Master's thesis proposal

Stijn Voss, s4150511

February 2017

1 Food journaling

Keeping track of food consumption can help for a variety of goals. For example identifying allergies, detecting deficiencies but probably most importantly achieving health and weight-loss goals. Existing applications used are MyFitnessPal[1], Lose It![2], FatSecret[3]. These applications depend on manually data entry. Either by searching in a database or scanning barcodes.

Research shows that people tend to underestimate their calorie intake(Schoeller, Bandini, & Dietz, 1990), moreover research clearly suggests that overweight and obese people tend to underestimate caloric intake more then non-overweight people(Pikholz, Swinburn, & Metcalf, 2004; Garriguet, 2008). Even more so this is true for children (Forrestal, 2011). This suggests that making people aware of their caloric intake, by using food diaries, can already help preventing overweight.

Not sure, look for more evidence

In controlled experiments it has been shown that in general web based intervention techniques can be effective in helping people obtain healthier lifestyles. Computer tailoring(personalized advices by a computer), expert advice or other more sophisticated behavior changing techniques results tend to increase the effectiveness (Lustria, Cortese, Noar, & Glueckauf, 2009; Kroeze, Werkman, & Brug, 2006; Gold, Burke, Pintauro, Buzzell, & Harvey-Berino, 2007; Webb, Joseph, Yardley, & Michie, 2010; Lustria et al., 2013). More specifically it has also shown to work for diet and physical activity (Turner-McGrievy et al., 2013; Wharton, Johnston, Cunningham, & Sterner, 2014)

However only a small part of the people that install food diaries keep using them or achieve their goals(Helander, Kaipainen, Korhonen, & Wansink, 2014) . Barriers identified include: too much effort, hard to determine portion size and which ingredients are used and loosing the habit(simply forgetting). Furthermore it was shown that users especially have difficulty tracking food that was not fast food or pre-packaged. For example home cooked meals, restaurant meals and meals prepared by friends. Other

[1] `https://www.myfitnesspal.com/`
[2] `https://www.loseit.com/`
[3] `https://www.fatsecret.com/`

problems include: feeling guilty about eating to much and a perceived stigma around tracking(people are ashamed of their tracking behaviour when with others) (Cordeiro et al., 2015).

2 Photo based food journaling

To make keeping track of food diary easier researchers have been looking into photo based approaches. Where the user makes a photo of their food and this is used to estimate calories and food intake.

One of the earliest applications is PlateMate(Noronha, Hysen, Zhang, & Gajos, 2011). Users can make a picture of their meal, using amazon machincal turk these images will then be segmented, labeled and measured. The resulting nutrition information is then shown to the user. The process is shown in figure 1. When manufacturer information is taken as a baseline, the application achieves error rates comparable or better then experts that had to determine intake based on pictures. Users generally found the application easier to use than manually tracking apps. Users however felt that the application didn't work very accurately. The PlateMate system indeed seemed to over estimate the caloric intake where as users tend to underestimate, explaining the perception of being inaccurate. (Schoeller et al., 1990).

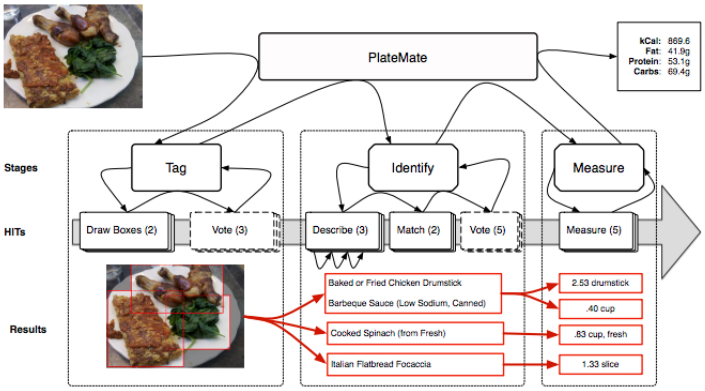

Figure 1: The PlateMate system. Work travels between stages and Human Intelligence Tasks (HITs) along the black arrows, starting from the input on the left and concluding with the output on the right. The system takes submitted photos and creates Tag tasks to annotate these photos with boxes. Each box becomes the input to a series of Identify tasks which end with a list of foods from a commercial food database. Each individual food is then input to a Measure task, which produces a unit and amount. Dashed boxes represent optional stages, which may be skipped during routing. Obtained from (Noronha et al., 2011)

2.1 Computer vision

Since this is of course quit a labour-intensive way of doing food journaling, lately research focused more on using computer vision based approaches. It's often claimed that, with the raise of deep learning and convolutional neural networks, computer vision now performs on a human level. However this mainly holds for the task of object recognition. The task of food recognition seems to be more challenging.

