2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Food Detection and Segmentation from Egocentric Camera Images

Ajay Ramesh[1], Viprav B. Raju[2], Madhav Rao[1], and Edward Sazonov[2]

Abstract— **Tracking an individual's food intake provides useful insight into their eating habits. Technological advancements in wearable sensors such as the automatic capture of food images from wearable cameras have made the tracking of food intake efficient and feasible. For accurate food intake monitoring, an automated food detection technique is needed to recognize foods from unstaged real-world images. This work presents a novel food detection and segmentation pipeline to detect the presence of food in images acquired from an egocentric wearable camera, and subsequently segment the food image. An ensemble of YOLOv5 detection networks is trained to detect and localize food items among other objects present in captured images. The model achieves an overall 80.6% mean average precision on four objects—Food, Beverage, Screen, and Person. Post object detection, the predicted food objects which are sufficiently sharp were considered for segmentation. The Normalized-Graph-Cut algorithm was used to segment the different parts of the food resulting in an average IoU of 82%.**

Clinical relevance— **The automatic monitoring of food intake using wearable devices can play a pivotal role in the treatment and prevention of eating disorders, obesity, malnutrition and other related issues. It can aid in understanding the pattern of nutritional intake and make personalized adjustments to lead a healthy life.**

I. INTRODUCTION

A healthy diet is essential for an individual's overall well being. Excessive energy intake is one of the biggest reasons for obesity, malnutrition, and related issues, especially among children [1]. Monitoring of food intake and the subsequent management of diet can play a crucial role in preventing such issues. Food intake monitoring is generally conducted using traditional self-monitoring as well as technology-aided methods. Self-monitoring may involve keeping a diary recording the kinds of food being eaten and their quantity [2]. However, such methods are prone to under-reporting and are not very accurate [3]. Thus, there is a need for other methods that can objectively and accurately monitor food intake.

Machine learning and deep learning methods have been used to detect and recognize food items from images. Kagaya *et al.* [4] used a convolutional neural network (CNN) which was trained on a custom dataset to recognize food in images. Pandey *et al.* [5] proposed FoodNet, a CNN-based ensemble model to recognize food and evaluate it on the Food-101 dataset [6]. A food classification and segmentation system called MyFood was proposed by Freitas *et al.* [7]. The above-mentioned methods, however, have a major drawback of using staged datasets that contain images captured from smartphones or other cameras. Such datasets require a conscious effort by a user to capture images using a smartphone, etc. to monitor food intake. In recent times, wearable devices have gained attention since they provide a good alternative to monitor food intake with minimal conscious effort [8]. Wearable devices can collect data about food consumption during meals without any additional intervention. Various kinds of wearable sensors and devices have been proposed to monitor food intake. Acoustic methods that detect chewing and swallowing sounds were used by Turan and Erzin [9], who developed a food detection system using a laryngeal microphone placed on the neck. Paßler and *et al.* [10] also developed a similar food intake activity detection scheme using an in-ear microphone. Food intake has also been monitored using piezo-electric sensors [11], [12], [13]. Another promising approach mentioned in [14], is to employ wearable cameras to record eating scenes, and was reported as an effective method to monitor food intake and eating behavior [14], [15]. Jia *et al.* [16] proposed a CNN-based method that can classify egocentric camera images that contain food items. Hossain *et al.* [17] proposed a real-time food monitoring system for egocentric camera images using the MobileNet classifier and implemented it on the Cortex-M7 micro-controller. However, these methods only classify the presence of food in an image but do not localize the food. Wearable camera images contain various background objects and to perform an accurate dietary assessment, it is essential to localize the food items. The localization can then help in estimating the quantity of food and its nutritional content.

The contributions of this work are as follows: (a) The application of the YOLOv5 algorithm to detect and localize food items from egocentric camera images that were captured dynamically during the eating process. Unlike existing datasets, these images are un-staged and hence represent the real eating environment. (b) The application of the latest data augmentation techniques to boost the performance of the detector on real-world data. (c) Segmentation of the detected food items using graph-cut based techniques that involve performing a Normalized-Graph Cut on a Region Adjacency Graph (RAG), which is constructed from the localized food item. The end goal of this work is to aid in the development of autonomous food intake monitoring systems using computer vision techniques.

