Camera-Based Meal Type and Weight Estimation in Self-Service Lunch Line Restaurants

Sarapisto, Teemu

IEEE
2022

Sarapisto , T , Koivunen , L , Makila , T , Klami , A & Ojansivu , P 2022 , Camera-Based Meal Type and Weight Estimation in Self-Service Lunch Line Restaurants . in 2022 12TH INTERNATIONAL CONFERENCE ON PATTERN RECOGNITION SYSTEMS (ICPRS) . IEEE , 12th International Conference on Pattern Recognition Systems (ICPRS) , Saint-Etienne , France , 07/06/2022 . https://doi.org/10.1109/ICPRS54038.2022.9854056

http://hdl.handle.net/10138/354517
https://doi.org/10.1109/ICPRS54038.2022.9854056

unspecified
acceptedVersion

Camera-Based Meal Type and Weight Estimation in Self-Service Lunch Line Restaurants

Teemu Sarapisto
Dept. of Computer Science
University of Helsinki
Helsinki, Finland
teemu.sarapisto@helsinki.fi

Lauri Koivunen
Dept. of Computing
University of Turku
Turku, Finland
lamkoi@utu.fi

Tuomas Mäkilä
Dept. of Computing
University of Turku
Turku, Finland
tuomas.makila@utu.fi

Arto Klami
Dept. of Computer Science
University of Helsinki
Helsinki, Finland
arto.klami@helsinki.fi

Pauliina Ojansivu
Functional Foods Forum
University of Turku
Turku, Finland
pauliina.ojansivu@utu.fi

***Abstract*—Individuals, restaurant owners and health organizations are all interested in accurate information about food intake, but collecting the information in sufficiently automated way remains a practical challenge. In controlled environments, such as lunch line restaurants, the food intake can be estimated by measuring the portions upon purchase and separately monitoring the food waste, but even this requires often complicated setups such as repeated weighing of the plate after every meal component. In this work we explore the feasibility of using a combination of ceiling-mounted cameras and computer vision for estimating both the types and weights of individual food items the customers are taking in a lunch line restaurant. We describe the imaging system and weighing-based sensing for obtaining ground truth training data, and develop and evaluate deep learning models for the computer vision tasks. We demonstrate high accuracy especially in meal type identification and hence validate the feasibility of the approach. We release the annotated dataset for further development of improved methods.**

***Index Terms*—self-serve lunch line, cafeteria, multispectral**

I. Introduction

Monitoring nutritional intake is important for both food intake studies and nutritional assessment and guidance work. However, current methods are laborious and subjective, limiting clinical inference and counselling capabilities [1]. Recent advances in automatic image-based food estimation have not yet been applied to long-term care facilities limiting nutritional assessment and counselling [2]. Here, we describe a novel tool for accurately estimating food intakes and dietary studies knowing that besides public health organizations food intake accurately can be important for customers and the restaurants. Interested customers may want to track the nutritional values of their food portions, such as vitamins and minerals, carbohydrates, proteins and fats. The restaurants want to minimize food waste, develop data-driven food production control, and offer health data as a service to the customers. The restaurants also want to predict customer behavior and to monitor how well the food they serve fills dietary recommendations provided by public officials. Similarly, public health organizations need accurate data about consumption statistics to better understand the behavior of individuals and their choices, required also as basis for developing recommendations for improving health and well-being at the population level.

Obtaining data about food intake is, however, extremely difficult in practice. In nutritional research slow and laborious data collection methods consisting of manual weight measurements, self-reporting questionnaires, and 24h dietary recall methods are used [3]. These techniques require high level of commitment, and consequently are not feasible solutions for the information needs of ordinary citizens or restaurant owners. Furthermore, misreporting and under-reporting are common in dietary assessment [4]–[6]. Making information on food intake available for restaurant owners and consumers requires considerably more automated means, where the effort required from the individual is minimized.

The best systems for estimating food intake are today based on weighing individual portions to allow measuring nutritional values and food intake, such as the service in use in the Flavoria test restaurant in Finland [7]. In self-service lunch restaurants, where the customer compiles the meal by selecting free amounts of individual food items available, this process is still tedious and unreliable since the plate needs to be measured several times (after each food item) to obtain information about individual food items and the customer needs to pay attention to measuring the weight properly. In this work we investigate the possibility of adding a more automatic camera-based system to complement (and possibly eventually replace) the one based on weighing, using computer vision for detecting meal types and weights from automatically captured images while using the weighing-based system for obtaining the ground truth labels for training the computer vision model.

