively assessed images from each of the three parts of the dataset. We then pass the images through the image altering module. These output is passed on to the ranking module. If the improved image has positive assessment, the image altering module is considered successful. We quantify the success by taking the percentage of negative quality images that were assessed positive after the image processing. This process is computationally cheap, and can be carried out using all negative looking images in the dataset.

4 BUDGET

All the software and datasets required for this proposal are open source and free. The only hardware required is a machine to train the machine learning algorithms on. The Earlham College cluster will be enough. No budget is needed.

5 TIMELINE

(1) $\boldsymbol{Week1 - Week2}$: Develop the grading algorithm of food using human curated dataset
(2) $\boldsymbol{Week3 - Week4}$: Develop food detection module.
(3) $\boldsymbol{Week5 - Week6}$: 1. Develop basic image processing methods and algorithms
(4) $\boldsymbol{Week7 - Week8}$: Improve the efficiency of the Image processing techniques.
(5) $\boldsymbol{Week9 - Week10}$: Tweak image altering module to figure out the most efficient and successful way of doing this, success being defined as most agreeability with humans.
(6) $\boldsymbol{Week11 - Week12}$: Develop the ranking algorithm to find the best image.
(7) $\boldsymbol{Week13 - Week14}$: Tweak ranking algorithm for best compromise between run speed and output quality.
(8) $\boldsymbol{Week15 - Week16}$: In this time period I will test the software, debug, optimize the project,. Depending on how well the previous work goes, I might implement different optimization levels were the machine will spend time according to how high quality the user wants. Otherwise, this time can be spent on working on other deliverables such as posters, presentations etc.

6 ACKNOWLEDGEMENTS

We would like to thank Dr. Xunfei Jiang and Dr. Igor Minevich for their mentoring and guidance over the course of writing this proposal. We would also like to extent our thanks to Dr. Kekai Sheng, who generously shared their dataset the Gourmet Dataset for Aestethic Assesment of Food Images with us.

REFERENCES

[1] Mei Chen, Kapil Dhingra, Wen Wu, and Rahul Sukthankar. 2009. PFID: Pittsburgh fast-food image dataset. 289-292 (11 2009), 289–292. https://doi.org/10.1109/ICIP.2009.5413511

[2] Yi-Ling Chen, Tzu-Wei Huang, Kai-Han Chang, Yu-Chen Tsai, Hwann-Tzong Chen, and Bing-Yu Chen. 2017. Quantitative Analysis of Automatic Image Cropping Algorithms: A Dataset and Comparative Study. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2017), 226–234.

[3] Yi-Ling Chen, Jan Klopp, Min Sun, Shao-Yi Chien, and Kwan-Liu Ma. 2017. Learning to Compose with Professional Photographs on the Web. In *Proceedings of the 25th ACM International Conference on Multimedia (MM '17)*. ACM, New York, NY, USA, 37–45. https://doi.org/10.1145/3123266.3123274

[4] Y. W. Guo, M. Liu, T. T. Gu, and W. P. Wang. 2012. Improving Photo Composition Elegantly: Considering Image Similarity During Composition Optimization. *Comput. Graph. Forum* 31, 7pt2 (Sept. 2012), 2193–2202. https://doi.org/10.1111/j.1467-8659.2012.03212.x

[5] Hokuto Kagaya, Kiyoharu Aizawa, and Makoto Ogawa. 2014. Food Detection and Recognition Using Convolutional Neural Network. In *Proceedings of the 22Nd ACM International Conference on Multimedia (MM '14)*. ACM, New York, NY, USA, 1085–1088. https://doi.org/10.1145/2647868.2654970

[6] Takao Kakimori, Makoto Okabe, Keiji Yanai, and Rikio Onai. 2016. A System to Help Amateurs Take Pictures of Delicious Looking Food. 456–461. https://doi.org/10.1109/BigMM.2016.47

[7] Yoshiyuki Kawano and Keiji Yanai. 2014. Automatic Expansion of a Food Image Dataset Leveraging Existing Categories with Domain Adaptation. In *ECCV Workshops*.

[8] Yoshiyuki Kawano and Keiji Yanai. 2015. FoodCam: A real-time food recognition system on a smartphone. *Multimedia Tools and Applications* 74, 14 (01 Jul 2015), 5263–5287. https://doi.org/10.1007/s11042-014-2000-8

[9] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1 (NIPS'12)*. Curran Associates Inc., USA, 1097–1105. http://dl.acm.org/citation.cfm?id=2999134.2999257

[10] Ligang Liu, Renjie Chen, Lior Wolf, and Daniel Cohen-Or. 2010. Optimizing Photo Composition. *Comput. Graph. Forum* 29 (2010), 469–478.

[11] Yuji Matsuda, Hajime Hoashi, and Keiji Yanai. 2012. Recognition of Multiple-Food Images by Detecting Candidate Regions. In *Proceedings of the 2012 IEEE International Conference on Multimedia and Expo (ICME '12)*. IEEE Computer Society, Washington, DC, USA, 25–30. https://doi.org/10.1109/ICME.2012.157

[12] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. *2012 IEEE Conference on Computer Vision and Pattern Recognition* (2012), 2408–2415.

[13] Kekai Sheng, Weiming Dong, Haibin Huang, Chongyang Ma, and Bao-Gang Hu. 2018. Gourmet Photography Dataset for Aesthetic Assessment of Food Images. In *SIGGRAPH Asia 2018 Technical Briefs (SA '18)*. Article 20, 4 pages. https://doi.org/10.1145/3283254.3283260

[14] Shulin Yang, Mei Chen, Dean Pomerleau, and Rahul Sukthankar. 2010. Food recognition using statistics of pairwise local features. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2249–2256. https://doi.org/10.1109/CVPR.2010.5539907

[15] Xin Zhang, Yee-Hong Yang, Zhiguang Han, Hui Wang, and Chao Gao. 2013. Object Class Detection: A Survey. *ACM Comput. Surv.* 46, 1, Article 10 (July 2013), 53 pages. https://doi.org/10.1145/2522968.2522978