II. APPROACH AND METHODOLOGIES

A. Data Collection

1) Wearable Equipment: The Automatic Ingestion Monitor device [18] was used to collect image data during eating episodes.The device consists of an STM32 processor, a camera: OV5640 along with an FPGA+1MB RAM as the

[1] IIIT-Bangalore, Bangalore-560100, India. (e-mail: ajayramesh.ranganathan@iiitb.org)

[2] Department of Electrical and Computer Engineering, The University of Alabama, Tuscaloosa, AL 35487 USA. (e-mail: esazonov@ua.edu)

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

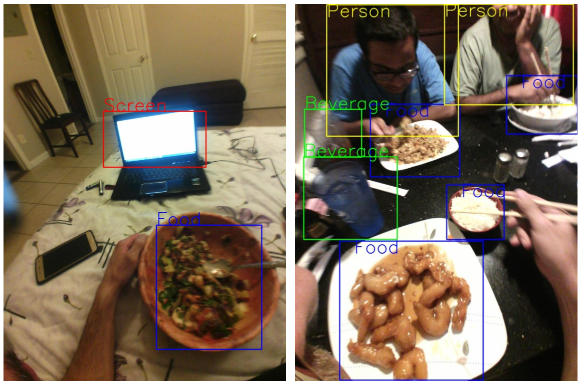


Fig. 1: Two sample images captured from the AIM 2.0 device.

frame buffer, an ADXL362 accelerometer, and a micro SD card to store the images. The camera is fitted with a 170-degree wide-angle lens. Images were captured automatically every 10 seconds [19]. The study was approved by the University of Alabama's Institutional Review board [18].

2) Dataset: The dataset consists of 5417 images in total and is annotated with four different objects—Food, Beverage, Screen, and Person. The annotations are in the form of class labels and rectangular bounding boxes. The dataset was split for training and testing phases according to different and independent eating episodes (4637 training images and 780 testing images). An eating episode refers to a continuous activity of food intake (e.g, a meal). The inclusion of independent datasets in the training and testing phase prevents over-fitting of the models. Sample images from the dataset are shown in the Figure 1.

B. Object Detection Network

The images captured from an egocentric camera include a variety of objects that may or may not be foods being consumed. Identifying the food that is being eaten is approached as an object detection task. Object detection techniques help identify and locate an object of interest in an image. Several deep learning algorithms such as R-CNN [20], Fast-RCNN [21], Faster-RCNN [22], YOLO [23], and SSD [24] have been successful in localizing objects in images and videos. Among these, the YOLO algorithm was chosen for this work. The YOLO algorithm is comparatively faster since it uses a single convolutional network to predict multiple bounding boxes and class probabilities [23]. This quick inference makes it an ideal candidate for potential real-time detection. Besides, the network is small in size, which enables inference on edge devices. The latest version of YOLO: YOLOv5 has shown state-of-the-art detection results, and the implementation provided by Ultralytics was used to develop our system [25].

Data augmentation has been shown to significantly improve the performance of the YOLO algorithm [26]. Augmentation also provides a strong regularization for small datasets such as described in this work and prevent over-fitting [27]. The augmentation techniques that were applied to the dataset include random scaling, x-flips, and y-flips, to incorporate variations in the pose and orientation of the objects in the dataset. Photometric distortions such as adjustments in hue, saturation, and brightness in images were also part of the applied augmentations. Additionally, mix-up [28] strategy in which objects are cropped out and pasted in random backgrounds were also applied for building a robust model. The new mosaic augmentation technique which was introduced in Yolov4 [26] was also attempted. This technique merges four training images into one. In other words, four different image contexts were mixed, which trained the model to detect objects out of their normal context. The above-mentioned augmentations form the "Bag of Freebies" [29] and "Bag of Specials" [26]. Transfer learning was applied to train the network. The network was initialized with the weights pre-trained on the COCO dataset [30]. The SGD with momentum and warm restarts algorithm [31] was used for optimization, and a cosine annealing scheduler [32] was used to decay the learning rate.