There is broad range of research on computer vision methods for food identification [8]–[12] using different kinds of imaging setups from mobile phones to multi-camera setups, but studies on practical systems remain limited. The closest work is *im2calories* [13] that studied calorie prediction in lunch restaurants where portions can be assumed to be of fixed form, weight and calorie content (e.g. burger in a fast food restaurant), so that predicted food items can directly be mapped to weights and calorie content based on a static menu. Compared to the previous research, our main contribution is considering learning tasks where the plate content is more diverse due to the customer deciding the specific food items

978-1-6654-6694-3/22/$31.00

Food item	True	Pred.
American salad	58	41
pickled cucumber slice	29	34
rice	125	136
chicken leg	227	232
grated carrot	77	0
carrot-cabbage	0	45
tomato-onion	32	0
tomato	0	58

Fig. 1. Example image of a lunch plate imaged with the system, true weights (in grams) and weight predictions provided by the computer vision model (ResNet101+Menu). While the model confuses some visually highly similar classes (indicated by the two colored pairs of lines), these mistakes would be largely insignificant for the customer because the nutritional content of the confused classes (e.g. 'grated carrot' and 'carrot-cabbage') is also similar.

and their amounts, and especially in providing labeled ground truth information for the weights of all items.

We collect ground truth data using the existing system in the restaurant's lunch line, while imaging the portions (at the end of the lunch line) with RGB cameras. We then train deep neural network for (a) identifying which foods the user has on their plate, and (b) estimating their weights. We solve the estimation tasks using a convolutional neural network built on the ResNet architecture [14] and study alternative ways of accounting for the daily menu (a list of items available during each day). We achieve high accuracy in recognizing individual food items and demonstrate that also the weights of the items can be estimated to a degree despite using simple 2D imaging with heavy occlusion. An example input image and predictions are shown in Figure 1, illustrating the complexity of the vision problem and the accuracy of the predictions. We make the dataset, consisting of roughly 1,700 images along with corresponding 7,890 separate food weight annotations, publicly available to encourage others to develop the computational methods further.

II. Background and Previous Work

The study was conducted in the Flavoria research restaurant. The restaurant includes multiple sensors and scales from tray identification to individual meal identification throughout the lunch line. The customer can also input additional data, such as drink contents via a mobile application, which links the customer and the meal trays together for a continuous view on, for example, nutritional values and biowaste amounts. Finally, the restaurant's biowaste stations can be used to infer the discarded food based on the previously measured food content identified by the customer's meal tray [7], and preliminary studies on using imaging for that have been conducted [15].

Computer vision offers a natural approach for analysis of food images, and several studies primarily using deep learning methods have been conducted; see [16] for a recent overview. However, all of the existing approaches have shortcomings, especially in the sense that they consider simplified problem instances and the solutions are not directly applicable for open-ended lunch restaurant lines. The previous approaches either

- use ground truths approximated by human experts instead of measured ones, with the ground truth describing only the total energy content of the whole plate [8];
- focus on addressing challenges caused by low-cost imaging (mobile and wearable cameras in bad lighting) [10];
- require fiducial markers such as QR codes, fingers or rice grains to estimate scale [17] [12]; or
- laboriously manually crop the images to get separate foods, and food items are pre-weighted by the staff [12].

Data availability is a severe challenge. Even the largest relevant data sets are small and simplified, such as the 96 image data of approximately 850 separate foods of pre-weighted portions used by [12], or the dataset of approximately 9,000 RGB-D images by [8] made up of only 60 different plates of food and the ground truths of only the total energy content of the plate. Larger food image datasets exist, such as the data of approximately 12,000 images by Myers et al. [13] that provides segmentation masks for every separate food on the plates, but they do not provide any data about weights or calories. Similarly, the Food-101 data [18] with more than 100,000 images only contains class information.

III. Methods

A. Problem formulation

Our task is to estimate all of the food types and their weights based on a single RGB image of the whole plate, and for training this model we have information on all types and their weights obtained by weighing the plate after each type. In addition, we have information about the daily menu, the subset of items available on the day the plate was imaged.

