SINGLE VIEW RECONSTRUCTION FOR FOOD PORTION ESTIMATION

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Shaobo Fang

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

May 2019

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Edward J. Delp, Co-chair

School of Electrical and Computer Engineering

Dr. Fengqing Zhu, Co-chair

School of Electrical and Computer Engineering

Dr. Amy R. Reibman

School of Electrical and Computer Engineering

Dr. Carol J. Boushey

Department of Nutrition Science

Approved by:

Dr. Pedro Irazoqui

Head of the School Graduate Program

This thesis is dedicated to my family
who has always been supporting me.

ACKNOWLEDGMENTS

First of all, I would like to express my sincere gratitude toward my major professor and co-chair, Professor Edward J. Delp. Professor Delp has provided me with both the opportunity to work in the Video and Image Processing Laboratory (VIPER lab), and his guidance for me to become an independent researcher. During the time pursuing my Ph.D., I have learned many insights that can not be easily learned elsewhere. He has taught me how to collaborate with my colleagues, attention to details, ownership of the work and most importantly, how to approach a challenging research problem. Professor Delp has been my role model for professionalism. His detailed comments and many iterations of communications when I first wrote a paper had helped me a lot, and I will forever be grateful for his efforts supervising me. He guided me through the process of breaking a challenging problem into different parts and making progresses step-by-step. And I can not express all my gratitude for all that in a few sentences.

I would also like to express my sincere gratitude toward my major professor and co-chair, Professor Fengqing Zhu. Professor Zhu has provided me with many detailed suggestions and advises. I deeply appreciate all the time and efforts she have spent discussing with me what is next for research and especially for her patience going through many revisions of manuscripts.

Again, I would like to express my sincere gratitude towards both Professor Edward J. Delp and Professor Fengqing Zhu, for their encouragements, supports, guidances, and being my role models holding the highest professional standard. The experience pursuing my Ph.D. at VIPER lab has been exceptional.

I am also truly grateful to the members of my dissertation committee: Professor Amy R. Reibman and Professor Carol J. Boushey for their valuable advices and guidances.

I am deeply grateful to Purdue University, the School of Electrical and Computer Engineering and the Graduate School. It is an amazing experience for me pursuing my Bachelor's, Master's and Doctorate's degrees at Purdue University. I am thankful for all the supports I have received from friends, staffs and faculty members at different stages of my journey. I appreciate all the advices and courses offered by professor Jan P. Allebach, professor Mike R. Melloch, professor Amy R. Reibman, professor Stanley Ho Chan and many other professors I met throughout the years at the school of Electrical and Computer Engineering.

It is a pleasure to meet all the colleagues at VIPER lab, for the collaborative working environment, the fun conversations and the supports for each other. I would like to thank all the members (last name in alphabetical order): Mr. Sriram BaireddyMr. Enyu Cai, Ms. Di Chen, Ms. Qingshuang Chen, Mr. Yuhao Chen, Ms. Jeehyun Choe, Ms. Dahjung Chung Ms. Blanca Delgado, Mr. Chichen Fu, Dr. Neeraj J. Gadgil, Mr. David Gera, Ms. Shuo Han, Mr. Hanxiang Hao Mr. Jiangpeng He, Mr. David Joon Ho, Mr. Jnos Horvth, Dr. Joonsoo Kim, Mr. Soonam Lee, Mr. He Li, Ms. Chang Liu, Mr. Daniel Mas, Mr. Runyu Mao, Dr. Albert Parra, Mr. Javier Ribera, Ms. Ruiting Shao, Mr. Zeman Shao, Dr. Khalid Tahboub, Dr. Yu Wang, Mr. Changye Yang, Mr. Sri Kalyan Yarlagadda, Mr. Jiaju Yue, Dr. Bin Zhao, Mr. Yifan Zhao and the lab visitor Ms. Thitiporn Pramoun, for the experiences traveling to conferences together and exchanging ideas. In particular, I would like to extend my gratitude toward Chichen Fu, who has been a good friend and a great colleague. I really appreciate all the discussions with him on research and exchange opinions on what could be the potential improvement late night in the lab.

Out of the all the members in the VIPER lab, I would like to highlight the colleagues on the food image analysis project. It always has to be a team effort to make progresses and achieve goals. I thank Ms. Chang Liu, Dr. Yu Wang and Mr. Sri Kalyan Yarlagadda for being wonderful teammate and the collaborative team efforts maintaining all the small aspects of projects on the server and on the front-end devices. Especially, I would like to thank Ms. Chang Liu, for the collaborative work in my early stage on research.

I would also like to thank the members on the food image analysis project who have recently joined, Mr. Jiangpeng He, Mr. Runyu Mao and Mr. Zeman Shao. They are exceptional colleagues and the experience working with them has been amazing. I appreciate their collaborative efforts on the project, the efficiency getting tasks done, and the innovative ideas brought up in research meetings.

Furthermore, I would also like to thank all who have provided support and help during my assignment as a teaching assistant. I would especially like to thank Dr. Matthew Swabey for his support when I was teaching assistant for "ECE208: Electronic Devices and Design Laboratory" and "ECE 207: Electronic Measurement Techniques" and assigning me to the lead teaching assistant role. The lead teaching assistant role has provided me with a unique experience examining the importance of team work from a different perspective and have a better understanding on planning and coordination. I have received the excellence in teaching award, but it really was the joint work done together by fellow graduate teaching assistants, the lab staffs who support daily operations and the undergraduate teaching assistants that made everything work out the right way.

Last but not the least, I would like to say "Thank you", to all who supported and helped.

This work was partially sponsored by National Science Foundation under grant 1657262, and US National Institutes of Health under grant NIH/NCI 1U01CA130784-01 and NIH/NIDDK 1R01-DK073711-01A1, 2R56DK073711-04, and a Healthway Health Promotion Research Grant from the Department of Health, Western Australia, and by the endowment of the Charles William Harrison Distinguished Professorship at Purdue University.

TABLE OF CONTENTS

[1] This section is in joint work with Mr. Zeman Shao

[2] This section is partially in joint work with Ms. Chang Liu

[3] This section is in joint work with Mr. Runyu Mao

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

ABSTRACT

Fang, Shaobo. Ph.D., Purdue University, May 2019. Single View Reconstruction for Food Portion Estimation. Major Professor: Edward J. Delp, Fengqing Zhu.

3D scene reconstruction based on single-view images is an ill-posed problem since most 3D information has been lost during the projection process from the 3D world coordinates to the 2D pixel coordinates. To estimate the portion of an object from a single-view requires either the use of priori information such as the geometric shape of the object, or training based techniques that learn from existing portion sizes distribution. In this thesis, we present a single-view based technique for food portion size estimation.

Dietary assessment, the process of determining what someone eats during the course of a day, provides valuable insights for mounting intervention programs for prevention of many chronic diseases such as cancer, diabetes and heart diseases. Measuring accurate dietary intake is considered to be an open research problem in the nutrition and health fields. We have developed a mobile dietary assessment system, the Technology Assisted Dietary Assessment$^{\text{TM}}$ (TADA$^{\text{TM}}$) system to automatically determine the food types and energy consumed by a user using image analysis techniques.

In this thesis we focus on the use of a single image for food portion size estimation to reduce a user's burden from having to take multiple images of their meal. We define portion size estimation as the process of determining how much food (or food energy/nutrient) is present in the food image. In addition to estimating food energy/nutrient, food portion estimation could also be estimating food volumes (in cm^3) or weights (in grams), as they are directly related to food energy/nutrient. Food portion estimation is a challenging problem as food preparation and consumption process can pose large variations in food shapes and appearances.

As single-view based 3D reconstruction is in general an ill-posed problem, we investigate the use of geometric models such as the shape of a container that can help to partially recover 3D parameters of food items in the scene. We compare the performance of portion estimation technique based on 3D geometric models to techniques using depth maps. We have shown that more accurate estimation can be obtained by using geometric models for objects whose 3D shape are well defined. To further improve the food estimation accuracy we investigate the use of food portions co-occurrence patterns. The food portion co-occurrence patterns can be estimated from food image dataset we collected from dietary studies using the mobile Food Record$^{\text{TM}}$ (mFR$^{\text{TM}}$) system we developed. Co-occurrence patterns is used as prior knowledge to refine portion estimation results. We have been shown that the portion estimation accuracy has been improved when incorporating the co-occurrence patterns as contextual information.

In addition to food portion estimation techniques that are based on geometric models, we also investigate the use deep learning approach. In the geometric model based approach, we have focused on estimation food volumes. However, food volumes are not the final results that directly show food energy/nutrient consumed. Therefore, instead of developing food portion estimation techniques that lead to an intermediate results (food volumes), we present a food portion estimation method to directly estimate food energy (kilocalories) from food images using Generative Adversarial Networks (GANs). We introduce the concept of an "energy distribution" for each food image. To train the GAN, we design a food image dataset based on ground truth food labels and segmentation masks for each food image as well as energy information associated with the food image. Our goal is to learn the mapping of the food image to the food energy. We then estimate food energy based on the estimated energy distribution image. Based on the estimated energy distribution image, we use a Convolutional Neural Networks (CNN) to estimate the numeric values of food energy presented in the eating scene.

1. INTRODUCTION

1.1 Problem Formulation and Challenges

Due to the growing concern of chronic diseases and other health problems related to diet, there is a need to develop accurate methods to estimate an individual's food and energy intake. Dietary assessment, the process of determining what someone eats during the course of a day, provides valuable insights for mounting intervention programs for prevention of many of the chronic diseases. Measuring accurate dietary intake is considered to be an open research problem in the nutrition and health fields. Traditional dietary assessment technique, for example the written dietary record, is a time consuming and tedious process, which requires individuals to keep detailed written reports for 3-7 days of all food or drinks consumed [1, 2].

By February 2016, 72% of American adults were smartphone owners and there has been a noticeable rise in mobile phone and the Internet usage in the past few years in the emerging and developing nations [3]. Smartphones provide a unique mechanism for collecting dietary information and monitoring personal health. With smartphones' capabilities of capturing images, connecting to the Internet and on-device computation ability, a user's burden to keep dietary record could be significantly reduced by using image-based dietary assessment techniques. For example, instead of keeping a detailed written record, a user can capture an eating occasion image using their mobile phones. Then, food types and food energy/nutrient amounts consumed can be estimated based on the captured eating occasion image. The importance of using eating occasion images to record and estimate dietary intake versus traditional approaches has been highlighted in [4, 5].

In recent years, several image-based dietary assessment systems that use food images acquired by mobile phones during eating occasions, have been developed to automatically determine the food types and energy consumed using image analysis and computer vision techniques. Such systems include the TADA$^{\text{TM}}$ system [6, 7], FoodLog [8], FoodCam [9], DietCam [10], and Im2Calories [11].

There are two important tasks in image-based automatic dietary assessment, one is food recognition/classification and the other is food portion estimation. Food recognition/classification determines food types. Food portion estimation determines how much energy/nutrient a user consumes. For food recognition/classification task, recent works focused on the use of Convolutional Neural Networks (CNN) have shown impressive results on benchmark food datasets [12]. Although CNN has become a common architecture for food classification, to date, food portion estimation techniques developed by different groups are based on fundamentally different approaches.

To obtain the food portions of energy/nutrients, food volumes could be useful intermediate results. Existing works in food volume estimation focused on investigating the geometric features of food objects. Based on the estimated food volumes, food weights (using food density) and food energy/nutrients are then obtained. Food volumes are not the necessary intermediate results for estimation of food energy/nutrients. Therefore, it is possible to skip the food volumes estimation step and directly estimate food energy/nutrients.

Furthermore, different dietary assessment systems may capture the eating scenes differently (for example, single-view versus multi-view, using RGB sensors versus using RGB-D sensors). Therefore, different food portion estimation techniques may have different requirements on input images. As a result, food portion estimation remains an open research problem. In this thesis we focus on addressing the challenge of estimating food portions in the Technology Assisted Dietary Assessment$^{\text{TM}}$ (TADA$^{\text{TM}}$) system that we developed.

We have been developing the Technology Assisted Dietary Assessment$^{\text{TM}}$ (TADA$^{\text{TM}}$) system as shown in Figure 1.1, to acquire and process food images [4, 6, 7]. The TADA$^{\text{TM}}$ system and the associated mobile Food Record$^{\text{TM}}$ (mFR$^{\text{TM}}$) application allows users to acquire food image using a mobile telephone. Our goal is to determine

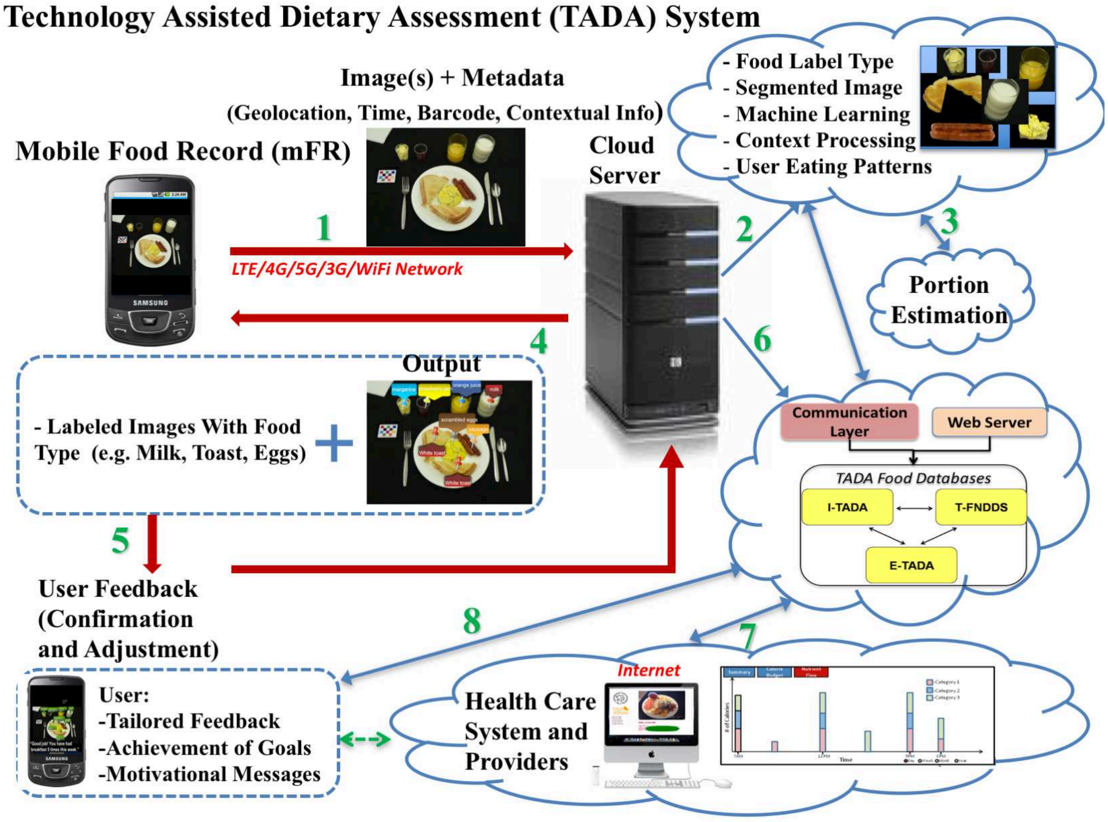


Fig. 1.1.: The Technology Assisted Dietary Assessment$^{\text{TM}}$ (TADA$^{\text{TM}}$) system.

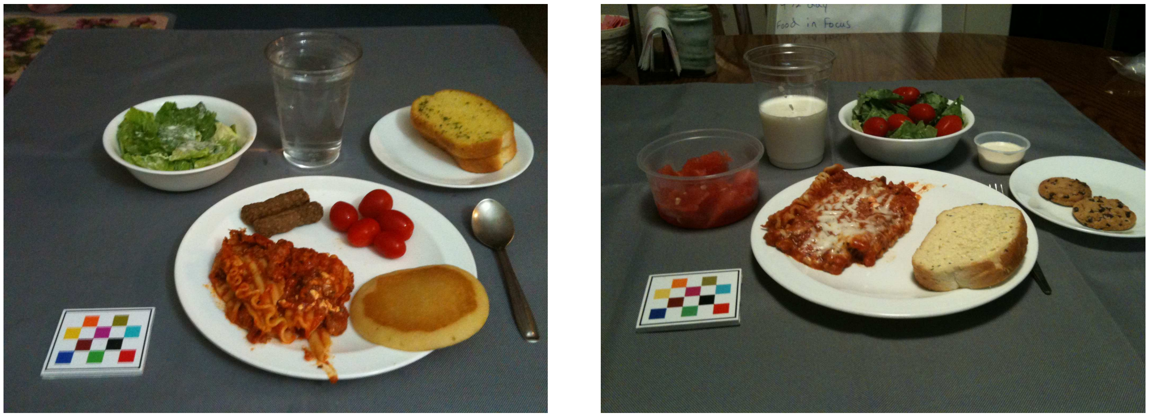


Fig. 1.2.: Examples of eating scene images.

what types of food and how much energy is consumed by a user during the course of a day. Image processing and computer vision analysis techniques have been used to determine the food type, portion size, the energy (kilocalories) and nutrients of the food [7,13,14]. As using the single image reduce a user burden capturing the eating scene, the food portion estimation technique we are developing is focused on the use of a single-view eating occasion image captured by using the mFR$^{\text{TM}}$ on a consumer mobile device (for example, iPhones and Android phones). To date the eating scenes are captured in RGB digital images.

Food portion estimation based on a single-view image is an ill-posed problem since most 3D information has been lost during the projection process from the 3D world coordinates onto the 2D pixel coordinates. To date, there is no 3D reconstruction techniques in computer vision that fully reconstruct the 3D structures of objects based on a single-view RGB image. Complete 3D reconstruction of the eating scene based on a single RGB image is not possible. For example, the food regions that are not visible from the viewing angle of the eating scene image can not be reconstructed. Furthermore, different food preparation and consumption processes impose large variations on food shapes and appearances that also add to challenges for food portion estimation. For example, as shown in Figure 1.2 spaghetti has different shapes and garlic bread has different cooking conditions.

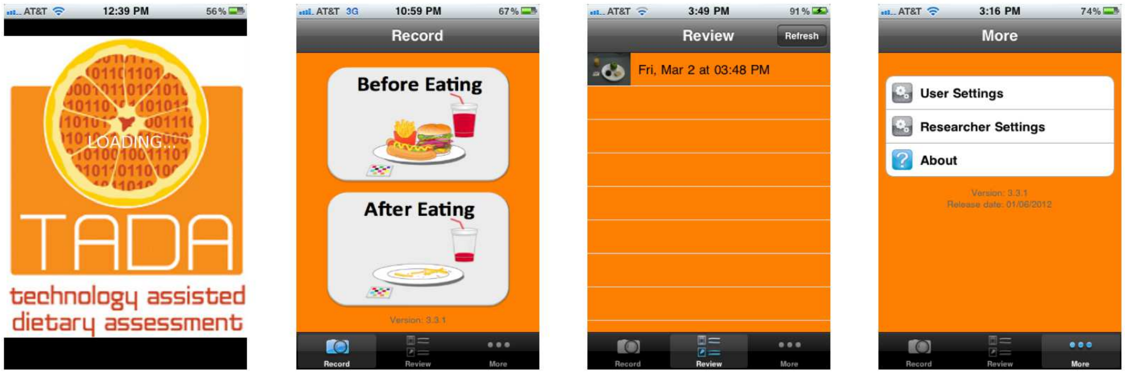


Fig. 1.3.: TADA^{TM} mFR^{TM} user interface.

To conclude, the 3D structure information of food objects is always limited and incomplete in the single-view eating scene image. The challenge of single-view food portion estimation is essentially obtaining an accurate food energy/nutrient estimate given that complete 3D food objects reconstruction is not possible. To obtain accurate estimates of food portions, the use of prior information is required. The prior information includes calibration object used as a reference in the eating scene, the food shapes, the food combination patterns and food portion distributions estimated from existing food image dataset.

1.2 Overview of the Technology Assisted Dietary Assessment^{TM} System

We have developed a mobile dietary assessment system, the Technology Assisted Dietary Assessment^{TM} (TADA)^{TM} system [1,6,15] as shown in Figure 1.1 to automatically determine the food types and energy consumed by a user using image analysis techniques [7, 13, 14].

The TADA^{TM} system consists of two main parts: a mobile application that runs on a mobile device, also know as the mobile food record^{TM} (mFR^{TM}), and the "backend" cloud-like system consisting of the database servers which includes the food image database and food nutrients database system, and the computational server for food-image analysis techniques [16].

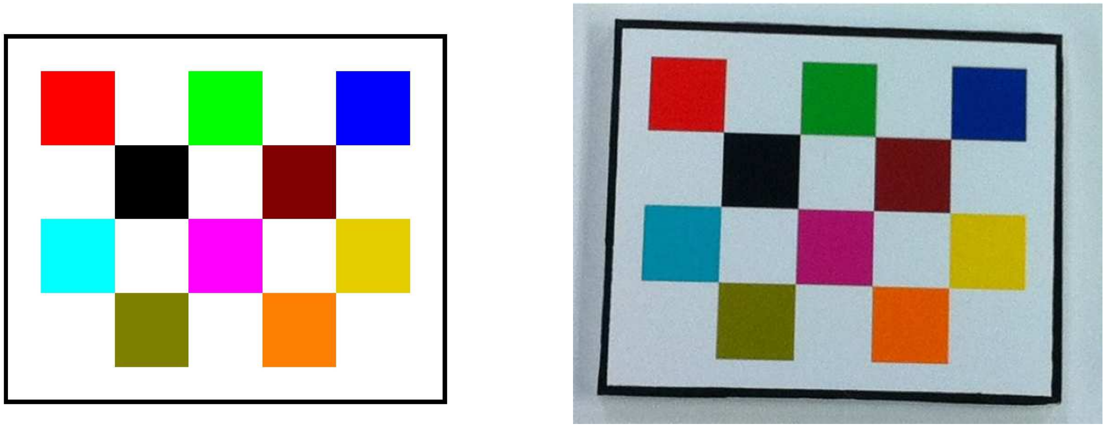
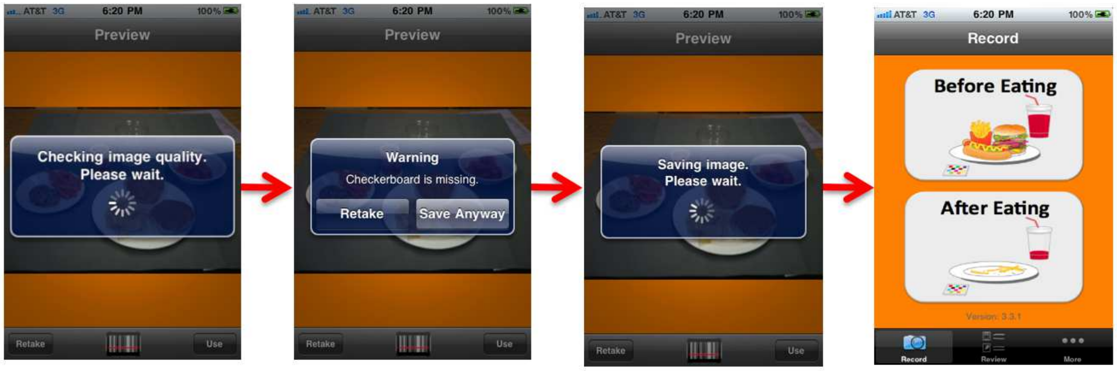


Fig. 1.4.: The fiducial marker pattern (left) and a cropped out fiducial marker (right) from the eating scene image captured using the TADA^{TM} app.

The TADA^{TM} mobile app is available for iOS 8 (iPhone, iPod and iPad) and above, and Android 4.3 and above at this moment. The user interface of TADA^{TM} system is shown in Figure 1.3. User can use the TADA^{TM} app to acquire before and after eating scenes images. The food images captured using TADA^{TM} mFR^{TM} are then uploaded to our server for processing. We implement image analysis and computer vision techniques for food segmentation, food classification and food portion estimation tasks on our servers.

We use a checkerboard pattern color fiducial marker (FM) shown in Figure 1.4, to provide essential information for color correction [17] and food portion estimation [14]. As most of the 3D information has been lost projecting food objects from 3D world coordinates onto the 2D image coordinates, the known size FM provides world scale reference and can also be used as a calibration target for camera calibration during food portion estimation using geometric models. To ensure that fiducial marker can be detected in the food image, we implement fiducial marker detection and blur detection on mobile app as shown in Figure 1.5.

In addition to the food images, metadata of the user can also be acquired. The metadata that will be sent to our server together with the before and after eating scene images include a user's ID (each participant has been assigned a unique user ID

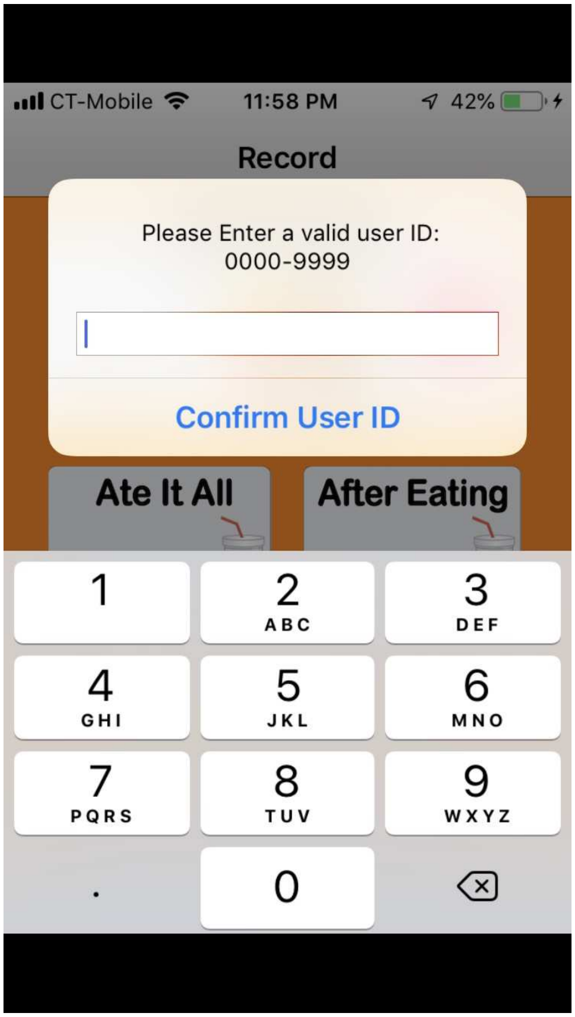


Checking image quality.
Please wait.

Warning
Checkerboard is missing.

Saving image.
Please wait.

Fig. 1.5.: Fiducial marker detection on TADA$^{\mathrm{TM}}$ app.



Please Enter a valid user ID:
0000-9999

Fig. 1.6.: A valid user ID must be entered prior to the first use of the mobile app.

in a dietary study), the time stamp of the food images taken and the GPS coordinates of the food images where they were taken (location information is optional and user can choose not to include GPS coordinates in the metadata).

The user ID is important for identifying which user has sent a specific eating scene images. To make sure that a user enter the assigned user ID before capturing food images, the first time a user downloads and uses the mFR$^{\mathrm{TM}}$, a notification will be popped up as a reminder as shown in Figure 1.6. A user will not be able to proceed using the mFR$^{\mathrm{TM}}$ unless the user ID has been entered.

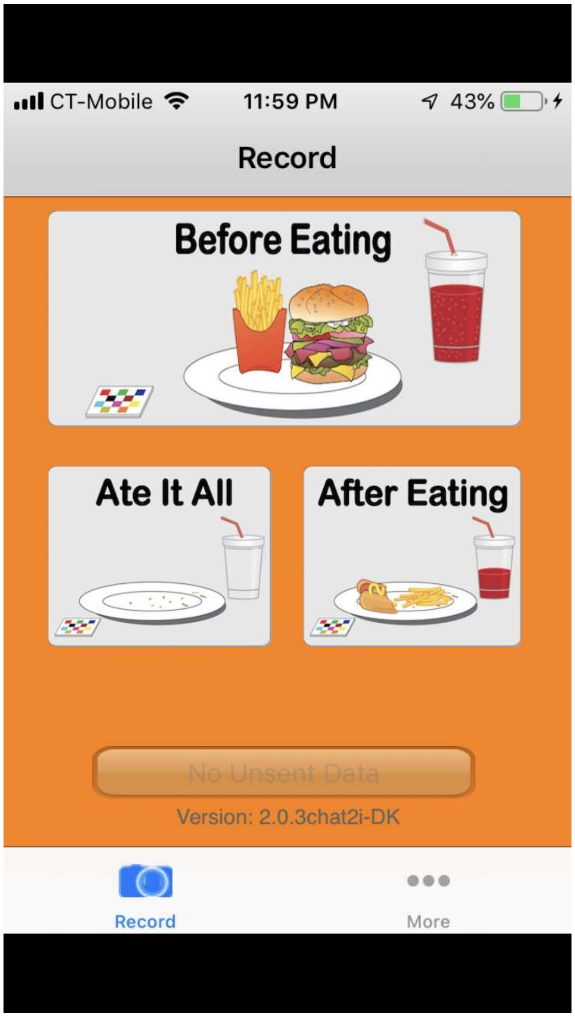
To assist the users using the mobile app, we have implemented the “Ate It All” function and the “Send Unsent Data” function. The system implementation for the above two functions is also part of the contribution for this thesis.

Sometimes a user consumes all food items that have been presented in the before eating image. Therefore, we have implemented the feature “Ate It All” button as shown in 1.7 to further simplify the process for users who consume all food presented. By clicking the “Ate It All” button, a user will no longer need to capture another image for the after eating scene for the estimation of food residue. The use of “Ate It All” button further reduce a user’s burden capturing the eating scenes.

As users will use the TADA$^{\text{TM}}$ mobile app in a free living condition, it is inevitable that occasionally the captured eating scene images could not be immediately sent to our server for image analysis due to no connection or weak signal. For the food images that are captured but not sent successfully, we will have the food images stored on a user’s mobile device and send the eating scenes images later when stable Internet connection becomes available. We have implemented the “Send Unsent Data” button as shown in Figure 1.8b so that a user can attempt to send captured images to our server. In addition, the “Send Unsent Data” button could also be used as a reminder to inform users of unsent data currently saved on their device. With all eating scene images have been successfully sent to our server, the “Send Unsent Data” button will be disabled and show “No Unsent Data” as shown in Figure 3.4.

The backend server processes food images captured using mFR$^{\text{TM}}$ and hosts a web interface for researchers in dietary study to examine eating occasion images captured by the participants. The main page of internal website (where only researchers can access using correct credentials) is shown as in Figure 1.9. The eating occasion images are organized in “before” and “after” image pairs as shown in Figure 1.10.

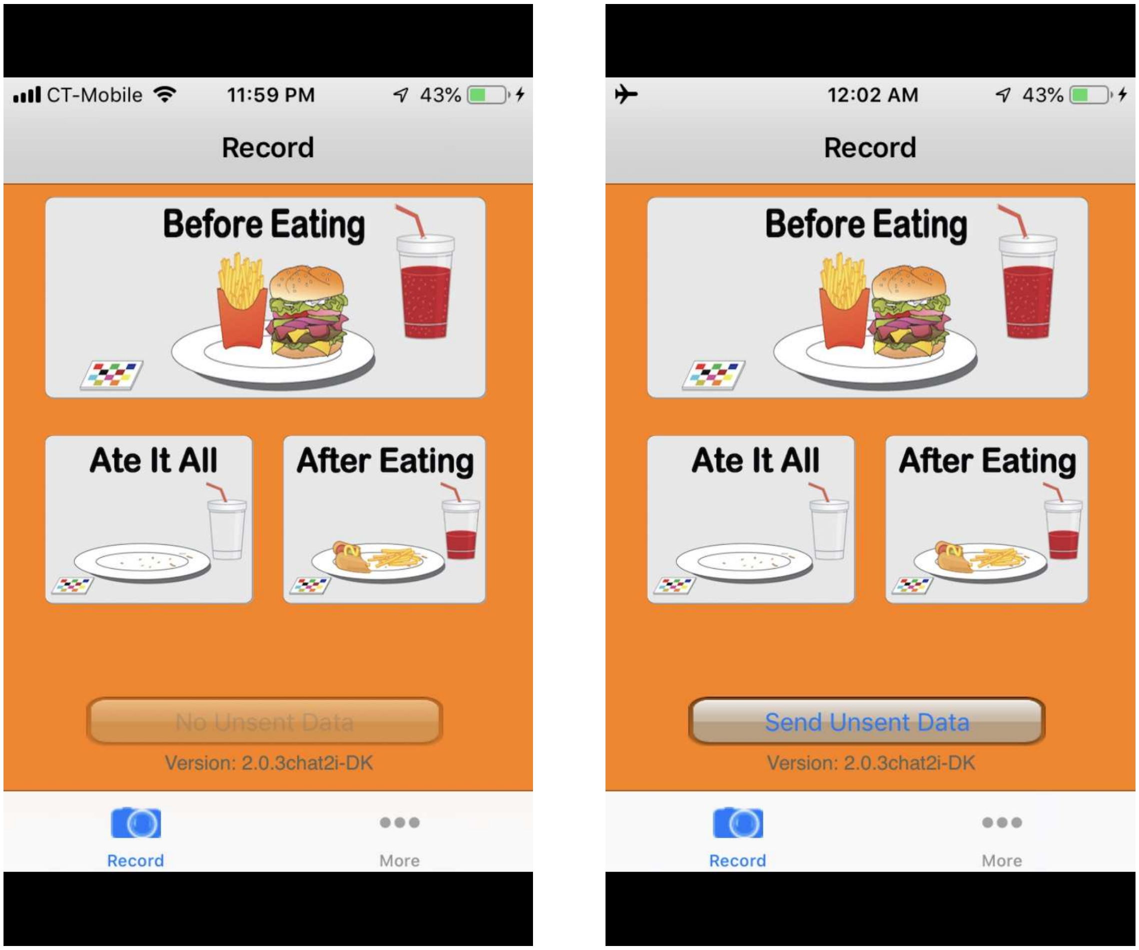
To date we have collected food images from different dietary studies. To organize the captured eating occasion images in different dietary studies, “Image Archive (I-TADA)” has been implemented. Researchers can examine eating occasion images based on dietary study tags through “I-TADA” as shown in Figure 1.11.



CT-Mobile 11:59 PM 43%

Record

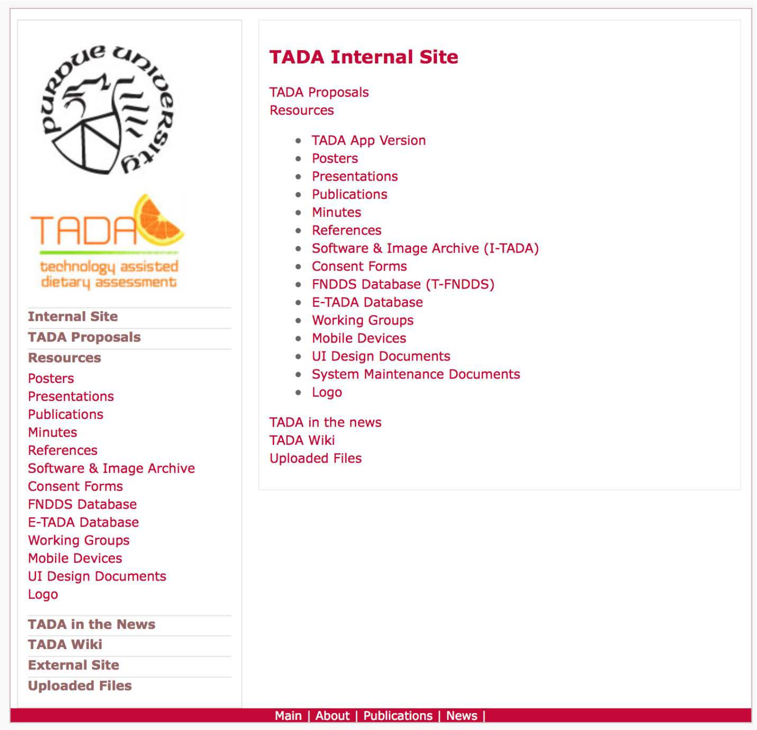
Fig. 1.7.: A user can choose to skip capturing the after eating image using “Ate It All” button.



(a) No unsent eating scene images. The “Send Unsent Data” button is disabled and shows the message: “No Unsent Data”.

(b) Unsent eating scene images saved in mFR^{TM}.

Fig. 1.8.: The “Send Unsent Data” button will be enabled and a user can click the button to send food images when stable Internet connectivity becomes available.



Internal Site
TADA Proposals
Resources
Posters
Presentations
Publications
Minutes
References
Software & Image Archive
Consent Forms
FNDDS Database
E-TADA Database
Working Groups
Mobile Devices
UI Design Documents
Logo

TADA in the News
TADA Wiki
External Site
Uploaded Files

TADA Internal Site

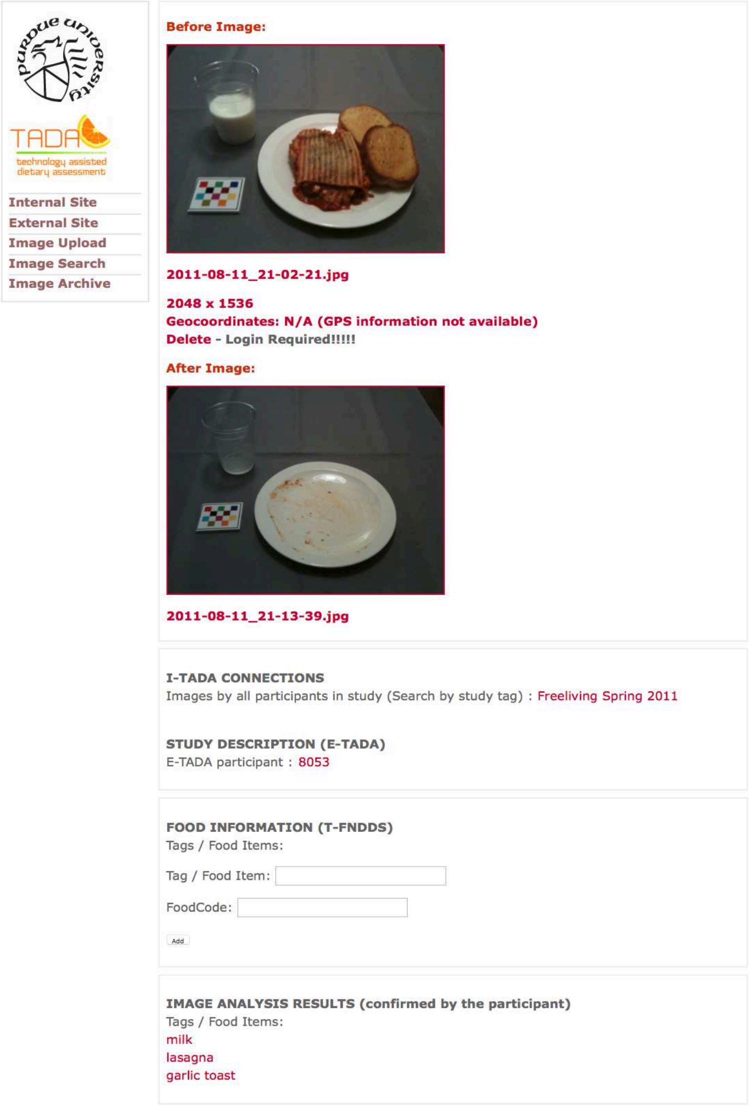
TADA Proposals
Resources

- TADA App Version
- Posters
- Presentations
- Publications
- Minutes
- References
- Software & Image Archive (I-TADA)
- Consent Forms
- FNDDS Database (T-FNDDS)
- E-TADA Database
- Working Groups
- Mobile Devices
- UI Design Documents
- System Maintenance Documents
- Logo

TADA in the news
TADA Wiki
Uploaded Files

Main | About | Publications | News |

Fig. 1.9.: The internal access user interface for researchers of dietary studies.



Before Image:

2011-08-11_21-02-21.jpg

2048 x 1536
Geocoordinates: N/A (GPS information not available)
Delete - Login Required!!!!!

After Image:

2011-08-11_21-13-39.jpg

I-TADA CONNECTIONS
Images by all participants in study (Search by study tag) : Freeliving Spring 2011

STUDY DESCRIPTION (E-TADA)
E-TADA participant : 8053

FOOD INFORMATION (T-FNDDS)
Tags / Food Items:

Tag / Food Item:

FoodCode:

IMAGE ANALYSIS RESULTS (confirmed by the participant)
Tags / Food Items:
milk
lasagna
garlic toast

Fig. 1.10.: Eating occasion images displayed in “before” and “after” image pairs.

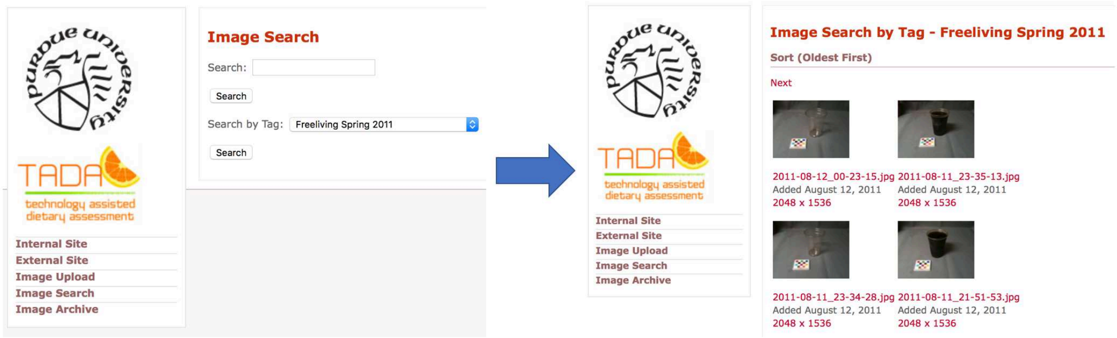


Image Search

Search:

Search by Tag: Freeliving Spring 2011

Image Search by Tag - Freeliving Spring 2011

Sort (Oldest First)

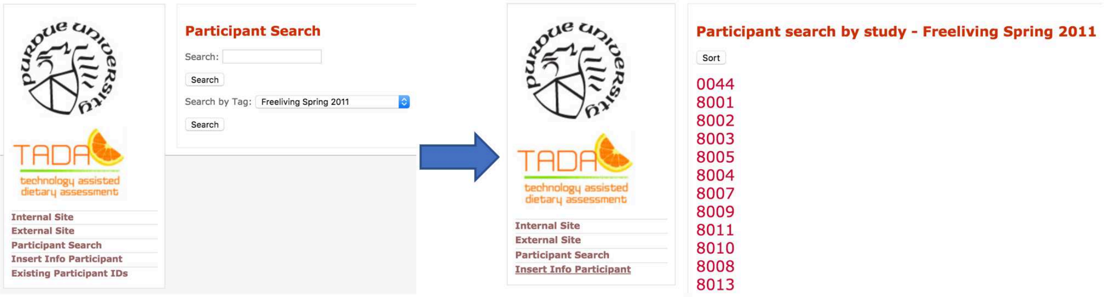
2011-08-12_00-23-15.jpg
Added August 12, 2011
2048 x 1536

2011-08-11_23-35-13.jpg
Added August 12, 2011
2048 x 1536

2011-08-11_23-34-28.jpg
Added August 12, 2011
2048 x 1536

2011-08-11_21-51-53.jpg
Added August 12, 2011
2048 x 1536

Fig. 1.11.: Eating occasion images indexed by dietary study tag through “I-TADA”.



