# Kenyan Food Type Recognition in Instagram Photos

Mona Jalal[1]*, Kaihong Wang[1]*, Jefferson Sankara[2], Yi Zheng[1], Elaine O. Nsoesie[3], Margrit Betke[1]
[1]Department of Computer Science, Boston University. [2]Lori Systems, LTD, Kenya. [3]Department of Global Health, Boston University.

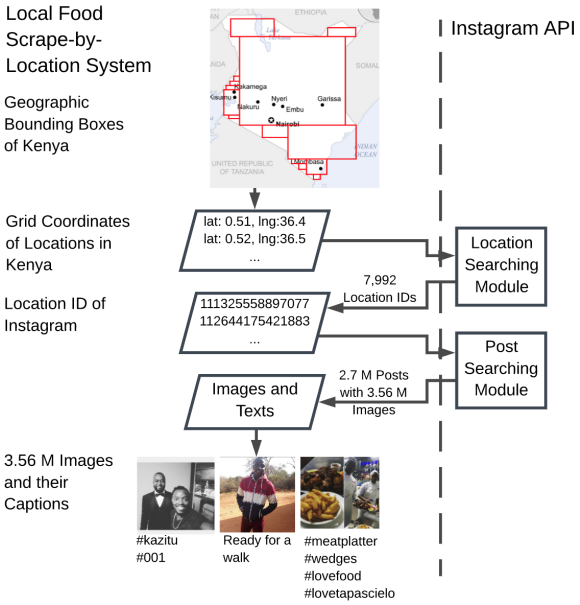


Figure 1: Overview of the proposed Scrape-by-Location System [1]

We propose a **scrape-by-location** methodology which applies Instagram API to collect Instagram-defined locations within hand-crafted geographic bounding boxes and finally collect images from these locations to create food image datasets from specific area through Instagram. The process is shown in Figures 1. We create a dataset, **Kenya104K**, using this method that contains 104,000 image/caption pairs and train a **Kenyan F**ood **C**lassifier (KenyanFC) to distinguish Kenyan food from non-food images posted in Kenya. We also propose a **scrape-by-keywords** methodology which applies Instagram API to search and collect posts given keywords and scrape ∼30,000 images along with their captions of 38 Kenyan food types.

Using this method, we create **KenyanFood13**, that contains 8,174 image/caption pairs to recognize 13 popular food types in Kenya using **Kenyan F**ood **T**ype **R**ecognizer (KenyanFTR), as shown in Figures 2, using a multimodal deep neural network using both images and their corresponding captions. Experiments show that the average top-1 accuracy of KenyanFC is 99% over 10,400 tested Instagram images and of KenyanFTR is 81% over 8,174 tested data points.

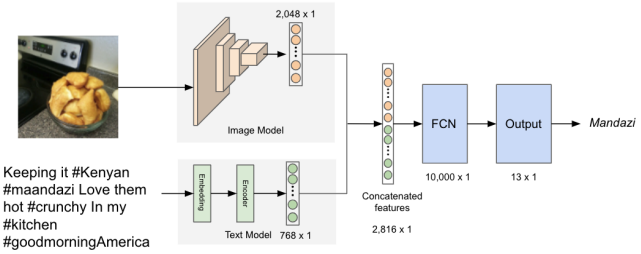


Figure 2: Architecture of food type recognition model (FCN stands for fully connected network)

In order to explore the value of taking advantage of the two modalities, image and text, in our KenyanFood13 dataset, we conducted ablation studies with models that take as input only images or only text (Table 1). For the former, we fine-tuned a ResNeXt101 (pre-trained on ImageNet) only with the images of KenyanFood13 and evaluated its performance, while for the latter, we fine-tuned a pre-trained BERT-based model using only the captions of KenyanFood13. Finally, we compared their performance with our KenyanFTR model, which takes both images and text as input.

We have a top-1 accuracy gain of more than 7 percent points when we add caption text as a modality to image modality for analysis. We also have a top-3 accuracy gain of more than 3 percent points when we add the caption text features to image features.

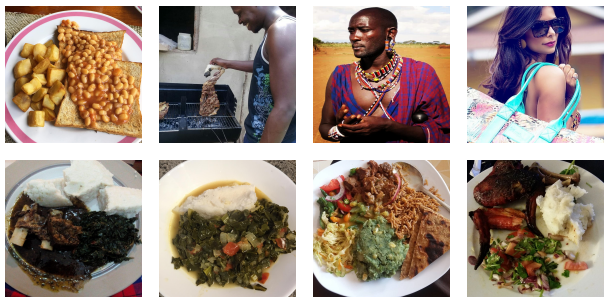

Figure 3: Sample images of the proposed Kenya104K dataset and KenyanFood13 dataset: the first two images at the first row are food images and next two images are non-food images from Kenya104K, while the images at the second row are food images from KenyanFood13 (ugali, sukuma wiki, mukimo, and kachumbari from left to right).

We investigated the performance of different image feature extractors. We compared the ResNeXt101 feature extractor used by our KenyanFTR with other popular pre-trained deep learning models, including InceptionV3, and DenseNet161. Our model comparison experiments reveal that all deep models fused with BERT generalized well on our dataset, with KenyanFRT performing the best (Table 2).

Table 1: Ablation Studies: Accuracy of different input settings on KenyanFood13.

| Method | Test Accuracy | |
|---|---|---|
| | Top-1 | Top-3 |
| Image only | 73.18%± 0.79% | 92.04%± 0.44% |
| Caption only | 65.30%± 1.70% | 83.68%± 1.55% |
| **Ours:** Image + Caption | 81.04%± 0.86% | 95.95%± 0.44% |

Table 2: Results of Comparison Experiments: Accuracy of different models on KenyanFood13.

| Method | Test Accuracy | |
|---|---|---|
| | Top-1 | Top-3 |
| InceptionV3+BERT | 71.92%± 1.52% | 88.57%± 0.68% |
| DenseNet161+BERT | 79.02%± 0.96% | 95.14%± 0.73% |
| **Ours:** ResNeXt101+BERT | 81.04%± 0.86% | 95.95%± 0.44% |

We further studied the remaining challenges as follows. For every image in a selected class, we computed their L2 distance to all other images not belonging to the current food type and found the pair of images with the smallest distance. Two examples of such similar image pairs are shown in Figure 4.


Figure 4: Examples of most confused food images in our KenyanFood13.

[1] Mona Jalal, Kaihong Wang, Jefferson Sankara, Yi Zheng, Elaine O. Nsoesie, and Margrit Betke. Scraping social media photos posted in kenya and elsewhere to detect and analyze food types. In *Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management, MADiMa @ ACM Multimedia 2019, Nice, France, October 21-25, 2019*, pages 50–59, 2019. doi: 10.1145/3347448.3357170. URL https://doi.org/10.1145/3347448.3357170.