A Neural network ensemble is a learning paradigm where multiple neural networks are jointly used to solve a problem [33]. Ensembles have shown to improve generalization and performance [34]. Hence an ensemble of three different networks which vary in depth —Yolov5 small (140 layers), Yolov5 medium (188 layers) and Yolov5 large (236 layers) [25] was utilized. The final prediction is computed as the mean of the individual predictions of each model (in terms of the bounding box coordinates and the confidence score). The results of ensemble modeling and its role in improving results are presented in Section III-A.

C. Unsupervised segmentation of food

The object detection network detects objects in terms of class probabilities and the corresponding bounding boxes. Since the proposed method intends to segment food, only the detected food objects were considered for segmentation. However, some images have serious object and camera motion blur due to being captured by a wearable camera. Hence, an image sharpness metric proposed in [19] and validated on images captured from the AIM 2.0 device, was incorporated to discard blurred food objects. The sharpness metric was computed from the mean amplitude of the highest 90 percent frequencies extracted from the Fast Fourier Transform (FFT) of the image. Then, a suitable threshold was configured to discard blurred food images. The sufficiently sharp images were considered for segmentation and processed to remove noise using a Gaussian filter. A blur test was performed only on the food object since images acquired by the wearable camera are likely to have blurred objects in the background, which are irrelevant to the context. Yet, a clear view of the food object being eaten is necessary. Since the ground truth dataset contained only bounding boxes and not mask annotations, an unsupervised method for segmenting different parts of food was required. Few popular methods such as K-means [35][36], thresholding methods [37], edge detection based methods, and graph-based methods [38] were

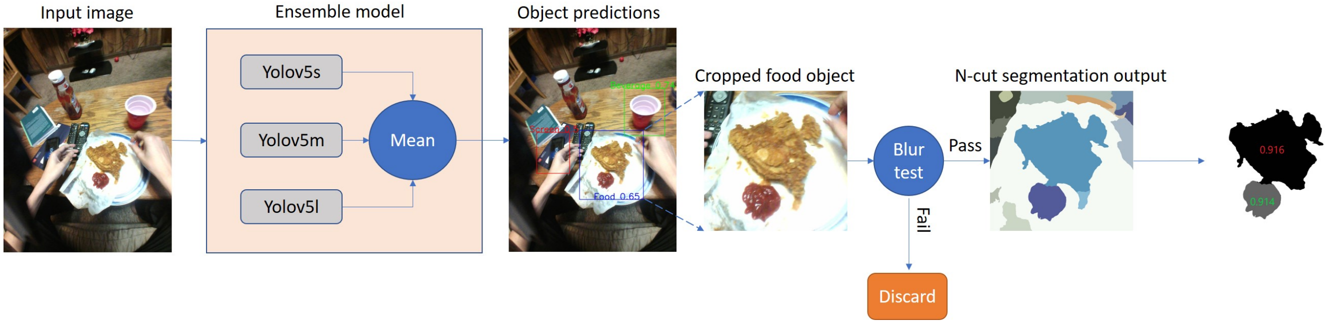


Fig. 2: Proposed workflow diagram to detect food objects and segment their parts. The resulting segmentation IoU with respect to the ground truth is indicated within the segment.

used in previous studies to segment different parts of the food object. In the K-means algorithm, the pre-defined value of K plays a major role in the segmentation outcome. Selecting the value of K is not feasible in the proposed real-life images, since the number of different food parts captured in the image may vary. Graph-cut based method with normalized-cut technique as proposed by Shi and Malik in [39], was found to perform the best when compared with other methods and was hence applied for food segmentation. The technique involves setting up a weighted graph from a given set of features and recursively partitioning the graph to obtain segments. Before applying the graph-cut method, the K-means algorithm was applied to obtain initial segments that were used to construct a Region Adjacency Graph (RAG). The RAG, $G = (V, E)$, consists of nodes representing the regions and edges representing the adjacency. Each node in the RAG represents a set of pixels within the image with the same label that was generated by the initial K-means clustering. The weight between two adjacent regions is a measure of the similarity of the two regions. A color similarity measure of the form proposed by Shi and Malik [39] was used in constructing the RAG. This is shown in Equation 1, where $d = |c_1 - c_2|$, and c_1 and c_2 are the mean colors of the two regions. This method of performing N-cuts on RAGs also reduces the time complexity.