Search by Tag: Freeliving Spring 2011

Participant search by study - Freeliving Spring 2011

0044
8001
8002
8003
8005
8004
8007
8009
8011
8010
8008
8013

Fig. 1.12.: User IDs indexed by dietary study tag through "E-TADA".

In addition to searching food images indexed by dietary studies, researchers can also search user IDs associated with a specific dietary study through "E-TADA" as shown in Figure 1.12.

Researchers can then examine all food images captured by a participant indexed by the user ID as shown in Figure 1.13.

The image analysis technique we implemented on the server include food region segmentation, food classification and food portion estimation as shown in Figure 1.1. In this thesis we focus on the food portion estimation of the TADA^{TM} system. We define portion size estimation as the process of determining how much food energy/nutrient is present in the food image. Food volumes or food weights (in cm^3 or grams) are useful intermediate results for food energy/nutrient estimation, therefore food volumes or food weights estimation can also be considered as food portion estimation. For example, after obtaining the volume of each food, we can estimate the weight using the food density (measured in grams/cubic centimeter [18]). The food energy (in kilocalories) can then be obtained from the United States Department of Agriculture (USDA) Food and Nutrient Database for Dietary Studies (FNDDS) [19].

1.3 Contributions of This Thesis

In this thesis we first investigate the use of geometric models for food portion estimation based on single-view eating occasion images. We focused primarily on

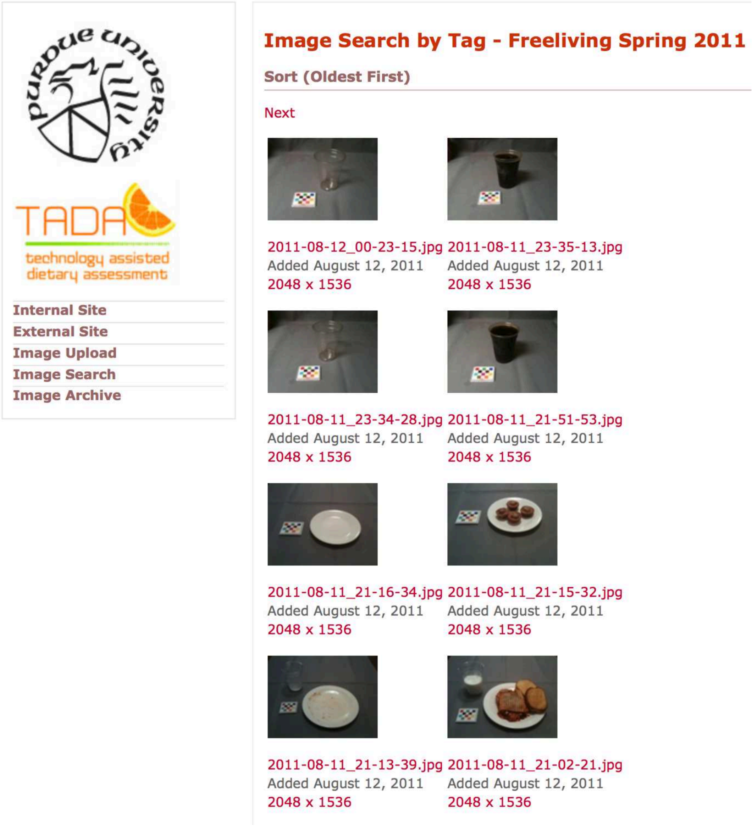


Image Search by Tag - Freeliving Spring 2011

Sort (Oldest First)

Next

2011-08-12_00-23-15.jpg
Added August 12, 2011
2048 x 1536

2011-08-11_23-35-13.jpg
Added August 12, 2011
2048 x 1536

2011-08-11_23-34-28.jpg
Added August 12, 2011
2048 x 1536

2011-08-11_21-51-53.jpg
Added August 12, 2011
2048 x 1536

2011-08-11_21-16-34.jpg
Added August 12, 2011
2048 x 1536

2011-08-11_21-15-32.jpg
Added August 12, 2011
2048 x 1536

2011-08-11_21-13-39.jpg
Added August 12, 2011
2048 x 1536

2011-08-11_21-02-21.jpg
Added August 12, 2011
2048 x 1536

Fig. 1.13.: Eating occasion images indexed by a user's ID.

cylinder model and prism model. The food portions are estimated in volumes (cm^3) using geometric models. We were able to obtain accurate estimates of food portions based on well-defined 3D models, camera calibration objects, correct food labels and correct food segmentation masks. We compared the accuracy between food portion estimation techniques using geometric models and using depth image. We show that portion estimation based on geometric models is more accurate for objects with well-defined 3D shapes compared to estimation using depth images. To further improve food portion estimation accuracy, we use co-occurrence patterns as prior knowledge to refine portion estimation results. In addition to food portion estimation using geometric models, we developed another approach based on the use of Generative Adversarial Networks (GAN). We introduce the concept of an "energy distribution" for each food image. We then estimate food energy based on the energy distribution.

Other than food portion estimation, we present a systematic design for a crowdsourcing tool aiming specifically for the task of online food image collection and annotations. Our goal is to fast expand food image dataset and to incorporate online food images into our dataset for training-based food classification techniques. In addition, we have developed a printer indexing system for color calibration with an application in image-based dietary assessment.

The main contributions of this thesis are listed as followed:

- Single-View Food Portion Estimation Based on Geometric Models

  We have developed a food portion estimation technique based on a single-view food image used for the estimation of the amount of energy (in kilocalories) consumed in a meal. Although single-view 3D scene reconstruction is in general an ill-posed problem, the use of geometric models such as the shape of a container can help to partially recover 3D parameters of food items in the scene. We are interested in 3D parameters that are essential determining food portions. Based on the estimated 3D parameters of each food item and a reference object in the scene, the volume of each food item in the image can be determined. We focused primarily on the use of cylinder model and prism model. The food portions are estimated in volumes (cm^3). Unlike previous methods, our technique is capable of estimating food portion without manual tuning of parameters. The weight of each food can then be estimated using the density of the food item. We were able to achieve an error of less than 6% for energy estimation of an image of a meal assuming accurate segmentation and food classification.

- A Comparison of Food Portion Estimation Using Geometric Models and Depth Images

  We compare two food portion estimation techniques. The two techniques are namely the geometric models based technique, and depth images based technique. An expectation-maximization based technique has been developed to detect the reference plane in depth images, which is essential for portion size estimation using depth images. We compare the accuracy of food portion estimation based on geometric models, to the accuracy based on high quality depth image. The depth image is obtained using structured light techniques. Our experimental results indicate that volume estimation based on geometric models is more accurate for objects with well-defined 3D shapes compared to estimation using depth images.

- The Use of Co-occurrence Patterns in Single Image Based Food Portion Estimation

  We use contextual information to further improve food portion estimation accuracy of geometric models based approach. We define contextual dietary information as the data that is not directly produced by the visual appearance of an object in the image, but provides information about a user's diet or can be used for diet planning. Food portion co-occurrence pattern is one type of contextual information. We estimate the patterns from food images we collected for dietary studies. We estimate the food portion co-occurrence patterns from food images we collected from dietary studies using the mobile Food Record$^{\text{TM}}$ (mFR$^{\text{TM}}$) system we developed. Co-occurrence patterns is used as prior knowledge to refine portion estimation results. We were able to improve the food portion estimation accuracy incorporating the co-occurrence patterns as contextual information.

- Learning Image-to-Energy Mappings Using Generative Adversarial Networks

  Accurate food portion estimation is challenging since the process of food preparation and consumption impose large variations on food shapes and appearances. In addition to our previous approach of geometric models based food portion estimation, we present a food portion estimation method to estimate food energy (kilocalories) from food images using Generative Adversarial Networks (GAN). We introduce the concept of an "energy distribution" for each food image. To train the GAN, we design a food image dataset based on ground truth food labels and segmentation masks for each food image as well as energy information associated with the food image. Our goal is to learn the mapping from the food image to the food energy. We can then estimate food energy based on the estimated energy distribution image.

- An End-to-end Image-Based Automatic Food Energy Estimation Technique Based on Learned Energy Distribution Images

  We proposed a novel end-to-end system to directly estimate food energy from a captured eating occasion image. Our system first estimated the image to energy mappings using a Generative Adversarial Networks (GAN) structure. Based on the estimated energy distribution images, we learned the food energy of the eating occasion image by training CNN based regression model. We are able to obtain accurate food energy estimation with an average error of 209.41 kilocalories for eating occasion images collected from a free-living dietary study. The training based technique for end-to-end food energy estimation no longer requires fitting geometric models onto the food objects that may have issues scaling up as we need a large amounts of geometric models to fit different food types in many food images.

- cTADA$^{\text{TM}}$: The Design of a Crowdsourcing Tool for Online Food Image Identification and Segmentation

  Training-based techniques have been widely used in recent years for developing automatic dietary assessment systems. For training-based techniques, increasing the training data size would in general improve the accuracy of the system, thus a larger image dataset is always preferred. Online image sharing is quickly gaining popularity in recent years (for example, through social networks such as Facebook and review orientated websites such as Yelp), and there are thousands of food images uploaded by smartphone users everyday. We believe online food images can be used as part of our training data developing automatic dietary assessment techniques and provide valuable contextual information such as users' dietary patterns and food co-occurrence patterns. We present a systematic design with a detailed description for a crowdsourcing tool aiming specifically for the task of online food image collection and annotations. This tool can be used to locate food items and obtaining groundtruth segmentation masks associated with all food objects presented in an image. The crowdsoucing tool we designed is tailored to meet the needs of building a large image dataset for developing automatic dietary assessment tools in the nutrition and health fields.

- A Printer Indexing System for Color Calibration

  In image based dietary assessment, color is a very important feature in food classification. One issue with using color in image analysis is the calibration of the color imaging system. We have implemented a color calibration system for food images using printed color checkerboards also known as fiducial markers (FMs). To use the FM for color calibration one must know which printer was used to print the FM so that the correct color calibration matrix can be used for calibration. We have designed an indexing scheme that allows one to determine which printer was used to print the FM based on a unique arrangement of color squares and binarized marks (used for error control) on the FM. Using normalized cross correlation and pattern detection, the index corresponding to the printer for a particular FM can be determined. We show the printer indexing scheme we developed is robust against most types of lighting conditions.

1.4 Publications Resulting From This Work

Journal Papers

1. **S. Fang**, Z. Shao, D. Kerr, C. Boushey, and F. Zhu, "An End-to-end Image-Based Automatic Food Energy Estimation Technique Based on Learned Energy Distribution Images", *To be submitted to Nutrients Special Issue on Advancement in Dietary Assessment and Self-Monitoring Using Technology.*

Conference Papers

1. **S. Fang**, Z. Shao, R. Mao, C. Fu, D. Kerr, C. Boushey, E. Delp and F. Zhu, "Single-View Food Portion Estimation: Learning Image-to-Energy Mappings Using Generative Adversarial Networks", *Proceedings of the IEEE International Conference on Image Processing*, Athens, Greece, to appear.

2. **S. Fang**, S. Yarlagadda, Y. Wang, F. Zhu, C. Boushey, D. Kerr and E. Delp, "Image Based Dietary Behavior and Analysis Using Deep Learning", *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society*, July 2018, Honolulu, HI

3. **S. Fang**, C. Liu, K. Tahboub, F. Zhu, C. Boushey and E. Delp, "cTADA: The Design of a Crowdsourcing Tool for Online Food Image Identification and Segmentation", *Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation*, April 2018, Las Vegas, NV.

4. **S. Fang**, F. Zhu, C. Boushey and E. Delp, "The Use of Co-occurrence Patterns in Single Image Based Food Portion Estimation", *Proceedings of the IEEE Global Conference on Signal and Information Processing*, pp. 462-466, November 2017, Montreal, Canada.

5. Y. Wang, **S. Fang**, C. Liu, F. Zhu, D. Kerr, C. Boushey and E. Delp, "Food Image Analysis: The Big Data Problem You Can Eat!", *Proceedings of the IEEE International Conference on Image Processing*, pp. 1263-1267, November 2016, Pacific Grove, CA.

6. **S. Fang**, F. Zhu, C. Jiang, S. Zhang, C. Boushey and E. Delp, "A Comparison of Food Portion Estimation Using Geometric Models and Depth Images", *Proceedings of the IEEE International Conference on Image Processing*, pp. 26-30, September 2016, Phoenix, AZ.

7. **S. Fang**, C. Liu, F. Zhu, C. Boushey and E. Delp, "Single-View Food Portion Estimation Based on Geometric Models", *Proceedings of the IEEE International Symposium on Multimedia*, pp. 385-390, December 2015, Miami, FL.

8. **S. Fang**, C. Liu, F. Zhu, C. Boushey and E. Delp, "A Printer Indexing System for Color Calibration with Applications in Dietary Assessment", *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops, Lecture Notes in Computer Science*, , Vol. 9281, Springer International, pp. 358-365, 2015.

2. SINGLE-VIEW FOOD PORTION ESTIMATION BASED ON GEOMETRIC MODELS

2.1 Overview of Image-Based Food Portion Estimation

Dietary assessment, the process of determining what someone eats during the course of a day, provides valuable insights for mounting intervention programs for prevention of many chronic diseases. Traditional dietary assessment techniques, such as dietary record, requires individuals to keep detailed written reports for 3-7 days of all foods or drink consumed [1] and is a time consuming and tedious process. Smartphones provide a unique mechanism for collecting dietary information and monitoring personal health. Several mobile dietary assessment systems, that use food images acquired during eating occasions, have been developed such as the TADA$^{\text{TM}}$ system [6, 7], FoodLog [8], FoodCam [9], DietCam [10], and Im2Calories [11] to automatically determine the food types and energy consumed using image analysis and computer vision techniques. Estimating food portion size/energy (kilocalories) is a challenging task since the process of food preparation and consumption impose large variations on food shapes and appearances. To date, several image-based food portion estimation techniques have been developed based on fundamentally different computer vision techniques.

2.1.1 Image-Based 3D Reconstruction

A camera provides the mapping from the 3D world coordinates onto the 2D pixel coordinates, as shown in Figure 2.1 where C is the camera center and p the principal point [20]. Understanding the 3D structure from images is a fundamental problem in computer vision [21].

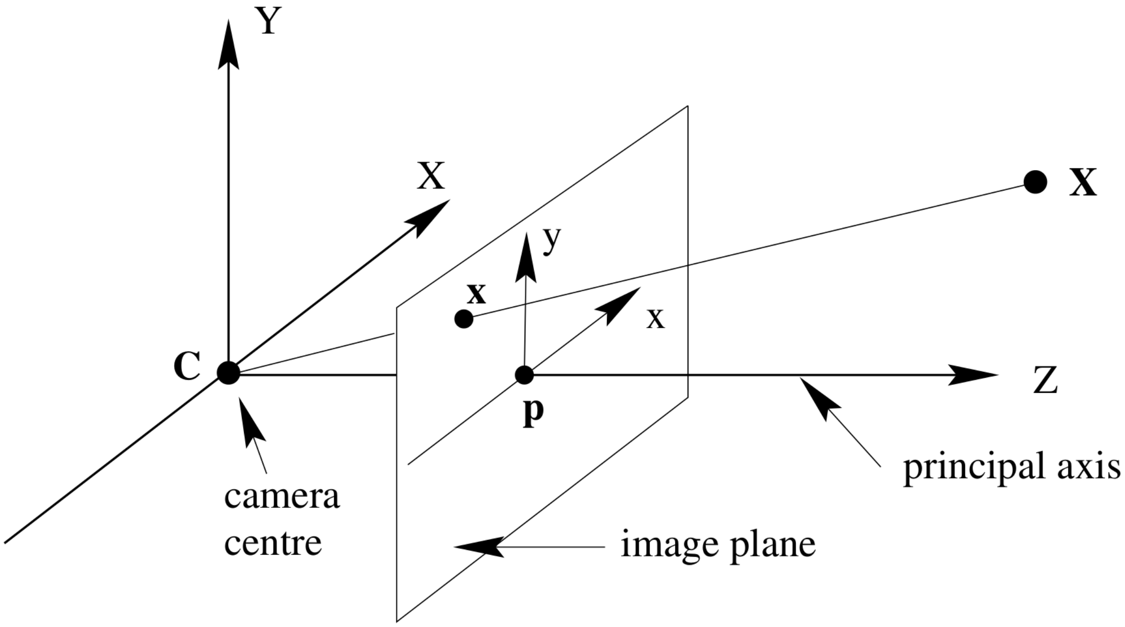


Fig. 2.1.: Pinhole camera geometry [20].

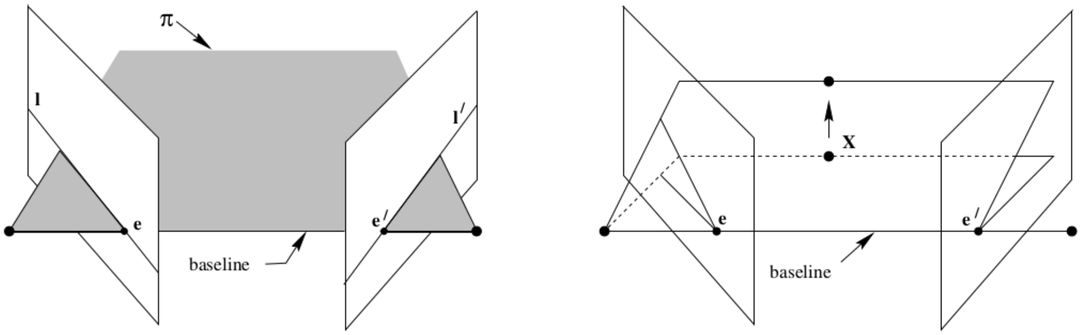


Fig. 2.2.: The epipolar geometry [20].

For 3D reconstruction, most existing works have been developed on multiple views of the scene using stereo vision techniques [20, 22]. Stereo correspondence has traditionally been and continues to be, one of the most heavily investigated topics in computer vision [23–25]. Based on the corresponded points from different views of the same scene, 3D structures can be reconstructed [20]. The epipolar geometry as shown in Figure 2.2 provides the intrinsic projective geometry between two views [20]. The epipoles e and e' are the intersection of camera baseline with each 2D image plane [20]. As stated in [20], any plane π containing the baseline is an epipolar plane, and intersects the image planes in corresponding epipolar lines l and l'. Therefore, as the position of the point X varies in 3D world coordinates, the epipolar planes rotate about the baseline [20]. With points correspondences, the scene can be reconstructed using epipolar geometry [20, 22, 26] as shown in Figure 2.2. Similarly, the recovery of scene structure can also be done using multiple images such as with sequence of motion images [27].

Epipolar geometry based techniques require at least two images of the same scene. For example, in dietary assessment applications, stereo vision techniques require users to take at least two images of the same eating scene from different angles. Therefore, stereo vision technique increases a user's burden [28] and is not desired for dietary assessment application.

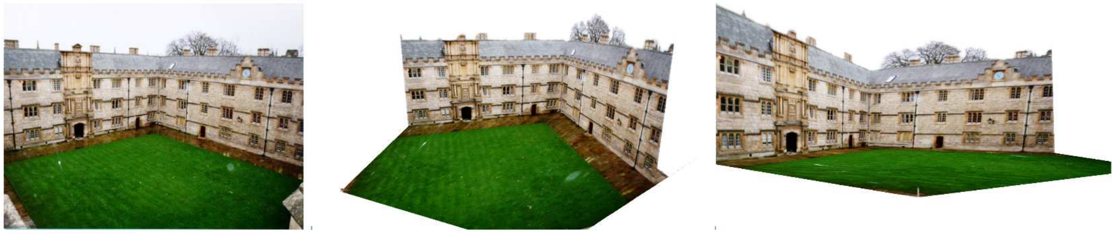


Fig. 2.3.: Example of automatically generated 3D model. (left: original image taken from similar work that requires manual specifying [31], middle, right: two novel views from the reconstructed 3D model).

Other commonly used multi-view 3D reconstruction techniques include shape carving [29] and structured light [30]. Shape carving requires multiple images to be captured that cover 360 degrees of the same object. Therefore, shape carving technique increases a user's burden even more comparing to epipolar geometry based approach. Structured light based technique requires a light projector to be built in a user's mobile device, that projects pre-defined light patterns onto the scene. To date, most mobile devices are not equipped with a structured light projector and sensor.

We focus on the use of a single-view RGB image for 3D reconstruction. The 3D scene reconstruction based on single-view images is in general an ill-posed problem since most 3D information has been lost during the projection process from 3D world coordinates to the 2D pixel coordinates. In [31], 3D scenes are reconstructed based on pre-defined features such as lines, planes, parallelism and orthogonality. However, the technique in [31] is based on visual cues and requires manual selections of vanishing points and lines, planar surfaces and occluding boundaries. Therefore, scenes where 3D structures are difficult to define or the 3D structure visual cues are not present (for example, no obvious lines, planes, parallelism and orthogonality present) are difficult to reconstruct. In addition, as manual selection of features is required, the technique in [31] is not automatic 3D reconstruction.

3D reconstruction based on hand-crafted features often assume constraints, characteristics, features and structures. A hand-crafted technique may work particularly well on one dataset, but failed another. To understand the scene from single monocular image, a multiple-hypothesis framework is developed for robustly estimating scene structure from a single image and obtaining confidences for each geometric label [32]. The geometric labels are the orientations of objects present in the scene. The technique presented in [32] is a learning-based technique learning the statistical distribution of 3D structure, where image was first converted to superpixels. Superpixels are then grouped into constellations (constellations are superpixels that likely share a common geometric label). Each superpixels label is then inferred from the likelihoods of the constellations that contain that superpixel [32]. However, the geometric labels in [32,33] are still manually defined, thereofore scaling with scenes that have ambiguous 3D structures is an issue. In [34] a dynamic Bayesian network model has been used for autonomous 3D reconstruction to infer 3D information of indoor scenes. The dynamic Bayesian network model was used to approximate a distribution over the possible structures of the scene.

The depth of the scenes does not represent the entire 3D structures of the scene (due to occulusion). However depth provides essential information and details for many tasks such as autonomous drive, 3D object modeling and augmented reality. Depth estimation from a single-view image is a difficult task and requires the use of prior knowledge of the scene. Depth prediction task is essentially learning the mapping of a RGB image to a depth image. In the work Make3D [21,35–37], monocular cues have been exploited to obtain 3D information such as predicting depths from single-view image. Markov Random Field (MRF) is used to infer a set of “plane parameters” that captures both the 3D location and 3D orientation of small homogeneous patch in the image [21]. The MRF is trained via supervised learning on ground truth depth dataset that is collected using a laser 3D scanner. Unlike previous techniques for 3D reconstruction based on a single-view RGB image, the model in [21,35–37] makes no explicit assumptions about the structure of the scene. Similar works in depth estimation from single monocular image include estimating depth using gradient-domain learning framework of visual-depth words [38] and by parameter transfer [39].

Convolutional Neural Networks (CNN) has achieved impressive results in many tasks such as object detection and image segmentation. CNN can assemble building blocks of very complex features, where the complex features could be very difficult for human to craft or design. Therefore, instead of defining statistical models to approximate 3D structures of the scenes, deep convolutional features can be learned to predict the depth of the scene from the RGB image [40, 41]. In [42], a single multiscale (coarse to fine) convolutional network architecture has been used for depth prediction based on a monocular image. In [43] the joint learning of depth and other image features (surface normal) has been used and shows improvement over the original architecture [42].

Image-based 3D reconstruction remains to be a challenging problems. Different 3D reconstruction techniques have been developed based on fundamentally different approaches. Similarly, as food portion estimation requires to understand the 3D information of eating scenes, it is a challenging problem. To date several food portion estimation techniques have been developed that based on fundamentally different approaches.

2.1.2 Multi-View Food Portion Estimation

Several image-based food portion estimation techniques are developed based on multi-view 3D reconstruction. Those techniques either require users to take multiple images/videos or modify the mobile device such as using multiple images [10, 44, 45], video [46] or 3D range finding [47].

In [47] a structured light based technique has been used to capture the 3D representation of a food item. However, a regular mobile phone is not equipped with the light projector that projects the pre-defined light patterns onto the scenes. Modifying a user's mobile phone is not feasible for a free living condition with many participants.

In [10,48–50] a technique based on multi-view images of the same eating scene has been used for food analysis. Although the 3D scene can be reconstructed based on multi-view images, the food portion/volume estimation result has not been reported. The features of foods from different views have only been used for food classification in [49, 50] instead of food portion estimation.

In [44, 45] stereo vision reconstruction based on points correspondences from different views have been used for food portion estimation. Multiple images are captured of the same eating occasion. A point clouds representing the eating scenes have been reconstructed based on the feature points correspondences from multiple views [44,45].

Although a dense point cloud can represent the 3D structure of the easting scene, accurate dense feature points correspondences are often difficult to obtain. To ensure the feature point correspondences are accurate for point cloud reconstruction, the eating scenes must remain same for different views. In addition, when using stereo vision techniques, it is often required that the view angles do not change significantly between different views. Therefore, the hidden requirements in geometric computer vision could significantly add to a regular user's burden to successfully capture the eating scene for point clouds reconstruction.

2.1.3 Single-View Food Portion Estimation

Estimating food portion size from a single-view RGB image is an ill-posed inverse problem. Most of the 3D information has been lost during the projection process from 3D world coordinates onto 2D camera sensor plane. Various approaches have been developed to estimate food portion size and energy information from a single-view food image based on fundamentally different approach.

In [51], a 3D model is manually fitted to a 2D food image to estimate the portion size. This approach is not feasible for automatic food portion analysis. The 3D models in [51] are pre-defined by researchers. Based on the eating scene image captures, researchers need to manually find the angle and volume size to best fit the 3D food model onto the food image. This approach is not feasible when there are many food images to be processed.

In [52], food image areas are used for portion size estimation based on user's thumbnail as a size reference. Therefore, the size of a user's thumbnail must be known in order to obtain accurate estimates of food portions. Different sizes of thumbnails cause errors when processing food images captured by different users.

In [53], the pixels in each corresponding food segment are counted to determine the portion sizes. The same food item captured from different angles or different distances have different counts of pixels of the food item. Therefore, the 3D structure of the food item has not been fully exploited in [53].

In [54], food image is divided into sub-regions and food portion estimation is done via pre-determined serving size classification.

Another approach for portion estimation is to utilize the depth information, where depth value is determined with respect to the camera sensor plane. The depth image is first converted to a voxel representation, then the volume for each object can be obtained by summing the voxels that belong to the same object [55]. The reference plane is critical for estimating volume using voxel representation since the height of each voxel cannot be determined without a reference plane. RANSAC [56] is used for reference plane detection in [55]. The supervised learning based depth prediction techniques require sufficient training data. In [55] the Convolutional Neural Network (CNN) architecture in [43] has been applied to food volume estimation and is trained on NYUv2 RGBD dataset [57] of indoor scenes obtained using Microsoft Kinect, and then fine tuned on a new 3D food dataset, **GFood3d** dataset, collected using Intel RealSense F200 depth sensor shown in Figure 2.4.

Depth prediction based on supervised learning techniques requires sufficient training data. Depth sensor is not available on most of the mobile telephones till this date, other than a few developer kit such as Google Project Tango.

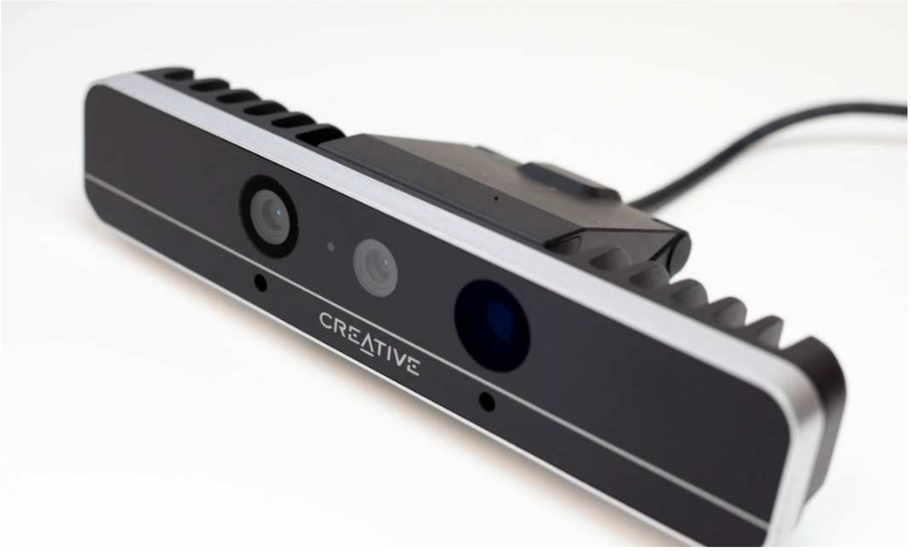


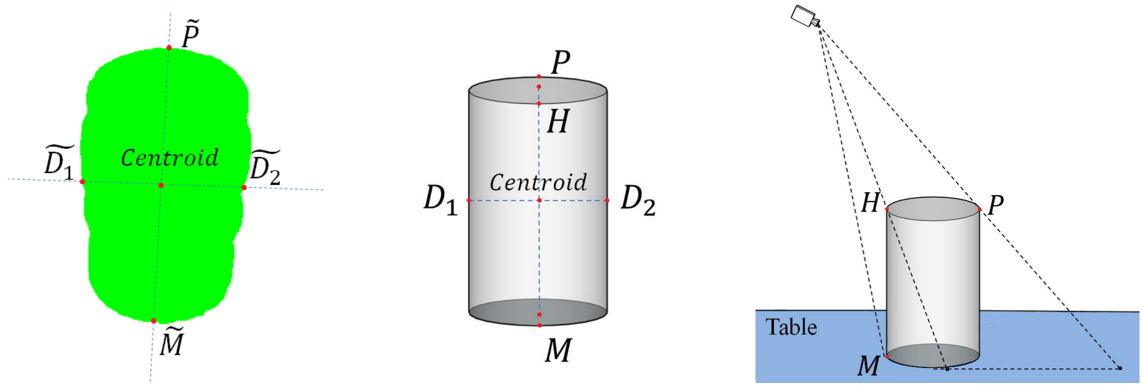
Fig. 2.4.: The Intel RealSense F200 3D Camera.

We have developed a food portion estimation technique based on single-view food image used for the estimation of the amount energy (in kilocalories) consumed at a meal. Our technique is capable of estimating food portion without manual tuning of parameters. Although single-view 3D scene reconstruction is in general an ill-posed problem, the use of geometric models such as the shape of a container can help to partially recover 3D parameters of food items in the scene. Based on the estimated 3D parameters of each food item and a reference object (a fiducial marker, as shown in Figure 1.4) in the scene, the volume of each food can then be estimated using the density of the food item.

The correct food classification label and segmentation mask in the image is alone insufficient for 3D reconstruction of a food item, hence the use of geometric models will allow for volume estimation where we can use the food label to index into a class of geometric models for single view volume estimation. The task then becomes finding the correct parameters for the selected geometric model.

2.2 Food Portion Estimation Using Geometric Models

3D reconstruction from a single image is an ill-posed problem and 3D objects in general can not be fully reconstructed from a single-view. However, since our goal is to estimate the volumes of foods in an image, it is not necessary to fully reconstruct the complete scene. Food volumes are an important intermediate results for estimations of food energy/nutrient. The use of geometric models will allow for volume estimation where we can use the food label to index into a class of geometric models for single view volume estimation. The 3D model for a food type (e.g. a banana) can be reconstructed based on multiple-views using shape from silhouettes [58]. We denote the 3D graphical model that is reconstructed from multiple-views as a pre-built 3D model [59]. In addition to pre-built 3D models we have added pre-defined 3D models for conventional shapes [60]. Using the camera parameters we can project both the pre-built and pre-defined 3D models of each food item back onto the image plane then



(a) Points of interest estimated from the segmentation mask in pixel coordinates.

(b) Front view of selected points of interest in world coordinates.

(c) Side view of selected points of interest in world coordinates.

Fig. 2.5.: Points of interest viewing from world coordinates and partial correspondences estimated in pixel coordinates.

the food volume can be estimated based on a similarity measure of the back-projected region overlaid on the food image segmentation mask. We have also examined the use of prism models (an area-based volume model) that either have non-rigid shapes or do not have significant 3D structures (e.g. scrambled eggs) [60,61]. Our previous portion estimation technique requires manual initialization of the parameters for different food types prior to use [60,62]. Although this approach has yielded reasonable results, the manual initialization can pose issues in scaling with many foods.

We develop a volume estimation technique that uses prior knowledge of the "container shape" as geometric contextual information. For example, the most commonly used containers that have significant 3D structures either can be modeled as cylinders or can be approximated to be cylinders. Knowing that a specific food is likely to be served in a cylindrical shaped container (e.g. milk served in a glass or lettuce in a bowl), using the estimated radius and height of the cylinder, the volume of the food can be obtained. Glasses, cups or even bowls can all be approximated as cylinders. More specifically we focus on estimating the locations of points of interest in 3D world coordinates of the container based on the projection of 3D containers onto a 2D image plane. The points of interest are selected so that they have sufficient information with respect to the radius and height of the food item as shown in Figure 2.5a. We use the prism model which is an area-based volume estimation method for food items that do not have significant 3D structure, such as scrambled egg on a plate with the plate size serving as a reference [62]. After obtaining the volume of each food, we can estimate the weight using the food density (measured in grams/cubic centimeter [18]). The food energy (in kilocalories) can then be obtained from the United States Department of Agriculture (USDA) Food and Nutrient Database for Dietary Studies (FNDDS) [19].

Since foods can have large variation in shapes, there does not exist a single geometric model that would be suitable for all types of foods. The correct food classification label and segmentation mask in the image is alone insufficient for 3D reconstruction of a food item, hence the use of geometric models will allow for volume estimation

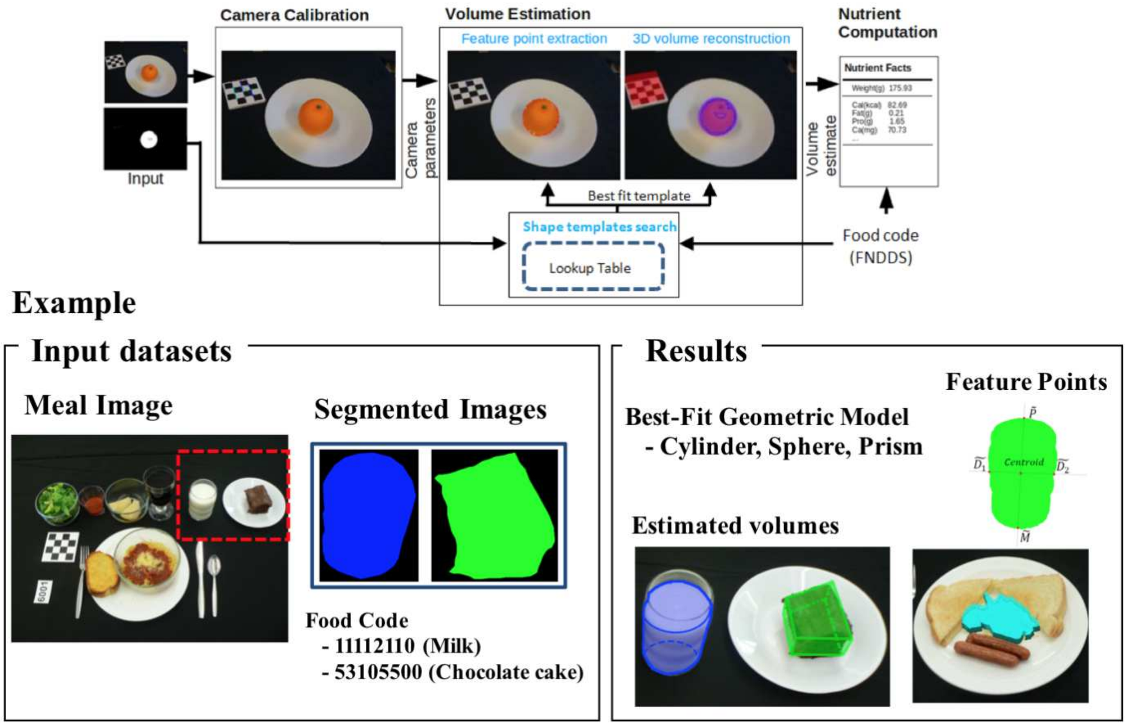


Fig. 2.6.: Portion estimation using geometric models.

where we can use the food label to index into a class of geometric models for single view volume estimation as shown in Figure 2.6.

2.2.1 The Cylinder Model

If we assume the food item is "cylinder-like" such as liquid in a glass or a bowl of lettuce then we know that the cylinder can be defined by its radius and height. We cannot estimate the radius and height of this cylinder solely based on the segmentation mask which is essentially a projection of a cylinder in world coordinates onto the camera sensor. Three coordinates systems are involved in the estimation of parameters for a cylinder model: the 3D world coordinates, the 2D pixel coordinates which is the original 2D image, and the 2D rectified image coordinates. The 2D rectified image coordinates have the projective distortion removed from the original image.

Camera Parameters and Coordinates Systems: Since the camera parameters are essential for both image rectification and 3D to 2D projection, the intrinsic parameters of the camera and the extrinsic parameters for a specific image must be known. This requires that we have some known structure in the scene. To provide essential reference information, we have designed a checkerboard pattern or fiducial marker (FM) in the TADA$^{\text{TM}}$ system. The fiducial marker is printed and is included in the scene by the user to serve as a reference for the estimation of scale and pose of the objects in the scene [17]. The FM is also used to estimate the camera parameters. Based on the detected corners on the checkerboard and their correspondences in world coordinates, the intrinsic and extrinsic parameters can be obtained [60, 63]. The intrinsic parameter **K** for a specific camera is in the following form:

$$\mathbf{K} = \begin{bmatrix} \alpha & \gamma & x_0 \\ 0 & \beta & y_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{2.1}$$

Based on the intrinsic camera parameters obtained for a specific camera the extrinsic camera parameters which include the rotational matrix $\vec{R}$ and displacement vector $\vec{t}$ can then be estimated accordingly for a specific image where we denote:

$$\mathbf{R} = \begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{32} & r_{33} \end{bmatrix} \tag{2.2}$$

and

$$\vec{t} = \begin{bmatrix} t_1 \\ t_2 \\ t_3 \end{bmatrix} \tag{2.3}$$

Using the intrinsic parameter **K**, extrinsic parameters of rotation matrix **R** and displacement vector $\vec{t}$, the 3D to 2D projection process for a given point in 3D world coordinates $X : (x_w, y_w, z_w, 1)^T$ to the corresponding point $\widetilde{X} : (\tilde{x}, \tilde{y}, 1)^T$ in the pixel coordinates in an image can be described as:

$$s \begin{bmatrix} \tilde{x} \\ \tilde{y} \\ 1 \end{bmatrix} = \mathbf{K}[\mathbf{R} \quad \vec{t}] \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \tag{2.4}$$

more specifically:

$$s \begin{bmatrix} \tilde{x} \\ \tilde{y} \\ 1 \end{bmatrix} = \begin{bmatrix} \alpha & \gamma & x_0 \\ 0 & \beta & y_0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_1 \\ r_{21} & r_{22} & r_{23} & t_2 \\ r_{31} & r_{32} & r_{33} & t_3 \end{bmatrix} \begin{bmatrix} x_w \\ y_w \\ z_w \\ 1 \end{bmatrix} \tag{2.5}$$

where (x_0, y_0) is the coordinates of the principal points, α, β are the scale factors of x and y axes and γ describes the skew between two image axes. $(\tilde{x}, \tilde{y}, 1)^T$ is homogeneous, and s is a scale factor. Based on the projection described above, although there is a unique projection in pixel coordinate $\widetilde{X} : (\tilde{x}, \tilde{y}, 1)^T$ for any point in 3D world coordinates $X : (x_w, y_w, z_w, 1)^T$, the converse is false.

A correspondence point that provides the reference location of the same object in the different coordinates must be defined in the segmentation mask. We denote such a reference point as locator M, as illustrated in Figure 2.5(b)(c). In world coordinates we define the locator M to be the closest point to the camera on the bottom surface of the cylinder, which has direct contact with the table. Furthermore, we define $z_w = 0$ for all the points in 3D world coordinates that are contacting the table directly or on the same elevation level. The locator M would be on the $z_w = 0$ surface accordingly. In order to detect the corresponding locator point $\widetilde{M}$ in pixel coordinates, we approximated it to be the lowest point in the column of pixels that is along the centroid of the segmentation mask as illustrated in Figure 2.5a. With

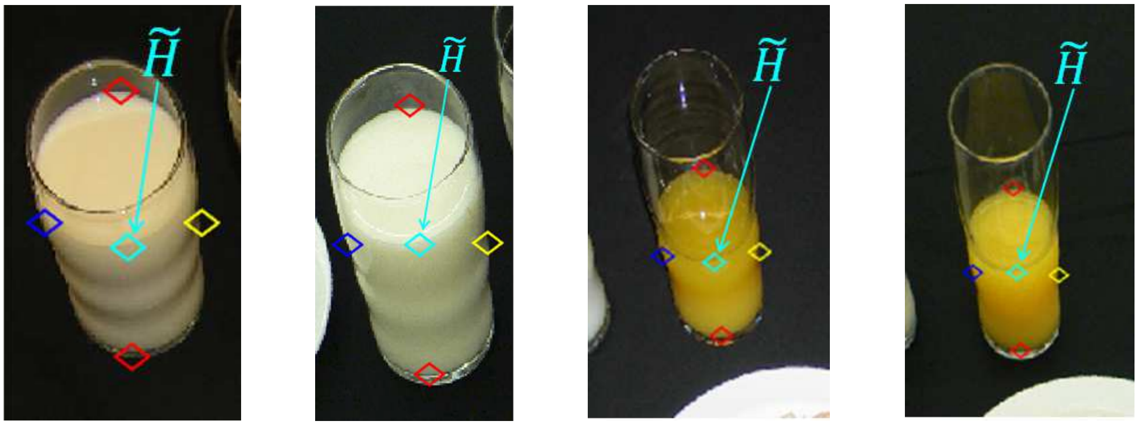


Fig. 2.7.: Examples of the estimated $\widetilde{H}$ (cyan $\diamond$) in rectified image coordinates.

the assumption that $z_w = 0$, the corresponding point M in world coordinates can be determined based on $\widetilde{M}$ using back projection from 2D to 3D as:

$$s \begin{bmatrix} \tilde{x} \\ \tilde{y} \\ 1 \end{bmatrix} = \mathbf{K}[\mathbf{R} \quad \vec{t}] \begin{bmatrix} x_w \\ y_w \\ z_w = 0 \\ 1 \end{bmatrix} = \underbrace{\mathbf{K}[\vec{r_1} \ \vec{r_2} \ \vec{t}]}_{\text{3 by 3 matrix}} \begin{bmatrix} x_w \\ y_w \\ 1 \end{bmatrix} \tag{2.6}$$

where $\vec{r_1}$ and $\vec{r_2}$ are the first and second column vectors of the rotation matrix $\mathbf{R}$. Back projection from 2D to 3D is only valid under the constraint where $z_w = 0$.

