IMAGE ANALYSIS OF PLANT BASED MEAT PRODUCTS

CS 499: Report 2

Authors
Badal Chowdhary - 20110034
Hitesh Jain - 20110077
Under the guidance of **Prof. Shanmuganathan Raman**

1 Introduction

In the ever-evolving culinary landscape and with changing dietary preferences, plant-based meat products have undoubtedly initiated a revolution in the food industry. These innovative products offer a sustainable and environmentally conscious substitute for conventional meat, catering to the discerning preferences of today's consumers. The goal of our study is to comprehensively explore the realm of plant-based meat products, with a focus on understanding the complexities of their relationship with traditional meat products. Ultimately, our aspiration is to develop a discerning model that can classify plant-based meat products uniquely when compared to their counterparts in the world of conventional meat.

1.1 Curation of Existing Dataset

The dataset retrieved from previous research poses significant challenges due to its noisy characteristics. The images frequently feature diverse background objects with contrasting colors. A prevalent problem observed is the misalignment of the patties, leading to tilted angles. Our initial attempts to extract the patties and eliminate the background from the images were based on employing various techniques from the cv2 library. This involved converting the original image to grayscale, applying Canny edge detection to clearly define the boundaries, and subsequently utilizing the Hough Circle Transform to identify circular shapes within the image, which correspond to the patties. Ultimately, a mask was applied to the original image in order to isolate and extract the patties.

Figure 1: Comparison of Images featuring Original Patties, Extracted Images using cv2 Techniques, and SAM Results for given prompts, highlighting various backgrounds and misalignments in the patties.

Following our initial attempts, it is evident from Figure 1 that the detection algorithm performs optimally only when the patty is correctly aligned. We have observed that the effectiveness of the current method is limited by the non-circular, elliptical shape of the patties. Consequently, we propose the adoption of a segmentation model to accurately extract images of the patties from the complex background. Implementing this approach is expected to significantly enhance the quality of the dataset and improve the precision of our research results. However, it is essential to acknowledge the labor-intensive nature of annotating images for training such a segmentation model. An estimated 1500 images would require manual annotation, which is a time-consuming task that demands careful consideration.

Furthermore, our experimentation with the Grounded-SAM by Idea-Research [1] state-of-the-art model demonstrated its remarkable potential in effectively segmenting various objects based on provided prompts, showcasing its versatility and robust performance. The notable success achieved during these trials as shown in Figure 1 has sparked our interest in further exploring the capabilities of SAM, specifically in the context of segmenting the intricate patties within our dataset. Currently, we have conducted preliminary experiments using the Hugging Face demo platform. However, we are planning to implement the model on Google Colab, provided that the GPU requirements are met and adequate resources are made available for this purpose.

1.2 Zero Shot inference on Foundational Model CLIP

The CLIP[2] model, short for "Contrastive Language-Image Pre-training," is an influential deep learning model created by OpenAI. It is specifically designed to comprehend and establish connections between images and text, enabling it to effectively accomplish diverse tasks. As a result, CLIP proves to be a versatile tool applicable in numerous domains and scenarios. One notable advantage of the CLIP model is its ability to perform zero-shot image classification effectively. Zero-shot learning refers to the capability of a model to classify images into classes that were not present during training. CLIP achieves this by leveraging the pre-training on a large dataset containing image-text pairs collected from the internet. This pre-training enables CLIP to learn the correspondence between natural language descriptions and images, allowing it to generalize to unseen classes during inference.

In this study, we employed zero-shot inference with CLIP to assess the accuracy of the 14 distinct classes. The code for this procedure can be found in the GitHub repository [3]. We established three metrics, namely R1, R3, and R5, as measurements of recall. The recall score, denoted as R@K, signifies the proportion of the top K retrieved captions that are relevant to the input image. For R1, the model associates one caption with a given image. Likewise, for R3 and R5, the model associates three and five captions, respectively, with a given image. An image-text pair is deemed correct if there is at least one predicted label that matches an actual label.

Table1 presents the inferences of each class on the R1, R3, and R5 metrics.