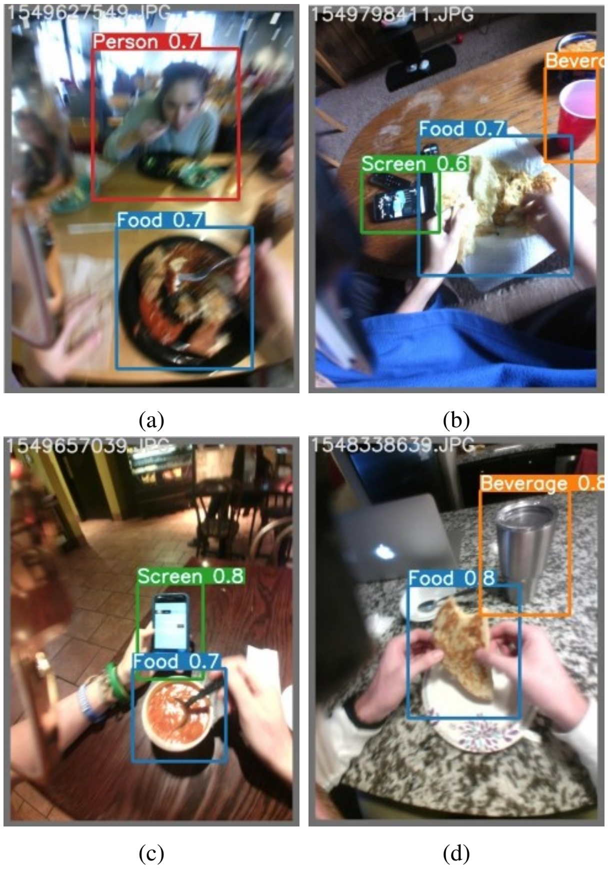
$$w_{ij} = e^{\frac{-d^2}{\sigma^2}} \tag{1}$$

For the initial clustering step, a sufficiently high value of $K = 100$ worked satisfactorily for all test cases, and hence, the same was configured in this step of the proposed workflow. Intersection Over Union (IoU), also called the Jaccardian Index, was computed for each generated food segment and compared with the ground truth annotation in order to evaluate the accuracy of segmentation. The entire workflow of the system is shown for a sample test image in Figure 2.

III. Results

A. Object Detection

The object detection result of the proposed ensemble model on the test set, for each object class, is shown in Table I. The detection results of the individual models that comprise the ensemble are also shown. The *mean average precision (map)* metric was used to evaluate the performance of each model. It was found that the performance improved when deeper networks were used, and the best results were obtained by ensembling the three individual models. The results of each model were aggregated by computing the mean of the predictions of each model in terms of the confidence scores and bounding box coordinates. When using an ensemble, objects that were undetected by one model were detected by another, thereby improving the detection *map*.



(a)

(b)

(c)

(d)

Fig. 3: Food detection results on four sample images taken from the test dataset, with prediction confidence scores specified alongside the predicted class labels in each image.

Example results are shown in Figure 3.

In Figure 3a, the food object has been successfully detected in the presence of camera motion blur from the wearable device. Figures 3b, 3c and 3d show the detection on food objects in different conditions. The model also detects food that is held in the hands (Figure 3d). Thus, the model is robust and resilient to variations in food orientation, shape, size, and color.