Height and Radius Estimation for Cylinder Model:

Knowing the locations of the locator alone is insufficient to estimate the radius and the height of the cylinder. Hence more points of interest must be selected and estimated on the segmentation mask in the pixel coordinates. Based the assumption that the food item is "cylinder-like" model, the points of interest are selected such that the line connecting D_1 and D_2 would represent the diameter and the line connecting H and locator M would represent the height in world coordinates as shown as in Figure 2.5(b)(c). D_1 and D_2 are defined to be on the same elevation level of cylinder's centroid.

Similar to the way we obtained the locator M, we estimate the diameter of the cylinder using the number of row pixels along the centroid of the segmentation mask in the pixel coordinates. The diameter can be determined based on the estimated $\widetilde{D_1}$ and $\widetilde{D_2}$ as shown in Figure 2.5a. However the point of interest $\widetilde{H}$ is lost in the 2D pixel coordinates. Instead of estimating $\widetilde{H}$ directly in the pixel coordinates, $\widetilde{P}$ can be estimated by assigning the highest point (away from M) in the column of pixels along the centroid in the segmentation mask (Figure 2.5a). We can infer the location of $\widetilde{H}$ by subtracting the diameter in the $\widetilde{P} \to \widetilde{M}$ direction from $\widetilde{P}$. The estimation of interest point $\widetilde{P}$ would be performed in the rectified image coordinates as illustrated in Figure 2.7, where the top of the cylinder is a circle with projective distortion removed.

The rectified image coordinates can be obtained by projecting the original image back to 3D world coordinates, under the assumption of $z_w = 0$, using the inverse projection operation of (2.6). All the points of interest estimated directly from the segmentation mask in 2D pixel coordinates can be projected onto rectified image coordinates. With the locations of the points of interest in both pixel coordinates and rectified image coordinates estimated, a points search process can be used in 3D world coordinates based on locator M to estimate the radius and height as shown in Figure 2.9.

The process of searching for points in 3D world coordinates whose projections are in 2D coordinates would correspondingly find the best match of $\widetilde{D}_1$, $\widetilde{D}_2$ and $\widetilde{H}$ in the segmentation mask (Figure 2.8).

Candidates sets are generated for the purpose of radius and height estimation in 3D world coordinates based on (2.4). A set of candidate points $\mathcal{H}$ can be obtained in 3D world coordinates by an incremental search along the vertical direction starting from M where each candidate point $H_h \in \mathcal{H}$ is associated with a specific height increment h. The candidates set $\mathcal{H}$ becomes $\widetilde{H}_h \in \widetilde{\mathcal{H}}$ when projected from 3D to 2D as shown in Figure 2.10(a). The estimated height is obtained by:

$$\hat{h} = \underset{H_h \in \mathcal{H}}{\operatorname{argmin}} ||\widetilde{H}_h - \widetilde{H}|| \tag{2.7}$$

Similarly, two sets of candidate points $\mathcal{D}_1'$ and $\mathcal{D}_2'$ that represent the vertical projections of points D_1 and D_2 onto $Z_w = 0$ plane can be obtained by incremental search

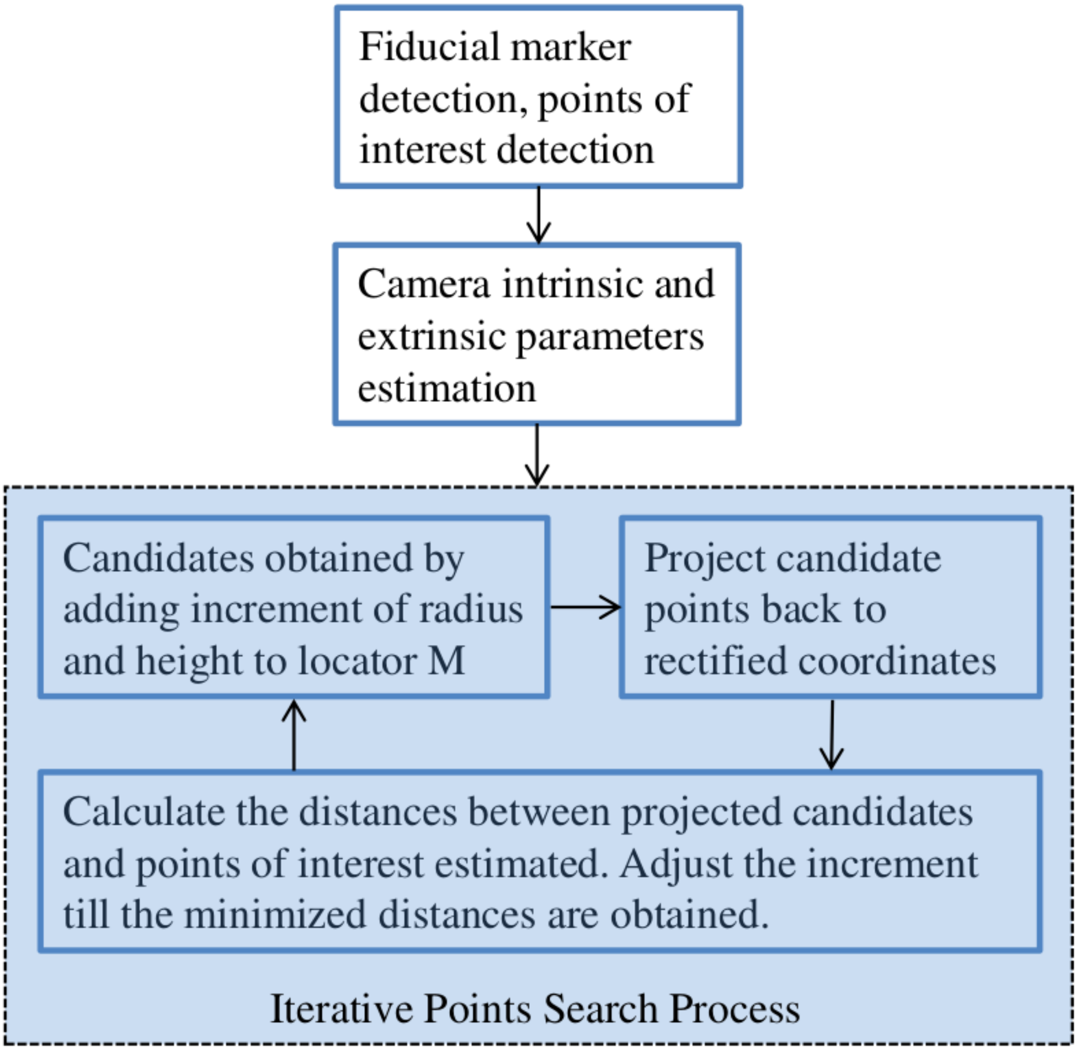

Fiducial marker detection, points of interest detection
Camera intrinsic and extrinsic parameters estimation
Candidates obtained by adding increment of radius and height to locator M
Project candidate points back to rectified coordinates
Calculate the distances between projected candidates and points of interest estimated. Adjust the increment till the minimized distances are obtained.
Iterative Points Search Process

Fig. 2.8.: The point of interest search process for radius and height estimation.

along horizontal direction of M (Figure 2.5), where $D'_{1_r} \in \mathcal{D}'_1$ and $D'_{2_r} \in \mathcal{D}'_2$ are points associated with a specific candidate radius r. The projected candidates sets in 2D pixel coordinates are denoted as $\widetilde{\mathcal{D}}'_1$, $\widetilde{\mathcal{D}}'_2$ and are shown in Figure 2.10(a). Therefore, the radius can be estimated based on the following:

$$\hat{r} = \underset{D'_{1r} \in \mathcal{D}'_1, D'_{2r} \in \mathcal{D}'_2}{\text{argmin}} \{\frac{1}{2}||\widetilde{D}'_{1_r} - \widetilde{D}_1|| + \frac{1}{2}||\widetilde{D}'_{2_r} - \widetilde{D}_2||\} \tag{2.8}$$

The errors in the estimated radius will be reflected in the estimating volume significantly, we propose a refinement method estimated radius. As shown in Figure 2.10(a),

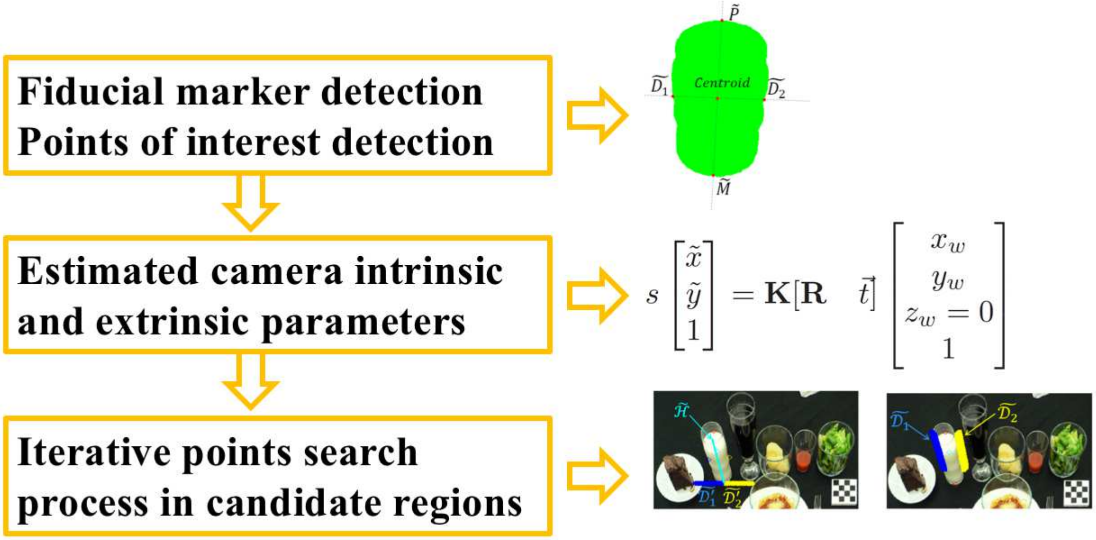
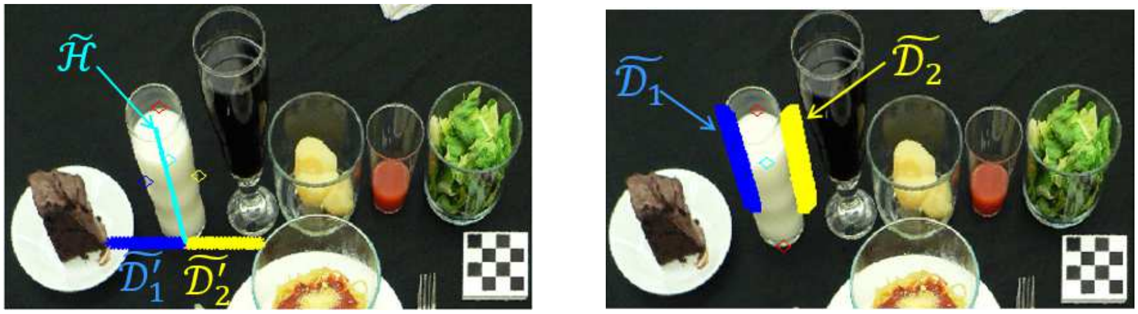

Fiducial marker detection
Points of interest detection
$\tilde{P}$
$\widetilde{D}_1$
Centroid
$\widetilde{D}_2$
$\widetilde{M}$
Estimated camera intrinsic and extrinsic parameters
$s\begin{bmatrix}\tilde{x}\\ \tilde{y}\\ 1\end{bmatrix} = \mathbf{K}[\mathbf{R} \quad \vec{t}]\begin{bmatrix}x_w\\ y_w\\ z_w=0\\ 1\end{bmatrix}$
Iterative points search process in candidate regions
$\widetilde{\mathcal{H}}$
$\widetilde{\mathcal{D}}_1'$
$\widetilde{\mathcal{D}}_2'$
$\widetilde{\mathcal{D}}_1$
$\widetilde{\mathcal{D}}_2$

Fig. 2.9.: The iterative point of interest search for cylinder model.


$\widetilde{\mathcal{H}}$
$\widetilde{\mathcal{D}}_1'$
$\widetilde{\mathcal{D}}_2'$
$\widetilde{\mathcal{D}}_1$
$\widetilde{\mathcal{D}}_2$

(a) Initial search region for radius and height in rectified image coordinates.

(b) Refined search region for radius in rectified image coordinates.

Fig. 2.10.: The projections of candidates sets from 3D world coordinates to rectified 2D coordinates.

the searching regions are the vertical projection of D_1 and D_2 onto $z_w = 0$ plane: $\mathcal{D}_1'$ and $\mathcal{D}_2'$. With the initial estimate of the radius $\hat{r}$ and height $\hat{h}$, we can refine our search region so the candidates sets match D_1 and D_2 in Figure 2.10(b).

2.2.2 The Prism Model

The prism model is an area-based volume estimation method that can be used for food types that do not have significant 3D structures such as scrambled eggs on a plate or toast. For the prism model we assume that the height is the same for the entire horizontal cross-section [62]. In order to accurately estimate the food areas, the original 2D image should be rectified so that the projective distortion can be removed. The fiducial marker can serve as a reference to obtain the 3×3 homography matrix $\mathbf{H}$ used for projective distortion removal. Denote the homogeneous coordinates of a corner $\vec{p}_i$ detected from fiducial marker is homogeneous: $(\tilde{x}, \tilde{y}, 1)^T$ and denote the corresponding world coordinates: $\vec{p}_w$: $(x, y, 1)^T$. The projective transformation matrix can be estimated using the Direct Linear Transform (DLT) method based on the estimated corners and correspondence pattern [20].

$$\begin{pmatrix} 0^T & -\vec{p}_w^T & \tilde{y} \cdot \vec{p}_w^T \\ \vec{p}_w^T & 0^T & -\tilde{x} \cdot \vec{p}_w^T \end{pmatrix} \begin{pmatrix} \vec{h}^1 \\ \vec{h}^2 \\ \vec{h}^3 \end{pmatrix} = 0 \tag{2.9}$$

The segmentation mask can be projected from the pixel coordinates of the original 2D image to the coordinates of rectified image as shown in Figure 2.11.

The area of segmentation mask $\hat{S}$ can be estimated in the rectified image. In order to have a better estimation of area of the food, we utilize the area of the plate. If the plate size is consistent across images, we choose the median of the estimated plate size $\hat{P}$ and use it as a scale reference in addition to fiducial marker. In our experimental data used here, the plate size is consistent and is estimated to be $441cm^2$ in world coordinates. The refined area estimation results demonstrated improvement with the estimated plate size serving as a reference:

$$\text{Refined } \hat{S} = \hat{S} \div \hat{P} \times 441cm^2. \tag{2.10}$$

The median height for each food item can be estimated based on the ground truth volume and median area estimated for the same type of food:

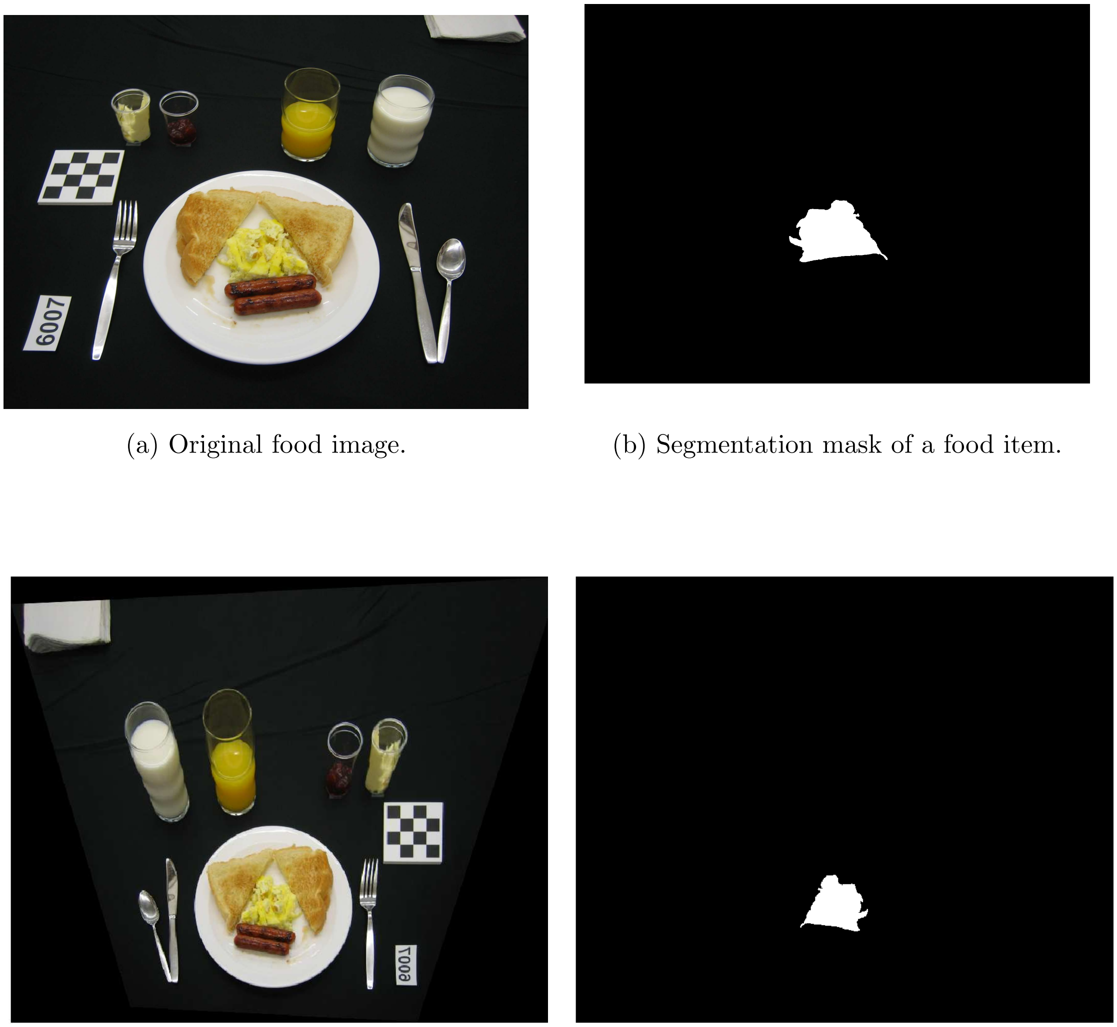
$$\text{Median Height } = \frac{\text{Ground Truth Volume}}{\text{Median Area}}. \tag{2.11}$$

2.3 Experimental Results

We used food images from various user studies we conducted as part of the TADA$^{\text{TM}}$ system to test our portion size estimation methods [28]. For these images we had ground truth information for the food types and portion sizes. We assume we have accurate segmentation and food classification. We used 19 food types in our experiments. We used the cylinder model for 9 types of food and the prism model

Table 2.1.: The estimated food weight (in grams ± standard deviation) using the cylinder and prism models.

Food Name[1]	n[2]	Estimated radius: $\hat{r}$ (mm ± SD)	Estimated height: $\hat{h}$ (mm ± SD)	Estimated weight: $\hat{W}$ (g ± SD)	Ground truth weight: W (g ± SD)	Ratio of estimates $\hat{W}$ to ground truth W[3]
Milk(C)	45	34.1 ± 1.6	66.0 ± 5.0	235.9 ± 26.8	220.0 ± 0.0	1.07
Orange Juice(C)	15	31.1 ± 1.3	40.1 ± 2.5	122.0 ± 10.6	124.0 ± 0.0	0.98
Strawberry Jam(C)	15	17.9 ± 0.8	18.2 ± 11.8	22.1 ± 15.3	21.1 ± 1.1	1.05
Margarine(C)	15	18.8 ± 2.2	29.4 ± 10.4	32.0 ± 13.1	27.8 ± 0.6	1.15
Lettuce(C)	15	51.1 ± 3.5	24.3 ± 13.0	61.1 ± 34.0	48.3 ± 4.8	1.26
Coke(C)	30	39.8 ± 2.5	64.8 ± 9.8	225.9 ± 43.5	227.2 ± 2.3	0.99
Chocolate Cake(C)	15	36.7 ± 4.3	28.3 ± 16.7	77.0 ± 41.1	81.5 ± 12.5	0.95
French Dressing(C)	15	22.6 ± 1.5	12.6 ± 4.7	22.1 ± 7.7	35.7 ± 1.0	0.62
Ketchup(C)	15	17.7 ± 1.1	9.6 ± 2.6	10.9 ± 3.7	15.5 ± 0.4	0.70
Food Name	n	Estimated area: $\hat{S}$ (cm^2 ± SD)	Median height: $\hat{h}$ (mm)	Estimated weight: $\hat{W}$ (g ± SD)	Ground truth weight: W (g ± SD)	Ratio of estimates $\hat{W}$ to ground truth W
Sausage(P)	15	32.5 ± 2.5	17.0	47.8 ± 3.6	41.5 ± 2.8	1.03
Scrambled Egg(P)	15	50.5 ± 4.5	10.8	61.3 ± 5.5	61.5 ± 0.7	1.00
White Toast(P)	15	141.2 ± 16.2	13.0	50.5 ± 5.8	47.7 ± 3.4	1.06
Garlic Bread(P)	15	79.8 ± 12.2	9.3	42.1 ± 6.4	41.1 ± 3.0	1.02
Sugar Cookie(P)	15	44.2 ± 5.4	7.1	26.8 ± 3.3	27.8 ± 1.9	0.97
Spaghetti(P)	15	137.0 ± 10.6	26.0	237.8 ± 18.4	240.3 ± 2.6	0.99
French Fries(P)	15	79.6 ± 6.6	37.8	72.5 ± 6.0	70.5 ± 4.3	1.03
Peaches(P)	15	62.2 ± 14.9	12.3	73.0 ± 17.5	69.3 ± 9.9	1.05
Pear Halves(P)	15	52.8 ± 11.5	13.5	74.5 ± 16.2	75.6 ± 4.9	0.99
Cheeseburger(P)	15	122.6 ± 16.9	26.1	191.7 ± 26.4	198.8 ± 11.5	0.96



(a) Original food image.

(b) Segmentation mask of a food item.

(c) Rectified food image.

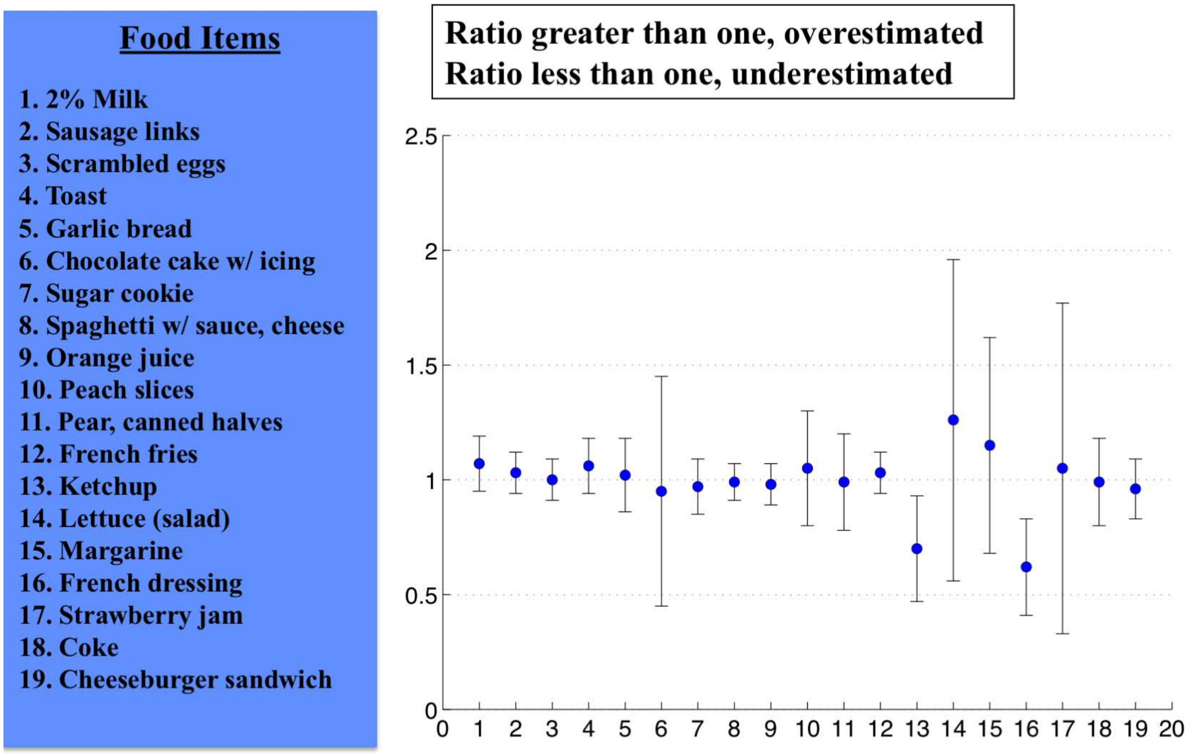
(d) Rectified segmentation mask of food item.

Fig. 2.11.: Removing projective distortion from original image.

for the rest of the 10 types of food. For the cylinder model with estimated radius $\hat{r}$ and height $\hat{h}$, the volumes $\hat{V}$ can be obtained by

$$\hat{V} = \pi \times \hat{r}^2 \times \hat{h}. \tag{2.12}$$

Although a glass containing a soft drink is more of a semi-cone in a single view than a cylinder, we use the radius and height to estimate the volume of the semi-cone. As



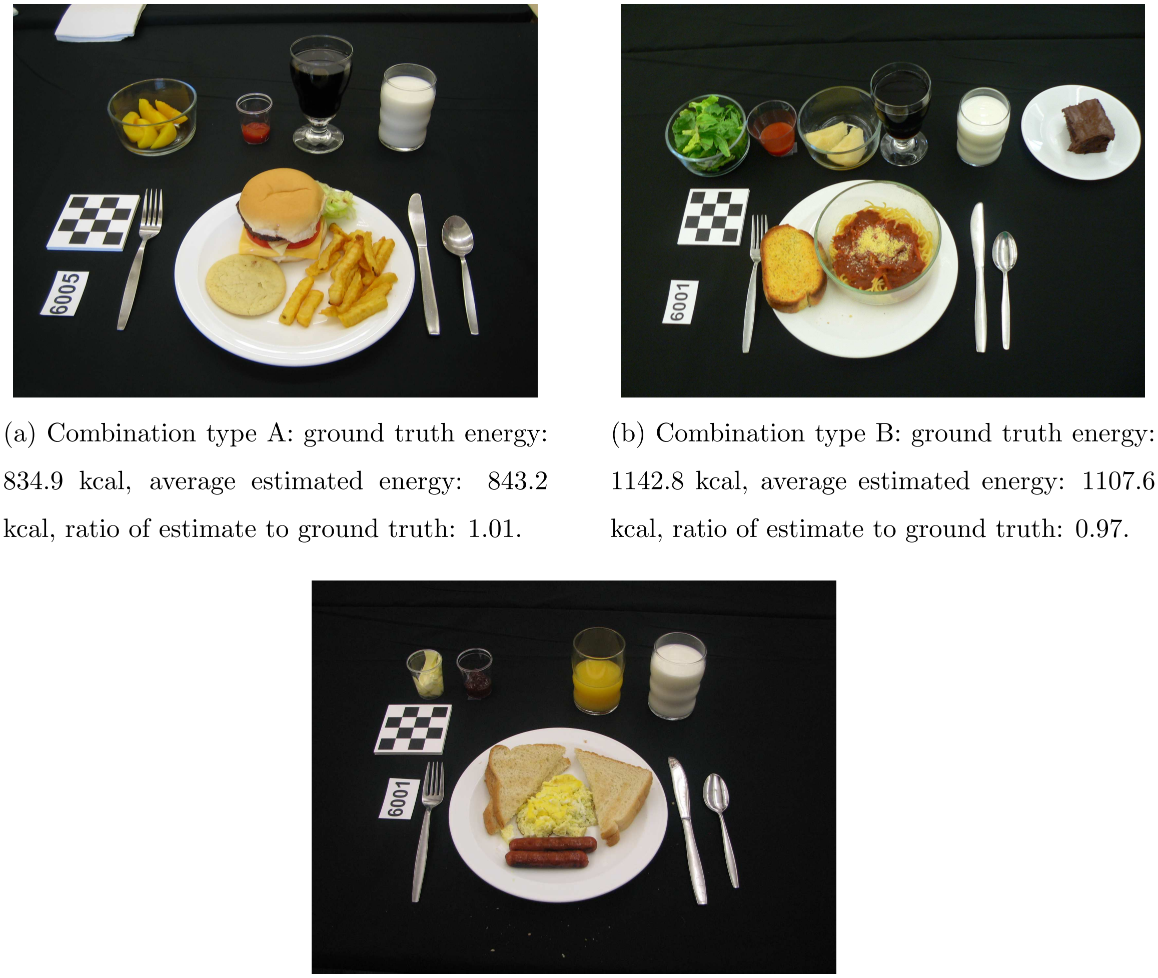
Fang *et al.* ISM 2015

Fig. 2.12.: Ratio of estimated food weights to ground truth.

another example, chocolate cake is not a cylinder, however since it has significant 3D structures we can approximately use the width and height of the cake to estimate the volume. For the prism model, the volume is the estimated area $\hat{S}$ of segmentation mask in the rectified image multiplied by the estimated median height $\hat{h}$ for the same type of food.

With the food density ρ (in grams/cubic centimeter), the food weight can be computed based on the volume as: $\hat{W} = \rho \times \hat{V}$ [18]. For our test data the same type of food has approximately the same ground truth weight [62]. We compare the estimated average weight for each type of food to the ground truth weight as shown in Table 2.1. The ratio of estimate food weight to ground truth food weight is used as an indicator to determine the accuracy of the estimates as shown in Figure 2.12. The ratios are obtained by dividing the mean of the estimated weight $\hat{W}$ to the mean of the ground truth weight W. We have compared our results here to those we previously reported [60,62]. We discussed in [62] that a 15% error or less (i.e. the ratio shown in Table 2.1 being from 0.85 to 1.15) would be considered to be an acceptable range for most foods. Out of the 19 food types, only 3 types of food: lettuce, French dressing and ketchup have estimated errors larger than 15%. Although lettuce has a ratio of 1.26 (26% error), it is an improvement compared to the results of 4.61 we presented in [62] and 1.70 we presented in [60]. Given the low energy density of lettuce, the error represents approximately 2 additional kilocalories. For the ketchup and French dressing, the errors generated are due to the height estimates using the cylinder model. Since one would not consume a large amount of ketchup or French dressing in a typical meal, the large errors would not result in a significant impact on the estimate of energy consumed for the entire meal.

We also estimated the energy for each meal as captured by the food images. There are a total of 45 images corresponding to 45 different individual eating occasions reported by participants. More specifically, for this particular dataset we only have 3 different combinations of food, with each combination having approximately the same energy for different images. Examples of each combination of food items are illustrated in Figure 2.13. For each image the total energy (kilocalories) can be obtained by summing the energy for each food item based on the estimated weight using the FNDDS database [19]. We compare the estimated energy to the ground truth energy (in kilocalories) and then determine the ratio of estimates to the ground truth, for each type of combination as shown by Figure 2.13. We were able to achieve an error of less than 6%. Therefore our method appears to be very promising for estimating the energy for a meal based on using a single image.



(a) Combination type A: ground truth energy: 834.9 kcal, average estimated energy: 843.2 kcal, ratio of estimate to ground truth: 1.01.

(b) Combination type B: ground truth energy: 1142.8 kcal, average estimated energy: 1107.6 kcal, ratio of estimate to ground truth: 0.97.

(c) Combination type C: ground truth energy: 745.9 kcal, average estimated energy: 788.3 kcal, ratio of estimates to ground truth: 1.06.

Fig. 2.13.: Examples of three combinations of food items. Ground truth energy is based on a single serve.

2.4 Conclusion and Future Work

We propose a method to estimate food portion size from a single-view image. Instead of relying on manual initialization of estimation parameters, our method can automatically do volume estimation using the geometric contextual information from the scene. We no longer have issues in scaling with many foods due to manual initialization of parameters. We plan to use more contextual information for volume estimation. We are also interested in developing a more robust scheme for energy estimation so that the impact of segmentation and food classification errors (or food portion estimation) can be minimized.

3. A COMPARISON OF FOOD PORTION ESTIMATION USING GEOMETRIC MODELS AND DEPTH IMAGES

3.1 A Comparison of Food Portion Estimation Using Geometric Models and Depth Images

Several image analysis based techniques have been developed for food portion estimation. 3D features are not fully exploited in some of the existing works. In [54] food portion estimation is done via pre-determined serving size classification. In [52] the food image area and the user's thumb are used as reference for estimation the portion size and in [53] the pixels in each corresponding food segment are counted to determine the portion. To better analyze the food eating scene, other methods attempt to recover 3D parameters of the scene including the use of mobile 3D range finding [47] and stereo vision techniques using multiple images [10, 44, 45].

We feel that either modifying the mobile device or acquiring multiple images of the eating scene is not desirable for users trying to collect information about their diets. Furthermore, a point cloud obtained from a few images using feature based stereo matching is sparse that cannot represent fine details on surfaces that are necessary with food images. We have investigated using stereo vision technique in TADA$^{\text{TM}}$ system [64]. With points correspondences the scene can be reconstructed using epipolar geometry [20, 22, 26] as shown in Figure 2.2. We use Scale-Invariant Feature Transform (SIFT) [65] keypoints for feature matching of different views [64]. To match the feature points we compare the Euclidean distances of feature vectors and find a pair of keypoints from each frame that are nearest measured in Euclidean distance. Epipolar constraint states that the correct match must lie on the epipolar line and we use the epipolar constraint to remove the false matching. We have shown in [64] that by removing the false matching we reduce the number of matched pairs as shown in

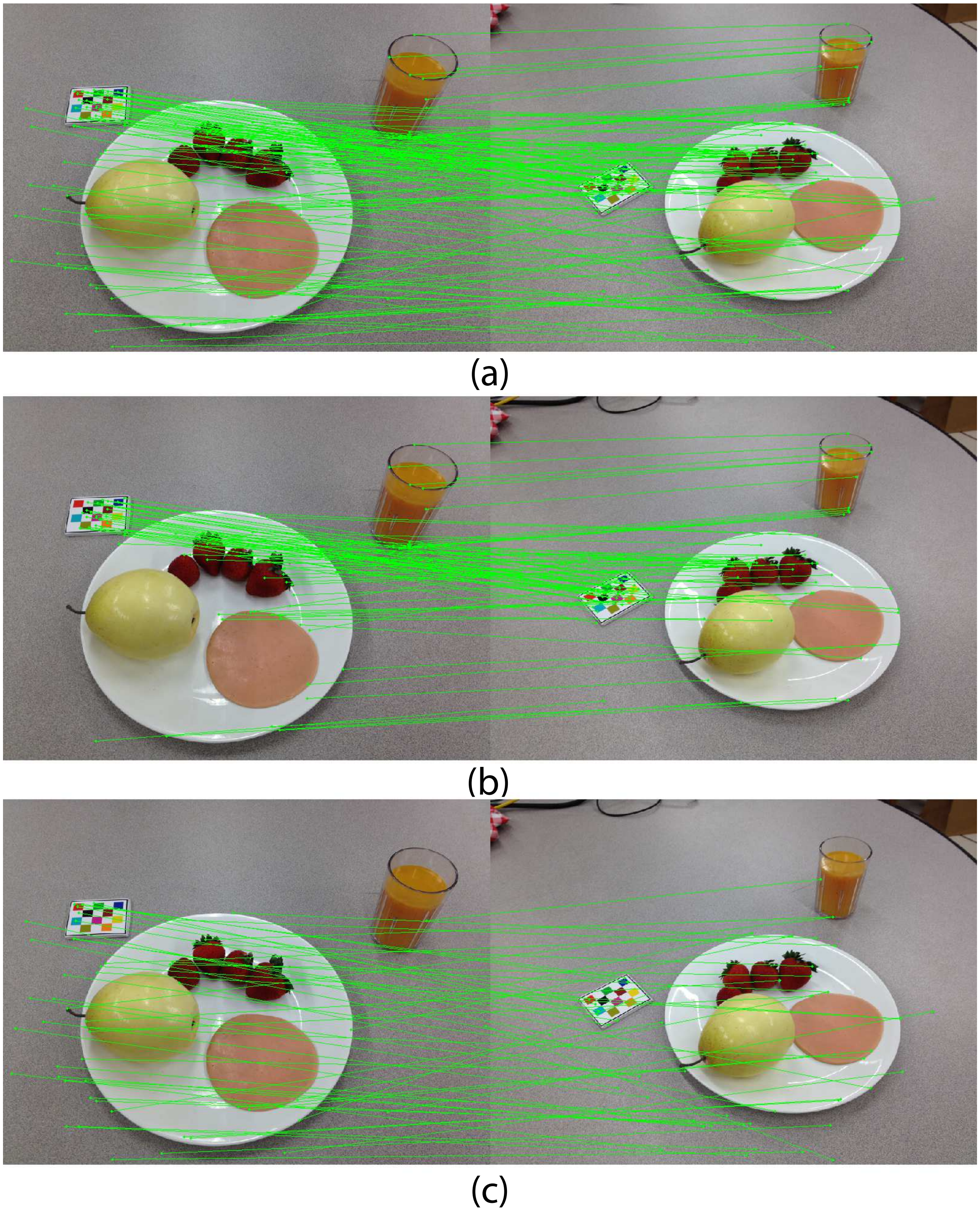


Fig. 3.1.: Feature matching using epipolar constraint. The corresponding points in left and right images are connected by green line. (a) shows all corresponding points, (b) shows matched pairs that satisfy epipolar constraint, (c) shows matched pairs that fail epipolar constraint.

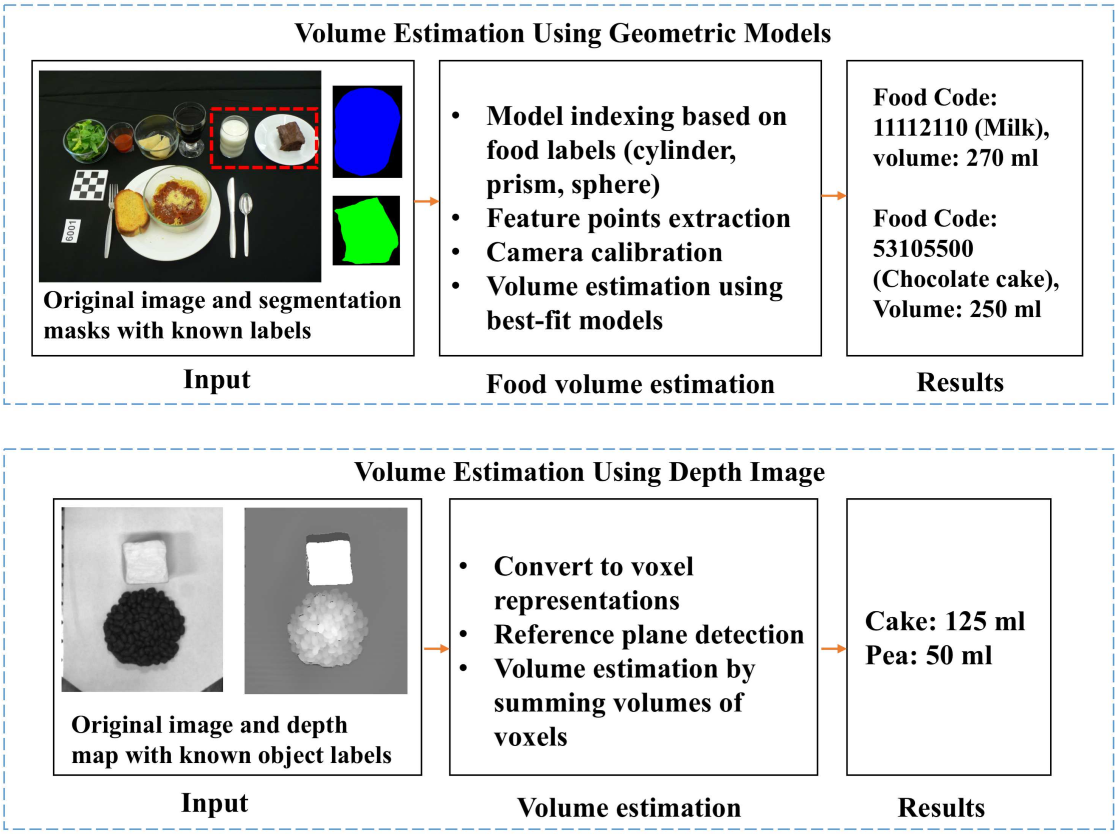

Volume Estimation Using Geometric Models
Original image and segmentation masks with known labels
Input
Model indexing based on food labels (cylinder, prism, sphere)
Feature points extraction
Camera calibration
Volume estimation using best-fit models
Food volume estimation
Food Code: 11112110 (Milk), volume: 270 ml
Food Code: 53105500 (Chocolate cake), Volume: 250 ml
Results
Volume Estimation Using Depth Image
Original image and depth map with known object labels
Input
Convert to voxel representations
Reference plane detection
Volume estimation by summing volumes of voxels
Volume estimation
Cake: 125 ml
Pea: 50 ml
Results

Fig. 3.2.: Food portion size estimation using geometric models and depth images.

Figure 3.1. With limited number of matched pairs it is challenging to reconstruct dense 3D shapes from stereo images. We have focused in our work on techniques that use single images for food portion size estimation [14].