Class Name	**R@1**	**R@3**	**R@5**
commercial_air_normal	0.0%	0.5%	16.5%
commercial_air_over	0.0%	2.0%	24.0%
commercial_deep_normal	0.0%	0.5%	56.5%
commercial_deep_over	0.0%	63.0%	76.5%
commercial_unbaked	98.0%	100.0%	100.0%
inhouse_air_normal	0.0%	0.0%	3.0%
inhouse_air_over	0.0%	4.0%	24.0%
inhouse_deep_normal	0.0%	1.0%	11.5%
inhouse_deep_over	0.0%	10.5%	74.5%
inhouse_old_air_normal	0.0%	0.0%	3.5%
inhouse_old_air_over	0.0%	4.5%	15.5%
inhouse_old_deep_normal	0.0%	1.0%	10.5%
inhouse_old_deep_over	0.0%	12.0%	44.0%
inhouse_unbaked	44.5%	98.0%	99.5%

Table 1: Zero shot inference for 14 Classes

In the current dataset, two different cooking methods, namely Deep Frying and Air Frying, were used for every three types of products. However, research presented in [4] suggests that the cooking method employed does not significantly impact critical parameters such as appearance, color, and texture of the plant-based meat products. To streamline the dataset, we have therefore reduced it to only one cooking method for all three types of products, resulting in a reduction of the number of classes from fourteen to eight. The inferences of the reduced eight classes on the R@1, R@3, and R@5 metrics are presented in Table2.

Class Name	**R@1**	**R@3**	**R@5**
commercial_deep_normal	0.0%	3.0%	89.5%
commercial_deep_over	0.0%	71.0%	88.5%
commercial_unbaked	98.5%	100.0%	100.0%
inhouse_deep_normal	0.0%	0.5%	15.0%
inhouse_deep_over	0.0%	18.0%	86.0%
inhouse_old_deep_normal	0.0%	2.0%	18.5%
inhouse_old_deep_over	0.0%	14.5%	53.0%
inhouse_unbaked	40.0%	100.0%	100.0%

Table 2: Zero shot inference for 8 Classes

Reducing the number of classes from 14 to 8 has contributed to minimizing noise and redundancy within the dataset, consequently enhancing the zero-shot inference accuracy of these 8 classes. Specifically, the zero-shot inference accuracy has experienced a notable increase of approximately 3% in the R@2 metric and 9% in the R@5 metric. Furthermore, the overall accuracy of these 8 classes has shown an approximate improvement of 4%.

The classification labels used in CLIP were the original class names in the dataset, which were imprecise and impeded the model's performance. Improved and more meaningful labels could lead to better results.

1.3 Quantifying Texture as a way to classify Image

Texture serves as a critical factor in understanding the visual and tactile properties of plant-based meat products. In the realm of texture-based classification for the plant based meat products, the application of the Kullback-Leibler (KL) divergence[5] proves to be valuable. This statistical measure facilitates the assessment of the disparity between the texture of a specific plant-based meat product and a reference distribution of textures. Establishing this reference level typically involves selecting the most representative image from each class. By utilizing these selected images as reference points, we can effectively quantify the differences in textures across the dataset.

In our approach, we employ the cv2 RGB2LAB function for converting the RGB color space to the CIE Lab color space as shown in Figure 2, facilitating a more comprehensive analysis of texture nuances. This conversion is particularly useful when attempting to capture subtle variations that may not be adequately represented in the RGB color space. To extract the patty shape accurately from each image, we are actively developing a segmentation model. Once this model is successfully implemented, we will be able to extract the patty shapes from the images, enabling further analysis using the KL divergence method.

Additionally, we have implemented the Histogram of Gradients (HoG)[6] algorithm, which proves to