> In contrast to scene classification or object detection, food typically does not exhibit any distinctive spatial layout: while we can decompose an outdoor scene with a ground plane, a horizon and a sky region, or a human as a trunk with a head and limbs, we cannot find similar patterns relating ingredients of a mixed salad. The point of view, the lighting conditions, but also (and not least) the very realization of a recipe are among the sources of high intra-class variations. On the bright side, the nature of dishes is often defined by the different colors and textures of its different local components, such that humans can identify them reasonably well from a single image, regardless of the above variations. Hence, food recognition is a specific classification problem calling for models that can exploit local information. ((Bossard, Guillaumin, & Van Gool, 2014)

This is quote from another paper, maybe not use it as such in my thesis

Furthermore even if we would achieve human level performance, it would probably still not be enough to accurately estimate calorie and nutrition intake. Since humans aren't always able to directly estimate all calories from a meal either, simply because it's not visible.

Applying computer vision to the problem of food recognition is quit a new field, that shows an increase in interest lately. However the challenges, approaches and datasets differ a lot. In the following section I try to describe the different challenges and perspectives i came across in the literature.

2.1.1 Food/nonfood detection

The simplest task is to try to differentiate images that contain food from images that do not. This is often solved by using a (pre-trained) convolutional neural network, and fine-tuning it on a binary classification task. Since this basically is a object recognition task performance is usually on a human level. Depending on the exact dataset and exact network, accuracy ranges from 90-99.5% for CNN bases approaches. In table 1 a overview of different approaches is given. Please note that datasets tend to differ a bit in their very nature(some datasets count images where people are holding food as food images too). Because of this the methods are hard to compare. Furthermore most

evidence suggests that fine-tuning pre-trained networks outperforms training networks from scratch.

Paper	**Method**(network)	**Dataset**	**Accuracy**
Aizawa et al. (2013)	SVM with hand-crafted features	own	89%
Kagaya et al. (2014)	CNN(own)	own(175,00 images from FoodLog)	93.8%
Kagaya and Aizawa (2015)	Pre-trained CNN network in network (Lin et al., 2013)	own(same as above)	99.1%
Meyers et al. (2015)	Pre-trained CNN(GoogleLeNet)	food101	99.02%
Singla et al. (2016)	Pre-trained CNN(GoogleLeNet)	food5k	99.2%
Ragusa et al. (2016)	SVM on extracted features of fine-tuned AlexNet	UNICT-FD889, NonFood-flickr	94.86%
Bolanos and Radeva (2016)	pre-trained GoogleLeNet, GAP layer in front of softmax	Constructed from Food101, ILSVRC, PASCAL	95.64%

Table 1: Overview approaches to nonfood/food classification

2.1.2 Food categorization

Other work involves trying to classify food into a certain set of categories, however it is not obvious how these categories should be determined. Certainly not with respect to food logging. Early approaches involved determining food balance(according to the food pyramid) estimation using hand crafted features (Kitamura, Yamasaki, & Aizawa, 2008). Other approaches categorize into a limited set of food types using SVM techniques(Aizawa et al., 2013) or more recently convolution neural networks (Meyers et al., 2015; J. Chen & Ngo, 2016). Usually these methods can perform reasonably well by somehow limiting the number of classes. For example by just differentiating between different meals in a single restaurant (Beijbom, Joshi, Morris, Saponas, & Khullar, 2015) or food categories (bread, pasta, dairy etc.). An overview of different approaches is given in table 2.

Other approaches make use of the fact that food categories can be represented as a hierachical structure (Wu, Merler, Uceda-Sosa, & Smith, 2016). Food is represented in a tree structure and the task of the network is to determine the category for each tree layer. The information incorperated in the tree structure is then enforced using a random walk approach.