We use deep neural networks for solving the identification and weight estimation problems. Given the collection of N pairs of images X_n and ground truth vectors y_n, we train a neural network $\hat{y} = f(X, w)$ parameterized by w to approximate y for new images X. The images X are in our case tensors of $(3 \times 512 \times 512)$ whereas the outcomes are C-dimensional vectors where C is the number of food items (130 in our experiments). For classification y is a binary vector, whereas for weight prediction it is a real vector.

a) Identification: The problem for food identification is an instance of *multi-label classification*; the output vector y indicates presence of all possible food categories, so that multiple labels can be simultaneously present. We solve this using a network that outputs C values, one for each class, with logistic function mapping the outputs to probability of presence of that class. The model is trained using weighted binary cross entropy (BCE) summing the loss

$$-\sum_{c=1}^{C} \frac{1}{N_c} \mathrm{BCE}(y_c, \mathrm{logistic}(z_c))$$

over individual data points. Here N_c is the number of instances of that class, y_c is the true label (binary indicator for the cth class) and z_c is the output of the network before the final logistic transformation. The classes are weighted with inverse frequency $\frac{1}{N_c}$ to encourage the model to pay attention

to all classes for our somewhat imbalanced data; without this weighting the overall accuracy would be substantially lower due to the network focusing on the most frequent classes.

b) Weight estimation: The weights are predicted for all food items separately, so that $y \in \mathbb{R}_+^C$ has weights for all of them. The model is trained by minimizing mean absolute error (again summer over individual data points)

$$\sum_{c=1}^{C} \|y_c - z_c\|$$

where y_k are now the ground truth weights. We directly train for absolute error instead of the more commonly used mean square error since we want to emphasize accuracy for typical portions rather than penalising heavily predictions for untypical portions with very low or high weight.

B. Network architecture

1) ResNet architecture: We address both tasks with the same basic neural network architecture, changing only the final layers and loss to match the specific task, as explained above. Our network uses ResNet [14] as the backbone architecture, as one of the most popular architectures for computer vision problems. The ResNet architecture makes it easier to train deeper convolutional networks by implementing residual learning. Models building on ResNet have also been previously used in food imaging tasks [11], [16], [19], [20].

The backbone consists of either 51 or 101 convolutional layers where the layers are organized as residual blocks chained one after each other. Each residual block consists of a convolutional layer followed up by a ReLU (Rectified Linear Unit) activation function $max(0, x)$, followed by a second convolutional layer and another ReLU activation. Finally the output of the second ReLU is summed together with the original input of the residual block. This summation of the outputs with the inputs facilitates residual learning in the residual blocks, since each block can either implement the identity function $f(x) = x + 0$ simply by passing the original input as is, or apply some transformation $f(x) = x + g(x)$ where $g(x)$ is some non-zero non-linear function implemented by the two convolutional layers and ReLUs. In addition, Batch Normalization [21] is applied after each convolutional layer and before the ReLU activation function as in the original ResNet implementation. The backbone outputs a $2048 \times \frac{W}{32} \times \frac{H}{32}$ tensor where W and H are the spatial width and height dimensions of the original input. This tensor is flattened to a vector of 2048 elements by taking the mean over the spatial dimensions for each channel, and then passed through two fully connected layers added after the backbone. The final output of the model is again a vector of 2048 elements.

2) Accounting for the daily menu: The total number of food items (classes) considered in this work is fairly high, 130 after filtering out classes with under 10 images, especially relative to the amount of training examples (1,700). The number of food items available during a specific day, however, is considerably smaller, on average around 11 for the filtered data. To account for the menu information already during training, we consider two alternative ways of using the information.

The first method, called `MenuConcat`, concatenates the menu information as a C-element binary vector m encoding the presence of each item in the menu with the hidden state of the network, before the final fully connected layers. This simple encoding provides the network information about the available items, but does it in somewhat naive manner.

The second method, called `MenuProd`, is designed to directly incorporate binary side information about the menu into any convolutional later. We denote by F the number of channels in a convolutional layer of the ResNet architecture. We then introduce a learnable $C \times F$ matrix W and compute a *menu weighting vector*

$$p = mW \in \mathbb{R}^{1 \times F}$$

to obtain scalar weights representing the relative importance of the different channels for this particular menu. We weight the channels with these weights, before passing them forward to successive layers. This way of accounting for the menu has the advantage that the menu information can directly zero out feature channels (by multiplication with zero) that are only needed for prediction of items not currently available. The menu information could in principle be applied to any layer in the convolutional architecture, but based on preliminary experimentation we only apply it on the final layer.

We also evaluate the model without accounting for the menu, calling it `NoMenu`, keeping the two fully connected layers after the backbone ResNet for this model as well.

C. Training

We use stochastic gradient descent for optimization, with learning rate of 0.01 for food classification and for 1.00 for weight estimation, chosen based on preliminary experimentation. Both models are trained for 4,000 epochs and use a validation set for selecting the best performing intermediate model. For classification we use the 'Micro F1' metric and for weight prediction the 'True items' metric as the validation criterion; see Section IV-D for details on the metrics.