Class	mean average precision (map)			
	Yolov5s	Yolov5m	Yolov5l	Ensemble
Food	0.6	0.624	0.615	0.666
Beverage	0.651	0.737	0.732	0.766
Screen	0.871	0.9	0.882	0.917
Person	0.847	0.839	0.878	0.876
All	0.742	0.772	0.784	0.806

TABLE I: Object detection results

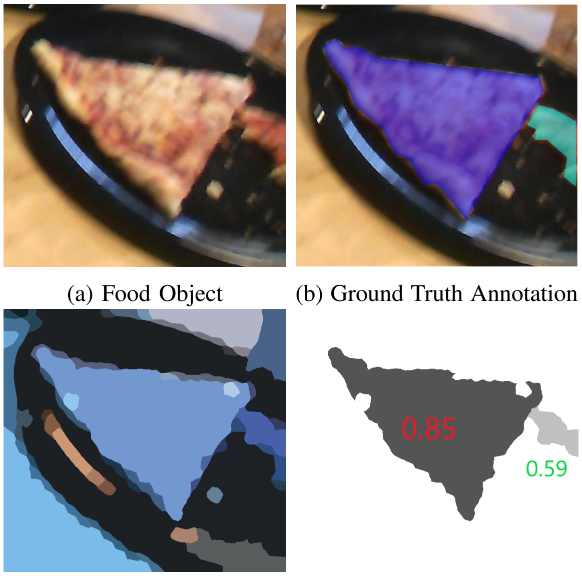
B. *Food Segmentation*

The segmentation result for a test image is shown in Figure 2. The detected food object was cropped and tested for blurness. The normalized graph cut method was applied if the image is sharp as described in Section II-C, to segment the parts of food successfully. The results of testing on an image captured while eating pizza is shown in Figure 4. The larger piece and the smaller piece are segmented at 0.849 and 0.589 IoU respectively. Another test case involving a bowl of soup is presented in Figure 5. The food parts are segmented at 0.9 and 0.6 IoU.

Experiments revealed that the segmentation technique performs reliably for solid food and soup type of items. The averaged results of segmentation on 20 such segments of food is shown in Table II. A segment of food is one part that is present in a food object. For example, in the test case in Figure 2 there are two segments of food, which have been segmented at 0.916 and 0.914 IoU respectively. The 20 test segments are present in 11 unique images picked randomly from 5 different eating episodes. The mean IoU of the predicted segments is 0.81 with respect to the ground truth, and the standard deviation of the IoUs is 0.1. The segmentation is hence fairly robust and accurate.

TABLE II: Averaged results of segmentation

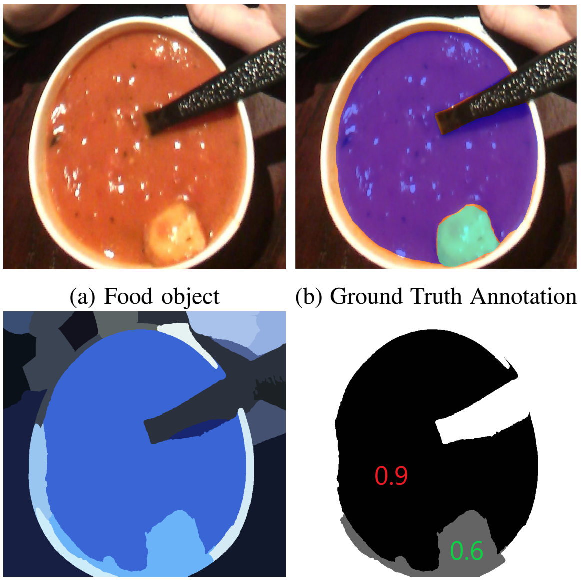
Number of food segments	**Mean IoU**	**Standard Deviation of IoU**
20	0.82	0.1



(a) Food Object (b) Ground Truth Annotation

(c) N-Cut segmentation (d) Result IoUs

Fig. 4: Segmentation results for two pieces of pizza on a plate. The IoUs for each piece of food is shown within/alongside the segment.



(a) Food object (b) Ground Truth Annotation

(c) N-Cut segmentation (d) Result IoUs

Fig. 5: Segmentation results for a bowl of soup. The IoUs for each piece of food is shown within the segment.