Paper	Challenge	Method	Performance
Kitamura et al. (2008)	Estimate food type according to food pyramid	SVM (handcrafted features)	Accuracy of 73%
Kagaya et al. (2014)	Dataset consists of 10 different meals	CNN (own)	Accuracy of 73.70%
J. Chen and Ngo (2016)	Categorizing for 172 categories	VIREO Food-172	Top-1 accuracy of 82.06% and a top-5 accuracy of 95.88%

Table 2: Overview of different challenges and methods in food categorization tasks

2.1.3 Recipe retrieval

Using categorization to recognize different meals can be very difficult since there might be possibly 10's of thousand categories. J. Chen and Ngo (2016) noted that instead of focusing on trying to find the dish directly by a picture, trying to to estimate the ingredients might be easier. Please note that their paper is mainly focused on the Chinese cuisine, where dishes often involve small portions with few ingredients(which are often pretty recognizable).

Based on a pre-trained VGG16 network they trained a multi-task deep neural network that had to predict the ingredients and the category of the dish. For the categorization task they achieved a top-1 accuracy of 82.06% and a top-5 accuracy of 95.88%(out of 172 categories).

The ingredients are predicted using a multi-label output layer with sigmoid activation. A conditional random field is used to account for inter-ingredient dependencies. For the ingredients they achieve a micro-F1 of 67.17% and a macro-f1 of 47.18%. It is clearly shown that the multi-task approach significantly increases their performance.

They also provided the network with images of recipes that the network had never seen before. Given the resulting ingredient probabilities for every recipes a score was then determined using eq 1.

$$S_i = \sum_{c \in O \cap c \in Q_i} x_c \tag{1}$$

Where O contains the ingredients of the recipe, Q_i contains the ingredients found in the network and X_c contains the probability for that ingredient. The network was able to predict the right recipe in the top-10 for about 50% of the cases.

In their next paper they extended their method with the concept of Stacked attention networks(SAN) (Yang, He, Gao, Deng, & Smola, 2016). SAN's are original introduced to tackle the challenge of answering questions that are related to the content of an image for example 'what is the color of the horns?'. A SAN uses the last features of a convolutional network where spatial information is still intact, just before the fully connected layers. Given the semantic representation of a query and the spatial features, a attention area is selected and a new query if formulated . They authors claim that this represents the normal reasoning process required behind such question. First search the horns and put your attention to this area, next determine the color.

J. Chen, Pang, and Ngo (2017) use the same idea but then try to locate the ingredients on a picture. The semantic information of the question is replaces by a vector representing the ingredients of the meal. The network job is to locate the ingredients.

I don't entirely understand their method. I don't understand their objective function and don't understand how they get actual ingredients from a to classify image. Since they use both the image and ingredients as input during training.
Furthermore i am not sure about their method either. Finding ingredients doesn't really seem to reflect a reasoning process.

2.1.4 Localization and segmentation

Another important aspect of computer vision for food seems to be locating food on a picture or segment if multiple food objects are being displayed in a picture. Bolanos and Radeva (2016) make use of foon/non-food classification network to address the challenge of food localization and segmentation. They used the GoogleLeNet network, replaced the last layers with a 'Global average pooling(GAP)'-layer such that spatial information is not lost. Using softmax they then fine-tuned the network on the binary food/nonfood identification task. Using the output from the GAP layer they now could create a food activation map for the image, indicating a probability of food being present for each pixel. This can then be used to draw bounding boxes. Testing the performance with a minimum IoU(intersection over union) of .5 to test each bounding box on the UECFood256 and EgocentricFood datasets they found a accuracy of 36.84% and 6.41% respectively.

Meyers et al. (2015) noted that instead of using bounding boxes it is better to use segmentation. Since food tend not to be rectangular in nature. They made use of the DeepLab(L.-C. Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2014) method. They achieve an accuracy of 76% and an average IoU of .25 on the Food101-segmented dataset.

I don't understand the DeepLab paper however (Noh, Hong, & Han, 2015) seem to be an improvement of that paper and i do understand that one.

2.1.5 Quantity estimation

Another important aspect of food logging is of course the volume of the meal. Meyers et al. (2015) did this by first predicting the pixel depth for each pixel using a CNN based on a simple RGB input image (Eigen & Fergus, 2015). They achieve an average error of 18 centimeters. Based on this they create a 3d voxel grid of 2mmx2mm each(where they used the average predicted height as height for each voxel), with respect to the surface(which they detect). By re-using the segmentation task they could then try to estimate the volume of the meal. The error differs a lot between the different meal classes. For most of the meal classes the volume error stayed below an average of 200ml.