We use two common computer vision techniques to improve the accuracy: pre-training and data augmentation. The model is pre-trained with images from the ImageNet data [22] to improve learning of the convolutional filters. Our own data is then augmented by rotating each image by a random degree, and by horizontally flipping each image with a 0.5 probability. These transformations are applied independently on each image on each epoch, to maximize the total variation seen during training. The validation and test is performed only on the original images, without augmentation.

IV. Data

A. Data collection

Our data was collected from volunteer customers in the Flavoria research restaurant at Finland, during approximately five weeks. The data consists of (a) the images X of plates taken at the end of the self-service line and (b) ground truth

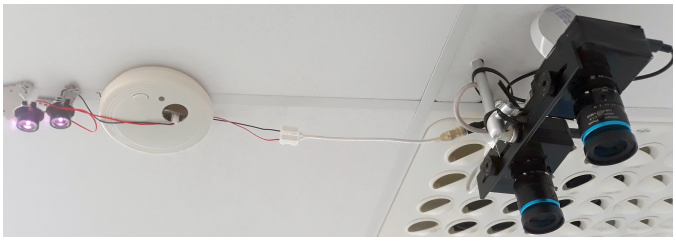

Fig. 2. Camera system used for imaging the plates.

information y obtained by weighing the plate after each food item picked up by the customer. The physical lunch line stations consist of RFID readers and weighing sensors along with customer displays to show lunch and weighing info. Trays are equipped with RFID tags for tracking.

For data collection we define a lunch line *session*, which stores information from each person's unique visit. The session can include, for example, meal types and weighings or waste data. Session data is collected/networked to a central server for storage from the physical devices in the lunch line. A session begins when a tray is first identified by any of the weighing systems in the lunch line. The session is set to expire at about 45 minutes after last identified action, so that trays can be reused after washing by any new customer. Waste points in the lunch line [7] expire the session almost immediately as we assume the customer will take a new tray or leaves the restaurant. Each used lunch component station adds weight information to the session along with what food component was in the station, set by the restaurant staff.

The plates were imaged with cameras mounted over the cash register area. We used separate RGB and infrared (IR) cameras illustrated in Figure 2, and pair of CCTV IR (840nm) emitters at a distance to avoid plate reflections. The IR and RGB cameras were spaced as near each other as physically possible. The plate was illuminated by existing LED-strips with some natural daylight coming through the windows. Despite imaging with both RGB and IR cameras, we eventually used only the RGB images for analysis since the information provided by the IR camera did not seem useful in preliminary experimentation, due to reasons explained in Section VI.

B. Preprocessing

The processing pipeline from raw measurements to the model inputs is shown in Figure 3. The weighing data and images are first paired based on temporal proximity, by comparing the image acquisition time with the session time stamps. The image data is then processed by extracting the plates. We use Hough circle transform to find the plates that are of near identical size due to constant imaging distance. The image is then automatically cropped to a rectangle the sides of which are the same length as the diameter of the plate, and the resulting corner's pixels outside the circular plates are zeroed out as can be seen in the black corner areas in Figure 3. Images for which the plate could not be identified were discarded. All of the images were then manually checked

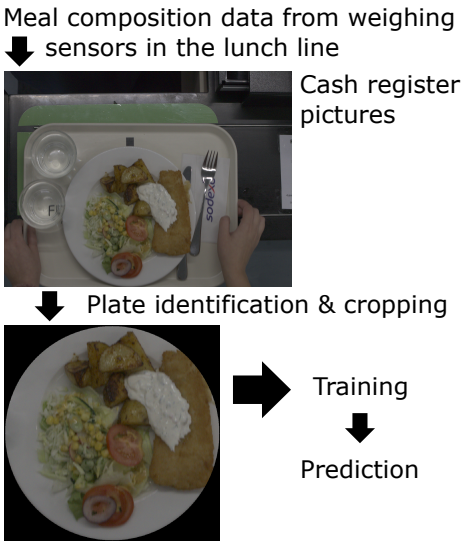

Fig. 3. Process flowchart. We gather images using the setup in Figure 2 and weights using existing series of scales. We automatically extract the plate from the original image containing the whole tray, and train the computer vision models only on the plate images. Privacy-related verification and data removal steps are not visualized here.

for any credit cards or other personal information, with images containing those removed from the dataset.