Estimating the volume of an object from a single view is an ill-posed inverse problem that requires the use of a priori information. The existing work using single images include the use of pre-defined 3D template matching [51,60], using prior knowledge of the geometric model [14] and depth map prediction based using a Convolutional Neural Network (CNN) [43,55]. Template matching using pre-defined 3D models involves manual tuning of model parameters which can cause scaling problems [51, 60]. The points search technique in 3D coordinates does not require manual tuning of parameters [14] hence scaling with many foods will not be an issue. Depth map prediction

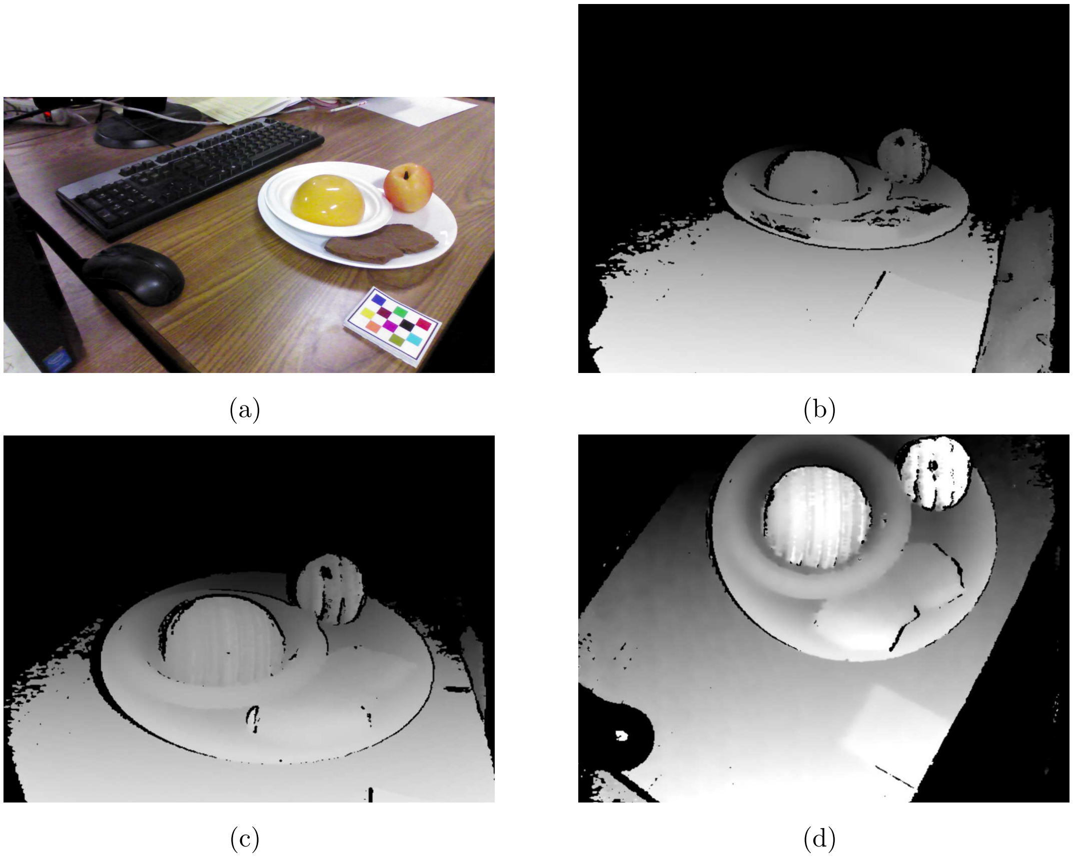


Fig. 3.3.: The RGB image (a) of the scene and depth images (b)(c)(d) of the same scene acquired with different angle, using Intel RealSense F200 3D camera.

using CNN requires sufficient images as training data and depth sensors that provide sufficient details. In [55] the Convolutional Neural Network (CNN) architecture in [43] is initially trained on NYUv2 RGBD dataset [57] of indoor scenes obtained using Microsoft Kinect, and then fine tuned on a new 3D food dataset, **GFood3d** dataset, collected using Intel RealSense F200 depth sensor shown in Figure 2.4.

In this work we examine food portion size estimation accuracy using geometric models and depth images as shown in Figure 3.2. The use of geometric models allow for volume estimation where we can use the food label to index into a class of pre-defined geometric models for single view portion size estimation. We have acquired sample depth images of food scene using Intel RealSense F200 depth sensor and decide the depth detail is not the finest quality shown in Figure 3.3. To acquire high quality depth maps, instead of using Intel RealSense F200 depth sensor, we use a structured light technique known as digital fringe projection [66]. Fine depth details are available in the depth image we collected shown in Figure 3.4b using the structured light system at Purdue University. The digital fringe projection technique is a special type of triangulation-based structured light method where variations in pattern intensity are sinusoidal. We adopt the binary defocusing method and phase-shifting-based fringe analysis technique for 3D shape measurement because of their high speed, high resolution, and high accuracy. In a phase-shifting technique, the sinusoidal fringe patterns are shifted spatially from frame to frame with a known phase shift. Analyzing a set of phase-shifted fringe images yields the wrapped phase, a distortion measurement, usually containing 2π discontinuities that are removed by employing a temporal phase-unwrapping algorithm [67]. The (x, y, z) coordinates are recovered from the unwrapped phase using the system parameters estimated from system calibration [68]. We were able to obtain the depth map based on the (x, y, z) coordinates recovered.

To best represent the 3D scene, we used an over-head view when acquiring the depth images. To estimate the volume for each food object from a depth image a reference plane is required such as the table surface. To detect the table surface, we developed an expectation-maximization (EM) [69, 70] based technique so that intra-class variations (such as different textures shown in Figure 3.4(a)) on the reference plane can be incorporated. To validate the two approaches we compared the estimated volumes of the same objects using the two methods to the ground-truth information.

3.2 Portion Size Estimation Using Depth Images

The depth maps, along with the grayscale images, are captured using a 3D sensor system designed by [67]. The depth map contains the distances of points on object's surfaces to the camera sensor. The depth map is converted to voxel representation.

We denote $\mathbb{V}$ the set of all voxels where for each voxel: $v_p \in \mathbb{V}$. For a known food image segmentation mask from the grayscale image, each voxel v_p can be associated with an object label: l_q for $l_q \in \mathbb{L}$, where $\mathbb{L}$ is the set of all object labels in the depth image. We use the mapping process: $\mathcal{L}(v_p) = l_q$ to determine which label l_q is the voxel v_p associated with. Assume $width_p$ and $length_p$ are the size of the base area of voxel v_p, if the $height_p$ is known, then the voxel volume can be determined. The volume V_{l_q} associated with each label l_q can than be obtained:

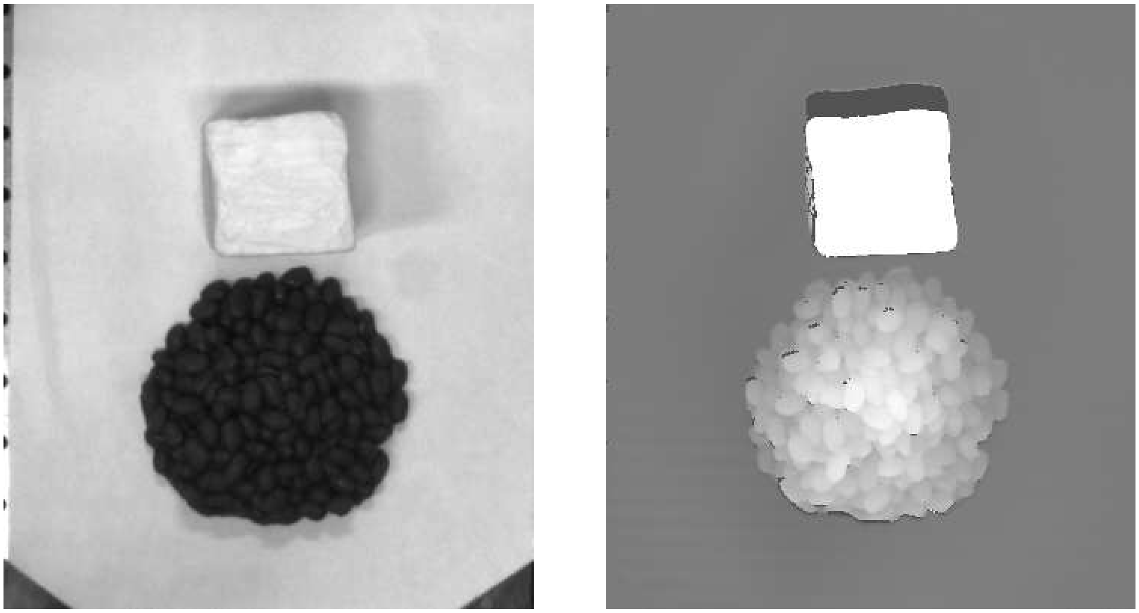
$$V_{l_q} = \sum_{v_p \in \mathbb{V}, \mathcal{L}(v_p) = l_q} width_p \times length_p \times height_p \tag{3.1}$$

The above estimation is based on the assumption we can align the voxel grid on the reference plane (e.g., table surface). If the reference surface cannot be correctly detected, we would not be able to obtain the volume for each voxel (the height for each voxel will be unknown).

3.3 Reference Plane Detection In Depth Images

Detecting the reference plane can be viewed as a clustering task. The goal is to cluster all pixels $\mathbb{P}$ into two subsets: $\mathbb{S}_{surf}$ which is associated with the reference plane, and $\mathbb{S}_{non-surf}$ which contains the rest. A Gaussian distribution is assumed for modeling the distribution of samples in $\mathbb{S}_{surf}$ and $\mathbb{S}_{non-surf}$. Such an assumption is made upon the characteristics of distribution of pixel and depth values. If the parameters for the Gaussian mixtures are known, we can cluster the pixels into different subsets.

Expectation-maximization (EM) can be used to estimate the above GMM parameters [69, 70]. EM has been used for image segmentation using multiple image features [71]. Our image features are the depth map and the grayscale pixels. An example of grayscale image of the scene is shown in Figure 3.4(a). The corresponding depth map associated with the grayscale image is shown in Figure 3.4(b). We have noticed that for a small portion of regions, valid depth values cannot be obtained due to the shadows generated by the structured light pattern projector.



(a) Gray scale image.

(b) Depth map (shadows do not have valid depths.

Fig. 3.4.: Gray scale image and the corresponding depth map.

We combine pixel and depth feature for surface detection. We denote $d_{i,j} \in \mathbb{D}$ the depth at image coordinates (i, j) where $\mathbb{D}$ is the set contains all valid depths. The size of the set $\mathbb{D}$ is N. Denote the $\vec{d} \in \mathbb{R}^{1\times N}$ as the vectorized representation of $\mathbb{D}$. We construct a vector $\vec{p} \in \mathbb{R}^{1\times N}$ consists of pixel values:

$$\forall i, j \quad s.t. \quad d_{i,j} \in \mathbb{D} : \quad p_{i,j} \in \vec{p} \tag{3.2}$$

where $\vec{p} \in \mathbb{R}^{1\times N}$ is of the same size as $\vec{d}$. Both the $\vec{d}$ and $\vec{p}$ are constructed using raster scan order. The set of all observations could then be obtained as:

$$Y = \begin{bmatrix} \vec{d} \\ \vec{p} \end{bmatrix} \tag{3.3}$$

where each observation is denoted as: $y_n \in Y$, $n = \{1, \cdots, N\}$.

As assumed Gaussian mixture the parameter $\theta_k = \{\pi_k, \vec{\mu}_k, \Sigma_k\}$ for each component $k \in \{1, \cdots, K\}$ would then be:

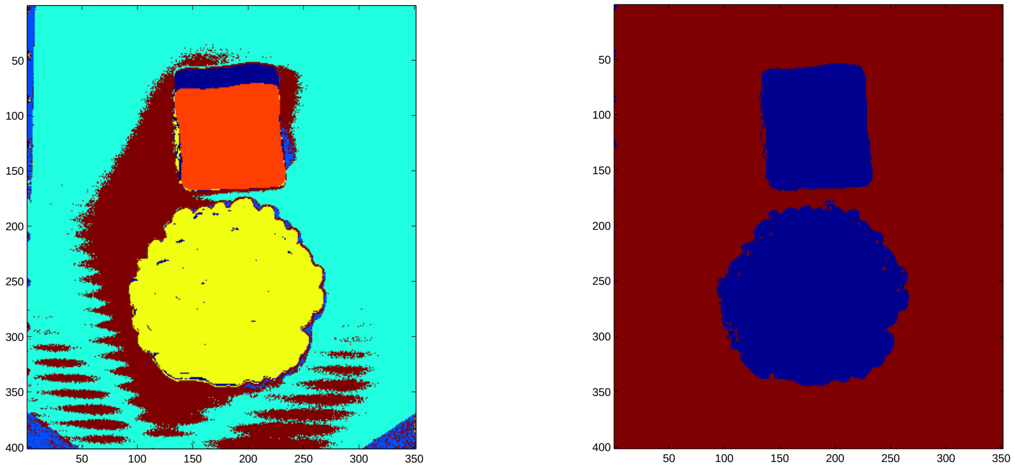
$$\begin{aligned} \pi_k &: \text{ The fraction of } k^{th} \text{ component in } Y \\ \vec{\mu}_k &: \text{ The mean vector of } k^{th} \text{ component} \\ \Sigma_k &: \text{ The covariance matrix of } k^{th} \text{ component} \end{aligned}$$

If parameter θ_k is known then the component's label $x_n \in \{1, \cdots, K\}$ corresponding to the observation y_n can then be determined:

$$\begin{aligned} x_n &= \underset{k\in\{1,\cdots,K\}}{\operatorname{argmax}} f(y_n|x_n = k) \\ &= \underset{k\in\{1,\cdots,K\}}{\operatorname{argmax}} \frac{1}{2\pi|\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(y_n-\vec{\mu}_k)'\Sigma_k^{-1}(y_n-\vec{\mu}_k)} \end{aligned}$$

We assume the observations are sampled from multivariate Gaussian mixtures which consist of K components where the parameters θ_k for each component is unknown, $k \in \{1, \cdots, K\}$. The task would then be to estimate the parameters $\Theta = \{\theta_1, \cdots, \theta_K\}$ based on observations $y_n \in Y$. The missing data would be the label $x_n \in X$ for each observation. Θ can be estimated recursively using the EM [70]. We denote the $\theta^{[m]}$ is the parameters estimated from m^{th} iteration and the Expectation-step is defined as:

$$\text{E-step: } Q(\theta|\theta^{[m]}) = \mathbb{E}[\log f(\mathbf{x}|\theta)|\mathbf{y}, \theta^{[m]}] \tag{3.4}$$



(a) Initially over clustered to incorporate intra-class variations, with $K = 5$.

(b) Clusters merged and reference surface detected (red area).

Fig. 3.5.: Reference plane detection with combined features using EM.

where $\mathbf{x}$ is the missing label associated with observation $\mathbf{y}$, and $\theta^{[m+1]}$ is the updated values of θ which maximizes $Q(\theta|\theta^{[m]})$. The Maximization-step is then [70]:

$$\text{M-step: } \theta^{[m+1]} = \underset{\theta}{\operatorname{argmax}}\, Q(\theta|\theta^{[m]}) \tag{3.5}$$

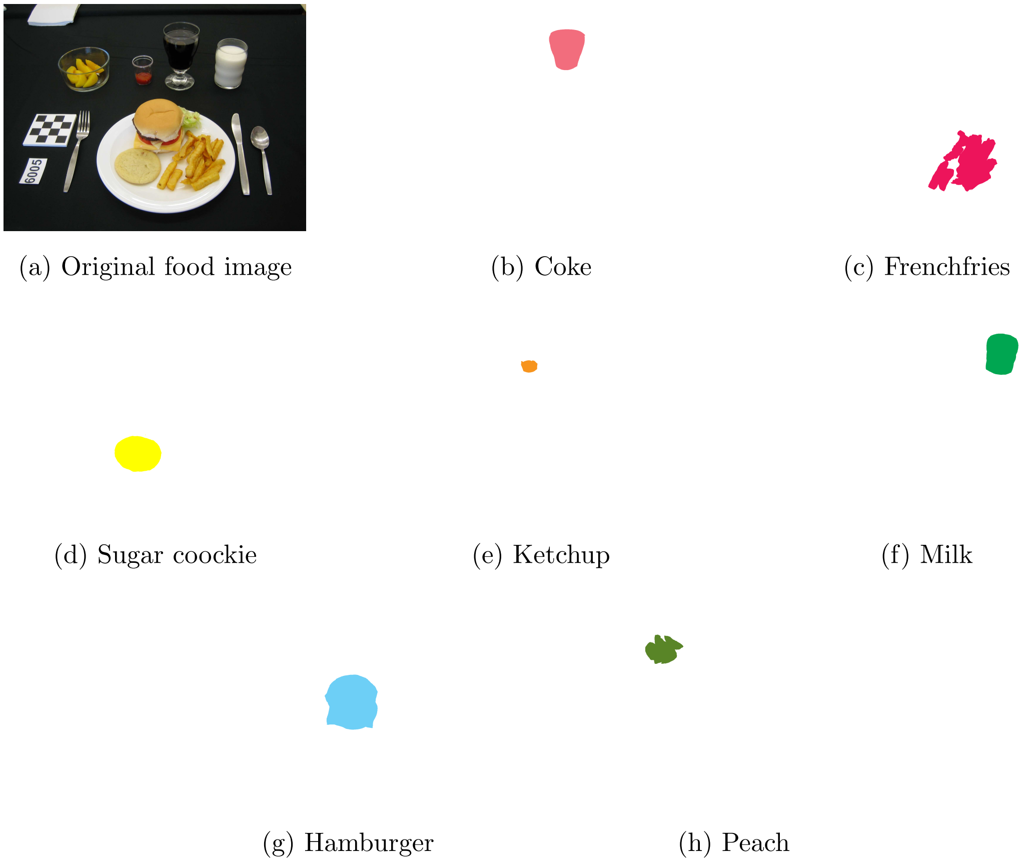
To better incorporate the intra-class variations we use $K > 2$ for table surface detection. We set $K = 5$ to over cluster the pixels initially, as shown in Figure 3.5(a). We then merge the clusters based on the Euclidean distance of θ_k using k-means. The reference plane can be detected based on the mean and variance of depth in a segment, as shown in Figure 3.5(b). Given the reference plane we can now estimate the height for each voxel, and the object's volume.

3.4 Experimental Results

We compare the estimated volumes of the same objects using the two methods to the ground-truth information. The ground truth volume for each object is obtained using water displacement [72]. We use 10 objects for testing, as listed in Table 3.1. Except for the paper cup, these objects are selected from NASCO food replicas made with plastic/rubber. The 9 plastic food replicas are selected to represent different shapes and features.

For the geometric models, we acquire test images using the TADA$^{\text{TM}}$ system [6] on an mobile device (an iPhone 6) as shown in Figure 3.7. Each test images contains $5 - 7$ objects from the objects listed in Table 3.1. We vary the layouts of the objects in the scene and the angles used to capture the images. A total of 36 test images with different combinations are captured. To avoid the errors propagated from automatic segmentation and classification, we use ground-truth segmentation masks and food labels shown in Figure 3.6, then compute the volume $\hat{V}_g$ for each object as shown in Table 3.1 using the appropriate geometric model.

For food portion size estimation using depth maps, we were able to acquire the depth images using a structured light system at Purdue University, more specifically using the digital fringe projection technique. Digital fringe projection (DFP) techniques have been extensively used for high-quality 3D shape measurement due to their speed, accuracy, and flexibility [30,66]. The system we used is able to obtain a root-mean-square error of about 73 μm with a calibration volume of 150 mm (height) $\times$ 250 mm (width) $\times$ 200 mm (depth) [68]. Since a grayscale image associated with depth map is available, we can obtain the pixel-wise alignment of the image with the depth map. For our system, we were able to obtain grayscale images and depth map at the resolution of 640×480 pixels. It is challenging to use structured light to reconstruct the 3D shape of an object with a large range of reflectivity. For our dataset we avoid using NASCO food replicas whose surfaces cannot be properly reconstructed due to issue cause by reflectivity. We converted the depth map into voxel representations based on the detection of the table surface. Similar to the test images acquired using mobile device, we use different layouts and combinations of objects. Each image contains 1–3 objects. We use ground truth label for each segment. The volume estimated using the depth map: $\hat{V}_d$ can than be obtained as shown in Table 3.1.



(a) Original food image

(b) Coke

(c) Frenchfries

(d) Sugar coockie

(e) Ketchup

(f) Milk

(g) Hamburger

(h) Peach

Fig. 3.6.: Original food image (a) and ground truth segmentation masks (b)-(h) and food labels.

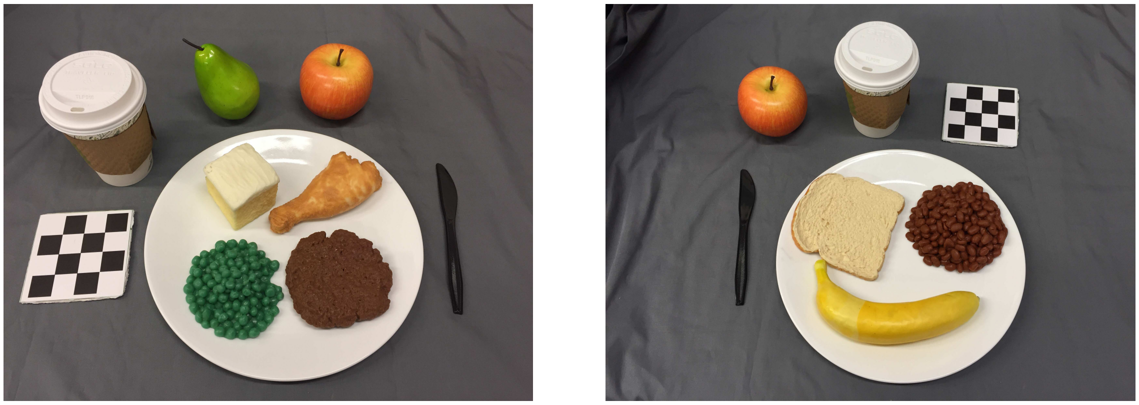


Fig. 3.7.: Sample test images captured using mobile device for geometric models-based portion estimation.

Based on our experimental results, we observed that the volume estimation using a depth map has a tendency to over estimate the food portion size. We observed that 9 out of 10 test objects are over-estimated, the coffee cup has a ratio of estimate to ground truth of 2.34 on average. The single image approach cannot fully represent the 3D shape since parameters on the surfaces that are not visible cannot be recovered. Hence there exists shape ambiguity in 3D coordinates. However, if we use the prior knowledge of the 3D shape, such as the cylinder model, we can obtain significantly better estimation for objects such as "cup". Similarly, for "apple", if we use a sphere model the volume estimates are better than those estimated using depth maps. We used the prism model for the other objects. As shown in Figure 3.8, we were able to obtain more accurate estimates using geometric models with well-defined 3D shapes compared to estimation using depth images. The number 1 – 10 on the horizontal axis in Figure 3.8 represents foods listed in Table 3.1.

3.5 Conclusions

We conduct a comparison of food portion estimation using two techniques: geometric models and depth images. We obtain more accurate volume estimation using geometric models for objects whose 3D shape are well-defined. We have noticed a tendency of over estimation using depth map.

Table 3.1.: Comparison of volume estimation using geometric models and depth maps.

Food Tag	V[a](ml)	N_d[b]	$\hat{V}_d$(ml): $\mu \pm$ SD	N_g[c]	$\hat{V}_g$(ml): $\mu \pm$ SD
1 – Apple	275	11	332±29	36	252±57
2 – Banana	200	11	183±29	19	197±26
3 – Cake	125	11	160±28	17	127±20
4 – Bean	100	11	124±16	19	98±10
5 – Pea	50	10	64±29	17	50±6
6 – Sausage	75	10	108±14	17	77±8
7 – Pear	180	11	241±20	17	186±51
8 – Chicken	80	9	95±13	17	83±13
9 – Bread	100	10	111±40	19	101±11
10 – Cup	450	23	1056±80	36	553±115

[a] Water displacement

[b] N_d is the number of images used for depth-based estimation

[c] N_g is the number of images used for geometric model-based estimation

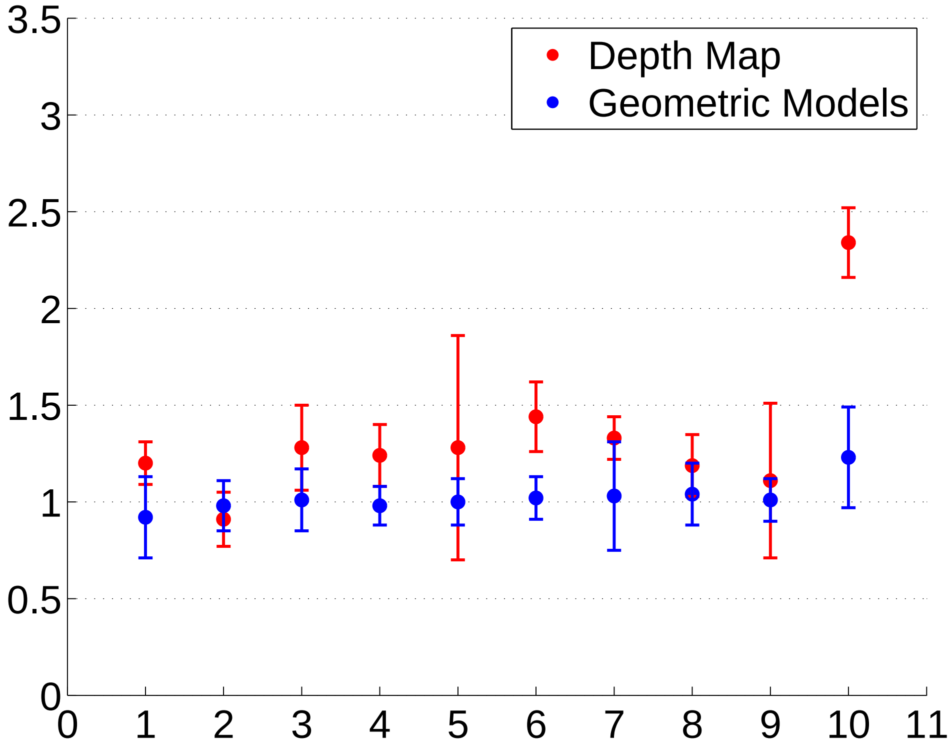


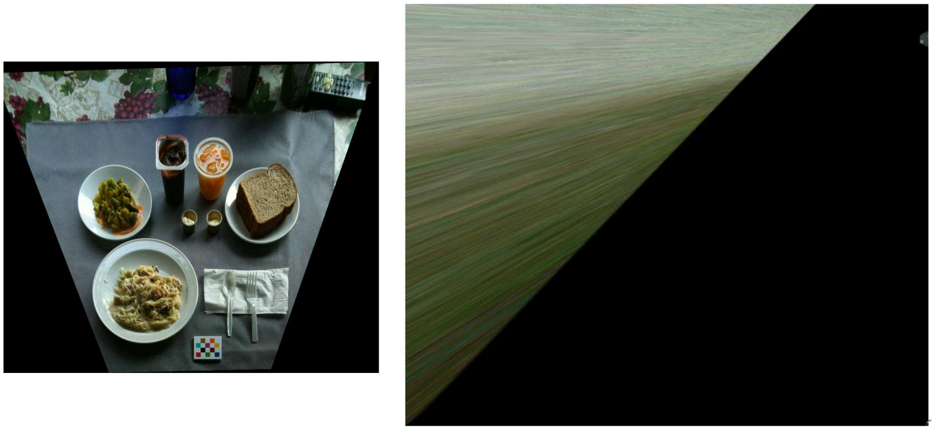
Fig. 3.8.: Comparison of the ratios of the estimate to ground truth. 1 – 10 on the horizontal axis represents foods listed in Table 3.1. A value '> 1' indicates the volume is overestimated, where a value '< 1' indicates the volume is underestimated.

4. THE USE OF CO-OCCURRENCE PATTERNS IN SINGLE IMAGE BASED FOOD PORTION ESTIMATION

4.1 Introduction

Estimating the portion of an object from a single-view image is an ill-posed problem. Most of the 3D information has been lost during the projection process from 3D world coordinates onto the 2D camera sensor plane. The use of the priori information is required to estimate the food portions. In [54] food portion estimation was converted into pre-determined serving size classification hence the technique could not be generalized. In [51, 60] pre-defined 3D template matching was used however it required manual tuning hence scaling with many foods became a problem. Food portion estimation using geometric models [14] and the approach based on predicted depth map using a Convolutional Neural Network (CNN) [11, 43] overcame the scaling issue with many foods. We compared the food portion size estimation accuracy using both geometric models and depth images [73]. We showed that geometric model based approach achieved higher accuracy compared to that using high quality depth images obtained by structured light technique [66]. In addition, the quality of depth map obtained using consumer level portable devices lead to even worse performance.

We have achieved accurate food portion estimation using geometric models, with ground truth food labels, segmentation masks been used for the experiments. In addition, accurate camera calibration that based on the fiducial marker included in the scene is critical for food portion estimation. As the fiducial marker included in the eating scene only occupy a small portion of the image area, the extrinsic camera parameters or the homography matrix to remove projective distortion of the scene may not be estimated accurately. We show a geometrically rectified image with projective distortion removed, and one that is not correctly rectified in Figure 4.1.



(a) A food image that is properly rectified to remove projective distortion

(b) A food image that is not properly rectified to remove projective distortion

Fig. 4.1.: Examples of food images that are properly rectified and food images that are not properly rectified.

The prism model estimate food areas based on the rectified food image [14, 60]. The inaccurate camera calibration could introduce errors for food portion estimation. To increase the portion estimation accuracy, we plan to incorporate contextual information of the scene such as food co-occurrence patterns for portion estimation. More specifically, food co-occurrence pattern may provide valuable insights for portion estimation.

4.2 The Use of Contextual Dietary Information

We use contextual information to further improve food portion estimation accuracy using geometric models based approach [14]. We define contextual dietary information as the data that is not directly produced by the visual appearance of an object in the image, but yields information about a user's diet or can be used for diet planning [74]. Food portion co-occurrence pattern is one type of contextual information. Other contextual information include the GPS coordinates associated with the food images, the temporal eating pattern and individual's eating pattern.

Such contextual information can not be determined by examining a single food image alone. In this work we develop a method to model the food portion co-occurrence patterns. We estimate the patterns from food images we collected for dietary studies. The patterns we estimated provide valuable insights about a user's eating behavior. We use the co-occurrence models to further refine the portion estimation results. We are able to obtain more accurate estimates of food portion sizes.

4.3 Estimating Food Portion Co-Occurrence Patterns for Portion Estimation Refinement

Food portion estimation based on a single-view image is an ill-posed problem and the 3D structure of the scene can not be fully reconstructed. The correct food classification label and segmentation mask in the image alone is insufficient for 3D reconstruction of a food item. The use of geometric models will allow for portion

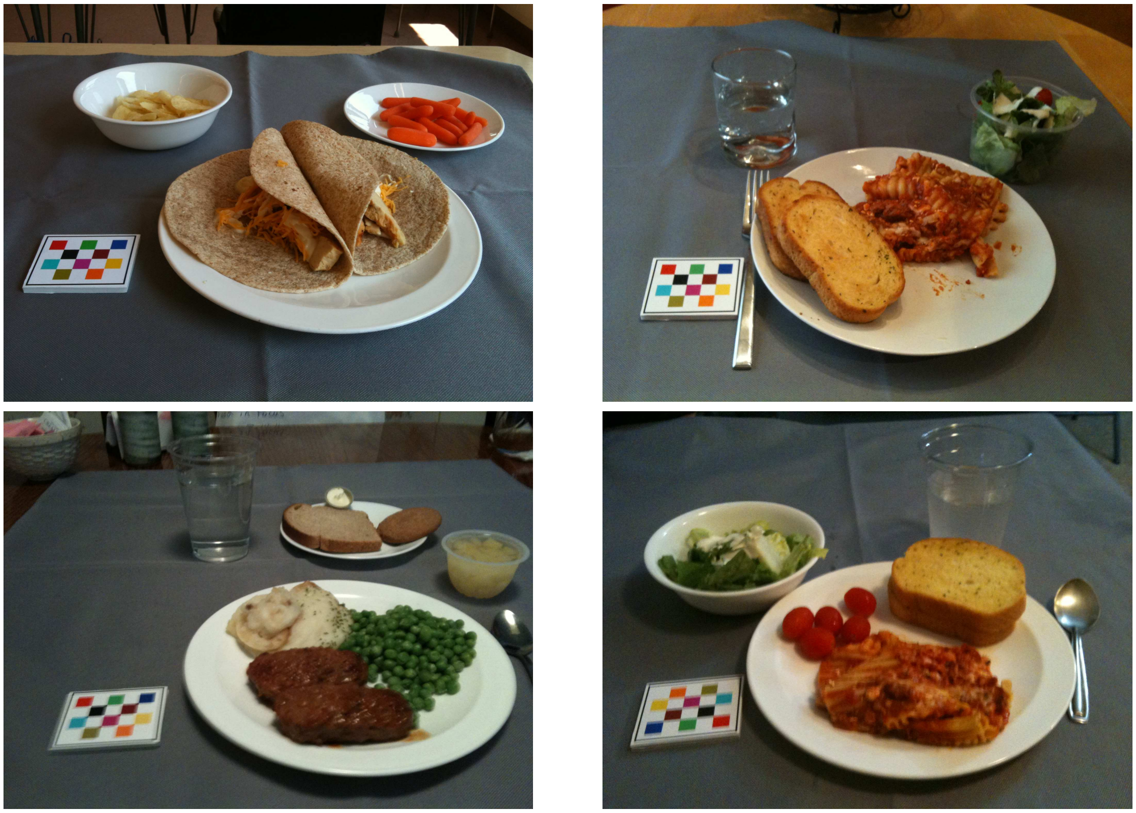


Fig. 4.2.: Sample food images collected by users using mFR with fiducial markers placed in the scenes.

estimation where food label is used to index into a proper class of a food type [14]. In this work we focus on the food classes that have varying shapes and appearances. We use prism model [14] to the food classes as the prism model is designed for food classes with varying shapes and appearances. We have designed a checkerboard pattern fiducial marker to be placed in the eating scene shown in Figure 4.2. The fiducial marker serves as a reference for both image rectification and food area sizes in world coordinates (in cm^2). We designed the fiducial marker to a credit card size for users to conveniently carry. The small size of the fiducial marker causes errors in the rectified image using computer vision techniques [20]. For example, if a food item is placed far away from the fiducial marker in the eating scene, the estimated portion (in cm^2) for such food item may be less accurate. To improve the accuracy of portion size estimation, we rely on a user's eating behavior modeled from food images of dietary studies. By proper modeling and incorporating the food portion co-occurrence patterns into portion estimation, we are able to improve the accuracy of portion estimation. Food portion co-occurrence patterns consist of the distributions of portion sizes and the associated weighting factor. We use Gaussian distributions as they best represent the characteristics of portion sizes distributions. We then refine the food portion estimates based on the models of food portion co-occurrence patterns.

4.3.1 Food Portion Estimation Using Prism Model

The prism model is an area-based food portion estimation technique based on the assumption that the height is the same for the entire horizontal cross-section of the food item. The 5 × 4 blocks color cherkerboard pattern fiducial marker is used as a reference for corner correspondences and the absolute size in world coordinates. The corners on the checkerboard pattern marker can be estimated using [75]. We obtain the 3×3 homography matrix $\mathbf{H}$ using Direct Linear Transform (DLT) [20]. Assume $\mathbf{I}$ is the original food image (as in Figure 4.2), the rectified image $\hat{\mathbf{I}}$ can then be obtained by: $\hat{\mathbf{I}} = \mathbf{H}^{-1}\mathbf{I}$. The segmentation mask S_j associated with food j in the original image can be projected from the pixel coordinates to rectified image coordinates. The area of segmentation mask $\hat{S}_j$ from the rectified image can then be estimated. We assume the height h_j for the entire horizontal cross-section. We use median height as the height of the same food class in our food image dataset. The portion of a food item V_j associated with segmentation mask S_j is then estimated: $V_j = \hat{S}_j \times h_j$.

4.3.2 Food Combination Patterns

Food combination pattern describes the frequencies of various food pairs present in the eating scenes. We use conditional probability of food items appearing in the same eating scene as the food combination patterns [74]. We estimate the food combination patterns from our food images collected from dietary studies. We define $c_{j,k}$ as the conditional probability of food category j appeared given that food category k is present as:

$$c_{j,k} = \frac{p(j,k)}{p(k)} = p(j|k) \tag{4.1}$$

The estimated food combination patterns represented by conditional probabilities is shown in Figure 4.3.

The food combination patterns only indicate whether two food items are likely to present in the same food image, hence it is insufficient to refine portion size estimation. We need to develop a technique to model the food portion co-occurrence patterns to refine portion estimation.

4.3.3 The Use of Food Portion Co-occurrence Patterns for Portion Estimation Refinement

The food portion co-occurrence patterns can help refining the portion estimates as they represent the insights reflected by the entire food image dataset rather than a single image. For example, if we know that food items j and k (e.g. fries and ketchup)

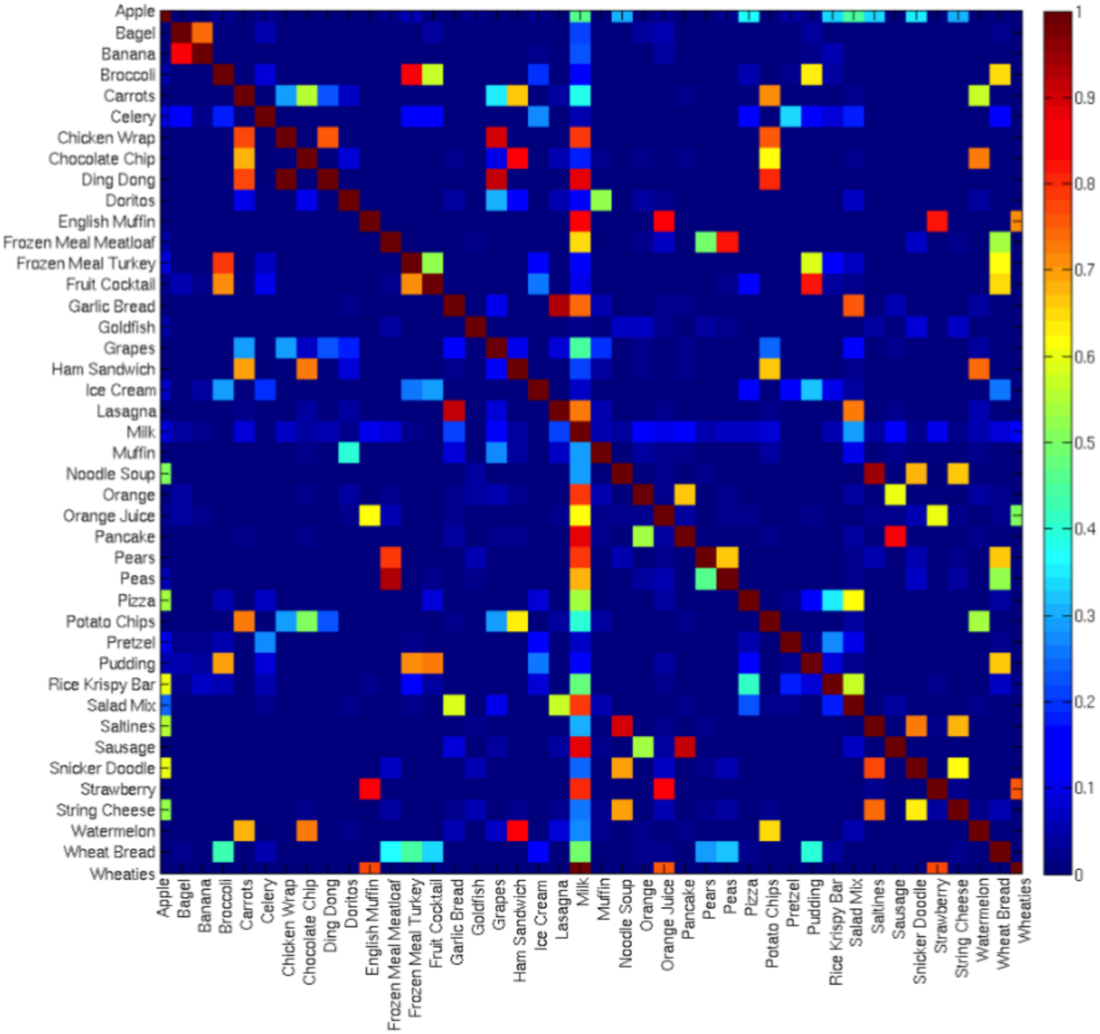


Fig. 4.3.: Food combination patterns represented by conditional probabilities.

usually appear in the same eating scene and the distribution of food portions j and k, we are able to refine the portion estimates based on such prior knowledge.

We use x_i^j, x_i^k to denote the food portions for food classes j, k estimated from food image with index i, where $i \in \{1, 2, 3, \cdots, N\}$. N is the size of our food image dataset, and $j, k \in \{1, 2, 3, \cdots, M\}$ where M is the number of the food classes we use. $\mathcal{C}_{j,k}$ is a combination pair that represents food items j and k are present in the same image. For the combination $\mathcal{C}_{j,j}$, the associated conditional probability is always $c_{j,j} = 1$. We denote $\mathcal{S}_i$ as the set containing all the combination pairs $\mathcal{C}_{j,k}$ exist in food image i. We use 2D Gaussian to model the distributions of portion sizes $\{x_i^j, x_i^k\}$ from our user food image data:

$$g_{j,k}(x_i^j, x_i^k) \sim N(\mu_{j,k}, \sigma_{j,k}) \tag{4.2}$$

Similarly, we use 1D Gaussian to model the distribution of portions x_i^j estimated for food class j:

$$g_{j,j}(x_i^j, x_i^j) = g_j(x_i^j) \sim N(\mu_j, \sigma_j) \tag{4.3}$$

where $\mu_{j,k}$, μ_j are the means and $\sigma_{j,k}$, σ_j are the standard deviations of the food estimates we obtained from our user food image data.

As the frequency of combination $\mathcal{C}_{j,k}$ appearing in our user food image data is different, we assign different weighting factors. For example, for the combination $\mathcal{C}_{j,k}$ that appears often across in food image dataset, we assign a heavier weight as it contributes more to the refinement of the portion estimates. Otherwise, we assign a lighter weight for combination $\mathcal{C}_{j,k}$. Furthermore, as the $\mathcal{S}_i$ is different for each food image i, the same $\mathcal{C}_{j,k}$ can carry different weight in different food image. We define the weighting factor $w_i^{j,k}$ of the combination $\mathcal{C}_{j,k}$ in image i as:

$$w_i^{j,k} = \frac{c_{j,k}}{\sum_{\forall \mathcal{C}_{j,k} \in \mathcal{S}_i} c_{j,k}} \tag{4.4}$$

Note that for each image i the weighting factor $w_i^{j,k}$ is different depending on the food combinations present. The food portion co-occurrence patterns consist of both the weighting factor $w_i^{j,k}$ and the distribution of the portion size estimates as shown

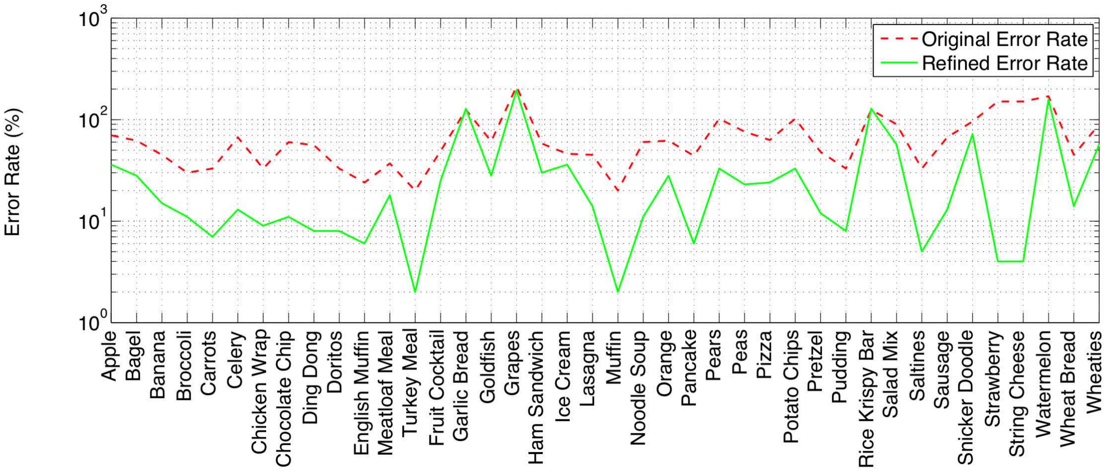


Fig. 4.4.: Average original errors vs. refined errors for portion estimates by food class.

in Equation 4.2 and 4.3. To refine the portion estimation results obtained using geometric models, we introduce a cost function in which we weight the probability of portion estimates: (x_i^j, x_i^k) in the image i based on co-occurrence patterns. The cost function is defined as:

$$f(x_i^j) = 1 - \sum_{\forall \mathcal{C}_{j,k} \in \mathcal{S}_i} w_i^{j,k} \cdot g_{j,k}(x_i^j, x_i^k) \tag{4.5}$$

Our goal is to minimize the cost such that the refined portion size best reflect the co-occurrence patterns we estimate from our food images collected in dietary studies. The refined food portion $\hat{x}_i^j$ in the eating scene can then be obtained by:

$$\hat{x}_i^j = \arg\min_{x_i^j} \{f(x_i^j)\} \tag{4.6}$$

4.4 Experimental Results

We divide our food image data into testing and training subsets. We tested on a total of 40 food classes. To reduce the errors propagate from the automatic classification and segmentation, we use ground truth food labels and segmentations masks. We use a subset of our food images for testing and leave the rest food images for training. Geometric model-based technique [14] is used to estimate the portion sizes from our food images. The median height of each food class is estimated from the training dataset. We model the food portion co-occurrence patterns based on the training subset for portion estimation refinement.