2.1.6 Context

Even with human level performance it would difficult to estimate the exact meal or food presented, so incorporating contextual knowledge seems crucial. Aizawa et al. (2013) used a Bayesian method to estimate food balance(according to the food pyramid) and incorporated user knowledge using priors. Other approaches assumed that the restaurant a user was visiting was known and limited the search space down this way (Joutou & Yanai, 2009). Meyers et al. (2015) trained a separate CNN network per restaurant.

2.1.7 Constant monitoring vs user photo

Research also makes a distinction between constant monitoring and pictures made by user. Wearable cams worn around the neck could constantly track what one is eating or even prepares while cooking (O'Loughlin et al., 2013; Liu et al., 2012; Jia et al., 2014). This could lead to less effort for the user and more accurate tracking, but also is technically way more challenging.

Mobile device/platform

(Horiguchi, Aizawa, & Ogawa, 2016)

3 Proposal

Since the litarature suggests that recognizing prepared meals is cause for the most difficulties i want to address that problem first. The problem of food categorization is difficult since the same category can have a lot of different spatial properties(the same meal or food doesn't always look the same, but sometimes it does) as compared to other visual tasks. Your model has to be invariant to these spatial features. Furthermore you need a large amount of categories in order to be useful, quickly leading to over-fitting certain of these specific spatial features. In short I can think of two ways that can help tackle this.

- Use more training data(especially more per category)
- Make your model less vulnerable to these spatial features anyway. Most models that i have seen fine-tune existing object recognition networks, which might be a disadvantage. Furthermore the network architecture itself might also be a problem.

Very speculative, should do more research

The recipe retrieval concept of (J. Chen & Ngo, 2016) seems very promising since, by predicting ingredients, they reduce the categories required and prevent having to define these categories. And thus the model will learn to be invariant more quickly. However I think it has some limitations. First of all if you try to recognize meals by just the ingredients you might lose quit some information that could help you predict the right meal. For example the relative ratio between ingredients or shapes of ingredients might actually tell us something about the actual meal. Furthermore the way they query for recipes now seems to require to calculate the distance to each recipe separately which can be too time consuming if you have lot's of recipes. Another disadvantage is that they not take into account that some ingredients will never be visible and actually might be quit important for caloric intake(for example: sugar, oils). Such ingredients will always have to inferred from something else: for example by asking the user or using context information.

Not very sure about this, also not really sure if my approach would improve on it

Instead I propose to model meals in a vector space('recipe space') that represents the relevant(for food logging) information for that meal. Analogous to representing the semantics of words in a vector space using word embeddings. . To make sure this vector does contain only relevant information. I would let the network predict metrics that are relevant for food journals. For example:

Look up some papers that do something like this, for other problems

- Ingredient and it's ratio prediction (it might also help to only predict ingredients that can be predicted by using vision)
- category prediction(to make sure meals of different categories, with similar ingredients are not mapped to the same vector space)
- Calorie(density) prediction

By forcing the network to go trough a small enough layer, before predicting the above classes and metrics, and using the activations of that layer. A vector could be created that contains all information that is required to predict these metrics for each picture of meal. Similar meals should then have a small distance to each other, where as un-similar meals would have a larger distance. Meals that are difficult to separate based on vision should also have a smaller distance.

Creating a vector space might seem a bit far fetched but i believe it would overcome the need to categorize food in finite set of categories, while still using all information that identifies a meal. In practice it might also allow to query for something like this: find recipes that lay close to this "visual recipe space" but have very different "invisible" ingredients? Providing a user with a set of options.

The research question I want to answer is: Can we represent meals in a vector space?

3.1 Approach

Dataset
Using the schema.org structured recipe format[4] combined with the common crawl[5] index I could obtain a very large dataset from a lot of sources that contain a huge variety of meals. The recipe format provides in a lot of different properties like ingredients, nutrition information and a image. However the properties are not always perfectly clean and probably need some additional processing. For example ingredients come with quantities and forms. Calories are defined for the whole meal or per serving etc. It could be interesting to try to apply word2vec on ingredients(like similar words will be used in the same context, same ingredients will be used in the same context). Double ingredients could than be found by looking for very similar vectors analogous to finding different spelling alternatives using word2vec values in NLP tasks.