The weighing process is manual and error-prone, and hence the ground truth weights were processed to reduce the errors. All measurements below 9 grams were rounded to zero due to likely being erroneous readings. In addition, for each food item separately we computed the mean and standard deviation of weights and replaced the weights of individual measurements exceeding the mean by more than two standard deviations with the mean. This was done do correct for failed readings where the customer is e.g. leaning to the scale. Figure 4 shows how the true weights are fairly concentrated around the mean and hence this preprocessing step is unlikely to influence real readings significantly. While training the models we further normalized the weights of individual items with the median weights of each class to standardize them, but the results are evaluated for the real weights.

The weight measurements for customers who had not identified themselves using the Flavoria app [7] were deemed to be too untrustworthy/noisy to be included automatically, and so they were only used for building the validation set by hand picking good quality photos with seemingly valid weights. Finally, food classes present in fewer than 10 images in the training dataset were completely discarded as there would not be enough information to learn a reasonable model for them.

C. Dataset

The dataset *FlavoriaFoodWeight1700* is made publicly available at https://doi.org/10.5281/zenodo.5850856 to encourage follow-up research and for validating the results. Some meal names have been translated to English compared to the the ones used in our results for easier data validation.

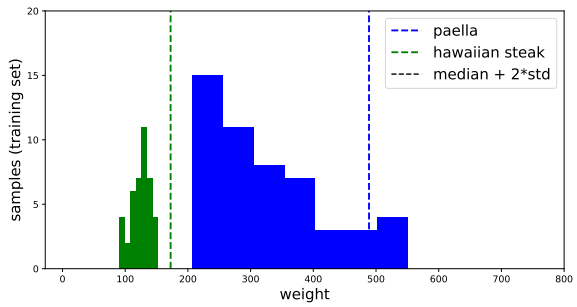

Fig. 4. Illustration of weight distributions and preprocessing. For some foods (hawaiian steak) the weights are concentrated on a narrow range, whereas for free-form foods (paella) the distribution is broader. Weights exceeding the mean by more than two standard deviations were assumed erroneous.

The non-cropped raw dataset (but with sensitive pictures removed/censored) is available on request, and for some validation and test samples we also have segmentation masks (not used in this study).

D. Learning setup and evaluation

Data for customers who used app identification is randomly split into a development and a test set, and the development set was further split into training and validation data. We use 1,716 images for training, and 104 and 193 for validation and testing, respectively. Some of the trays were photographed twice, both when the tray entered the checkout area and when the customer checked in with their personal QR code. For these cases both images were put in the same side of the development/test split.

We evaluate classification using *F1-score*, the harmonic mean of precision and recall. F1-score is more descriptive than accuracy, since prediction of a vector of zeroes (no food items) would here have very high accuracy due to average plate only having on average four food items out of the 130 possible ones. We use both macro and micro averages of F1, corresponding to average F1 over all classes and average over the F1 for the combination of all problems pooled together.

The weight prediction is evaluated using mean absolute error and compared against the baselines of (a) predicting the median weight of each food type (estimated from the training data) and (b) predicting all weights to be zero. The former is a strong baseline that for some food items (e.g. a sausage) is accurate due to their uniform weight, whereas the latter provides a simple yardstick as the error becomes the median weight itself. We evaluate the metric in three different ways to investigate the behavior of the model in more detail:

- `All`: Error for all 130 classes
- `Menu`: Error for the items appearing in the daily menu
- `True items`: Error for the true items on the plate

Note that median weight predictions evaluated only for the true items is extremely strong baseline that relies on oracle providing information on the items present.

TABLE I
CLASSIFICATION RESULTS

Model name	Macro F1	Micro F1
ResNet50 + MenuProd	0.874	0.871
ResNet50 + MenuConcat	0.888	0.881
ResNet50 + NoMenu	0.890	0.870
ResNet101 + MenuProd	**0.907**	0.881
ResNet101 + MenuConcat	0.902	**0.886**
ResNet101 + NoMenu	0.897	0.879

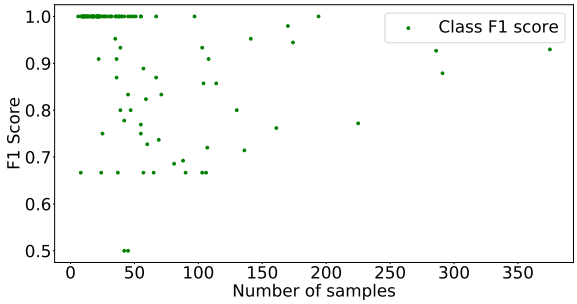

Fig. 5. Number of samples per class and the corresponding F1 score for each class for test set using the ResNet101 + Menu classification model.