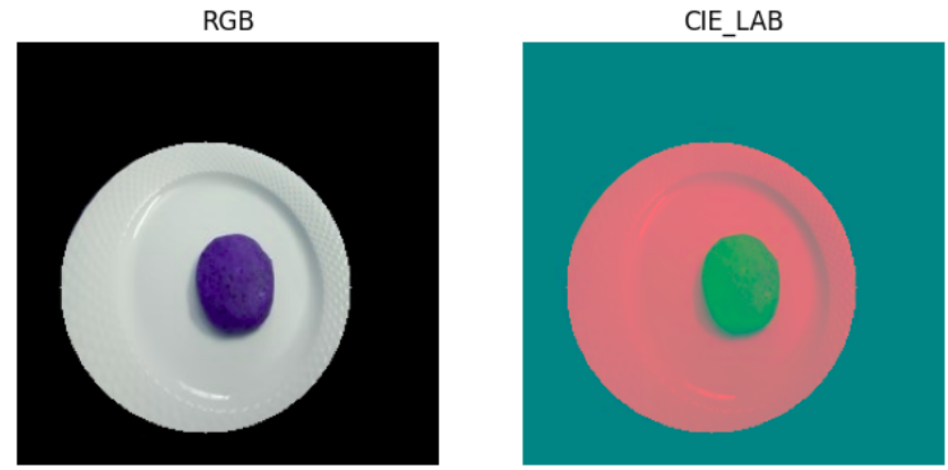

Figure 2: Original RGB vs CIE-Lab Conversion Image of a Patty

be instrumental in classifying plant-based meat product images based on texture as shown in Figure 3. This algorithm effectively captures crucial texture information by counting gradient orientation occurrences in localized image areas. The resulting HoG vector serves as a representative feature for each image, enabling subsequent classification using techniques such as support vector machines (SVM). We are currently working on extending this method to process and extract HoG features from a broader set of images. This expansion will allow us to leverage the unique ability of the HoG technique to discern and represent texture information accurately, thereby enhancing the overall effectiveness of the classification process.

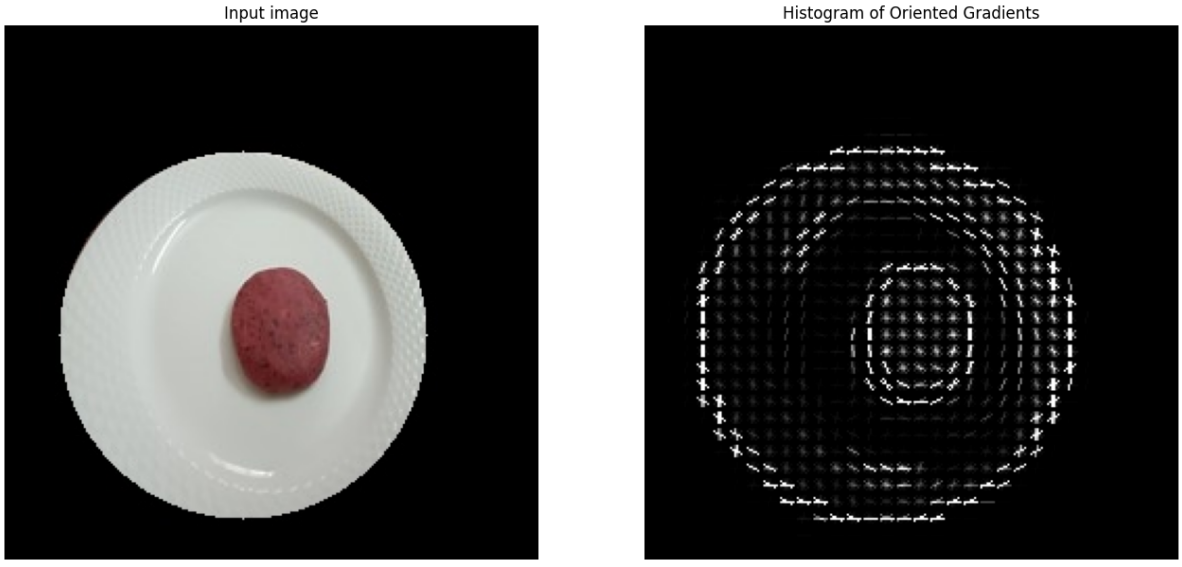

Figure 3: Visualization of the input image and its corresponding Histogram of Oriented Gradients (HoG).

References

[1] Y. Zhang, X. Huang, J. Ma, Z. Li, Z. Luo, Y. Xie, Y. Qin, T. Luo, Y. Li, S. Liu, Y. Guo, and L. Zhang, "Recognize anything: A strong image tagging model," 2023.

[2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[3] "Image analysis of plant based meat products." [Online]. Available: https://github.com/badalchowdhary/Food_Recognition

[4] G. Vu, H. Zhou, and D. J. McClements, "Impact of cooking method on properties of beef and plant-based burgers: Appearance, texture, thermal properties, and shrinkage," *Journal of Agriculture and Food Research*, vol. 9, p. 100355, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666154322000886

[5] S. Fekri-Ershad and F. Tajeripour, "Impulse-Noise Resistant Color-Texture Classification Approach Using Hybrid Color Local Binary Patterns and Kullback–Leibler Divergence," *The Computer Journal*, vol. 60, no. 11, pp. 1633–1648, 04 2017. [Online]. Available: https://doi.org/10.1093/comjnl/bxx033

[6] H. Demir, "Classification of texture images based on the histogram of oriented gradients using support vector machines," *Electrica*, vol. 18, pp. 90–94, 2018. [Online]. Available: https://electricajournal.org/Content/files/sayilar/28/90-94.pdf