We refine the portion estimation results using food portion co-occurrence patterns. We compare the refined portion estimation error of each food class obtained using geometric models to the portion size errors without refinement. The errors in portion size estimation in Figure 4.4 are defined as:

$$Error = \frac{|\text{Estimated Portion Size} - \text{Ground Truth Portion Size}|}{\text{Ground Truth Portion Size}} \tag{4.7}$$

The ground truth portion sizes for each food item are provided by nutrient professionals.

The average errors per food class are obtained based on average error of 20 trials. In each trial we randomly sample 5% of our food images as testing subset. We use sampling with replacement technique so that the sizes of the training and testing subsets are the same for each trial. The original error in Figure 4.4 is the error of food portion estimates obtained using geometric model based [14] approach where the refined error is obtained by incorporating food portion co-occurrence patterns. For most food classes, we are able to improve portion estimation accuracy significantly using refinement technique, such as turkey meal. For some food classes the refinement technique was not sufficient to improve the estimation accuracy significantly (such as grapes). For a few food classes (garlic bread and rice krispy bar) the portion estimates become less accurate with refinement. This is due to the co-occurrence patterns we estimated from training dataset do not generalize well for these specific food classes. Such issue can be addressed by increasing the size of the food image dataset collected from future dietary studies as refinement is fundamentally adding biasness to our system based on past observations. If the past observations include variety of scenarios for most user behavior patterns, we can further improve the estimation accuracy. Our geometric model based portion estimation technique becomes less sensitive to noise by incorporating co-occurrence patterns. It has been shown in Figure 4.4 that the co-occurrence patterns we estimated generalize well for most of the food classes in our dietary studies. We define the improvement rate as:

$$Improvement = \frac{\text{Original Error} - \text{Refined Error}}{\text{Ground Truth Portion Size}} \tag{4.8}$$

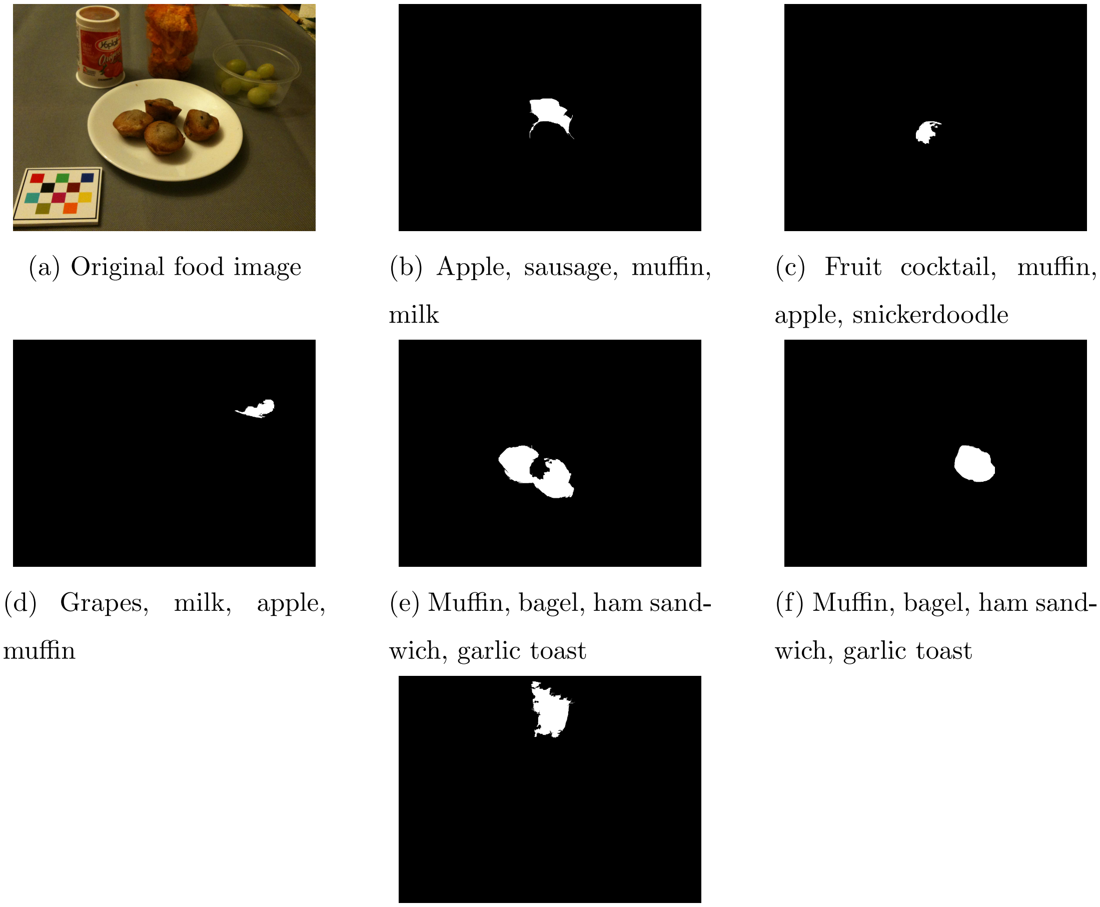
The overall improvement rate for our dataset is 36.9%.

4.5 Conclusion

We model the food portion co-occurrence patterns based on the food images we collected in dietary studies. The food portion estimation is refined by incorporating the portion co-occurrence patterns. We have shown that with the refinement we significantly improve the estimation accuracy for most of the food classes.

Single view food portion estimation technique requires the correct food label as we use food label to index into the correct class of geometric model. Although we have achieved accurate food portion estimation using geometric models, ground truth food labels, segmentation masks have been used for the experiments. Inaccurate segmentation mask leads to errors estimating the food portion sizes.

As our goal is to automatically analyze food images, ground truth food segmentation masks and food labels are not available in the automatic image analysis. Therefore, errors generated in automatic food region segmentation and food type classification shown in Figure 4.5 will propagate into portion estimation. It remains a challenge estimating food portion sizes with inaccurate food labels and segmentation masks.



(a) Original food image

(b) Apple, sausage, muffin, milk

(c) Fruit cocktail, muffin, apple, snickerdoodle

(d) Grapes, milk, apple, muffin

(e) Muffin, bagel, ham sandwich, garlic toast

(f) Muffin, bagel, ham sandwich, garlic toast

(g) Pizza, apple, turkey, carrots

Fig. 4.5.: Original food image (a) and automatic segmentation masks (b)-(g) and top 4 candidate food labels associated with segmentation mask.

5. LEARNING IMAGE-TO-ENERGY MAPPINGS USING GENERATIVE ADVERSARIAL NETWORKS

5.1 Introduction

We have previously developed a 3D geometric-model based technique for portion estimation [14, 76] which incorporates the 3D structure of the eating scene and use geometric models for food objects. We showed that more accurate food portion estimates could be obtained using geometric models for food objects whose 3D shape can be well-defined compared to a high resolution RGB-D images [73]. Geometric-model based techniques require accurate food labels and segmentation masks. Errors from these steps can propagate into food portion estimation.

More recently, several groups have developed food portion estimation methods using deep learning [77] techniques, in particular, Convolutional Neural Networks (CNN) [78]. In [11], a food portion estimation method is proposed based on the prediction of depth maps [57] of the eating scene. However, we have shown that the depth based technique is not guaranteed to produce accurate estimation of food portion [73]. In addition, energy/nutrient estimation accuracy was not reported in [11]. In [79], a multi-task CNN [80] architecture was used for simultaneous tasks of energy estimation, food identification, ingredient estimation and cooking direction estimation. Food calorie estimation is treated as a single value regression task [79] and only one unit in the last fully-connected layer (FC) in the VGG-16 [81] is used for calorie estimation.

Although CNN techniques have achieved impressive results for many computer vision tasks, they depend heavily on well-constructed training datasets and proper selection of the CNN architecture. We propose in this work to use generative models to estimate the food energy distribution from a single food image. We construct a

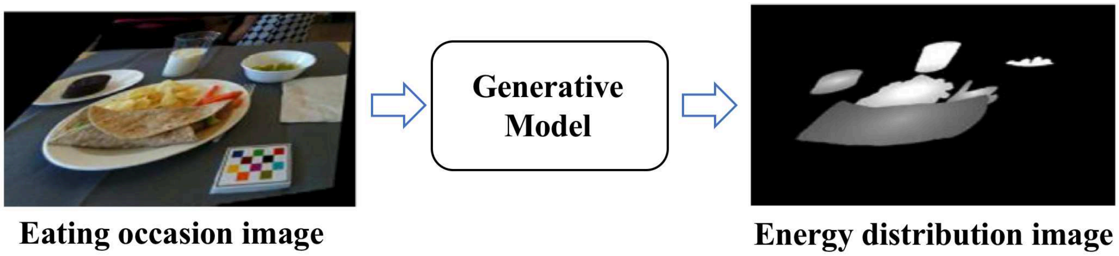


Fig. 5.1.: Learning image-to-energy mappings using Generative Adversarial Nets.

food energy distribution image that has a one-to-one pixel correspondence with the food image. Each pixel in the energy distribution image represents the relative spatial amount (or weight) of food energy at the corresponding pixel location. Therefore, a food energy distribution image provides insight not only on where the food items are located in the scene, but also reflects the weights of energy in different food regions (for example, regions of the image containing broccoli should have smaller weights due to lower energy (kilocalories) compared to regions of the image containing steak). The energy distribution image is one way that we can visualize these relationships.

More specifically, the generative model is trained on paired images [82] mapping a food image to its corresponding energy distribution image. Our goal is to learn the mapping of the food image to the food energy distribution image so that we can construct an energy distribution image for any eating occasion and then use this energy distribution to estimate portion size as shown in Figure 5.1.

The weights in food energy distribution image for the training data are assigned based on ground truth energy using a linear transform described in Section 5.2.1. We use a Generative Adversarial Networks (GAN) architecture [83] as GAN has shown impressive success in training generative models [82,84–87] in recent years. Currently, no publicly available food image dataset meets all of the requirements for training our generative model that learns the “image-to-energy mapping.” We constructed our own dataset based on ground truth food labels, segmentation masks and energy information for training the generative model. In this section, we first show that the proposed method can obtain accurate estimates of food energy from a single food image. Secondly, we introduce a method for modeling the characteristics of energy distribution in an eating scene.

5.2 Learning Image-to-Energy Mappings

Here we will initially discuss the requirements of the training dataset and then we will describe in more details how we construct the energy distribution image from the training data. Image pairs consisting of the food image and corresponding energy distribution image are required to train the GAN. There are a several publicly available food image datasets such as the PFID [88], UEC-Food 100/256 [89] and Food-101 [90]. However, none of these dataset contains sufficient information required for training a generative model that we can use to learn the "image-to-energy mapping". We created our own paired image dataset for training the GAN with ground truth food labels, segmentation masks and energy/nutrient information from a food image dataset we have collected from dietary studies. This is described in more details in Section 5.2.1. We use the conditional GAN architecture [82] for training our generative model.

5.2.1 The Image-to-Energy Dataset

The generative model is designed to best capture the characteristics of the energy distribution associated with food items in an eating scene. For food types that have different energy distribution (such as broccoli versus steak), the differences should be reflected in the energy distribution image. For constructing the image-to-energy training dataset, we use food images collected from a free-living dietary TADA$^{\text{TM}}$ study [91]. We manually generated the ground truth food label and segmentation mask associated with each food item in the user food image dataset. The ground truth energy information (in kilocalories) for each food item was provided by registered dietitians. For these food images we have a fiducial marker with known dimension that is located in each eating scene to provide references for worlds coordinates, camera pose, and color calibration. The fiducial marker is a 5 × 4 color checkerboard pattern as shown in Figure 5.2a. The food energy distribution image we construct from the above ground truth information needs to reflect the differences in spatial energy distribution for food regions in the scene. For example, for French fries stacked in pyramid shape, the center region of French fries should have more relative energy weight compared to the edge regions in the energy distribution image.

To construct the energy distribution image we first detect the location of the fiducial marker using [63]. We then obtain the 3 × 3 homography matrix $\mathbf{H}$ using the Direct Linear Transform (DLT) [20] to rectify the image and remove projective distortion. Assume $\mathbf{I}$ is the original food image, the rectified image $\hat{\mathbf{I}}$ can then be obtained by: $\hat{\mathbf{I}} = \mathbf{H}^{-1}\mathbf{I}$. The segmentation mask S_k associated with food k can then be projected from the original pixel coordinates to the rectified image coordinates as $\hat{S}_k = \mathbf{H}^{-1}S_k$. At each pixel location $(\hat{i}, \hat{j}) \in \hat{S}_k$, we assign a scale factor $\hat{w}_{\hat{i},\hat{j}}$ reflecting the distance of the pixel location $(\hat{i}, \hat{j})$ to the centroid of the segmentation mask $\hat{S}_k$. The scale factor $\hat{w}_{\hat{i},\hat{j}}$ is defined as:

$$\hat{w}_{\hat{i},\hat{j}} = \frac{1}{\sqrt{(\hat{i}-\hat{i}_c)^2 + (\hat{j}-\hat{j}_c)^2 + \phi_{\hat{S}_k}^{0.5}}}, \quad \forall(\hat{i}, \hat{j}) \in \hat{S}_k, \tag{5.1}$$

where $(\hat{i}_c, \hat{j}_c)$ is the centroid of $\hat{S}_k$ and the regularization term, $\phi_{\hat{S}_k}$, is defined as:

$$\phi_{\hat{S}_k} = (\sum_{\forall(\hat{i},\hat{j})\in\hat{S}_k} \mathbf{1}). \tag{5.2}$$

If the pixel location $(\hat{i}, \hat{j})$ is outside of the segmentation mask $\hat{S}_k$, then $\hat{w}_{\hat{i},\hat{j}} = 0, \forall(\hat{i}, \hat{j}) \notin \hat{S}_k$. With the scale factor $\hat{w}_{\hat{i},\hat{j}}$ assigned to each pixel location in $\hat{S}_k$, we can project the weighted segmentation masks $\hat{S}_k$ back to the original pixel coordinates as $\bar{S}_k = \mathbf{H}\hat{S}_k$, and learn the parameter ρ_k such that:

$$c_k = \rho_k \sum_{\forall(\bar{i},\bar{j})\in\bar{S}_k} \bar{w}_{\bar{i},\bar{j}}, \tag{5.3}$$

where c_k is the ground truth energy associated with food k, ρ_k is the energy mapping coefficient for $\bar{S}_k$ and $\bar{w}_{\bar{i},\bar{j}}$ is the energy weight factor at each pixel that makes up the ground truth energy distribution image. We then update the energy weight factors in $\bar{S}_k$ as:

$$\bar{w}_{\bar{i},\bar{j}} = \rho_k \cdot \bar{w}_{\bar{i},\bar{j}}, \quad \forall(\bar{i},\bar{j}) \in \bar{S}_k. \tag{5.4}$$

We repeat the process following Equation 5.1 and 5.3 for all $k \in \{1,\ldots,M\}$ where M is the number of food items in the eating scene image. We can then construct a ground truth energy distribution image $\bar{\mathcal{W}}$ of the same size as $\bar{\mathbf{I}}$: $\bar{\mathbf{I}} = \mathbf{H}\hat{\mathbf{I}}$, by overlaying all segments $\bar{S}_k$, $k \in \{1,\ldots,M\}$ onto $\bar{\mathcal{W}}$. Thus, we obtain the paired images of an eating scene: the image $\bar{\mathbf{I}}$ and the energy distribution image $\bar{\mathcal{W}}$ with one-to-one pixel correspondence as shown in Figure 5.2a and 5.2b.

5.2.2 Generative Adversarial Nets

In GANs, two models are trained simultaneously: a generative model G that captures the data distribution, and a discriminative model D that determines the probability that a sample came from the training data rather than G [83]. The common analogy for the GANs architecture is a game between producing counterfeits (generative models) and detecting counterfeits (discriminative model) [83].

To formulate the GANs, we specify the cost functions. We use $\theta^{(G)}$ to denote the parameters of generative model G and $\theta^{(D)}$ to denote the parameters of discriminative model D. The generative model G attempts to minimize the cost function:

$$J^{(G)}(\theta^{(D)},\theta^{(G)}) \tag{5.5}$$

where the discriminative model D attempts to minimize the cost function:

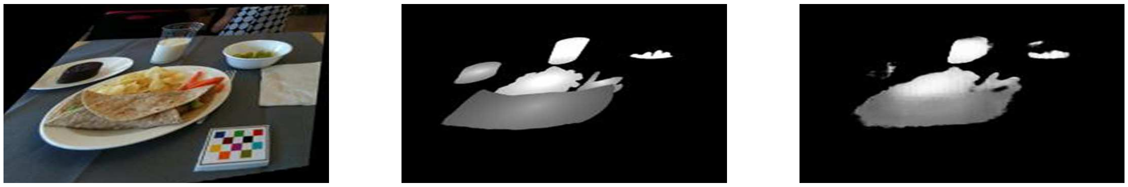
$$J^{(D)}(\theta^{(D)},\theta^{(G)}) \tag{5.6}$$

In a zero-sum game, we have:

$$J^{(G)}(\theta^{(D)},\theta^{(G)}) = -J^{(D)}(\theta^{(D)},\theta^{(G)}) \tag{5.7}$$

Therefore, the overall cost can be formulated as:

$$J^{(D)}(\theta^{(D)},\theta^{(G)}) = -\frac{1}{2}E_{x\sim p_{data}}(x)[\log D(x)] - \frac{1}{2}E_{z\sim p_z(z)}[\log D(1-(G(z)))] \tag{5.8}$$



(a) Eating occasion image $\bar{\mathbf{I}}$.

(b) Ground truth energy distribution image $\bar{\mathcal{W}}$.

(c) Estimated energy distribution image $\tilde{\mathcal{W}}$.

Fig. 5.2.: Learning image-to-energy translation using generative models.

where x is sampled from the true data p_{data} and z is random noise generated by distribution p_z. The generative model takes z and generate fake sample $G(z)$. The goal of the minimax game would then be:

$$\min_{\theta^{(G)}} \max_{\theta^{(D)}} -J^{(D)}(\theta^{(D)}, \theta^{(G)}) \tag{5.9}$$

During each update on the generative model G, the generated fake sample $G(z)$ will become more like the true sample x. Therefore, eventually after sufficient epochs of training, the discriminator D is unable to differentiate between the two distributions x and $G(z)$ [83].

The GANs takes adversarial training samples by its nature therefore could hugely reduce the adversarial space for the generative models to make mistakes. Therefore, the use of GANs architecture can greatly reduce the training samples needed and model the statistical insights of the true data.

5.2.3 Learning The Image-to-Energy Mappings

We use a Conditional GAN (cGAN) [82] that learns a generative model under conditional setting based on an input image. A cGAN is a natural fit for our "image-to-energy mapping" task since we want to predict the energy distribution image based on a food image.

More specifically, the cGAN attempts to learn the mapping from a random noise vector $\mathbf{z}$ to a target image $\mathbf{y}$ conditioned on the observed image $\mathbf{x}$: $G(\mathbf{x}, \mathbf{z}) \rightarrow \mathbf{y}$. The objective function of a conditional GAN is expressed as:

$$\begin{aligned} \mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{\mathbf{x},\mathbf{y} \sim p_{data}(\mathbf{x},\mathbf{y})}[\log D(\mathbf{x}, \mathbf{y})] + \\ \mathbb{E}_{\mathbf{x} \sim p_{data}(\mathbf{x}), \mathbf{z} \sim p_z(\mathbf{z})}[\log(1 - D(\mathbf{x}, G(\mathbf{x}, \mathbf{z}))]. \end{aligned} \tag{5.10}$$

An additional conditional loss $\mathcal{L}_{conditional}(G)$ is added [82] that further improves the generative model's mapping $G(\mathbf{x}, \mathbf{z}) \rightarrow \mathbf{y}$:

$$\mathcal{L}_{conditional}(G) = \mathbb{E}_{\mathbf{x},\mathbf{y} \sim p_{data}(\mathbf{x},\mathbf{y}), \mathbf{z} \sim p_z(\mathbf{z})}[\mathcal{D}(\mathbf{y}, G(\mathbf{x}, \mathbf{z}))], \tag{5.11}$$

where $\mathcal{D}(\mathbf{y}, G(\mathbf{x}, \mathbf{z}))$ measure the distance between $\mathbf{y}$ and $G(\mathbf{x}, \mathbf{z})$. Commonly used criteria for $\mathcal{D}(\cdot)$ are the L_2 distance [92]:

$$\mathcal{D}(\mathbf{y}, G(\mathbf{x}, \mathbf{z})) = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{y}_i - G(\mathbf{x}_i, \mathbf{z}_i))^2, \tag{5.12}$$

the L_1 distance [82]:

$$\mathcal{D}(\mathbf{y}, G(\mathbf{x}, \mathbf{z})) = \frac{1}{n} \sum_{i=1}^{n} |\mathbf{y}_i - G(\mathbf{x}_i, \mathbf{z}_i)|, \tag{5.13}$$

and a smooth version of the L_1 distance:

$$\mathcal{D}(\mathbf{y}, G(\mathbf{x}, \mathbf{z})) = \frac{1}{n} \sum_{i=1}^{n} \begin{cases} \frac{(\mathbf{y}_i - G(\mathbf{x}_i, \mathbf{z}_i))^2}{2} & \text{if } |\mathbf{y}_i - G(\mathbf{x}_i, \mathbf{z}_i)| < 1 \\ \\ |\mathbf{y}_i - G(\mathbf{x}_i, \mathbf{z}_i)| & \text{otherwise.} \end{cases} \tag{5.14}$$

The final objective for both the cGAN and the conditional terms is defined as [82,83]:

$$G^* = \arg \min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{conditional}(G). \tag{5.15}$$

The generative model G^* obtained from Equation 5.15 is then used to predict the energy distribution image $\tilde{\mathcal{W}}$ (Figure 5.2c) based on the food image (Figure 5.2a).

5.2.4 Experimental Results

We have 202 food images that have been manually annotated with ground truth segmentation masks and labels as training samples. All the food images are collected from a free-living (in the wild) TADA$^{\mathrm{TM}}$ dietary study [91]. Registered dietitians provided the ground truth energy information for each food item in the images. We constructed a dataset of paired images based on the 202 food images. Data augmentation techniques such as rotating, cropping and flipping were used to further expand our training dataset so that a total of 1875 paired images were used to train the cGAN. We used 220 paired images for testing.

We believe the training dataset size is sufficient for our task of predicting the energy distribution image because the cGAN is a mapping of a higher dimensional food image to a lower dimensional energy distribution image. In addition, since all food images are captured by users sitting naturally at a table, there is no drastic changes in viewing angles (for example, from wide angle to close up). In other image-to-image mapping tasks, a training dataset size of 400 has been used [82] for architectural labels (simple features) to photo translation (complex features) [93].

In testing, once the cGAN estimates the energy distribution image $\tilde{\mathcal{W}}$, we can then determine the energy for a food image (portion size estimation) as:

$$\text{Estimated energy} = \sum_{\forall (i,j) \in \bar{\mathbf{I}}} (\tilde{w}_{i,j}). \tag{5.16}$$

We compared the estimated energy image $\tilde{\mathcal{W}}$ (Figure 5.2c) to the ground truth energy image $\bar{\mathcal{W}}$ (Figure 5.2b), and define the error between $\bar{\mathcal{W}}$ and $\tilde{\mathcal{W}}$ as:

$$\text{Energy Estimation Error Rate} = \frac{\sum_{\forall (i,j) \in \bar{\mathbf{I}}} (\tilde{w}_{i,j} - \bar{w}_{i,j})}{\sum_{\forall (i,j) \in \bar{\mathbf{I}}} (\bar{w}_{i,j})} \tag{5.17}$$

To compare different cGAN models, we used the encoder-decoder architecture [94] as shown in Figure 5.3 and the U-Net architecture [95] as shown in Figure 5.4. We compared the energy estimation error rates at different epochs for both architectures. We observed that the U-Net architecture (Figure 5.5b) is more accurate in energy estimation and more stable compared to the encoder-decoder architecture (Figure 5.5a).

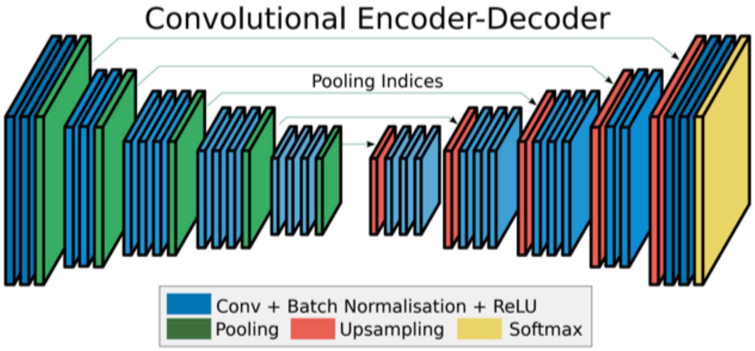

Convolutional Encoder-Decoder
Pooling Indices
🟦 Conv + Batch Normalisation + ReLU
🟩 Pooling
🟥 Upsampling
🟨 Softmax

Fig. 5.3.: The convolutional encoder-decoder architecture.

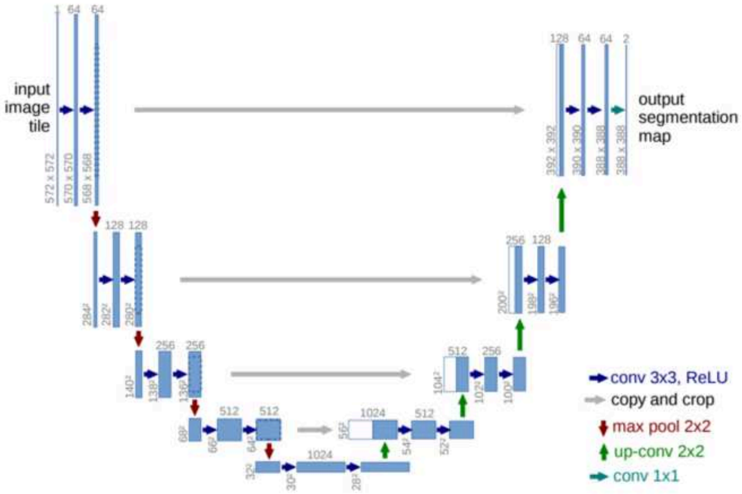
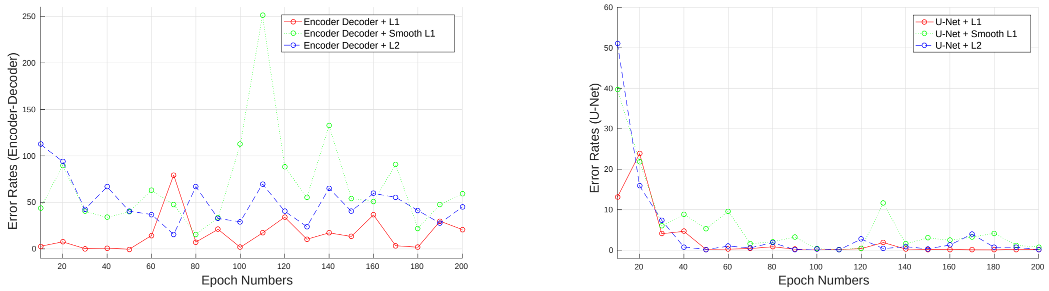


Fig. 5.4.: The U-Net architecture.



(a) Encoder-decoder.

(b) U-Net.

Fig. 5.5.: Comparison of error rates of different generative models: encoder-decoder versus U-Net.

This is due to the fact that the U-Net can copy information from the "encoder" layers directly to the "decoder" layers to provide precise locations [95], an idea similar to ResNet [96].

We also compared the energy estimation error rates under different conditional loss settings: $\mathcal{L}_{conditional}(G)$ using U-Net. We used the batch size of 16 with $\lambda = 100$ in Equation 5.15, the Adam [97] solver with initial learning rate $\alpha = 0.0002$, and momentum parameters $\beta_1 = 0.5$, $\beta_2 = 0.999$ as in [82]. Based on our experiments, distance measure $\mathcal{D}(\cdot)$ using the L_1 or L_2 norms is better than using smoothed L_1 norm. At epoch 200, the energy estimation error rates are 10.89% (using L_1 criterion) and 12.67% (using L_2 criterion), respectively. Using geometric-models [14] techniques, the energy estimation error was 35.58%. In the experiments, we included food types whose shapes are difficult to define (for example, fries). Estimating those food types is very challenging using geometric-model based approach [14].

5.3 Food Energy Estimation Based on Energy Distribution Mappings [1]

Single-view based food portion estimation is a challenging problem. To estimate food portions (in energy), we introduce the energy distribution image. Energy distribution image is a way we visualize where foods are in the image and how much relative energy is presented at different food regions. We use the GAN architecture to train the generative model that predicts the food energy distribution images based on eating occasion images. We have built a food image dataset with paired images [98] for the training of the GAN. To complete the end-to-end task of estimating food energy value based on a single-view eating occasion image, we use a CNN based regression model to estimate the numeric food energy value using the learned energy distribution mappings.

5.3.1 System Architecture for Food Energy Estimation Based on Energy Distribution Mappings

We are able to obtain the energy distribution image [98] for each RGB eating occasion image using generative model G trained by GAN. An example original food image and estimated energy distribution image are shown in Figure 5.2a and Figure 5.2c. Energy distribution images represents how food energy is distributed in the eating scene. Our goal is to estimate food energy (a numerical value), based on the estimated energy distribution image. This is essentially a regression task as shown in Figure 5.6. We use a Convolutional Neural Networks (CNN) based regression model to conduct the task of estimating energy from energy distribution image. For the regression model, we use a VGG-16 [81] based architecture as shown in Figure 5.7. As VGG-16 has shown impressive results on object detection tasks, VGG-16 is sufficient for learning complex image features. We modified the original VGG-16 architecture and added an additional linear layer as shown in Figure 5.7 so that the CNN based architecture is suitable for energy value regression task. Instead of using random

[1] This section is in joint work with Mr. Zeman Shao

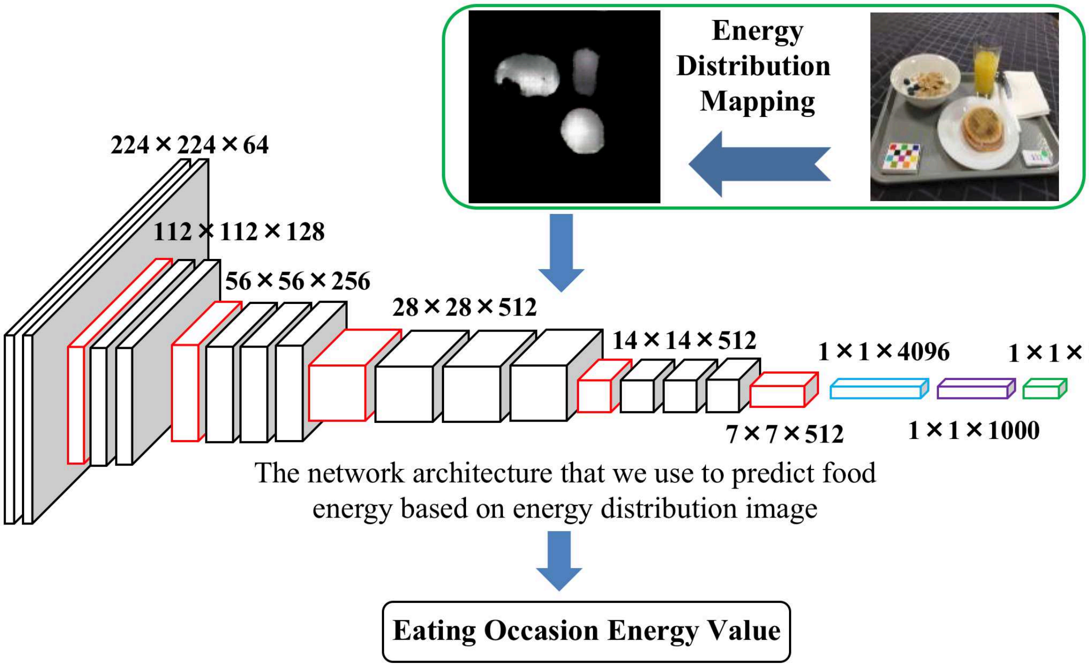


Fig. 5.6.: Estimating food energy of a meal based on predicted energy distribution image.

initialization for VGG-16 and training from scratch, we use pre-trained weights of VGG-16 architecture on ImageNet [99]. The pre-trained weights are indicated in the dash bounding box in Figure 5.7. We use random initialization for the linear layer. We then fine tune the pre-trained weights of VGG-16 network for energy value prediction task based on the building blocks of complex features originally learned from ImageNet [99]. With the regression model, we can predict the energy of the foods in a single-view eating occasion image.

5.3.2 Experimental Results for Food Energy Estimation Based on Energy Distribution Mappings

We predict the food energy of each eating occasion image based on its energy distribution generated by generative model. We then compare the food energy estimation to the ground truth food energy provided by the registered dietitians. We

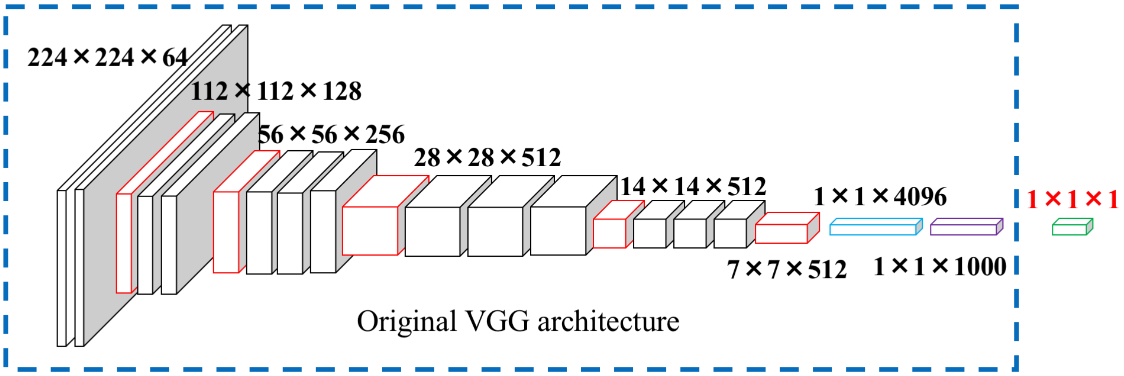


Fig. 5.7.: Using pre-trained weights to further fine tune for food energy estimation .

use 1390 eating occasion images collected from a free living dietary study [91], with ground truth food energy (kilocalories) for each food item in the eating occasion image provided by registered dietitians. We use 1043 of these eating occasion images for training and 347 of them for testing. All of eating occasion images are captured by the users sitting naturally at a table. There is no drastic changes in viewing angle. The errors for estimated food energy in Figure 5.9 are defined as:

$$\text{Error} = \text{Estimated Food Energy} - \text{Ground Truth Food Energy} \tag{5.18}$$

Figure 5.8 shows the relationship between the ground truth food energy and the food energy estimation of the eating occasion images in the testing dataset. The dash line in Figure 5.8 indicates the ground truth and estimated energy are the same, i.e., estimation error is equal to zero. Therefore, the points above this line are over estimated, and the points below this line are under estimated. Figure 5.10 and Figure 5.11 show examples of food energies have been over and under estimated, we use “+” and “-” to indicate over and under estimation, respectively. The average ground truth of eating occasion image in the testing dataset is 538.56 kilocalories. We observed that the estimation is more accurate for the eating occasion image with ground truth energy around average, compared to those with extremely high or low ground truth energy such as zero kilocalories. This is due to the fact that there are not sufficient eating occasion images in our dataset with very high or low ground truth energy provided to the neural networks for training.

The error distribution of estimated food energies for 347 eating occasion images is shown in Figure 5.9. We found that the average energy estimation error is 209.41 kilocalories.

5.3.3 Incorporating Depth Features for Energy Distribution Mappings

Previously, we directly train the food image to energy distribution mappings using generative models with random initialization. As food objects have 3D structures, the “energy distribution image” should incorporate how food energy is distributed

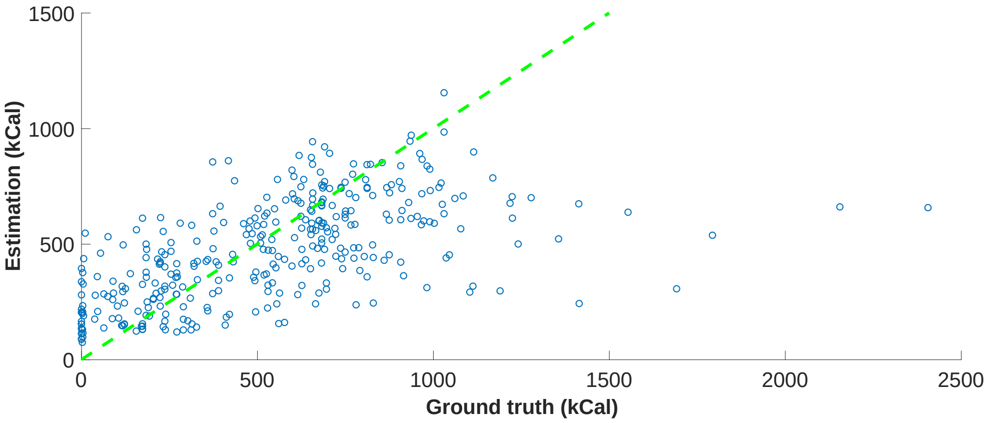


Fig. 5.8.: Relationship between the ground truth food energy and the food energy estimation.

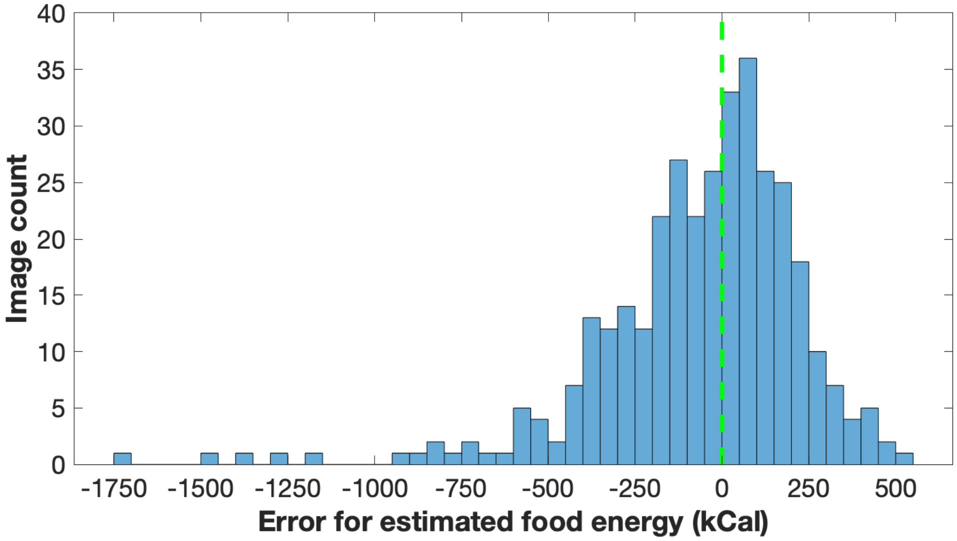
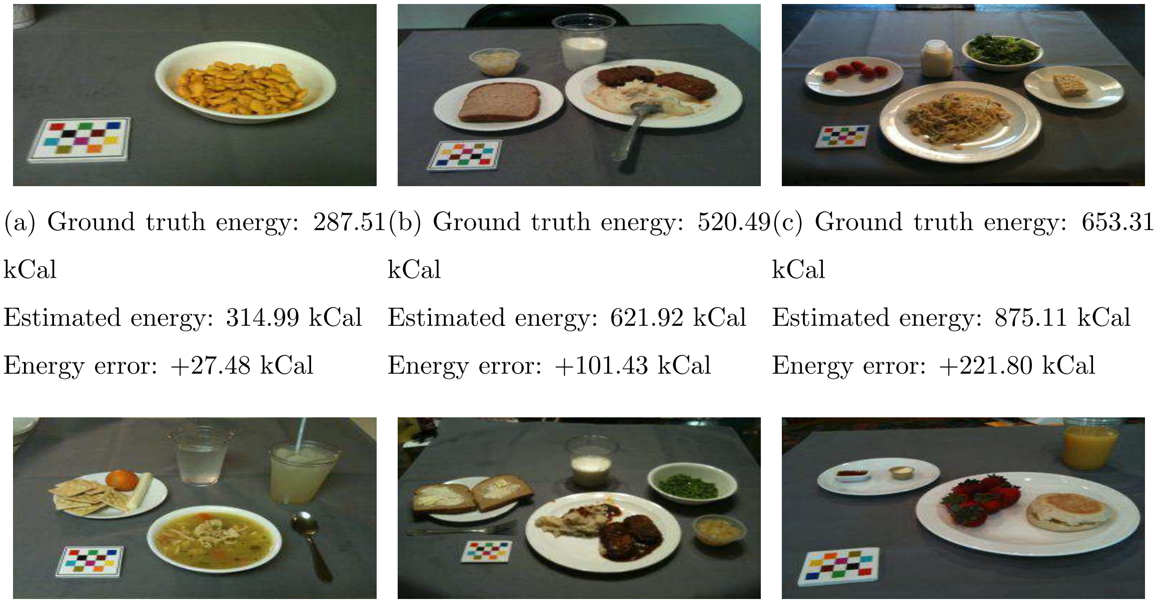


Fig. 5.9.: Error of food energy for each eating occasion images.