V. RESULTS

A. Food recognition

We evaluate six model variants for solving the classification task, combining two ResNet backbones and three ways of accounting for the daily menu (one of which ignores it) described in Section III-B. The results are reported in Table I. The deeper model, ResNet101, outperforms the more shallow model, ResNet50, and incorporating the menu information improves the accuracy. Figure 5 illustrates the results on the level of individual food items, showing that for many of the classes we reach essentially perfect accuracy. For some of the smaller classes, however, the accuracy remains lower.

B. Weight Estimation

Table II shows the results for weight prediction, again for the six different network architectures. In addition, the test set errors are illustrated at the level of individual food items in Figure 6. We also evaluated an additional model variant that first predicts the items using the classifier model and then outputs median weight for each predicted class, which provides the best overall accuracy when measuring accuracy over all 130 food items. For the more relevant metrics measuring accuracy over the items in the daily menu or the ones the customer had on the plate, the direct weight prediction using the deeper ResNet with MenuProd is the best.

The differences between the model variants are, however, small, and they all solve the problem approximately as well for practical purposes. We reach approximately 15g error per food item over all items available in the menu, which is dramatic reduction from the baselines of 36g (zero prediction) and 84g (median prediction). Even when evaluating the accuracy only on the items the customer took (`True items`) the accuracy

TABLE II
WEIGHT ESTIMATION RESULTS

Model name	All	Menu	True items
Proposed models			
ResNet50 + MenuProd	1.32	14.83	25.85
ResNet50 + MenuConcat	1.49	15.11	26.47
ResNet50 + NoMenu	1.60	15.22	28.477
ResNet101 + MenuProd	1.306	**14.708**	**24.750**
ResNet101 + MenuConcat	1.497	15.215	25.865
ResNet101 + NoMenu	1.579	14.742	26.873
Classifier + MenuProd + median	**1.285**	14.745	25.781
Baselines			
Zero prediction	3.164	36.451	93.071
Median prediction	7.263	83.667	*23.329*

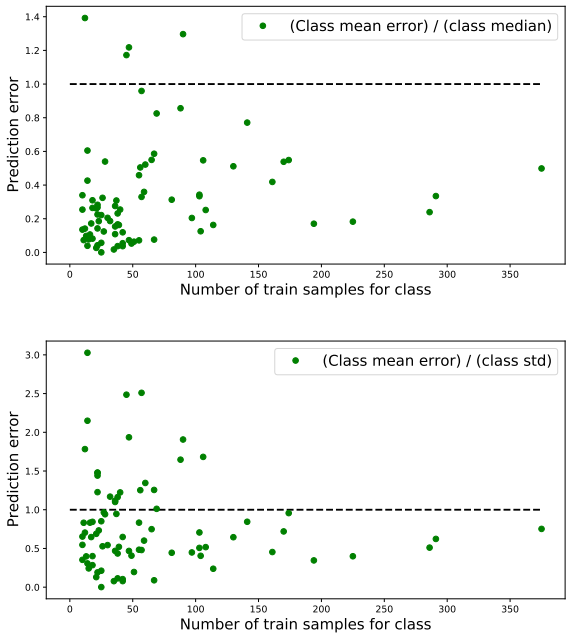

Fig. 6. Weight prediction errors by ResNet101 + MenuProd for individual classses, with two alternative scalings and baselines indicated by dashed lines. The top plot shows prediction errors divided by median weight of each class, and the line refers to error that matches the median weight. The bottom plot normalizes the errors with the standard deviation of each class, and the line indicates error made by naive mean predictor.

is close to the strong oracle that knows in advance which items the customer took and predicts their median weight, which is strong indication of accuracy.

VI. DISCUSSION

The proposed neural network architecture achieved high accuracy in estimating presence of individual food items, and in particular has F1 score above 0.65 for all but two classes. The performance is likely already sufficient for partial automation of food category identification and the accuracy could be increased further by collecting a larger training data, or by merging classes with high visual similarity and nutritional content (see Figure 1 for an example).

We can also predict weights of individual food items relatively well, but conditional on already knowing which items are on the plate the result is similar to simple mean prediction. This can be explained by two reasons: (1) For many of the food items, the weights are fairly closely concentrated as illustrated in Figure 4 and hence it is difficult to outperform the mean prediction, and (2) The task of estimating the weight from 2D image is fundamentally ill-posed due to severe occlusion, and has only been solved in earlier works by either using depth-based imaging or simplified learning problems with fixed-weight food items as discussed in Chapter II.