IV. Conclusions

A robust food detection model and segmentation pipeline were developed for the first time with an intention to monitor food intake from egocentric wearable cameras. An ensemble of Yolov5 networks of varying depth was found accurate for the custom dataset collected from the AIM 2.0 wearable device. The proposed workflow caters to blurred and unstaged images and is applicable in real-life scenarios. A normalized-Graph Cut method, post the blurriness test on the detected food object, was designed in the proposed workflow to segment different parts of the food accurately. The novel implemented workflow is a significant step towards building food monitoring systems that are completely autonomous, and available at a relatively low cost. In the future, further steps on extending the dataset to include different kinds of food need to be explored. The segmentation method will also be extended to other food types. The extension of various food data in the custom datasets would enable precise estimation of the nutritional content of the food and thereby aid in not only automated monitoring of food intake but also the consumed nutritional index.

REFERENCES

[1] K. Kuźbicka and D. Rachoń, "Bad eating habits as the main cause of obesity among children," *Pediatric endocrinology, diabetes, and metabolism*, vol. 19, pp. 106–10, 01 2013.

[2] F. E. Thompson and A. F. Subar, "Chapter 1 - dietary assessment methodology," in *Nutrition in the Prevention and Treatment of Disease (Fourth Edition)*, fourth edition ed., A. M. Coulston, C. J. Boushey, M. G. Ferruzzi, and L. M. Delahanty, Eds. Academic Press, 2017, pp. 5 – 48. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780128029282000011

[3] A. H. C. Goris, E. P. Meijer, and K. R. Westerterp, "Repeated measurement of habitual food intake increases under-reporting and induces selective under-reporting," *British Journal of Nutrition*, vol. 85, no. 5, p. 629–634, 2001.

[4] H. Kagaya, K. Aizawa, and M. Ogawa, "Food detection and recognition using convolutional neural network," 11 2014.

[5] P. Pandey, A. Deepthi, B. Mandal, and N. B. Puhan, "Foodnet: Recognizing foods using ensemble of deep networks," *IEEE Signal Processing Letters*, vol. 24, no. 12, p. 1758–1762, Dec 2017. [Online]. Available: http://dx.doi.org/10.1109/LSP.2017.2758862

[6] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – mining discriminative components with random forests," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 446–461.

[7] C. N. C. Freitas, F. R. Cordeiro, and V. Macario, "Myfood: A food segmentation and classification system to aid nutritional monitoring," 2020.

[8] J. M. Fontana, M. Farooq, and E. Sazonov, "Chapter 20 - detection and characterization of food intake by wearable sensors," in *Wearable Sensors (Second Edition)*, second edition ed., E. Sazonov, Ed. Oxford: Academic Press, 2021, pp. 541 – 574. [Online]. Available: http://www.sciencedirect.com/science/article/pii/B9780128192467000206

[9] M. A. T. Turan and E. Erzin, "Food intake detection using autoencoder-based deep neural networks," in *2018 26th Signal Processing and Communications Applications Conference (SIU)*, 2018, pp. 1–4.

[10] S. Paßler and W. Fischer, "Food intake activity detection using a wearable microphone system," in *2011 Seventh International Conference on Intelligent Environments*, 2011, pp. 298–301.

[11] G. HUSSAIN, K. JAVED, J. CHO, and J. YI, "Food intake detection and classification using a necklace-type piezoelectric wearable sensor system," *IEICE Transactions on Information and Systems*, vol. E101.D, no. 11, pp. 2795–2807, 2018.

[12] H. Kalantarian, N. Alshurafa, T. Le, and M. Sarrafzadeh, "Monitoring eating habits using a piezoelectric sensor-based necklace," *Computers in Biology and Medicine*, vol. 58, 01 2015.

[13] E. Sazonov and J. Fontana, "A sensor system for automatic detection of food intake through non-invasive monitoring of chewing," *IEEE sensors journal*, vol. 12, pp. 1340–1348, 05 2012.

[14] L. Gemming, E. Rush, R. Maddison, A. Doherty, N. Gant, J. Utter, and C. Mhurchu, "Wearable cameras can reduce dietary under-reporting: Doubly labelled water validation of a camera-assisted 24 h recall," *The British journal of nutrition*, vol. 113, pp. 1–8, 11 2014.