(a) Ground truth energy: 287.51 kCal
Estimated energy: 314.99 kCal
Energy error: +27.48 kCal

(b) Ground truth energy: 520.49 kCal
Estimated energy: 621.92 kCal
Energy error: +101.43 kCal

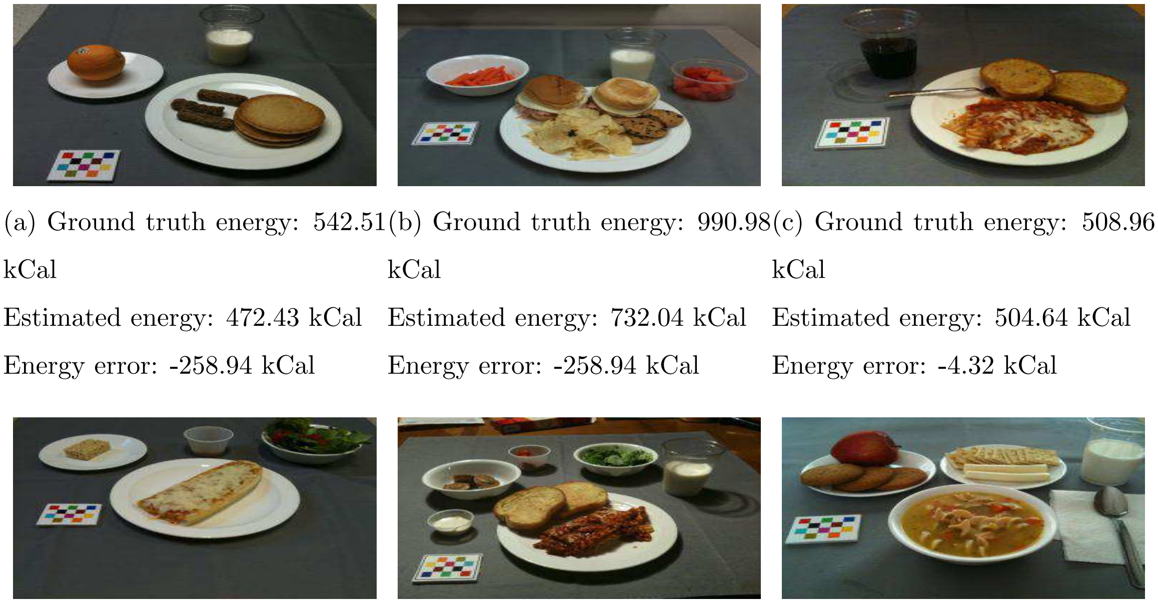
(c) Ground truth energy: 653.31 kCal
Estimated energy: 875.11 kCal
Energy error: +221.80 kCal

(d) Ground truth energy: 498.92 kCal
Estimated energy: 579.65 kCal
Energy error: +80.72 kCal

(e) Ground truth energy: 705.04 kCal
Estimated energy: 893.22 kCal
Energy error: +188.18 kCal

(f) Ground truth energy: 354.14 kCal
Estimated energy: 425.75 kCal
Energy error: +71.61 kCal

Fig. 5.10.: Examples of food energies been over estimated.



(a) Ground truth energy: 542.51 kCal
Estimated energy: 472.43 kCal
Energy error: -258.94 kCal

(b) Ground truth energy: 990.98 kCal
Estimated energy: 732.04 kCal
Energy error: -258.94 kCal

(c) Ground truth energy: 508.96 kCal
Estimated energy: 504.64 kCal
Energy error: -4.32 kCal

(d) Ground truth energy: 508.96 kCal
Estimated energy: 474.06 kCal
Energy error: -34.90 kCal

(e) Ground truth energy: 749.17 kCal
Estimated energy: 629.01 kCal
Energy error: -120.16 kCal

(f) Ground truth energy: 1084.61 kCal
Estimated energy: 708.83 kCal
Energy error: -375.78 kCal

Fig. 5.11.: Examples of food energies been under estimated.

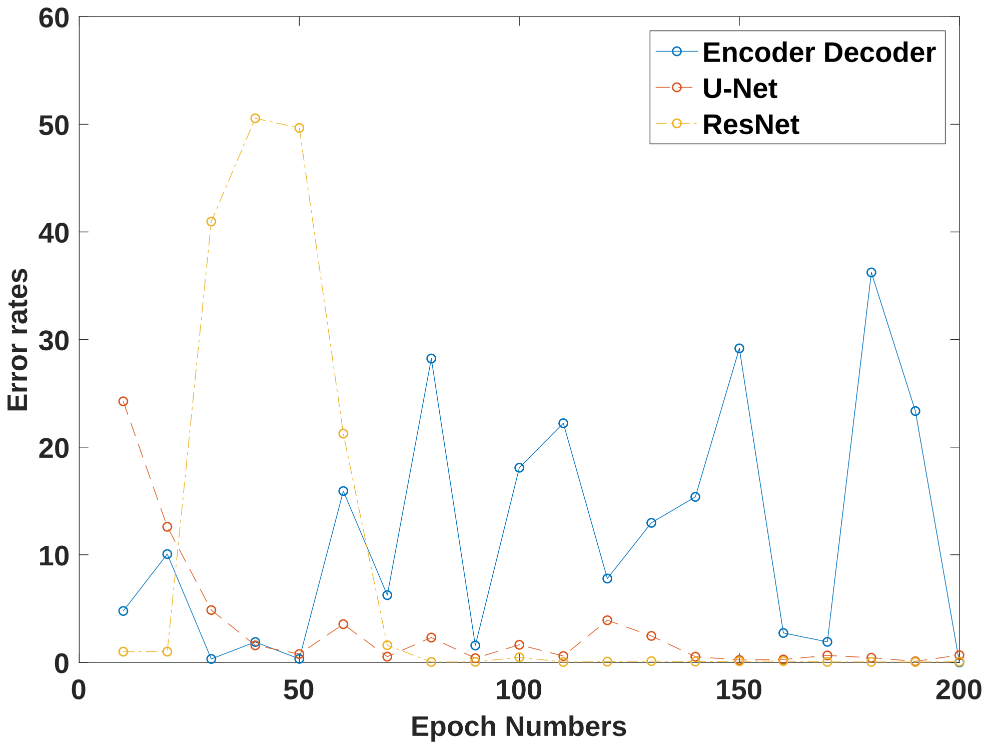


Fig. 5.12.: Incorporating depth features for energy distribution prediction.

spatially in the eating scene. Using the depth image of the scene is one way we can model the 3D structures of the scene. In addition, the depth images have more details/features for objects' structures and surfaces. Therfore, to better incorporate the 3D structure of food objects, instead of directly training on food image - energy distribution image pairs, we propose to first train the generative model using RGB-depth images pairs in [57]. Furthermore, mapping the RGB image to the depth images is a more complex task than mapping the RGB to energy distribution as there are more object categories, surfaces, normals and 3D structures.

With the trained generative model using GAN architecture that accurately estimate depth images on monocular RGB image [57], the complex features that map RGB to the depth have been learned by the generative model. We then train the same GAN architecture using learned weights for the generative model and discriminative

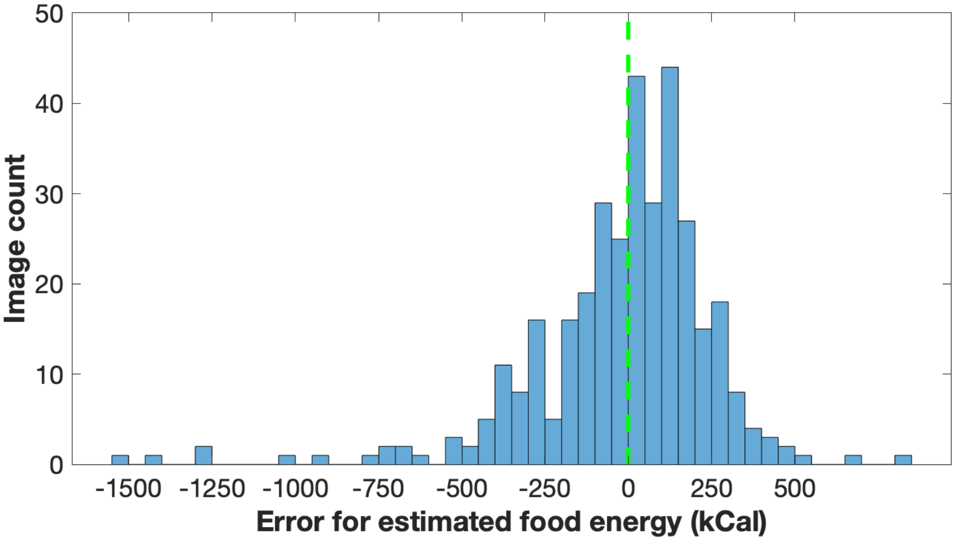


Fig. 5.13.: Error of food energy for each eating occasion images.

model for the task of image-to-energy mappings We show in Figure 5.12 that the use of pre-trained generative model on RGB-depth mappings increase the accuracy of the energy distribution mappings.

We use 9 layers of the networks block of ResNet [96] on [57] dataset. We use the weights of generative model trained for 200 epochs on depth prediction and fine tune on the task of energy estimation. We show that by incorporating depth information and using transfer learning, we have achieved more accurate energy distribution estimates using ResNet blocks as generative model shown in Figure 5.12 comparing to the previous approach using U-Net [95] and Encoder-Decoder architecture [94]. As the weights for the generative models was learned on RGB-D dataset, as early epochs the error increases as shown in Figure 5.12 but the error is consistently smaller than the previous approach [98] directly training on image-to-energy mappings in the later epochs.

We then estimate the food energy numeric value following the approach as shown in Section 5.3.1. Similarly to Section 5.3.1, we show the error distribution of estimated food energies for 347 eating occasion images in Figure 5.13 and the relationship

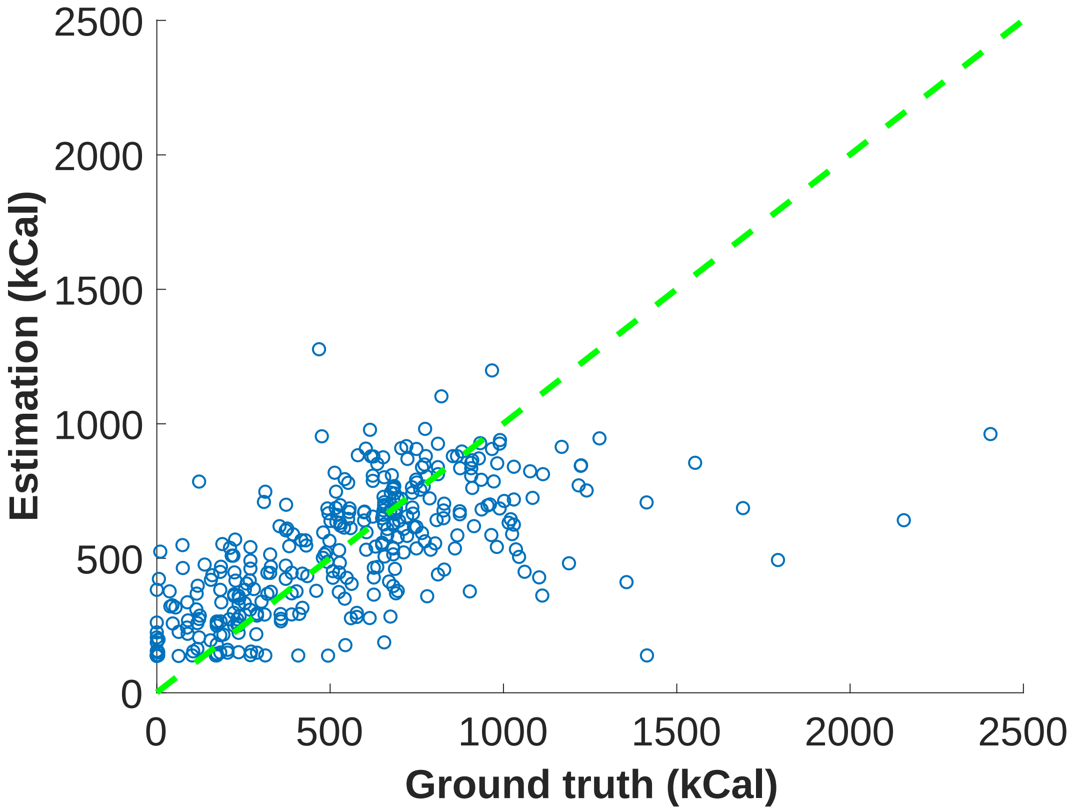


Fig. 5.14.: Relationship between the ground truth food energy and the food energy estimation.

between the ground truth food energy and the food energy estimation in Figure 5.14 using generative models pre-trained on depth. We found that the average energy estimation error is 191 kilocalories (comparing to 209.41 kilocalories without pre-trained on depth in Section 5.3.1).

5.4 Conclusion

We proposed a novel end-to-end system to directly estimate food energy from a captured eating occasion image. Our system first estimated the image to energy mappings using a Generative Adversarial Networks (GAN) structure. Based on the estimated energy distribution images, we learned the food energy of the eating occasion image by training CNN based regression model. We are able to obtain accurate food energy estimation with an average error of 209.41 kilocalories for eating occasion images collected from a free-living dietary study. The training based technique for end-to-end food energy estimation no longer requires fitting geometric models onto the food objects that may have issues scaling up as we need a large amounts of geometric models to fit different food types in many food images. In the future, combining automatically detected food labels, segmentation masks, and contextual dietary information has the potential to further improve the accuracy of such end-to-end food portion estimation system.

6. THE DESIGN OF A CROWDSOURCING TOOL FOR ONLINE FOOD IMAGE IDENTIFICATION AND SEGEMENTATION

6.1 Introduction

Training-based techniques have been widely used in recent years for developing automatic dietary assessment systems [7, 55, 100]. For training-based techniques, increasing the training data size would in general improve the accuracy of the system, thus a larger image dataset is always preferred. To date we have a food image dataset with more than 60,000 food images all collected from scientific studies that can be possibly used as training data for our system. The food images are collected using mFR^{TM} we developed from more than 14 scientifically implemented user studies, including environments in the wild, by more than 800 users. We have groundtruth food labels, segmentation masks and portion sizes information for thousands of the food images. In addition to the food images we have collected, a few other food image datasets are available, namely the PFID: Pittsburgh fast-food image dataset [88], UEC-Food 100/256 [89] and Food-101 [90]. The images in [88] are collected under laboratory set-up and only with fast food. Thus the categories and the appearances of the eating scenes do not best suit our use to examine realistic, diverse eating occasions. Furthermore, although both [90] and [89] contain a large amount of food images and a decent range of food types, we feel a detailed description for systematic design of food images collection and annotation is not revealed.

Without a well-designed user interface, removing the noisy images from candidate sets and generating the groundtruth segmentation masks are inefficient and not feasible. In addition, many food tags in [90] and [89] are dish names instead of individual foods (in [89] many are Asian style cuisine), we feel the datasets do not meet all of our needs. As our goal is not only to identify the food items but also to estimate the energy/nutrient information from the food images, we are interested in food items that have nutrient information made available by standard food nutrient databases, such as the United States Department of Agriculture (USDA) Food and Nutrient Database for Dietary Studies (FNDDS) [101].

Online image sharing is quickly gaining popularity in recent years (for example, through social networks such as Facebook and review orientated websites such as Yelp), and there are hundred-thousands of food images uploaded by smartphone users. We believe online food images can be used as part of our training data developing automatic dietary assessment techniques and provide valuable contextual information such as users' dietary patterns and food co-occurrence patterns. We define the contextual information as the data that is not directly produced by the visual appearance of an object in the image, but yields information about users' diet pattern or can be used for diet planning [100]. Collecting food images with proper annotations in a systematic way is a challenging task and requires systematic designs [99]. "Crowdsourcing", as defined in [102], also referred to as the collective intelligence, the wisdom of the crowd or human computation, is often considered as an effective solution to problems that involve cognitive tasks. Amazon Mechanical Turk (AMT) has been used in the past for food image collection and annotation tasks [89,103] however the AMT is not tailored for the needs emerged from our research of building a large food image dataset efficiently with food items labeled, localized, and segmented.

We present a crowdsourcing tool, namely the crowdsourcing TADA$^{\text{TM}}$ (cTADA$^{\text{TM}}$), that is tailored to address our needs of online food image collection and annotation. In addition to label and localize the target objects in the images [99], the cTADA$^{\text{TM}}$ is also capable of generating accurate segmentation masks for food objects based on users' input. To generate the segmentation masks, both the user input and automatic segmentation technique [104] are required. We used a programming interface to collect a large amount of online food images. We designed criteria for the removal of noise from images. Similar to [99], we are able to label and localize the food objects in images. In addition, the cTADA^{TM} tool allows us to identify all the food items in an image (located by bounding boxes) and generate associated segmentation masks for each food item.

6.2 The Design of the Food Image Crowdsourcing Tool [1]

Various food websites (such as `foodspotting.com`, `foodgawker.com`) contain large amounts of food images. Many food images are uploaded by users on reviews-oriented websites (such as Yelp) and image sharing/social networks (such as Flickr, Instagram, Pinterest, Facebook). We believe many of those food images can be used as the training data in our TADA^{TM} image analysis system. We define a set of criteria for a food image to be included in our dataset. In addition, the crowdsourcing tool must be efficient and effective as each of the crowd members will go through thousands of food images.

6.2.1 Obtaining Online Food Images

Manually downloading thousands of online food images is not feasible. We use Application Programming Interface (API) made available by image website or the search engine for image collection. The APIs we used were Flickr API [105] and Google Custom Search Engine (CSE) API [106]. The APIs allow us to obtain the food images based on the search terms (food tags) we are interested in. Existing datasets frequently use dish names as food tags. The disadvantage of using dish name is that the same type of dish posts very large variation by the look, ingredients and layouts as they were prepared by different people/restaurants. We use the food categories that are frequently present in our existing food image dataset collected from users in nutrition/health studies. The advantage of using such food categories is the energy and nutrient information is made available by the FNDDS database [101].

[1] This section is partially in joint work with Ms. Chang Liu

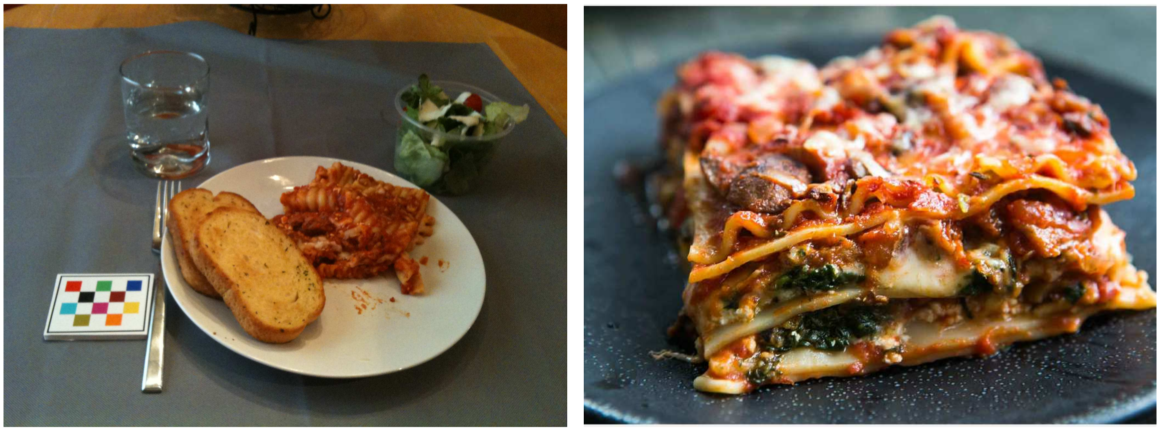


Fig. 6.1.: Examples of food images we collected for the nutrition scientific studies (left) v.s. food images collected online with aesthetic appearances (right).

The food images obtained based on the tags will inevitably contain noisy images that we can not use. We define the noisy images as those that either contain irrelevant content, or have significant different appearances compared to our existing food images collected from scientific studies. A crowdsourcing process is required to remove the noisy images from the candidate food images collected.

6.2.2 Noisy Image Removal Using Crowdsourcing

We first remove images that contain irrelevant contents. The irrelevant content means no food item in the image, images with logos/watermarks/texts and images containing faces. As our goal is to incorporate the online food images collected as part of the training dataset, we want to only include the images that are taken by actual users and exclude those images with aesthetic appearances (a comparison as shown in Figure 6.1). Food images with aesthetic appearances are likely captured and/or retouched by professional photographers and have fundamental differences compared to the images taken by average users regarding textures, colors, angles and layouts. To guide the crowds to successfully remove images with such aesthetic appearances, we define clear criteria for crowd members with image examples that show different lightings (e.g. professional lighting versus environment light), colors (e.g. vivid and saturated color versus natural color), textures (e.g. very smooth and reflective surface versus regular surface), angles (e.g. close-up or other creative angles versus common camera poses).

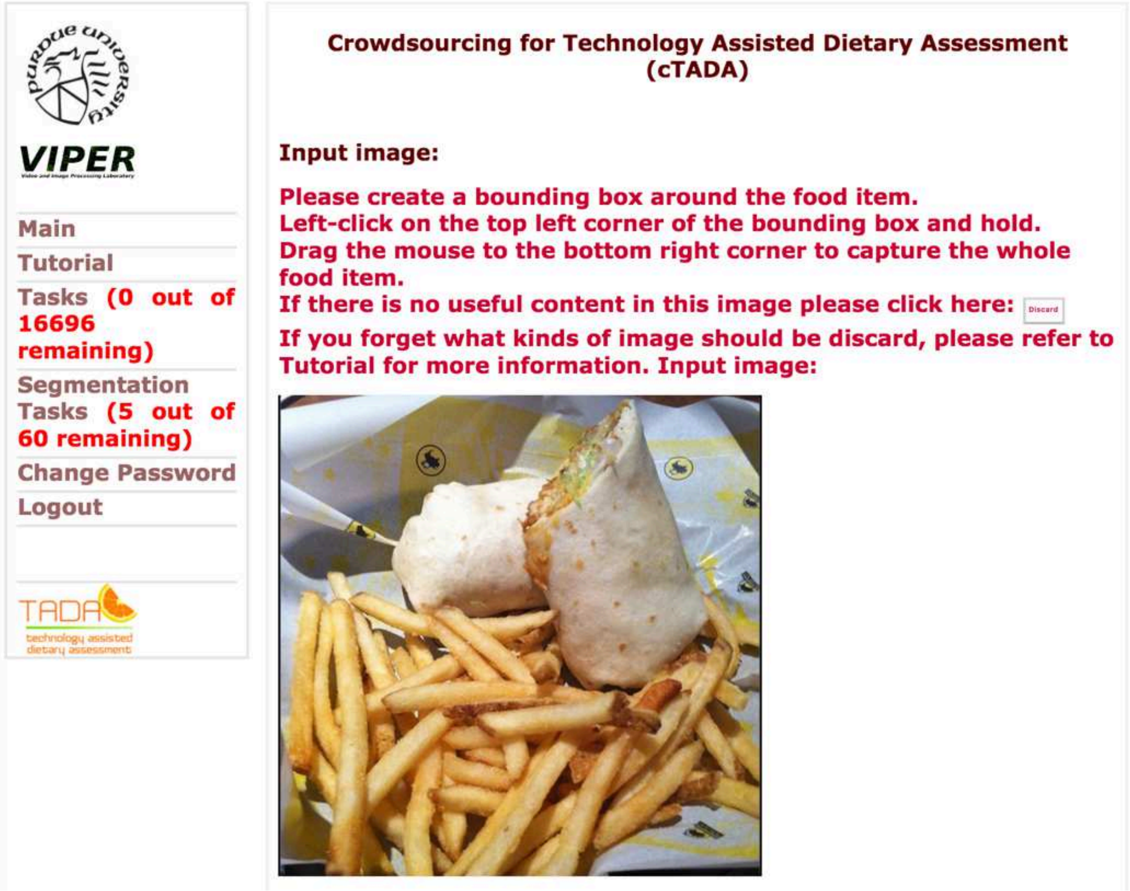
We do not exclude the food images that contain multiple food items. In fact, we believe food images that contains multiple food items will help us better understand the users' diet patterns and food co-occurrence patterns. Such patterns can provides us with important insights that can help dietary assessment.

6.2.3 Food Item Localization and Segmentation

In addition to removing noisy images, we also want to be able to efficiently have crowds locating and obtaining the segmentation masks associated with the food items in an image. We only assign food images that passed the noisy image removal step to crowds for food item localization and segmentation. Users can still discard noisy images as shown in Figure 6.2 in case of a false positive (where the image should be neglected in a noisy image removal step).

To locate the food item, we ask the users to first draw a bounding box around one food item. This task can be performed easily and efficiently by click-and-drag using a computer mouse on our web interface. The bounding box drawn is then cropped out of the original image as preparation for generating the segmentation mask. Users can then select a food tag associated with the bounding box from the hierarchical drop-down food list. The hierarchical drop-down list is designed to best incorporate users' intuitions, for example, we use "meats", "beverages", "green vegetables", "red and orange vegetables" as top level entries where more food categories are available once a top level entry is selected.

The hierarchical drop-down list for food category selection is shown in Figure 6.3.

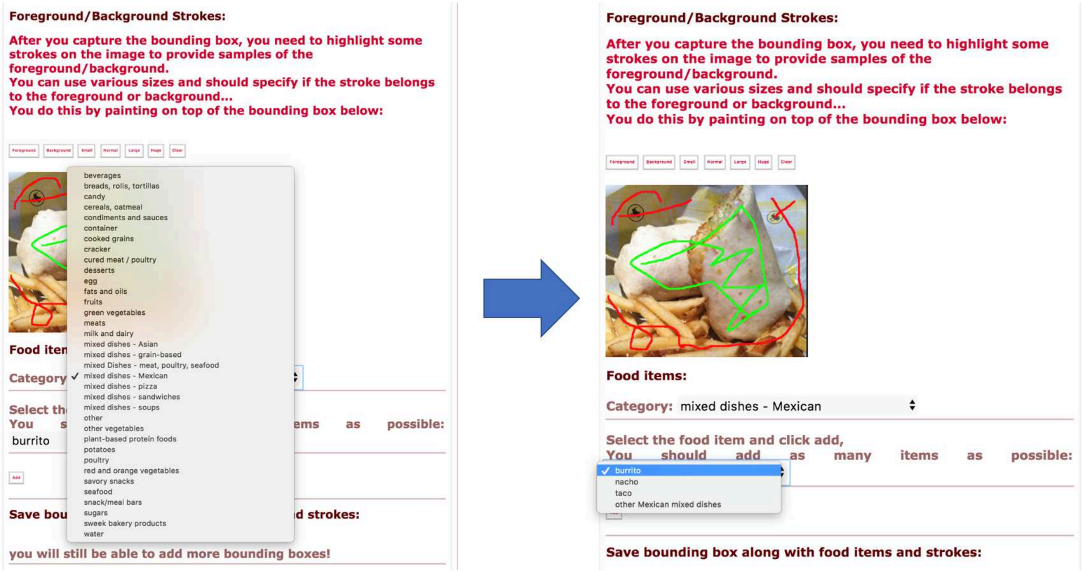

VIPER
Video and Image Processing Laboratory
Main
Tutorial
Tasks (0 out of 16696 remaining)
Segmentation Tasks (5 out of 60 remaining)
Change Password
Logout
TADA
technology assisted dietary assessment
Discard

Crowdsourcing for Technology Assisted Dietary Assessment (cTADA)

Input image:

Please create a bounding box around the food item.
Left-click on the top left corner of the bounding box and hold.
Drag the mouse to the bottom right corner to capture the whole food item.
If there is no useful content in this image please click here: Discard
If you forget what kinds of image should be discard, please refer to Tutorial for more information. Input image:

Fig. 6.2.: Image can still be discarded in food item localization and segmentation step.



Foreground/Background Strokes:

**After you capture the bounding box, you need to highlight some strokes on the image to provide samples of the foreground/background.
You can use various sizes and should specify if the stroke belongs to the foreground or background...
You do this by painting on top of the bounding box below:**

Food iten

Category

Select th
You s ems as possible:

burrito

Save bou d strokes:

you will still be able to add more bounding boxes!

Foreground/Background Strokes:

**After you capture the bounding box, you need to highlight some strokes on the image to provide samples of the foreground/background.
You can use various sizes and should specify if the stroke belongs to the foreground or background...
You do this by painting on top of the bounding box below:**

Food items:

Category: mixed dishes - Mexican

**Select the food item and click add,
You should add as many items as possible:**

Save bounding box along with food items and strokes:

Fig. 6.3.: Hierarchical food tag selection user interface.

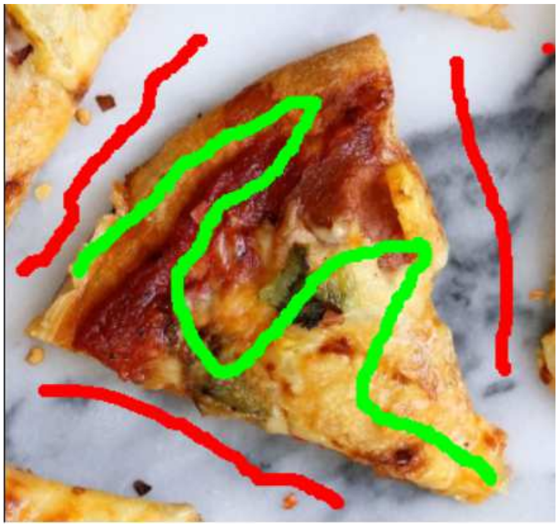


Fig. 6.4.: Defining the foreground (green) and background (red).

To segment the food items, we implemented a stroke tool for users to define foreground and background. Foreground is the area that is associated with the food item, otherwise it will be defined as background. Users do not need to cover all areas of foreground nor background. Drawing lines (the traces of the stroke) across the foreground and background (shown in Figure 6.4) is sufficient.

Similar to many drawing softwares, users can select the linewidth of the stroke tool. With foregrounds and backgrounds defined within the bounding boxes, we use automatic segmentation technique to generate the segmentation mask within the bounding box using the grab cut technique [75, 104]. For the food images that contains multiple food items (shown in Figure 6.5), the above procedures are repeated till bounding boxes associated with all food tags are located and a segmentation mask is generated for each food item in the images as shown in Figure 6.6.

6.3 Experimental Results for cTADA^{TM} Crowdsourcing Tool

For the initial crowdsourcing experiment we recruited the crowds from graduate school students pool all with engineering background in the field of image processing

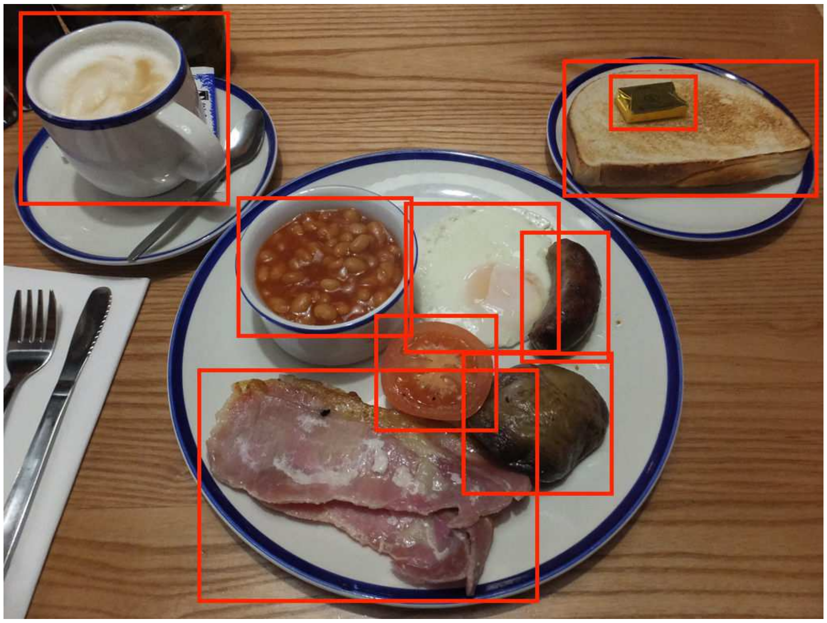
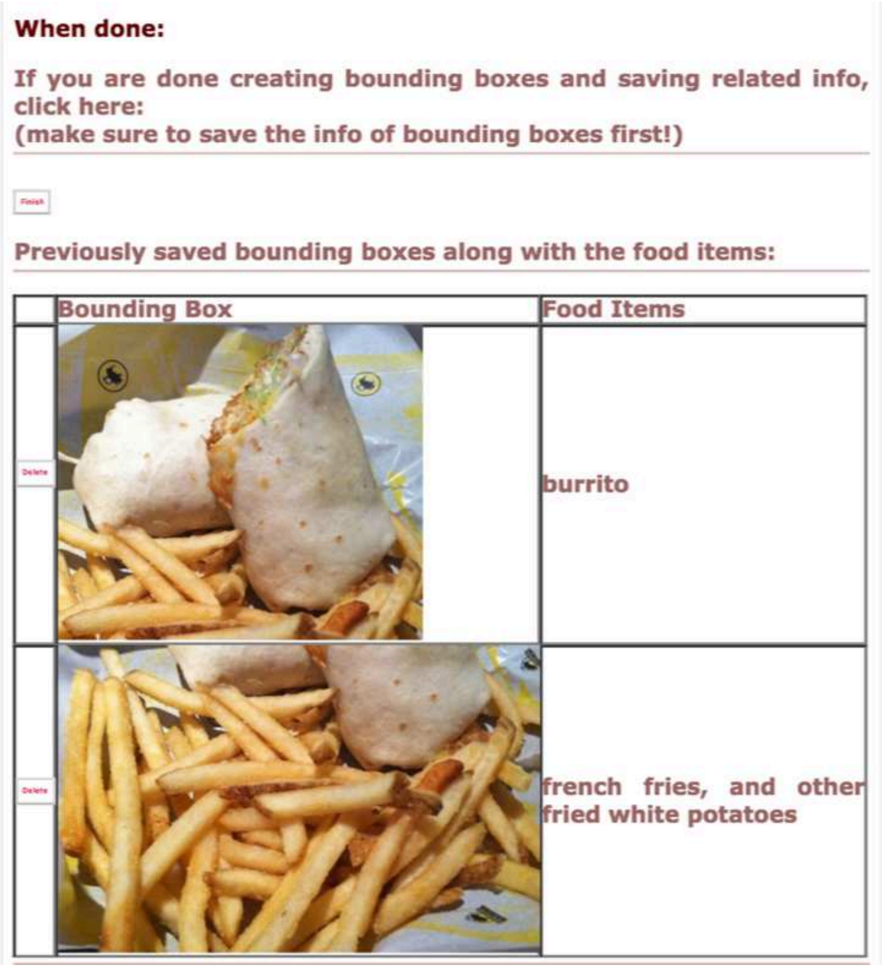


Fig. 6.5.: An example of online food image that contains multiple food items.



When done:

If you are done creating bounding boxes and saving related info, click here:
(make sure to save the info of bounding boxes first!)

Previously saved bounding boxes along with the food items:

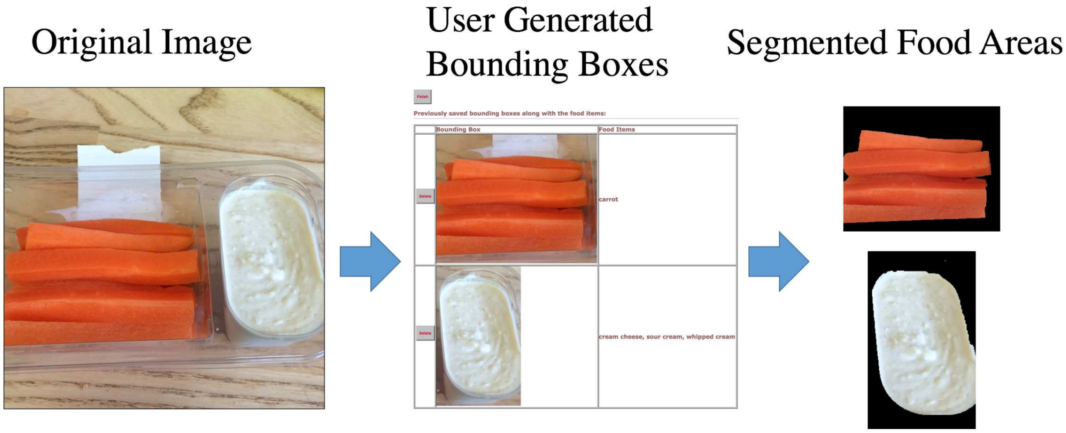
	Bounding Box	Food Items
Delete		burrito
Delete		french fries, and other fried white potatoes

Fig. 6.6.: Localizing multiple food objects in the same image.

and computer vision. Our crowds are able to give valuable feedback on improving the cTADA^{TM} crowdsourcing tool at initial design stage. The crowd users can only use our web interface, and were not involved in any of the programming tasks.

For noisy image removal, we implemented a one-click confirmation and short-cut keys on the keyboard, so users can even skip the point-and-click using the computer mouse. The confirmation is then saved in our database and the next image will be automatically present to users to minimize a user's effort. We provide a tutorial on the criteria of noisy image removal to the users. In tutorials, we provide side-to-side comparisons of images and a descriptions for the criteria we designed. We found that users can easily adapt to our set of noisy image removal criteria. With the tutorials, identifying aesthetic appearances is no longer a challenging task even for the crowd members lacking experiences in photography. Based on our observation, we find examining one image takes one second on average for the user, and a maximum of a few seconds. The cTADA^{TM} system has shown great efficiency in the task of noise image removal and we were able to obtain almost 40,000 food images that can be added to our dataset.

The process of localizing and obtaining the segmentation masks associated with all the food items in an image is shown in Figure 6.7. Users work on one food item at a time. For example, a user will first obtain the bounding box associated with one food item, then identify the food type and define the foreground and background using a 'stroke' tool and 'save' the action performed using the user interface. If there is more than one food item present, an 'add' button can be clicked to repeat the above procedures till all food items are done. The procedure is straight forward and minimizes users' efforts. We do not require users to manually crop out the segmentation masks as it is time consuming and not feasible when working with a large image dataset. Instead, the automatic segmentation tool [75, 104] we implemented on our server will generated very accurate segmentation masks from the bounding boxes and foregrounds and backgrounds defined, as shown in Figure 6.7.



Previously saved bounding boxes along with the food items:

	Bounding Box	Food Items
		carrot
		cream cheese, sour cream, whipped cream

Fig. 6.7.: Locating the food items and obtaining the segmentation masks.

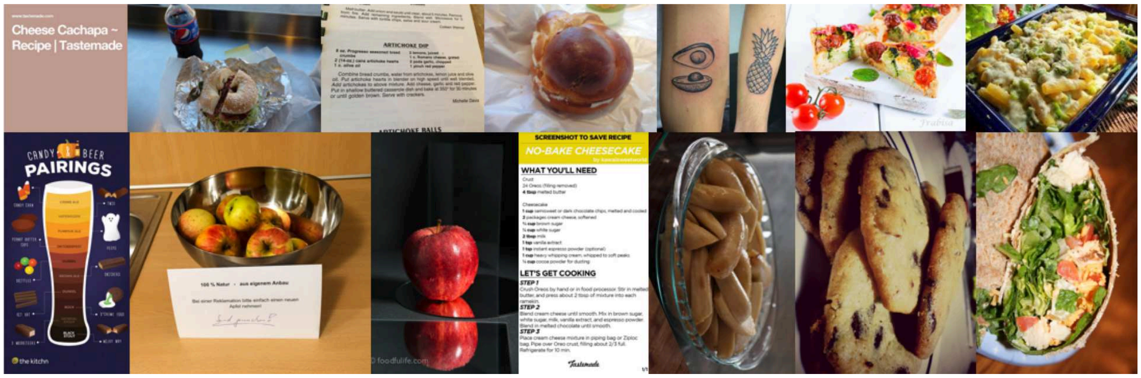


Fig. 6.8.: The images downloaded from the Internet contains many noises.

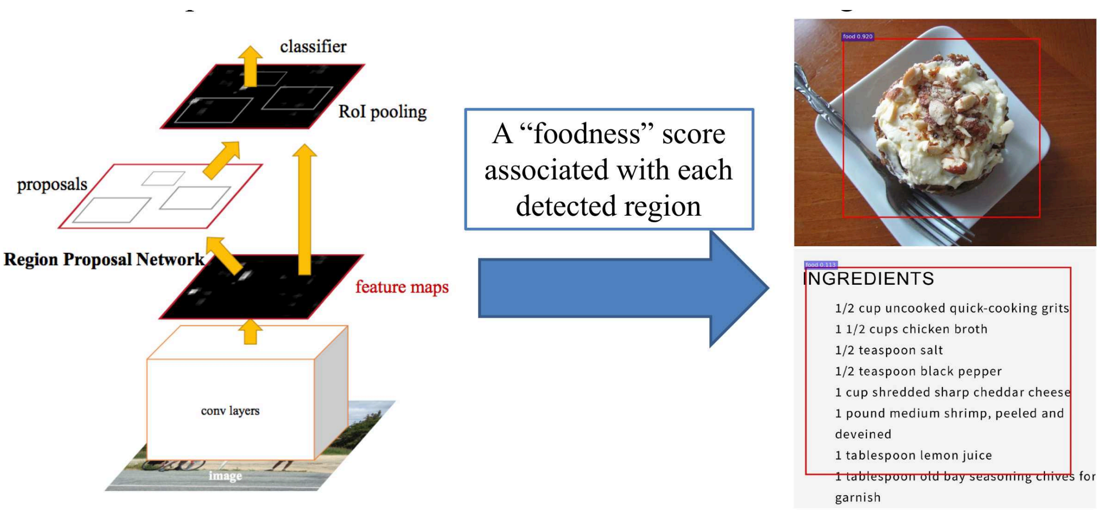
6.4 Automated Noisy Image Removal For Online Food Image Collection

6.4.1 Motivation for Automated Noisy Image Removal

We have developed learning-based methods for automatic dietary assessment, that require sufficient training data to achieve high accuracy. We incorporate the online food images into our existing food image dataset, to increase the size of training data. However, many food images retrieved from the Internet are considered noisy to us as they may either have no food item presents the image, or contain irrelevant content such as logos/watermarks/texts and images with human faces as shown in Figure 6.8.

In the original cTADA^{TM} design, we have used crowdsourcing tools that completely rely on human annotators to verify each downloaded image. Although we have developed efficient tools for crowdsoucing task, it is always preferred that fewer noisy images would be passed to the crowds to speed up the process of image annotation. Therefore, we propose a technique to automatically remove non-food images instead of relying on human annotators which are expensive and time consuming. By removing many noisy images automatically, we can speed up and improve the quality of subsequent crowdsourcing tasks.