The results reported here are for RGB images alone, and hence provide information about the practical value of the imaging configuration that is easiest to set up. We also experimented with infrared imaging as shown in Figure 2, but terminated the data collection after observing the additional data source did not provide useful information in our setup. Even though infra-red imaging is likely to be useful in identification of food items, as suggested e.g. by [23], our technical setup aiming for simplicity of deployment was too limited. We used regular Raspberry Pi high-quality camera with the infrared filter removed, but the camera sensor was not very sensitive in the infrared areas of interest. In addition, our low-cost off-the-shelf LED CCTV illuminator was relatively weak and likely did not provide sufficient IR illumination. Better lighting and cameras should be used to produce a more definitive answer on whether the models can be improved by incorporating the infrared spectrum.

VII. CONCLUSION

We described an automated imaging platform and computer vision method for collecting information about food intake in lunch restaurants. We described the system for acquiring and preprocessing the images, as well as the ground truth information for training the models by weighing the plates after picking up each individual food item. We also described a deep neural network model for estimating both the types of different food items and their weights on individual plates. This model reaches 90% F1 score (averaged over classes) in a multi-label classification task of detecting the food items, and many of the misclassifications are between classes of similar appearance and nutritional content. The model is also able to estimate the weights to a degree, but the accuracy would not yet be sufficient for monitoring individual food intake in detail.

We release the data for public development and evaluation of better computer vision solutions for both tasks.

ACKNOWLEDGEMENTS

We thank Business Finland, grant *Mobile Spectral Imaging and Computer Vision Platform*, for funding part of this research. We also thank Huld Oy for the lunch line software and Elomatic Oy for the hardware development. We would also like to thank all our colleagues and the restaurant operator Sodexo Oy. As a research artifact we acknowledge the *Flavoria® multidisciplinary research platform* and extend our gratitude towards everyone involved in its development.

REFERENCES

[1] A. Doulah, M. A. Mccrory, J. A. Higgins, and E. Sazonov, "A systematic review of technology-driven methodologies for estimation of energy intake," *IEEE Access*, vol. 7, pp. 49 653–49 668, 2019.

[2] K. J. Pfisterer, R. Amelard, A. G. Chung, B. Syrnyk, A. MacLean, H. H. Keller, and A. Wong, "Automated food intake tracking requires depth-refined semantic segmentation to rectify visual-volume discordance in long-term care homes," *Scientific reports*, vol. 12, no. 1, pp. 1–16, 2022.

[3] N. Kaartinen, H. Tapanainen, H. Reinivuo, H. Pakkala, S. Aalto, S. Raulio, S. Männistö, T. Korhonen, S. Virtanen, K. Borodulin *et al.*, "The finnish national dietary survey in adults and elderly (findiet 2017) finnish institute for health and welfare, finland," *EFSA Supporting Publications*, vol. 17, no. 8, p. 1914E, 2020.

[4] L. Garden, H. Clark, S. Whybrow, and R. J. Stubbs, "Is misreporting of dietary intake by weighed food records or 24-hour recalls food specific?" *European journal of clinical nutrition*, vol. 72, no. 7, pp. 1026–1034, 2018.

[5] K. Poslusna, J. Ruprich, J. H. de Vries, M. Jakubikova, and P. van't Veer, "Misreporting of energy and micronutrient intake estimated by food records and 24 hour recalls, control and adjustment methods in practice," *British Journal of Nutrition*, vol. 101, no. S2, pp. S73–S85, 2009.

[6] E. Foster, C. Lee, F. Imamura, S. E. Hollidge, K. L. Westgate, M. C. Venables, I. Poliakov, M. K. Rowland, T. Osadchiy, J. C. Bradley *et al.*, "Validity and reliability of an online self-report 24-h dietary recall method (intake24): a doubly labelled water study and repeated-measures analysis," *Journal of nutritional science*, vol. 8, 2019.

[7] L. Koivunen, S. Laato, S. Rauti, J. Naskali, P. Nissilä, P. Ojansivu, T. Mäkilä, and M. Norrdal, "Increasing customer awareness on food waste at university cafeteria with a sensor-based intelligent self-serve lunch line," in *2020 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, 2020, pp. 1–9.