[15] T. Vu, F. Lin, N. Alshurafa, and W. Xu, "Wearable food intake monitoring technologies: A comprehensive review," *Computers*, vol. 6, p. 4, 01 2017.

[16] W. Jia, Y. Li, R. Qu, T. Baranowski, L. Burke, H. Zhang, Y. Bai, J. Mancino, G. Xu, Z.-H. Mao, and M. Sun, "Automatic food detection in egocentric images using artificial intelligence technology," *Public Health Nutrition*, vol. 22, pp. 1–12, 03 2018.

[17] D. Hossain, M. H. Imtiaz, T. Ghosh, V. Bhaskar, and E. Sazonov, "Real-time food intake monitoring using wearable egocnetric camera," in *2020 42nd Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, 2020, pp. 4191–4195.

[18] A. B. M. S. U. Doulah, T. Ghosh, D. Hossain, M. H. Imtiaz, and E. Sazonov, ""automatic ingestion monitor version 2" — a novel wearable device for automatic food intake detection and passive capture of food images," *IEEE Journal of Biomedical and Health Informatics*, pp. 1–1, 2020.

[19] V. Raju and E. Sazonov, "Processing of egocentric camera images from a wearable food intake sensor," in *2019 SoutheastCon*, 2019, pp. 1–6.

[20] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580–587.

[21] R. Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.

[22] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[23] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.

[24] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," *CoRR*, vol. abs/1512.02325, 2015. [Online]. Available: http://arxiv.org/abs/1512.02325

[25] G. Jocher, A. Stoken, J. Borovec, NanoCode012, ChristopherSTAN, L. Changyu, Laughing, A. Hogan, lorenzomammana, tkianai, yxNONG, AlexWang1900, L. Diaconu, Marc, wanghaoyang0106, ml5ah, Doug, Hatovix, J. Poznanski, L. Y. , changyu98, P. Rai, R. Ferriday, T. Sullivan, W. Xinyu, YuriRibeiro, E. R. Claramunt, hopesala, pritul dave, and yzchen, "ultralytics/yolov5: v3.0," Aug. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3983579

[26] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," 2020.

[27] B. Zoph, E. D. Cubuk, G. Ghiasi, T.-Y. Lin, J. Shlens, and Q. V. Le, "Learning data augmentation strategies for object detection," 2019.

[28] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," 2018.

[29] Z. Zhang, T. He, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of freebies for training object detection neural networks," 2019.

[30] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, and P. Dollár, "Microsoft coco: Common objects in context," 2015.

[31] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," 2017.

[32] T. He, Z. Zhang, H. Zhang, Z. Zhang, J. Xie, and M. Li, "Bag of tricks for image classification with convolutional neural networks," 2018.

[33] Z.-H. Zhou, J. Wu, and W. Tang, "Ensembling neural networks: Many could be better than all," *Artificial Intelligence*, vol. 137, no. 1, pp. 239 – 263, 2002. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S000437020200190X

[34] L. K. Hansen and P. Salamon, "Neural network ensembles," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 10, pp. 993–1001, 1990.

[35] R. K. R. G. T. R. J. Mahesh Kumar, Rupalin Nanda, "Image segmentation using k-means clustering," *International Journal of Advanced Science and Technology*, vol. 29, no. 6s, pp. 3700 – 3704, May 2020. [Online]. Available: http://sersc.org/journals/index.php/IJAST/article/view/23282

[36] X. Zheng, Q. Lei, R. Yao, Y. Gong, and Q. Yin, "Image segmentation based on adaptive k-means algorithm," *EURASIP Journal on Image and Video Processing*, vol. 2018, 08 2018.

[37] S. Abdullah, H. Hambali, and N. Jamil, "An accurate thresholding-based segmentation technique for natural images," 06 2012, pp. 919–922.

[38] T. Zuva, O. Olugbara, and S. Ngwira, "Image segmentation, available techniques, developments and open issues," *Canadian Journal on Image Processing and Computer Vision*, vol. 2, 01 2011.

[39] Jianbo Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.