Object detection is one of common tasks in computer vision research and recent techniques based on Convolutional Neural Networks (CNNs) have shown impressive



INGREDIENTS

1/2 cup uncooked quick-cooking grits
1 1/2 cups chicken broth
1/2 teaspoon salt
1/2 teaspoon black pepper
1 cup shredded sharp cheddar cheese
1 pound medium shrimp, peeled and deveined
1 tablespoon lemon juice
1 tablespoon old bay seasoning chives for garnish

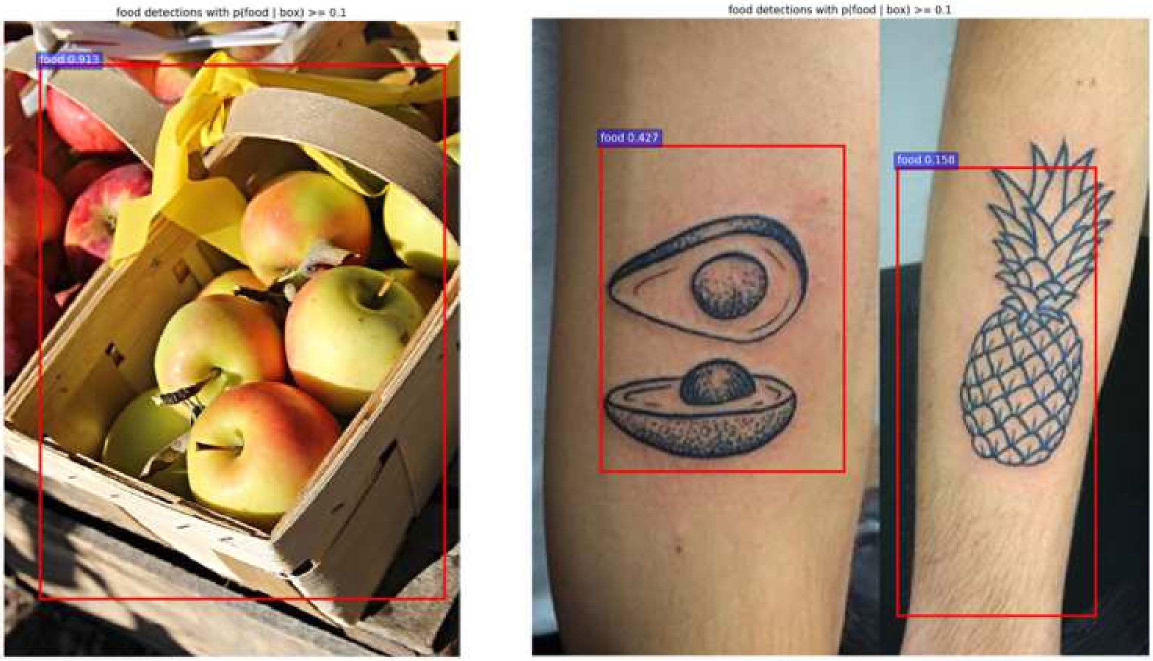
Fig. 6.9.: Region proposal networks.

results in object detection task [107–111]. In order to automatically remove noisy images, we need a system that can accurately propose food regions. Then, based on the proposed food regions and their associated confidence scores, we can decide either to keep or discard the food image.

We train a neural network based on the Faster R-CNN [109] architecture for food region proposal as shown in Figure 6.9.

For each proposed food region, a confidence score has been assigned [109]. We denote the confidence score associated with each proposed food region as the "foodness" score shown in Figure 6.10a and Figure 6.10b. The "foodness" score has the range of $[0, 1]$. The higher the confidence score is, the more likely the region proposal network believe there is food present in the proposed region. Based on the highest "foodness" score in an image, we then decide whether or not to keep the food image for the subsequent crowdsourcing tasks.

Therefore, we need to determine the threshold value for the "foodness" score for the decision of "keep/discard" for each image. We want to discard as many noisy images as possible to speed up the subsequent crowdsourcing task, while not discarding too many real food images that are good to be used later for training


food detections with p(food | box) >= 0.1
food 0.913
food detections with p(food | box) >= 0.1
food 0.427
food 0.158

(a) Real food object associated with high "foodness" score.

(b) Non-food object associated with low "foodness" score.

Fig. 6.10.: Comparison of "foodness" scores scores associated with food and non-food regions.

food classification networks. The best "foodness" score threshold should be obtained based on the characteristics of "foodness" scores' distribution.

6.4.2 Experimental Results for Automated Noisy Image Removal [2]

To obtain the most suitable threshold for the "foodness" score, we first build a dataset that contains both food images and non-food images downloaded from the Internet. We use 1,000 food images from 50 food categories and 1,000 non-food images in our experiment. We have manually verified all 2,000 images used in our experiment. In each food image there may be multiple regions proposed as food regions. We keep the detected food bounding box that has the highest "foodness" score in each image for the decision of "keep/discard".

It is unknown what is the ratio of non-food vs. food images (for example, 40% food and 60% non-food, or 50% food and 50% non-food) in the downloaded images. Therefore, we tested image mixtures which have 50 90% food images with our trained networks. For each mixture ratio, we sampled 1000 images and repeat the process for 1000 trials.

As we later will make the "keep/discard" decision based on the "foodness" score, we require the region proposal network to achieve high statistical accuracy. We examine the the Precision-Recall (PR) curve as shown in Figure 6.12, Receiver Operation Characteristics (ROC) curve as shown in Figure 6.11 and Average Precision as shown in Table 6.1.

Based on the results shown in Figure 6.12, Figure 6.11, and Table 6.1 we show that the trained region proposal network is accurate for food region proposal.

We obtain the most suitable "foodness" score threshold based on the trade-off between precision and recall [112] for different mixture ratios. A good "foodness" score threshold should result in both high precision and high recall. Based on the work [113], the minimum values we aim to achieve for both precision and recall

[2] This section is in joint work with Mr. Runyu Mao

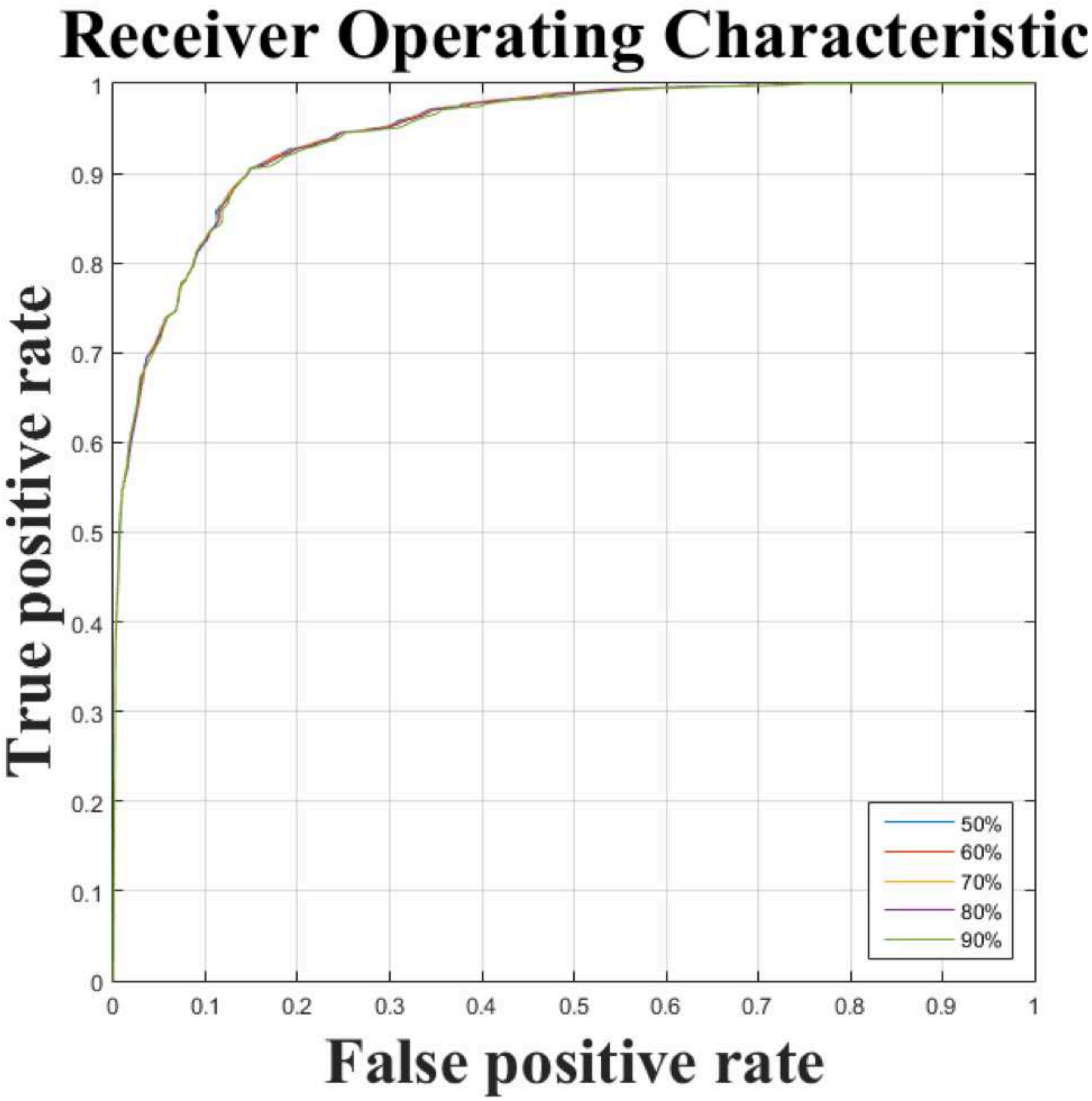


Fig. 6.11.: The Receiver Operating Characteristics (ROC) for different mixtures.

Table 6.1.: Average precision for each food image mixture ratio

Food Image Portion	Average Precision (AP)
50%	0.9473
60%	0.9625
70%	0.9746
80%	0.9843
90%	0.9923

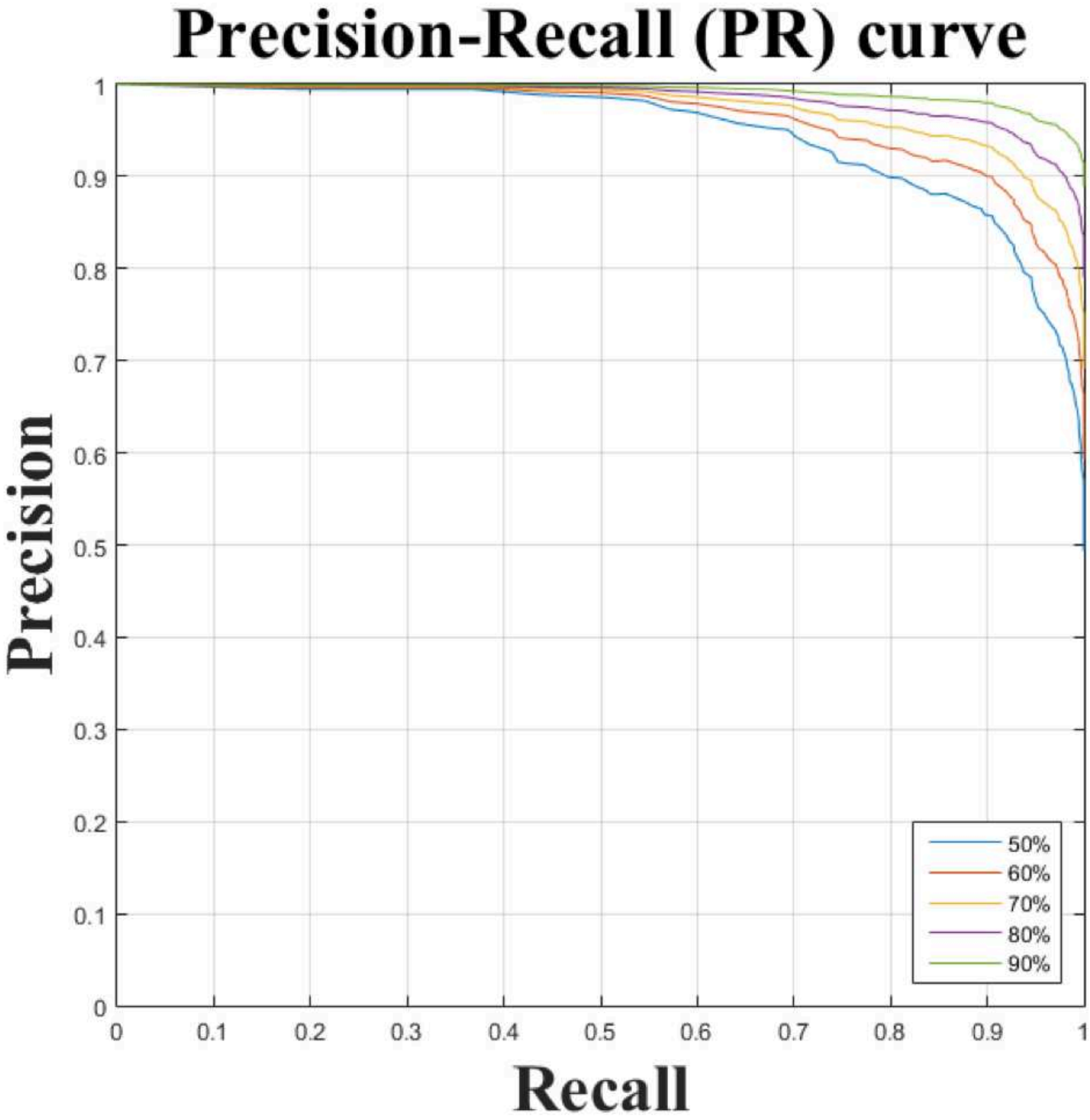


Fig. 6.12.: The Precision-Recall for different mixtures.

Table 6.2.: Acceptable Threshold Ranges and Discarded Images Portions for Different Food/Nonfood Mixtures

Food Images Portions	50%	60%	70%	80%	90%
Acceptable Threshold Ranges	$[0.57, 0.77]$	$[0.45, 0.77]$	$[0.32, 0.77]$	$[0.07, 0.77]$	$[0.00, 0.77]$
Discarded Images Portions	$[42.6\%, 55.3\%]$	$[28.5\%, 48.1\%]$	$[14.3\%, 41.1\%]$	$[1.3\%, 34.0\%]$	$[0.0\%, 26.8\%]$

are set to be 0.8. Since we want to build a large food image dataset, the networks may falsely discard no more than 20% correct images (recall ≥ 0.8). To improve the quality of subsequent crowdsourcing tasks, the remaining images should contain more than 80% correct images (precision ≥ 0.8).

"Foodness" score threshold values that meet our criteria of achieving high precision and high recall are considered to be within the acceptable range. A preliminary result of the acceptable ranges for different food/non-food image mixtures is shown in the Table 6.2, where a "keep/discard" action threshold within the range $[0.57, 0.77]$ can guarantee high precision and high recall while discarding a large portion of noisy images to improve the quality of the data collection.

More specifically, we highlight the acceptable range for each of the food mixtures as shown in Figure 6.13, Figure 6.14, Figure 6.15, Figure 6.16 and Figure 6.17.

6.5 Summary and Discussion

We have designed and implemented the cTADA$^{\text{TM}}$ crowdsourcing tool tailored for the task of incorporating online food images into our food image dataset. We

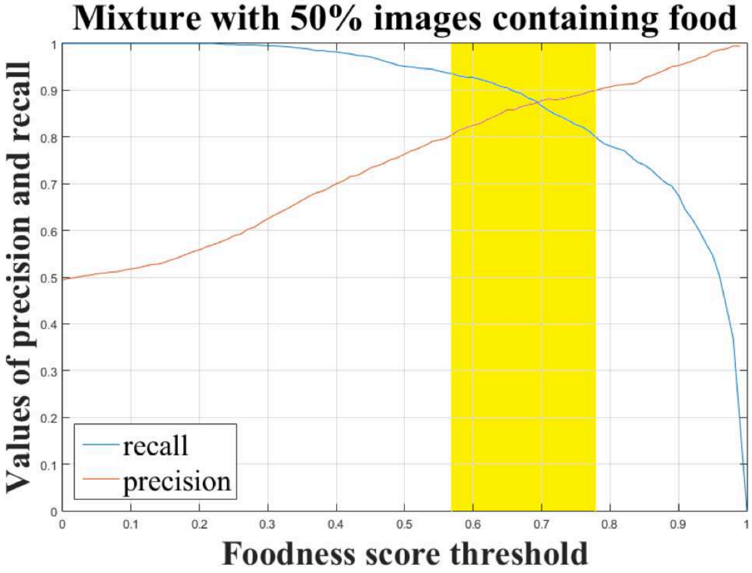


Fig. 6.13.: Acceptable range (highlighted in yellow) for mixture of 50% food images.

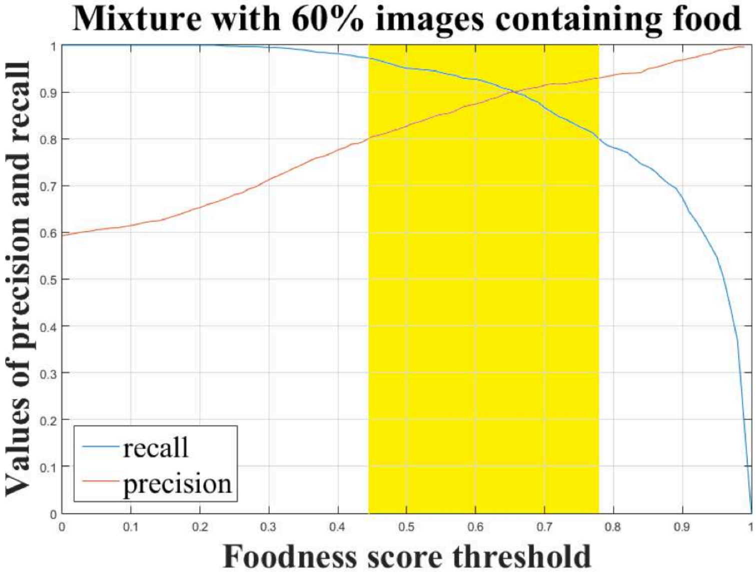


Fig. 6.14.: Acceptable range (highlighted in yellow) for mixture of 60% food images.

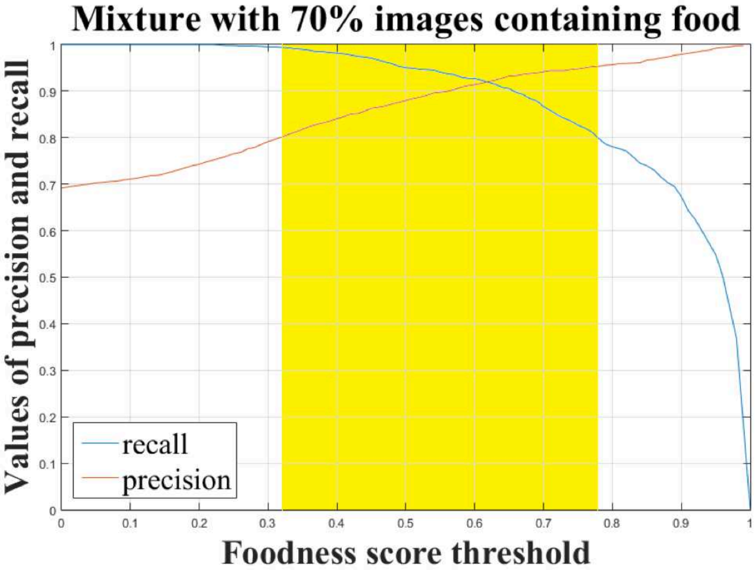


Fig. 6.15.: Acceptable range (highlighted in yellow) for mixture of 70% food images.

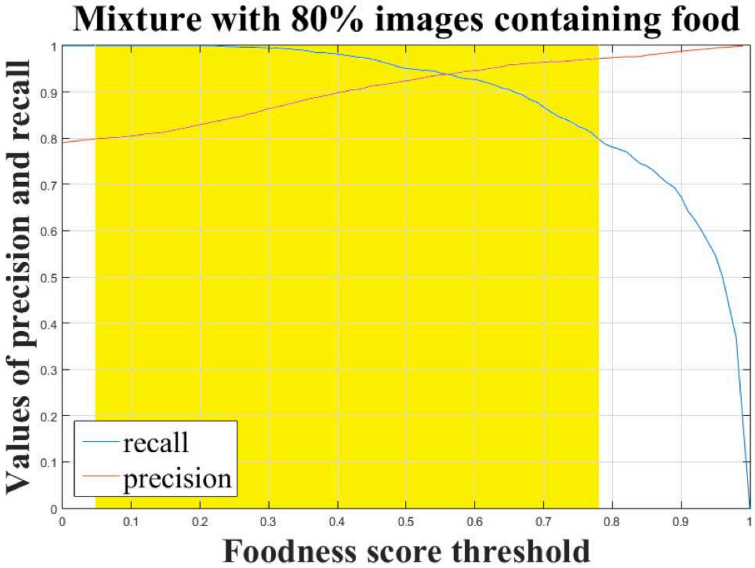


Fig. 6.16.: Acceptable range (highlighted in yellow) for mixture of 80% food images.

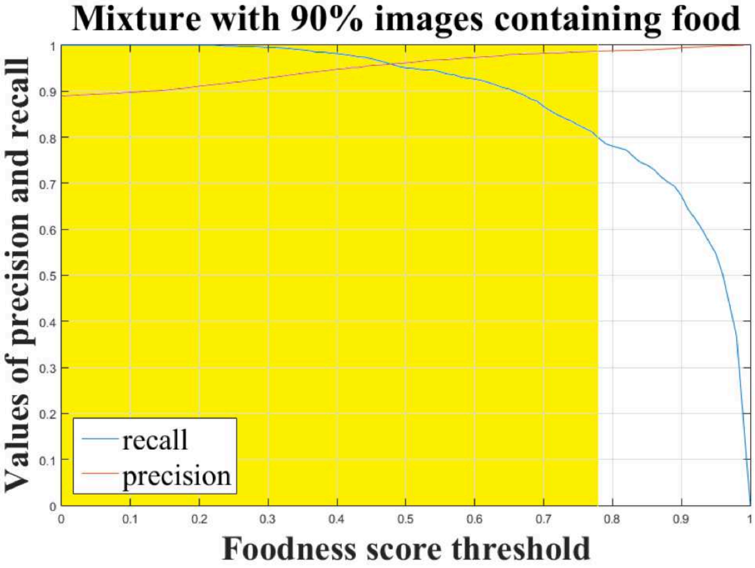


Fig. 6.17.: Acceptable range (highlighted in yellow) for mixture of 90% food images.

show that cTADA^{TM} is efficient and effective in removing noisy images, locating the bounding boxes containing the food items and obtaining segmentation masks associated with all the food items in the image. However, we have noticed some mistakes are made unwillingly by the users, especially for the noisy image removal step as each task is done on the scale of a few seconds. In order to minimize or avoid mistakes made unwillingly by the users in noisy image removal step, we developed a technique that automatically remove some of the noisy images.

We have gained valuable insights from our experiments on the design of cTADA^{TM} crowdsourcing tool for online food image collection. For example, which food tags to use as search entries and common appearances of food images taken by the users. Online food images introduce new perspectives as how we can collect and work on food images that are captured by users with no specific instructions. With the cTADA^{TM} tool, we are capable of expanding our food image dataset with online food images based on the food tags. We no longer have the issue of lacking training images for new food categories in our TADA^{TM} image analysis system.

In the future we are also interested in relating texts (e.g. recipes/comments on the same webpage) to food images as more nutrient or contextual information can be revealed and used. It still remains a challenging task to estimate valuable information from the large amount of image data generated by numerous users which can potentially contribute to research in the health and nutrient fields.

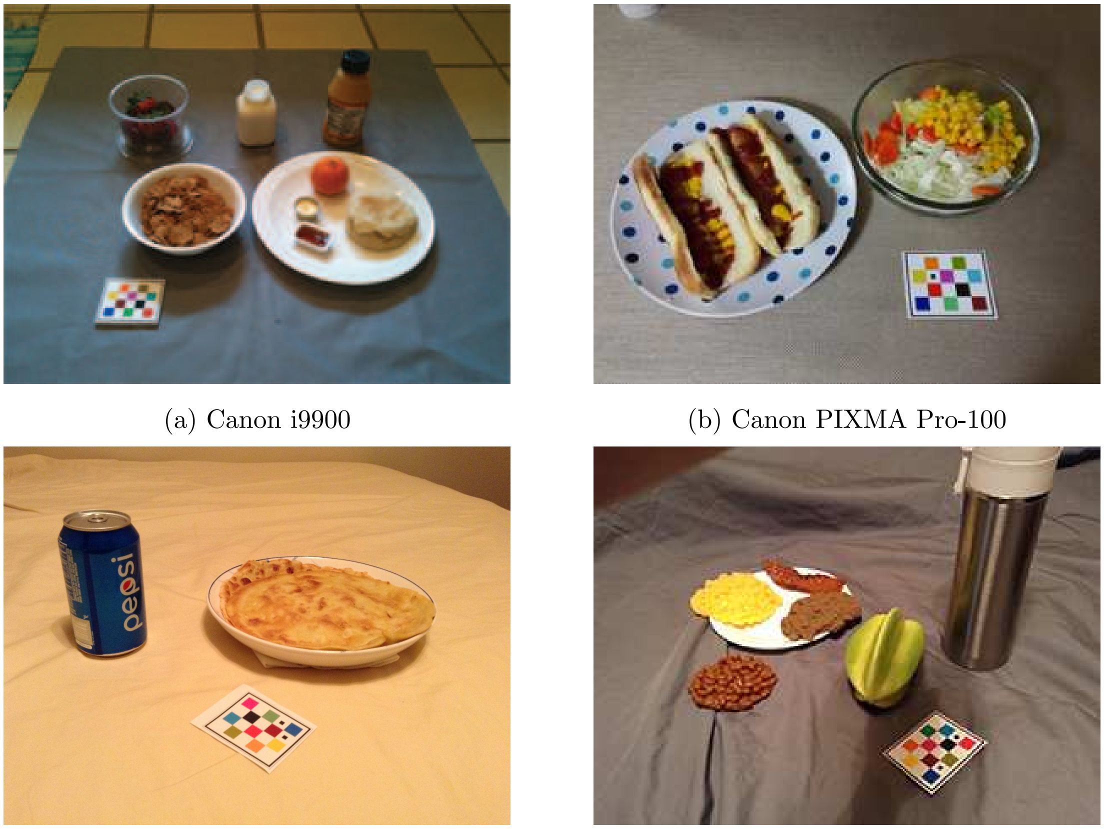
7. A PRINTER INDEXING SYSTEM FOR COLOR CALIBRATION WITH APPLICATIONS IN DIETARY ASSESSMENT

7.1 Introduction

Color is an important feature for identifying food types in current system, therefore it is crucial to maintain the consistency of color for accurate food classification [7]. A color calibration process based on reference information is required prior to food classification to eliminate the influences of varying lighting conditions and mitigate variations in camera sensor response.

To provide reference information, we have designed a color checkerboard pattern or a fiducial marker (FM) as illustrated in Figure 1.4. This color checkerboard consists of M colors where $M = 11$ including background “white” for our current version of the FM. The fiducial marker is included in the scene by the user to serve as a reference for the estimation of scale and pose of the objects in the scene and to provide reference information for color calibration [17]. Our research group has generated and distributed all the FMs used in our previous studies by printing the FMs on the same printer (a Canon i9900). As the number of users in our studies increases, we need to develop a method for the users to generate the FM themselves.

The issue of reproducing colors is a fundamentally difficult problem [114]. We have tested printing the FMs using various printers and significant color mismatch can be observed based on both the perception of a human observer (see Figure 7.3) and our estimates of the sRGB values [115]. Therefore we must design a system that allows us to know which printer was used to print the FM so that we can properly color calibrate the images. Our goal is to design the FM so that we can determine the printer by extracting the printer index from an image of the FM. Note that by



(a) Canon i9900

(b) Canon PIXMA Pro-100

(c) HP Color LaserJet 4700

(d) HP LaserJet M551

Fig. 7.1.: Food images from our TADA$^{\text{TM}}$ system with the fiducial marker (FM) present.

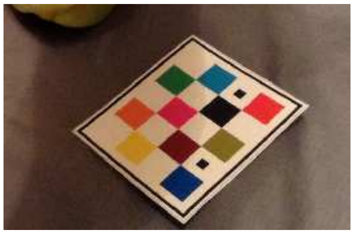


Fig. 7.2.: Magnified FM in Figure 7.1(d).

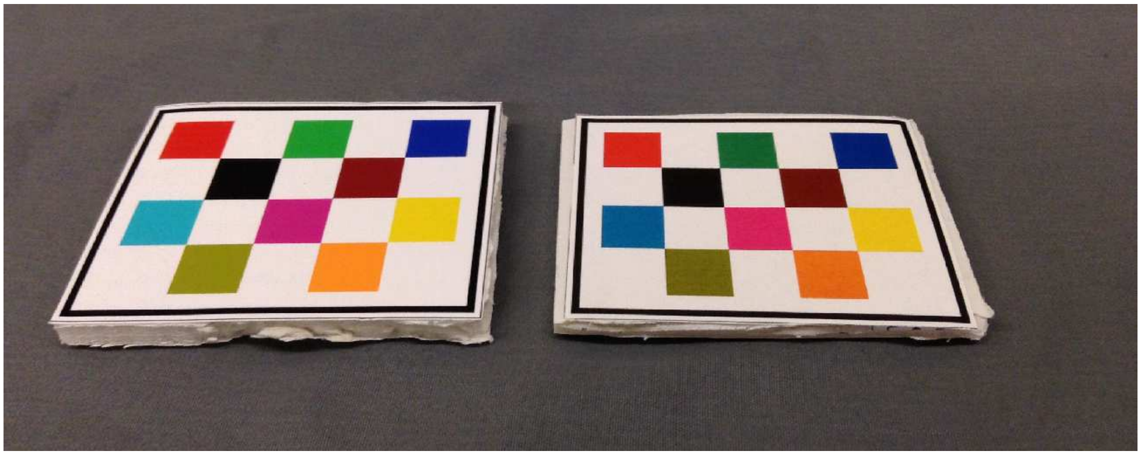


Fig. 7.3.: FM color differences using two printers.

“printer index” we mean a number that we can use to associate to a particular printer used to print the FM. We assume for this work that we know the color calibration matrix for the printer. We are in the process of planning a large study whereby the number of simultaneous users of the mFR^{TM} will be in the 100s. For this study we are designing a process where a user will be sent an FM as a digital file (e.g. a pdf file) with the indexing described in this work that assigns that FM to the particular user. They will be asked to print the FM and send the printed FM back to us for extraction of the color calibration matrix. In this work we are describing the FM indexing scheme and its relative robustness. One approach for printer indexing is to add a QR code containing the printer index in addition to the FM. However, this would require an additional step for the user to scan the QR code, thus increasing user burden [28]. Another approach we have used in the past for printer identification is based on texture features [116]. Unfortunately a texture-based printer identification technique will have issues with insufficient texture details in the FMs we use. We are interested in developing a method that will embed the index information in the fiducial marker without interfering with the color calibration and image analysis processes.

7.2 Color Correction and Printer Indexing

7.2.1 Color Correction

Color is one of the key variables in imaging [114, 115]. It is difficult to maintain color consistency due to illumination of the scene and camera settings such as auto exposure and auto white balance. Existing approaches attempt to increase the robustness of color descriptors based on features such as the RGB histogram, color moments and C-SIFT [117, 118]. Our approach for color correction is a linear RGB mapping based on the von-Kries model [119]. Before we can calibrate the colors in an image acquired with our mobile food record$^{\text{TM}}$ system (mFR$^{\text{TM}}$) one needs to calibrate the printer and determine its calibration matrix. This is done for a specific printer by printing the FM assigned to that printer and using a spectral radiometer to determine the sRGB estimates under CIE standard illuminant D65 [120].

A $M \times 3$ color reference matrix, where M is the number of checkerboard colors, is constructed and denoted as C_{ref}.

C_{ref} is constructed by assigning color sRGB estimates to the rows of C_{ref} based on the appearances of colors using a raster scan order, with the background color "white" last. The location associated with "red" square on the fiducial marker as indicated in Figure 7.3 will always be used as the starting point for raster scan order.

For color calibration, both the presence of the FM and the pixel coordinates of vertices for each color square in the FM need to be detected in the image we want to color correct.

In our system, the detection is done on a gray scale version of the image [121]. Once the pixel coordinates of the vertices of the color squares are obtained, the sRGB values for each of the M colors in FMs are then estimated by examining the pixels in each color square.

A color matrix for the image to be color corrected, denoted as C_{test}, is then constructed similarly to C_{ref}. Note that C_{test} is constructed using the lighting conditions in the scene and not the CIE standard illuminant D65.

Color calibration is conducted using C_{ref} based on linear least squares [17]. The color correction matrix $D \in \mathbb{R}^{3\times3}$ is:

$$\hat{D} = \underset{D\in\mathbb{R}^{3\times3}}{\operatorname{argmin}} \sum_{j=1}^{M} ||(\vec{C}_{ref_j})^t - D(\vec{C}_{test_j})^t||^2 \tag{7.1}$$

where $\hat{D}$ is the estimated color correction matrix, and $\vec{C}_{ref_j}, \vec{C}_{test_j} \in \mathbb{R}^{1\times3}$ are the j^{th} color, $j \in \{1, \cdots, M\}$. The image can be color corrected pixel by pixel as:

$$\vec{C}^t_{corrected} = \hat{D}\vec{C}^t_{original} \tag{7.2}$$

where $\vec{C}_{original} \in \mathbb{R}^{1\times3}$ is the original uncorrected sRGB values at any pixel location and $\vec{C}_{corrected} \in \mathbb{R}^{1\times3}$ is the color corrected result. Thus, equivalently for red, green and blue channels:

$$\begin{bmatrix} R_{corrected} \\ G_{corrected} \\ B_{corrected} \end{bmatrix} = \hat{D} \begin{bmatrix} R_{original} \\ G_{original} \\ B_{original} \end{bmatrix} \tag{7.3}$$

7.2.2 Printer Indexing System

We are interested in constructing an indexing system that allows us to identify from an image containing an FM which printer is used to print the FM. We designed the indexing system by associating each printer with a unique FM with different color squares arrangement. Rearranging the color squares on the checkerboard with no constraint yields a theoretical maximum of N_{max} permutations (or printers) where $N_{max} = 10! = 3,628,000$ and it is sufficient to address our needs. Denote i as the index for the i^{th} FM (or the i^{th} printer) and its corresponding color reference matrix is $C^{(i)}_{ref} \in \mathbb{R}^{M\times3}$.

C_{test} is the color matrix estimated from an image to be color corrected, we shall refer to this image as the test image. Denote the lighting condition as I, the conditional probability that the FM with assigned index i is in the test image is defined as:

$$p(C_{test}|C^{(i)}_{ref}, I) \tag{7.4}$$

We want to estimate the index $\hat{\imath}$ based on (7.4) such that:

$$\hat{\imath} = \underset{i \in \{1,\ldots,N\}}{\operatorname{argmax}} \{p(C_{test} | C_{ref}^{(i)}, I)\} \tag{7.5}$$

We will find $\hat{\imath}$ using normalized cross correlation (NCC).

Normalized cross correlation is a method for template or image matching [122, 123]. Our experimental results indicate that NCC can minimize the impact of the external lighting condition when estimating the printer index. The printer index is estimated using the NCC score between C_{test} and $C_{ref}^{(i)}$ where the NCC score, $f(C_{test}, C_{ref}^{(i)})$, is defined as:

$$f(C_{test}, C_{ref}^{(i)}) = \frac{1}{3 \cdot M - 1} \frac{\sum_{k=1}^{3 \cdot M} \{(\vec{C}_{test}(k) - \mu_{\vec{C}_{test}})(\vec{C}_{ref}^{(i)}(k) - \mu_{\vec{C}_{ref}^{(i)}})\}}{\sigma_{\vec{C}_{test}} \sigma_{\vec{C}_{ref}^{(i)}}} \tag{7.6}$$

where C_{test} and $C_{ref}^{(i)}$ are vectorized as $\vec{C}_{test}$, $\vec{C}_{ref}^{(i)} \in \mathbb{R}^{1 \times 3 \cdot M}$, $M = 11$, μ and σ^2 are mean and sample standard deviation, respectively. Based on the above definition, we have the estimated index as:

$$\hat{\imath} = \underset{i \in \{1,\ldots,N\}}{\operatorname{argmax}} \{f(C_{test}, C_{ref}^{(i)})\} \tag{7.7}$$

After we obtain the estimate $\hat{\imath}$, we use the reference information associated with this specific FM for color calibration as described in Section 7.2.1.

7.2.3 Error Control Using Binarized Marks

From our experiments, we observed that similar FM colors (such as red, orange and brown) may be very difficult to differentiate under certain lighting conditions (e.g. a dim restaurant) or due to poor printing quality. For these similar colors that are likely to cause incorrect index decisions using NCC, we define “similar colors” sets. For example, red, orange and brown can form a “similar colors” set. If similar colors can not be differentiated, the assumption that each FM has a unique arrangement of color squares can no longer hold true. To address this issue, we propose the use of a "binarized mark" that we can add to the FM to serve as an error control method in addition to NCC. Binarized marks are combinations of small black squares placed at the center of one or more white squares as illustrated in Figure 7.2.

A numeric value can be generated from the binarized mark based on the detection of the black square. Following raster scan order begins at the second white square in the FM, each subsequent white square represents a "bit" in the binary sequence starting from the least significant bit. A bit is assigned a "1" when a black square is present and "0" otherwise. The corresponding numeric value in decimal can be obtained by converting the binary sequence. For the FM shown in Figure 7.2, the binary sequence is "000000011" and accordingly the numeric value in decimal is "3". Since the length of the binary sequence is 9, only $2^9 = 512$ binarized marks can be generated. However, the theoretical maximum number of printers we can index is $N_{max} = 3,628,800$. Since we "assign" a binarized mark to every FM, we will quickly run out of binarized marks without having identical binarized marks assigned to each FM. Since we cannot assign a unique binarized mark to each FM, we need to define a criteria for assigning binarized marks to the FMs.

We define a threshold T to activate the error control. For a given printer index $\hat{i}$ obtained from (7.7), if there is no other $i \in \{1, \cdots, N\}$ and $i \neq \hat{i}$ such that the NCC score defined in (7.6) satisfies the following:

$$f(C_{test}, C_{ref}^{(\hat{i})}) - f(C_{test}, C_{ref}^{(i)}) < T \tag{7.8}$$

we can safely assume that using the NCC score is sufficient for indexing the printer, hence no binarized marks are needed in this case. We set $T = 0.01$ based on our experimental results. Otherwise, error control method is activated and printer index can be corrected based on the detection of the binarized mark. As a result, we only need to assign a unique binarized mark to each FM that meets the criteria of (7.8). Based on our experimental results, we have observed that the number of FMs with "similar colors" swapped generally do not exceed the maximum number of binarized marks that can be generated, which is 512. For example, in the case where similar

FM colors such as red, orange and brown are swapped, we only need $3! = 6$ binarized marks to guarantee the correct printer indexing.

7.3 Experimental Results

The initial evaluation of our printer indexing system is based on FMs we printed using several printers. Our test images contain various foods images with different FMs taken under several lighting conditions. There is no other information in the test images that can indicate which FM is used in a specific test image. After extracting the color information from the FMs and the binarized marks, we estimated the index i using the methods described in Section 7.2.2. The indexing is estimated primarily using the NCC score. Binarzied marks will be used only when error control is activated as described in (7.8). The ground truth is obtained by a human observer examining the arrangement of color squares. The accuracy of indexing decisions can then be obtained by comparing to the ground truth.

We are interested in testing the NCC-based method for a variety of FMs. To conduct such a test we generated 9 FMs, where 1 had no color swapped (original FM with the following colors in raster scan order: red, green, blue, black, brown, cyan, magenta, yellow, dark green and orange), 7 had two colors swapped compared to the original one (red and brown, green and dark green, green and yellow, red and magenta, yellow and orange, blue and cyan, red and orange), and 1 had three colors swapped compared to the original one (green, yellow and dark green). Figure 7.1(b) and (c) show examples of images with FMs that have color swapped and binarized marks.

We have obtained 579 test images with different lighting conditions using 9 models of laser and inkjet printers. These lighting conditions include incandescent and fluorescent lightings with various luminance, sunlight, shadows and more complex lighting conditions in the restaurants. The index decisions can be made accurately as reflected by the average NCC scores illustrated in Figure 7.4. The average NCC

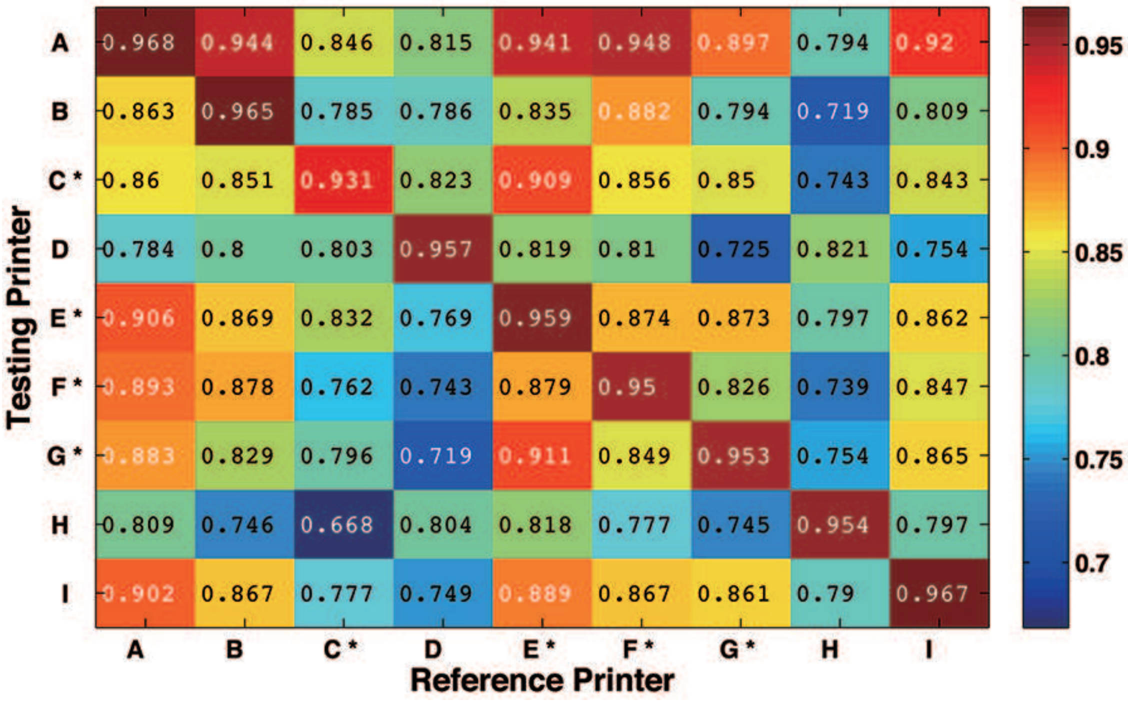


Fig. 7.4.: Accuracy of estimated printer index based on average NCC scores from 9 printers, with each printer associated with a unique FM as listed below: (A) Canon i9900 (original FM), (B) Canon PIXMA Pro-100 (red and brown swapped), (C*) HP LaserJet M551 (green and dark green swapped), (D) Canon PIXMA Pro-10 (green and yellow swapped), (E*) HP Color LaserJet 4700 (red and magenta swapped), (F*) Ricoh Aficio MP C6501 (yellow and orange swapped), (G*) TOSHIBA e-STUDIO 3530c (blue and cyan swapped), (H) Epson WF3540 (green, yellow and dark green swapped), (I) HP D110a (red and orange swapped)

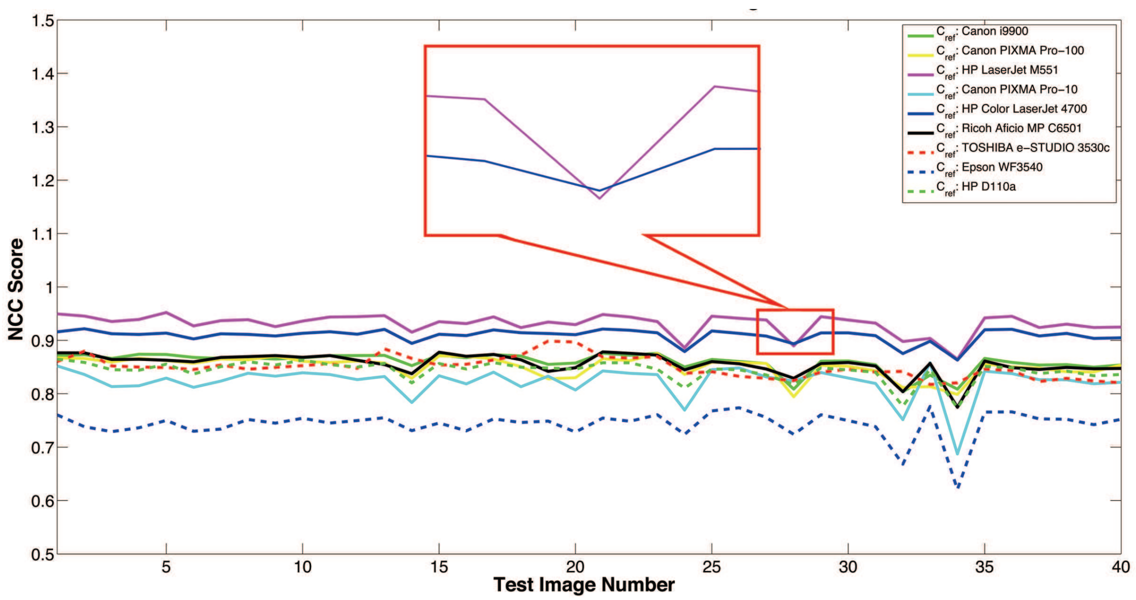


Fig. 7.5.: A test set of 40 images from 9 printers. The two lines inside the red rectangular area show an example where a wrong printer indexing decision is made based on NCC scores alone.

scores for each FM are obtained from test images containing the same FM. Note that printers with "*" in Figure 7.4 are laser printers.