[8] P. F. Christ, S. Schlecht, F. Ettlinger, F. Grün, C. Heinle, S. Tatavarty, S. Ahmadi, K. Diepold, and B. H. Menze, "Diabetes60 - inferring bread units from food images using fully convolutional neural networks," in *2017 IEEE International Conference on Computer Vision Workshops, ICCV Workshops 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 2017, pp. 1526–1535. [Online]. Available: https://doi.org/10.1109/ICCVW.2017.180

[9] S. Fang, Z. Shao, R. Mao, C. Fu, E. J. Delp, F. Zhu, D. A. Kerr, and C. J. Boushey, "Single-view food portion estimation: Learning image-to-energy mappings using generative adversarial networks," in *2018 IEEE International Conference on Image Processing, ICIP 2018, Athens, Greece, October 7-10, 2018*. IEEE, 2018, pp. 251–255. [Online]. Available: https://doi.org/10.1109/ICIP.2018.8451461

[10] F. P. W. Lo, M. L. Jobarteh, Y. Sun, J. Qiu, S. Jiang, G. Frost, and B. Lo, "An intelligent passive food intake assessment system with egocentric cameras," *CoRR*, vol. abs/2105.03142, 2021. [Online]. Available: https://arxiv.org/abs/2105.03142

[11] W. Min, S. Jiang, L. Liu, Y. Rui, and R. C. Jain, "A survey on food computing," *ACM Comput. Surv.*, vol. 52, no. 5, pp. 92:1–92:36, 2019. [Online]. Available: https://doi.org/10.1145/3329168

[12] J. He, Z. Shao, J. Wright, D. A. Kerr, C. J. Boushey, and F. Zhu, "Multi-task image-based dietary assessment for food recognition and portion size estimation," in *3rd IEEE Conference on Multimedia Information Processing and Retrieval, MIPR 2020, Shenzhen, China, August 6-8, 2020*. IEEE, 2020, pp. 49–54. [Online]. Available: https://doi.org/10.1109/MIPR49039.2020.00018

[13] A. Myers, N. Johnston, V. Rathod, A. Korattikara, A. N. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. Murphy, "Im2calories: Towards an automated mobile vision food diary," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015, pp. 1233–1241. [Online]. Available: https://doi.org/10.1109/ICCV.2015.146

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. IEEE Computer Society, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[15] T. Sormunen, S. Järvinen, A. Lämsä, K. Immonen, J. Mannila, and J. Peltola, "Material sorting using hyperspectral imaging for biocomposite recycling," in *4th International Conference on Natural Fibers–Smart Sustainable Solutions, ICNF 2019: Book of Abstracts*, 2019, pp. 250–251.

[16] F. P. W. Lo, Y. Sun, J. Qiu, and B. Lo, "Image-based food classification and volume estimation for dietary assessment: A review," *IEEE J. Biomed. Health Informatics*, vol. 24, no. 7, pp. 1926–1939, 2020. [Online]. Available: https://doi.org/10.1109/JBHI.2020.2987943

[17] T. Ege, W. Shimoda, and K. Yanai, "A new large-scale food image segmentation dataset and its application to food calorie estimation based on grains of rice," in *Proceedings of the 5th International Workshop on Multimedia Assisted Dietary Management, MADiMa @ ACM Multimedia 2019, Nice, France, October 21-25, 2019*, S. G. Mougiakakou, G. M. Farinella, and K. Yanai, Eds. ACM, 2019, pp. 82–87. [Online]. Available: https://doi.org/10.1145/3347448.3357162

[18] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 – mining discriminative components with random forests," in *European Conference on Computer Vision*, 2014.

[19] N. Martinel, G. L. Foresti, and C. Micheloni, "Wide-slice residual networks for food recognition," in *2018 IEEE Winter Conference on Applications of Computer Vision, WACV 2018, Lake Tahoe, NV, USA, March 12-15, 2018*. IEEE Computer Society, 2018, pp. 567–576. [Online]. Available: https://doi.org/10.1109/WACV.2018.00068

[20] J. Qiu, F. P. W. Lo, Y. Sun, S. Wang, and B. Lo, "Mining discriminative food regions for accurate food recognition," in *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*. BMVA Press, 2019, p. 158. [Online]. Available: https://bmvc2019.org/wp-content/uploads/papers/0839-paper.pdf

[21] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning, ICML 2015, Lille, France, 6-11 July 2015*, ser. JMLR Workshop and Conference Proceedings, F. R. Bach and D. M. Blei, Eds., vol. 37. JMLR.org, 2015, pp. 448–456. [Online]. Available: http://proceedings.mlr.press/v37/ioffe15.html

[22] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[23] R. Lapcharoensuk, A. Malithong, D. Thappho, and P. Phonpho, "Discrimination of vegetable oil types using fourier transforms near infrared spectroscopy coupled with pattern recognition techniques," in *IOP Conference Series: Earth and Environmental Science*, vol. 301, no. 1. IOP Publishing, 2019, p. 012067.