We also test the accuracy of printer indexing based on NCC scores for each test image from a subset of the 579 test images used above. This subset of test images contains 40 images with the same FM printed from a HP LaserJet M551 printer. For this particular example, color squares green and dark green are swapped compared to the original FM (no color swapped). A wrong indexing decision is made based on NCC scores alone as shown in the zoomed in red rectangular area of Figure 7.5. The printer is indexed to HP Color LaserJet 4700 instead of HP LaserJet M551 without activating error control for one of the test images. The test image that generates the wrong indexing decision is shown in Figure 7.1(c), with a zoomed in image of the fiducial marker shown in Figure 7.2. However, the criteria for activating the error control method as defined in (7.8) is satisfied for this test image. Therefore, by detecting the binarized marks, the correct indexing decision can be made.

7.4 Summary and Discussion

We have described a printer indexing system for use in color image correction. We show that the printer index can be accurately estimated. Our experimental results show this scheme is robust against most types of lighting conditions.

8. SUMMARY AND FUTURE WORK

8.1 Summary

In this thesis we first investigate the use of geometric models for food portion estimation based on single-view eating occasion images. We focused primarily on cylinder model and prism model. The food portions are estimated in volumes (cm^3). We were able to obtain accurate estimates of food portions based on well-defined 3D models, camera calibration objects, correct food labels and correct food segmentation masks. We then compared the accuracy of different food portion estimation techniques using geometric models and using depth image. We show that portion estimation based on geometric models is more accurate for objects with well-defined 3D shapes compared to estimation using depth images. To further improve food portion estimation accuracy, we use co-occurrence patterns as prior knowledge to refine portion estimation results. In addition to food portion estimation approaches based on geometry computer vision, we developed another approach based on the use of Generative Adversarial Networks (GAN). We introduce the concept of an "energy distribution" for each food image. We can then estimate food energy based on the energy distribution.

Other than food portion estimation, we have also present a systematic design for a crowdsourcing tool aiming specifically for the task of online food image collection and annotations. Our goal is to fast expand food image dataset and incorporate online food image into our dataset for training-based food classification techniques. IN addition, we have also developed a printer indexing system for color calibration with an application in image-based dietary assessment.

The main contributions of this thesis are listed as follows:

- Single-View Food Portion Estimation Based on Geometric Models

We developed a food portion estimation technique based on a single-view food image used for the estimation of the amount of energy (in kilocalories) consumed at a meal. Although single-view 3D scene reconstruction is in general an ill-posed problem, the use of geometric models such as the shape of a container can help to partially recover 3D parameters of food items in the scene. Based on the estimated 3D parameters of each food item and a reference object in the scene, the volume of each food item in the image can be determined. We focused primarily on the use of cylinder model and prism model. The food portions are estimated in volumes (cm^3). Unlike previous methods, our technique is capable of estimating food portion without manual tuning of parameters. The weight of each food can then be estimated using the density of the food item. We were able to achieve an error of less than 6% for energy estimation of an image of a meal assuming accurate segmentation and food classification.

- A Comparison of Food Portion Estimation Using Geometric Models and Depth Images

  We compare two techniques of estimating food portion size from images of food. The techniques are based on 3D geometric models and depth images. An expectation-maximization based technique is developed to detect the reference plane in depth images, which is essential for portion size estimation using depth images. We compare the accuracy of food portion estimation based on geometric-model, to the accuracy estimated based on high quality depth map obtained using structured light techniques. Our experimental results indicate that volume estimation based on geometric models is more accurate for objects with well-defined 3D shapes compared to estimation using depth images.

- The Use of Co-occurrence Patterns in Single Image Based Food Portion Estimation

  We use contextual information to further improve food portion estimation accuracy using geometric models based approach. We define contextual dietary information as the data that is not directly produced by the visual appearance of an object in the image, but yields information about a user's diet or can be used for diet planning. Therefore, food portion co-occurrence pattern is one type of contextual information. We estimate the patterns from food images we collected for dietary studies. We estimate the food portion co-occurrence patterns from food images we collected from dietary studies using the mobile Food Record$^{\text{TM}}$ (mFR$^{\text{TM}}$) system we developed. Co-occurrence patterns is used as prior knowledge to refine portion estimation results. We were able to improve the food portion estimation accuracy incorporating the co-occurrence patterns as contextual information.

- Learning Image-to-Energy Mappings Using Generative Adversarial Networks

  Accurate food portion estimation is challenging since the process of food preparation and consumption impose large variations on food shapes and appearances. In addition to our previous approach of geometric models based food portion estimation, we present a food portion estimation method to estimate food energy (kilocalories) from food images using Generative Adversarial Networks (GAN). We introduce the concept of an "energy distribution" for each food image. To train the GAN, we design a food image dataset based on ground truth food labels and segmentation masks for each food image as well as energy information associated with the food image. Our goal is to learn the mapping from the food image to the food energy. We can then estimate food energy based on the estimated energy distribution image.

- An End-to-end Image-Based Automatic Food Energy Estimation Technique Based on Learned Energy Distribution Images

  We proposed a novel end-to-end system to directly estimate food energy from a captured eating occasion image. Our system first estimated the image to energy mappings using a Generative Adversarial Networks (GAN) structure. Based on the estimated energy distribution images, we learned the food energy of the eating occasion image by training CNN based regression model. We are able to obtain accurate food energy estimation with an average error of 209.41 kilocalories for eating occasion images collected from a free-living dietary study. The training based technique for end-to-end food energy estimation no longer requires fitting geometric models onto the food objects that may have issues scaling up as we need a large amounts of geometric models to fit different food types in many food images.

- cTADA$^{\text{TM}}$: The Design of a Crowdsourcing Tool for Online Food Image Identification and Segmentation

  Training-based techniques have been widely used in recent years for developing automatic dietary assessment systems. For training-based techniques, increasing the training data size would in general improve the accuracy of the system, thus a larger image dataset is always preferred. Online image sharing is quickly gaining popularity in recent years (for example, through social networks such as Facebook and review orientated websites such as Yelp), and there are hundred-thousands of food images uploaded by smartphone users. We believe online food images can be used as part of our training data developing automatic dietary assessment techniques and provide valuable contextual information such as users' dietary patterns and food co-occurrence patterns. We present a systematic design for a crowdsourcing tool aiming specifically for the task of online food image collection and annotations with a detailed description. This tool can be used to locate food items and obtaining groundtruth segmentation masks associated with all the foods presented in an image. The crowdsoucing tool we designed is tailored to meet the needs of building a large image dataset for developing automatic dietary assessment tools in the nutrition and health fields.

- A Printer Indexing System for Color Calibration

  In image based dietary assessment, color is a very important feature in food identification. One issue with using color in image analysis is the calibration of the color imaging capture system. We have designed a reference indexing system for color camera calibration using printed color checkerboards also known as fiducial markers (FMs). To use the FM for color calibration one must know which printer was used to print the FM so that the correct color calibration matrix can be used for calibration. We have designed an indexing scheme that allows one to determine which printer was used to print the FM based on a unique arrangement of color squares and binarized marks (used for error control) on the FM. Using normalized cross correlation and pattern detection, the index corresponding to the printer for a particular FM can be determined. We show the printer indexing scheme we developed is robust against most types of lighting conditions.

8.2 Future Work

We are able to estimate the food energy value using a single-view food image as input. Currently we do not have models in the energy estimation system that classify food types and segment food areas. With models that can accurately classify food types and segment food regions, the food label and region information could potentially be combined with the food energy distribution estimation. In addition, with the development of camera sensors on mobile, additional image information (for example, depth) may become available without increasing a user's burden capturing the eating scene. The future work is as followed:

- Depth sensor and dual camera configuration are quickly gaining popularity on consumer mobile devices. More 3D information can be collected without significantly adding to a user's burden capturing the eating scene. For example, if a mobile phone is equipped with a depth sensor then the depth image along with the RGB image can be capture simultaneously. For dual camera systems, at least two images are captured from slightly different angles and therefore enables multi-view 3D reconstruction techniques such as stereo vision. The additional information captured by the mobile devices could potentially improve the accuracy of food portion estimation by providing additional 3D information on food objects.

- Currently, we use synthetic energy distribution images for the training of energy distribution mappings. We crafted the synthetic energy distribution images with self-defined functions. In the future, we would like to incorporate real 3D features in the synthetic energy distribution image. For example if depth image could be available for the eating scene, food objects' 3D models can be reconstructed. By incorporating more 3D information the accuracy of the synthetic energy distribution image will increase. Therefore, a more accurate mappings of image to energy distribution can be subsequently learned.

- We are interested in developing accurate techniques for food type classification. Existing techniques based on Convolutional Neural Networks (CNN) have shown impressive tasks on object detection and classification tasks. However, the lack of training data for food categories has become the bottle neck for object detection/classification task where a large public available dataset is not available. In the past we rely on dietary studies to obtain food images used for training. In this thesis we have developed the crowdsourcing tool that enabled us to quickly obtain online food images. With the crowdsouring tool, we no longer have the issue of expanding food image dataset. However, even with the online food images added to the training dataset, the size of training dataset may still not be sufficient for certain food classes using flat structure classifier (classify all categories all at once). Therefore, we are also interested in the use of a hierarchical structure for food classification.

8.3 Publications Resulting From This Work

Journal Papers

1. **S. Fang**, Z. Shao, D. Kerr, C. Boushey, and F. Zhu, "An End-to-end Image-Based Automatic Food Energy Estimation Technique Based on Learned Energy Distribution Images", *To be submitted to Nutrients Special Issue on Advancement in Dietary Assessment and Self-Monitoring Using Technology.*

Conference Papers

1. **S. Fang**, Z. Shao, R. Mao, C. Fu, D. Kerr, C. Boushey, E. Delp and F. Zhu, "Single-View Food Portion Estimation: Learning Image-to-Energy Mappings Using Generative Adversarial Networks", *Proceedings of the IEEE International Conference on Image Processing*, Athens, Greece, to appear.

2. **S. Fang**, S. Yarlagadda, Y. Wang, F. Zhu, C. Boushey, D. Kerr and E. Delp, "Image Based Dietary Behavior and Analysis Using Deep Learning", *Proceedings of the International Conference of the IEEE Engineering in Medicine and Biology Society*, July 2018, Honolulu, HI

3. **S. Fang**, C. Liu, K. Tahboub, F. Zhu, C. Boushey and E. Delp, "cTADA: The Design of a Crowdsourcing Tool for Online Food Image Identification and Segmentation", *Proceedings of the IEEE Southwest Symposium on Image Analysis and Interpretation*, April 2018, Las Vegas, NV.

4. **S. Fang**, F. Zhu, C. Boushey and E. Delp, "The Use of Co-occurrence Patterns in Single Image Based Food Portion Estimation", *Proceedings of the IEEE Global Conference on Signal and Information Processing*, pp. 462-466, November 2017, Montreal, Canada.

5. Y. Wang, **S. Fang**, C. Liu, F. Zhu, D. Kerr, C. Boushey and E. Delp, “Food Image Analysis: The Big Data Problem You Can Eat!”, *Proceedings of the IEEE International Conference on Image Processing*, pp. 1263-1267, November 2016, Pacific Grove, CA.

6. **S. Fang**, F. Zhu, C. Jiang, S. Zhang, C. Boushey and E. Delp, “A Comparison of Food Portion Estimation Using Geometric Models and Depth Images”, *Proceedings of the IEEE International Conference on Image Processing*, pp. 26-30, September 2016, Phoenix, AZ.

7. **S. Fang**, C. Liu, F. Zhu, C. Boushey and E. Delp, “Single-View Food Portion Estimation Based on Geometric Models”, *Proceedings of the IEEE International Symposium on Multimedia*, pp. 385-390, December 2015, Miami, FL.

8. **S. Fang**, C. Liu, F. Zhu, C. Boushey and E. Delp, “A Printer Indexing System for Color Calibration with Applications in Dietary Assessment”, *New Trends in Image Analysis and Processing – ICIAP 2015 Workshops, Lecture Notes in Computer Science*, , Vol. 9281, Springer International, pp. 358-365, 2015.

REFERENCES

REFERENCES

[1] B. Six, T. Schap, F. Zhu, A. Mariappan, M. Bosch, E. Delp, D. Ebert, D. Kerr, and C. Boushey, “Evidence-based development of a mobile telephone food record,” *Journal of the American Dietetic Association*, vol. 110, no. 1, pp. 74–79, January 2010. [Online]. Available: http://dx.doi.org/10.1016/j.jada.2009.10.010

[2] T. Aflague, C. Boushey, R. Guerrero, Z. Ahmad, D. Kerr, and E. Delp, “Feasibility and use of the mobile food record for capturing eating occasions among children ages 3–10 years in guam,” *Nutrients*, vol. 7, no. 6, pp. 4403–4415, 2015. [Online]. Available: http://www.mdpi.com/2072-6643/7/6/4403

[3] “Smartphone ownership and internet usage continues to climb in emerging economies,” Pew Research Center. [Online]. Available: http://www.pewglobal.org/2016/02/22/smartphone-ownership-and-internet-usage-continues-to-climb-in-emerging-economies/

[4] C. Boushey, D. Kerr, J. Wright, K. Lutes, D. Ebert, and E. Delp, “Use of technology in children’s dietary assessment,” *European Journal of Clinical Nutrition*, vol. 63, pp. S50–S57, February 2009. [Online]. Available: http://dx.doi.org/10.1038/ejcn.2008.65

[5] T. Schap and F. Z. E. Delp, “Merging dietary assessment with the adolescent lifestyle,” *Journal of Human Nutritino and Dietetics*, vol. 27, no. s1, pp. 82–88, January 2014. [Online]. Available: http://dx.doi.org/10.1111/jhn.12071

[6] F. Zhu, M. Bosch, I. Woo, S. Kim, C. Boushey, D. Ebert, and E. J. Delp, “The use of mobile devices in aiding dietary assessment and evaluation,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 4, pp. 756 –766, August 2010. [Online]. Available: http://dx.doi.org/10.1109/JSTSP.2010.2051471

[7] F. Zhu, M. Bosch, N. Khanna, C. Boushey, and E. Delp, “Multiple hypotheses image segmentation and classification with application to dietary assessment,” *IEEE Journal of Biomedical and Health Informatics*, vol. 19, no. 1, pp. 377–388, January 2015. [Online]. Available: http://10.1109/JBHI.2014.2304925

[8] K. Kitamura, T. Yamasaki, and K. Aizawa, “Foodlog: Capture, analysis and retrieval of personal food images via web,” *Proceedings of the ACM multimedia workshop on Multimedia for cooking and eating activities*, pp. 23–30, November 2009, Beijing, China. [Online]. Available: http://doi.dx.org/0.1145/1630995.1631001

[9] T. Joutou and K. Yanai, “A food image recognition system with multiple kernel learning,” *Proceedings of the IEEE International Conference on Image Processing*, pp. 285–288, October 2009, Cairo, Egypt. [Online]. Available: http://dx.doi.org/10.1109/ICIP.2009.5413400

[10] F. Kong and J. Tan, “Dietcam: Automatic dietary assessment with mobile camera phones,” *Pervasive and Mobile Computing*, vol. 8, pp. 147–163, February 2012. [Online]. Available: http://dx.doi.org/10.1016/j.pmcj.2011.07.003

[11] A. Meyers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. P. Murphy, “Im2calories: Towards an automated mobile vision food diary,” *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1233–1241, December 2015, Santiago, Chile. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.146

[12] W. Min, S. Jiang, L. Liu, Y. Rui, and R. Jain, “A survey on food computing,” *arXiv:1808.07202*.

[13] Y. He, C. Xu, N. Khanna, C. Boushey, and E. Delp, “Analysis of food images: Features and classification,” *Proceedings of the IEEE International Conference on Image Processing*, pp. 2744–2748, October 2014, Paris, France. [Online]. Available: http://dx.doi.org/10.1109/ICIP.2014.7025555

[14] S. Fang, C. Liu, F. Zhu, E. Delp, and C. Boushey, “Single-view food portion estimation based on geometric models,” *Proceedings of the IEEE International Symposium on Multimedia*, pp. 385–390, December 2015, Miami, FL. [Online]. Available: http://dx.doi.org/10.1109/ISM.2015.67

[15] “The TADA project.” [Online]. Available: http://tadaproject.org

[16] Z. Ahmad, M. Bosch, N. Khanna, D. A. Kerr, C. J. Boushey, F. Zhu, and E. J. Delp, “A mobile food record for integrated dietary assessment,” *Proceedings of the 2nd International Workshop on Multimedia Assisted Dietary Management*, pp. 53–62, October 2016, amsterdam, Netherlands. [Online]. Available: http://dx.doi.org/10.1145/2986035.2986038

[17] C. Xu, F. Zhu, N. Khanna, C. J. Boushey, and E. J. Delp, “Image enhancement and quality measures for dietary assessment using mobile devices,” *Proceedings of the IS&T/SPIE Conference on Computational Imaging X*, pp. 82 960Q–1–82 960Q–10, January 2012, San Francisco, CA. [Online]. Available: http://dx.doi.org/10.1117/12.909949

[18] S. Kelkar, S. Stella, C. Boushey, and M. Okos, “Developing novel 3D measurement techniques and prediction method for food density determination,” *Procedia Food Science*, vol. 1, pp. 483 – 491, May 2011. [Online]. Available: http://dx.doi.org/10.1016/j.profoo.2011.09.074

[19] “USDA food and nutrient database for dietary studies, 3.0.” Beltsville, MD: Agricultural Research Service, Food Surveys Research Group, 2008.

[20] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004.

[21] A. Saxena, M. Sun, and A. Ng, “Make3D: Learning 3d scene structure from a single still image,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 824 – 840, May 2009. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2008.132

[22] Z. Zhang, "Determine the epipolar geometry and its uncertainty: A review," *International Journal of Computer Vision*, vol. 27, no. 2, pp. 161 – 195, March 1998. [Online]. Available: http://dx.doi.org/10.1023/A:1007941100561

[23] D. Marr and T. Poggio, "A computational theory of human stereo vision," *Proceedings of the Royal Society of London. Series B, Biological Sciences*, vol. 204, no. 1156, pp. 301 – 328, May 1979. [Online]. Available: http://dx.doi.org/10.1098/rspb.1979.0029

[24] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," *Proceedings of the International Joint Conference on Artificial Intelligence*, pp. 674 – 679, August 1981, Vancouver, Canada. [Online]. Available: http://dx.doi.org/10.1002/scj.4690211209

[25] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *International Journal of Computer Vision*, vol. 47, no. 1-3, pp. 7–42, April 2002. [Online]. Available: http://dx.doi.org/10.1109/SMBV.2001.988771

[26] C. Loop and Z. Zhang, "Computing rectifying homographies for stereo vision," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 125 – 131, June 1999, Fort Collins, CO. [Online]. Available: http://dx.doi.org/10.1109/CVPR.1999.786928

[27] D. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*, 2003.

[28] B. L. Daugherty, T. E. Schap, R. Ettienne-Gittens, F. Zhu, M. Bosch, E. J. Delp, D. S. Ebert, D. A. Kerr, and C. J. Boushey, "Novel technologies for assessing dietary intake: evaluating the usability of a mobile telephone food record among adults and adolescents," *Journal of Medical Internet Research*, vol. 14, no. 2, p. e58, April 2012. [Online]. Available: http://dx.doi.org/10.2196/jmir.1967

[29] C. Foshee, "Goal-driven three-dimensional object inspection from limited view backprojection reconstruction," *Ph.D. Dissertation*, December 1991, purdue University, West Lafayette, IN, USA.

[30] J. Geng, "Structured-light 3d surface imaging: a tutorial," *Advances in Optics and Photonics*, vol. 3, no. 2, pp. 128–160, June 2011.

[31] D. Liebowitz, A. Criminisi, and A. Zisserman, "Creating architectural models from images," *Proc. EuroGraphics (EG)*, vol. 18, p. 3950, September 1999. [Online]. Available: http://dx.doi.org/10.1111/1467-8659.00326

[32] D. Hoiem, A. Efros, and M. Hebert, "Geometric context from a single image," *Proceedings of the IEEE International Conference on Computer Vision*, vol. 1, pp. 654 – 661, October 2005, Beijing, China. [Online]. Available: gttp://dx.doi.org/10.1109/ICCV.2005.107

[33] ——, "Automatic photo pop-up," *ACM Transactions on Graphics - Proceedings of ACM SIGGRAPH*, vol. 24, no. 3, pp. 577 – 584, July 2005. [Online]. Available: http://dx.doi.org/10.1145/1186822.1073232

[34] E. Delage, H. Lee, and A. Ng, "A dynamic bayesian network model for autonomous 3d reconstruction from a single indoor image," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 2418 – 2428, June 2006, New York, NY. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2006.23

[35] A. Saxena, S. Chung, and A. Ng, "Learning depth from single monocular images," *Advances in Neural Information Processing Systems*, pp. 1161 – 1168, 2006. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2015.2505283

[36] ——, "3-D depth reconstruction from a single still image," *International Journal of Computer Vision*, vol. 76, no. 1, pp. 53 – 69, 2008.

[37] A. Saxena, J. Schulte, and A. Y. Ng, "Depth estimation using monocular and stereo cues," *Proceedings of the International Joint Conference on Artifical Intelligence*, pp. 2197–2203, 2007, Hyderabad, India.

[38] S. Kim, S. Choi, and K. Sohn, "Learning depth from a single image using visual-depth words," *Proceedings of the IEEE International Conference on Image Processing*, pp. 1895 – 1899, September 2015, Quebec City, Canada. [Online]. Available: http://dx.doi.org/10.1109/ICIP.2015.7351130

[39] X. Liu, H. Qin, Y. Wang, Y. Zhang, and Q. Dai, "DEPT: Depth estimation by parameter transfer for single still image," *Proceedings of the Asian Conference on Computer Vision*, pp. 45 – 58, November 2014, Singapore. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-16808-1_4

[40] B. Li, C. Shen, Y. Dai, A. van den Hengel, and M. He, "Depth and surface normal estimation from monocular images using regression on deep features and hierarchical CRFs," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1119 – 1127, June 2015, Boston, MA. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2015.7298715

[41] F. Liu, C. Shen, G. Lin, and I. Reid, "Learning depth from single monocular images using deep convolutional neural fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2015.2505283

[42] E. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," *Advances in Neural Information Processing Systems*, 2014.

[43] D. Eigen and R. Fergus, "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture," *Proceedings of the IEEE International Conference on Computer Vision*, December 2015, Santiago, Chile. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.304

[44] M. Puri, Z. Zhu, Q. Yu, A. Divakaran, and H. Sawhney, "Recognition and volume estimation of food intake using a mobile device," *Proceedings of the IEEE Workshop on Applications of Computer Vision*, pp. 1–8, December 2009, Snowbird, UT. [Online]. Available: http://dx.doi.org/10.1109/WACV.2009.5403087

[45] J. Dehais, S. Shevchik, P. Diem, and S. Mougiakakou, “Food volume computation for self dietary assessment applications,” *Proceedisng of the IEEE International Conference on Bioinformatics and Bioengineering*, pp. 1–4, November 2013, Chania, Greece.

[46] M. Sun, J. Fernstrom, W. Jia, S. Hackworth, N. Yao, Y. Li, C. Li, M. Fernstrom, and R. Sclabassi, “A wearable electronic system for objective dietary assessment,” *Journal of the American Dietetic Association*, p. 110(1): 45, January 2010. [Online]. Available: http://dx.doi.org/10.1016/j.jada.2009.10.013

[47] J. Shang, M. Duong, E. Pepin, X. Zhang, K. Sandara-Rajan, A. Mamishev, and A. Kristal, “A mobile structured light system for food volume estimation,” *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pp. 100–101, November 2011, Barcelona, Spain. [Online]. Available: http://dx.doi.org/10.1364/AOP.3.000128

[48] F. Kong and J. Tan, “Dietcam: regular shape food recognition with a camera phone,” *Proceedings of the IEEE International Conference on Body Sensor Networks*, pp. 127–132, May 2011, dallas, TX. [Online]. Available: http://dx.doi.org/10.1109/BSN.2011.19

[49] F. Kong, H. He, H. A. Raynor, and J. Tan, “Dietcam: Multiview regular shape food recognition with a camera phone,” *Pervasive and Mobile Computing*, vol. 19, pp. 108–121, 2015. [Online]. Available: http://dx.doi.org/10.1016/j.pmcj.2011.07.003

[50] H. He, F. Kong, and J. Tan, “Dietcam: Multiview food recognition using a multikernel svm,” *IEEE Journal of Biomedical and Health Informatics*, vol. 20, no. 3, pp. 848–855, May 2016. [Online]. Available: http://dx.doi.org/10.1109/JBHI.2015.2419251

[51] H. Chen, W. Jia, Z. Li, Y. Sun, and M. Sun, “3D/2D model-to-image registration for quantitative dietary assessment,” *Proceedings of the IEEE Annual Northeast Bioengineering Conference*, pp. 95–96, March 2012, Philadelphia, PA. [Online]. Available: http://dx.doi.org/10.1109/NEBC.2012.6206979

[52] P. Pouladzadeh, S. Shirmohammadi, and R. Almaghrabi, “Measuring calorie and nutrition from food image,” *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 8, pp. 1947–1956, August 2014. [Online]. Available: http://dx.doi.org/10.1109/TIM.2014.2303533

[53] W. Zhang, Q. Yu, B. Siddiquie, A. Divakaran, and H. Sawhney, “Snap-n-Eatfood recognition and nutrition estimation on a smartphone,” *Journal of Diabetes Science and Technology*, vol. 9, no. 3, pp. 525–533, April 2015. [Online]. Available: http://dx.doi.org/10.1177/1932296815582222

[54] K. Aizawa, Y. Maruyama, H. Li, and C. Morikawa, “Food balance estimation by using personal dietary tendencies in a multimedia Food Log,” *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 2176 – 2185, December 2013. [Online]. Available: http://dx.doi.org/10.1109/TMM.2013.2271474

[55] A. Myers, N. Johnston, V. Rathod, A. Korattikara, A. Gorban, N. Silberman, S. Guadarrama, G. Papandreou, J. Huang, and K. Murphy, “Im2Calories: towards an automated mobile vision food diary,” *Proceedings of the IEEE International Conference on Computer Vision*, December 2015, Santiago, Chile. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.146

[56] M. Fischler and R. Bolles, “Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography,” *Communications of the ACM*, vol. 24, no. 6, pp. 381–395, June 1981. [Online]. Available: http://dx.doi.org/10.1016/B978-0-08-051581-6.50070-2

[57] N. Silberman, P. Kohli, D. Hoiem, and R. Fergus, “Indoor segmentation and support inference from RGBD images,” *Proceedisng of the European Conference on Computer Vision*, pp. 746–760, October 2012, Florence, Italy. [Online]. Available: http://dx.doi.org/0.1007/978-3-642-33715-4_54

[58] K. Kutulakos and S. Seitz, “A theory of shape by space carving,” *International Journal of Computer Vision*, vol. 38, no. 3, pp. 199–218, 2000. [Online]. Available: http://dx.doi.org/10.1023/A:1008191222954

[59] C. Xu, Y. He, N. Khanna, C. Boushey, and E. Delp, “Model-based food volume estimation using 3D pose,” *Proceedings of IEEE International Conference on Image Processing*, pp. 2534 – 2538, September 2013, Melbourne, Australia. [Online]. Available: http://dx.doi.org/10.1109/ICIP.2013.6738522

[60] C. Xu, Y. He, N. Khannan, A. Parra, C. Boushey, and E. Delp, “Image-based food volume estimation,” *Proceedings of the International Workshop on Multimedia for Cooking & Eating Activities*, pp. 75–80, 2013. [Online]. Available: http://dx.doi.org/10.1145/2506023.2506037

[61] Y. He, C. Xu, N. Khanna, C. Boushey, and E. Delp, “Food image analysis: Segmentation, identification and weight estimation,” *Proceedings of IEEE International Conference on Multimedia and Expo*, pp. 1–10, July 2013, San Jose, CA. [Online]. Available: http://dx.doi.org/10.1109/ICME.2013.6607548

[62] C. Lee, J. Chae, T. Schap, D. Kerr, E. Delp, D. Ebert, and C. Boushey, “Comparison of known food weights with image-based portion-size automated estimation and adolescents’ self-reported portion size,” *Journal of Diabetes Science and Technology*, vol. 6, no. 2, pp. 428–434, March 2012. [Online]. Available: http://dx.doi.org/10.1177/193229681200600231

[63] Z. Zhang, “A flexible new technique for camera calibration,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, November 2000. [Online]. Available: http://dx.doi.org/10.1109/34.888718

[64] C. Xu, “Volume estimation and image quality assessment with application in dietary assessment and evaluation,” Ph.D. dissertation, Purdue University, West Lafayette, IN, May 2014.

[65] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision*, vol. 2, no. 60, pp. 91–110, January 2004. [Online]. Available: http://dx.doi.org/10.1023/B:VISI.0000029664.99615.94

[66] S. Gorthi and P. Rastogi, "Fringe projection techniques: Whither we are?" *Optics and Lasers in Engineering*, vol. 48, no. 2, pp. 133–140, Feburuary 2010.

[67] S. Zhang, "Flexible 3d shape measurement using projector defocusing: extended measurement range," *Optical Letter*, vol. 35, no. 7, pp. 934–936, April 2010. [Online]. Available: http://dx.doi.org/10.1364/OL.35.000934

[68] B. Li, N. Karpinsky, and S. Zhang, "Novel calibration method for structured-light system with an out-of-focus projector," *Applied Optics*, vol. 53, no. 16, pp. 3415 – 3426, 2014. [Online]. Available: http://dx.doi.org/10.1364/AO.53.003415

[69] L. Baum, T. Petrie, G. Soules, and N. Weiss, "A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains," *The Annals of Mathematical Statistics*, vol. 41, no. 1, pp. 164–171, 1970. [Online]. Available: http://dx.doi.org/10.1214/aoms/1177697196

[70] T. Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, vol. 13, no. 6, pp. 47–60, November 1996. [Online]. Available: http://dx.doi.org/10.1109/79.543975

[71] S. Belongie, C. Carson, H. Greenspan, and J. Malik, "Color- and texture-based image segmentation using EM and its application to content-based image retrieval," *Proceedisng of the IEEE International Conference on Computer Vision*, pp. 675 – 682, January 1998, Bombay, India. [Online]. Available: http://dx.doi.org/10.1109/ICCV.1998.710790

[72] T. Heath, *The works of Archimedes.* London, UK: Cambridge University Press, 1897.

[73] S. Fang, F. Zhu, C. Jiang, S. Zhang, C. Boushey, and E. Delp, "A comparison of food portion size estimation using geometric models and depth images," *Proceedings of the IEEE International Conference on Image Processing*, pp. 26 – 30, September 2016, Phoenix, AZ. [Online]. Available: http://dx.doi.org/10.1109/ICIP.2016.7532312

[74] Y. He, C. Xu, N. Khanna, C. Boushey, and E. Delp, "Context based food image analysis," *Proceedings of IEEE International Conference on Image Processing*, pp. 2748–2752, September 2013, Melbourne, Australia. [Online]. Available: http://dx.doi.org/10.1109/ICIP.2013.6738566

[75] G. Bradski and A. Kaehler, *Learning OpenCV: Computer vision with the OpenCV library.* O'Reilly Media, Inc., 2008.

[76] S. Fang, F. Zhu, C. Boushey, and E. Delp, "The use of co-occurrence patterns in single image based food portion estimation," *Proceedings of the IEEE Global Conference on Signal and Information Processing*, pp. 462 – 466, November 2017, Montreal, Canada. [Online]. Available: http://dx.doi.org/10.1109/GlobalSIP.2017.8308685

[77] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[78] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998. [Online]. Available: http://dx.doi.org/10.1109/5.726791

[79] T. Ege and K. Yanai, "Image-based food calorie estimation using knowledge on food categories, ingredients and cooking directions," *Proceedings of the Workshops of ACM Multimedia on Thematic*, pp. 367–375, 2017, Mountain View, CA. [Online]. Available: http://dx.doi.org/0.1145/3126686.3126742

[80] A. H. Abdulnabi, G. Wang, J. Lu, and K. Jia, "Multi-task cnn model for attribute prediction," *IEEE Transactions on Multimedia*, vol. 17, no. 11, pp. 1949–1959, Nov 2015. [Online]. Available: http://dx.doi.org/10.1109/TMM.2015.2477680

[81] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[82] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5967–5976, July 2017, Honolulu, HI. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2017.632

[83] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems 27*, pp. 2672–2680, December 2014, Montreal, Canada.

[84] T. Wang, M. Liu, J. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional gans," *arXiv preprint arXiv:1711.11585*, 2017.

[85] J. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," *Proceedings of the International Conference on Computer Vision*, pp. 2223–2232, 2017, Venice, Italy. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2017.244

[86] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *arXiv preprint arXiv:1511.06434*, 2015.

[87] M. Liu, T. Breuel, and J. Kautz, "Unsupervised image-to-image translation networks," *Advances in Neural Information Processing Systems*, pp. 700–708, 2017, Long Beach, CA.

[88] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, and J. Yang, "Pfid: Pittsburgh fast-food image dataset," *Proceedings of the IEEE International Conference on Image Processing*, pp. 289–292, November 2009, Cairo, Egypt. [Online]. Available: http://dx.doi.org/10.1109/ICIP.2009.5413511

[89] Y. Kawano and K. Yanai, "Automatic expansion of a food image dataset leveraging existing categories with domain adaptation," *Proceedings of European Conference on Computer Vision Workshops*, pp. 3–17, September 2014, Zurich, Switzerland.

[90] L. Bossard, M. Guillaumin, and L. V. Gool, “Food-101 – mining discriminative components with random forests,” *Proceedings of European Conference on Computer Vision*, vol. 8694, pp. 446–461, September 2014, Zurich, Switzerland. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-10599-4_29

[91] C. J. Boushey, M. Spoden, E. J. Delp, F. Zhu, M. Bosch, Z. Ahmad, Y. B. Shvetsov, J. P. DeLany, and D. A. Kerr, “Report energy intake accuracy compared to doubly labeled water and usability of the mobile food record among community dwelling adult,” *Nutrients*, vol. 9, no. 3, 2017. [Online]. Available: https://dx.doi.org/10.3390/nu9030312

[92] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, and A. Efros, “Context encoders: Feature learning by inpainting,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2536–2544, June 2016, Las Vegas, NV. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2016.278

[93] R. Tylecek and R. Sara, “Spatial pattern templates for recognition of objects with regular structure,” *Proceedings of the German Conference on Pattern Recognition*, pp. 364–374, 2013, Saarbrucken, Germany. [Online]. Available: http://dx.doi.org/10.1007/978-3-642-40602-7_39

[94] V. Badrinarayanan, A. Kendall, and R. Cipolla, “Segnet: A deep convolutional encoder-decoder architecture for image segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, Dec 2017. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2016.2644615

[95] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” *Proceedings of the Medical Image Computing and Computer-Assisted Intervention*, pp. 231–241, October 2015, Munich, Germany. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-24574-4_28

[96] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” *Proceedisng of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, June 2016, Las Vegas, NV. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2016.90

[97] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” *arXiv preprint arXiv:1412.6980*, 2014.

[98] 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,” *Proceedings of the IEEE International Conference on Image Processing*, pp. 251–255, October 2018, athens, Greece.

[99] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. Berg, and F. Li, “Imagenet large scale visual recognition challenge,” *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211 – 252, 2015. [Online]. Available: http://dx.doi.org/10.1007/s11263-015-0816-y

[100] Y. Wang, Y. He, C. Boushey, F. Zhu, and E. Delp, “Context based image analysis with application in dietary assessment and evaluation,” *Multimedia Tools and Applications*, pp. 1–26, November 2017. [Online]. Available: http://dx.doi.org/10.1007/s11042-017-5346-x

[101] “USDA food and nutrient database for dietary studies, 1.0.” Beltsville, MD: Agricultural Research Service, Food Surveys Research Group, 2004.

[102] J. Howe, “The rise of crowdsourcing,” *Wired Magazine*, vol. 14, no. 6, pp. 1–4, 2006, Dorsey Press.

[103] M. Rabbi, J. Costa, F. Okeke, M. Schachere, M. Zhang, and T. Choudhury, “An intelligent crowd-worker selection approach for reliable content labeling of food images,” *Proceedings of the conference on Wireless Health*, pp. 9:1–9:8, October 2015, bathesda, MD. [Online]. Available: http://dx.doi.org/10.1145/2811780.2811955

[104] C. Rother, V. Kolmogorov, and A. Blake, “Grabcut: Interactive foreground extraction using iterated graph cuts,” *ACM Transactions on Graphics*, vol. 23, no. 3, pp. 309–314, 2004. [Online]. Available: http://dx.doi.org/10.1145/1015706.1015720

[105] “Flickr: The app garden,” [Online]. Available: https://www.flickr.com/services/api/.

[106] “Google: Custom search engine,” [Online]. Available: https://cse.google.com/cse/.

[107] J. Redmon and A. Farhadi, “Yolo9000: Better, faster, stronger,” *arXiv preprint arXiv:1612.08242*, 2016. [Online]. Available: http://dx.doi.org/10.1109/CVPR.2017.690

[108] R. Girshick, “Fast r-cnn,” *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448, December 2015. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.169

[109] S. Ren, K. He, R. Girshick, and J. Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” *Advances in Neural Information Processing Systems 28*, pp. 91–99, December 2015, Montreal, Canada. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2016.2577031

[110] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, “Focal loss for dense object detection,” *arXiv preprint arXiv:1708.02002*, 2017. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2017.324

[111] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, “SSD: single shot multibox detector,” *European conference on computer vision*, pp. 21–37, October 2016, Amsterdam, The Netherlands. [Online]. Available: https://dx.doi.org/10.1007/978-3-319-46448-0_2

[112] D. M. Powers, “Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation,” *Journal of Machine Learning Technologyies*, vol. 2, pp. 37–63, 2011.

[113] B. Baldwin, “Cogniac: High precision coreference with limited knowledge and linguistic resources,” *Proceedings of the Workshop on Operational Factors in Practical, Robust Anaphora Resolution for Unrestricted Texts*, pp. 38–45, 1997.

[114] B. Wandell, *Foundations of Vision.* Sinauer Associates, Inc., 1995, Sunderland, MA.

[115] G. Sharma, *Digital Color Imaging Handbook.* Boca Raton, Florida: CRC Press, 2002.

[116] A. Mikkilineni, P. Chiang, G. Ali, G. Chiu, J. Allebach, and E. Delp, “Printer identification based on graylevel co-occurrence features for security and forensic applications,” *Proceedings of the SPIE Security, Steganography, and Watermarking of Multimedia Contents VII*, vol. 5681, pp. 430–440, January 2005, San Jose, CA. [Online]. Available: http://dx.doi.org/10.1117/12.593796

[117] K. van de Sande, T. Gevers, and C. Snoek, “Evaluating color descriptors for object and scene recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1582–1596, September 2010. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2009.154

[118] F. Mindru, T. Tuytelaars, L. Van Gool, and T. Moons, “Moment invariants for recognition under changing viewpoint and illumination,” *Computer Vision and Image Understanding*, vol. 94, no. 1-3, pp. 3–27, April-June 2004. [Online]. Available: http://dx.doi.org/10.1016/j.cviu.2003.10.011

[119] J. von Kries, “Chromatic adaptation,” *Festschrift der Albrecht-Ludwigs-Universit*, pp. 145–158, 1902.

[120] M. Stokes, M. Anderson, S. Chandrasekar, and R. Motta, “A standard default color space for the internet-srgb,” *Microsoft and Hewlett-Packard Joint Report*, 1996.

[121] C. Xu, N. Khanna, C. J. Boushey, and E. J. Delp, “Low complexity image quality measures for dietary assessment using mobile devices,” *Proceedings of the IEEE International Symposium on Multimedia*, Dana Point, CA, December 2011, pp. 351–356. [Online]. Available: http://dx.doi.org/10.1109/ISM.2011.64

[122] K. Briechle and U. Hanebeck, “Template matching using fast normalized cross correlation,” *Proceedings of the SPIE Optical Pattern Recognition XII*, vol. 4387, pp. 95–102, April 2001, Orlando, FL. [Online]. Available: http://dx.doi.org/10.1117/12.421129

[123] F. Zhao, Q. Huang, and W. Gao, “Image matching by normalized cross-correlation,” *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 729–732, May 2006, Toulouse, France. [Online]. Available: http://dx.doi.org/10.1109/ICASSP.2006.1660446

VITA

VITA

Shaobo Fang was born in Zhengzhou, China. He received the B.S. in Electrical Engineering (EE) from Purdue University, West Lafayette. He also received M.S. in Electrical and Computer Engineering (ECE) from Purdue University, West Lafayette, USA. Mr. Fang joined the Ph.D. program at the School of Electrical and Computer Engineering, Purdue University, West Lafayette, Indiana in August 2013.

He has worked as Research Assistant in the Video and Image Processing Laboratory (VIPER) under the supervision of Professor Edward J. Delp and Professor Fengqing Zhu since May 2014. He is a student member of the IEEE and the IEEE Signal Processing Society.