An alternative approach could be to write a site specific scraping script by hand for a website and use these. However using the schema.org approach might be useful since it can be handy in a real practical application. For example logging your meal by providing an url could be a relatively easy and useful feature. Information about how to use a wide variety of web sources can be useful.

General steps
Once I have a dataset I would start by re-using an existing network(like GoogleNet, this one seems to be successful in other papers). Remove the last layers and replace it with my multi-task layers. First i could train it without fine-tuning the first layers, speeding up the process. This would allow me to find out what the dimensions of the recipe space could work and determine other parameters/choices. Later on i could try to improve my performance by fine-tuning or using other architectures etc.

[4] `http://schema.org/Recipe`
[5] `http://commoncrawl.org/`

One open question is how to evaluate my method..?

4 Appendix

4.1 Datasets

While reading the papers I came across a lot of datasets. I sum them up here for later reference.

- **Food101** Contains 101,000 images of 101 different meals/categories. Taken from `foodspotting.com` `http://www.vision.ee.ethz.ch/datasets/food-101/.` (Bossard et al., 2014)
- **UEC-FOOD-100** Contains 9060 images of 100 categories mainly japanese recipes. Focused on so called multi-food meals. (Matsuda, Hoashi, & Yanai, 2012)
- **UEC-FOOD-256** 256 food categories, extension of UEC-FOOD-100. Build by searching for the original categories on flickr and instagram. Based on first dataset, check if indeed food and assign some category. Using crowdsourcing improve further(Kawano & Yanai, 2014)
- **FID dataset** obtained by looking for the #food tag on instagram manually labeled. Resulting in 4,230 food images and 5,428 non-food images. Based on a small sub sample it seems that this dataset hash quit a broad definition of food images. Some food images contain humans as the main topic while enjoying food. These images are labeled as food images. (Kagaya & Aizawa, 2015)
- **food 5k and 11k** These data-sets where assembled using the Food101 datasets and UEC-FOOD-* datasets. 5k is focused on non-food/food classification, 11k divides the images into 11 categories. Optionally some images from social media where added (Singla et al., 2016) `http://mmspg.epfl.ch/food-image-datasets`
- **VIREO Food-172** (J. Chen & Ngo, 2016) 110,241 images divided into 172 chinese food categories and assigned an average of 3 ingredients per recipe out of 353 possibles recipes.
- **UNICT-FD889** 3583 images of 889 different plates (Farinella, Allegra, & Stanco, 2014) `http://iplab.dmi.unict.it/UNICT-FD889/`
- **Flickr-Food, FlickrNonFood** Obtained from flickr and manually labeld 4805 of food and 8005 not food (Farinella, Allegra, Stanco, & Battiato, 2015) `http://iplab.dmi.unict.it/m`
- **EgocentricFood** contains images from a wearable cam (Bolanos & Radeva, 2016) `http://www.ub.edu/cvub/egocentricfood/`

Not entirely clear to me what they did, but not so relevant now

References

Aizawa, K., Maruyama, Y., Li, H., & Morikawa, C. (2013). Food balance estimation by using personal dietary tendencies in a multimedia food log. *IEEE Transactions on multimedia*, *15*(8), 2176–2185.

Beijbom, O., Joshi, N., Morris, D., Saponas, S., & Khullar, S. (2015). Menu-match: restaurant-specific food logging from images. In *Applications of computer vision (wacv), 2015 ieee winter conference on* (pp. 844–851).

Bolanos, M., & Radeva, P. (2016). Simultaneous food localization and recognition. *arXiv preprint arXiv:1604.07953*.

Bossard, L., Guillaumin, M., & Van Gool, L. (2014). Food-101–mining discriminative components with random forests. In *European conference on computer vision* (pp. 446–461).

Chen, J., & Ngo, C.-W. (2016). Deep-based ingredient recognition for cooking recipe retrieval. In *Proceedings of the 2016 acm on multimedia conference* (pp. 32–41).

Chen, J., Pang, L., & Ngo, C.-W. (2017). Cross-modal recipe retrieval: How to cook this dish? In *International conference on multimedia modeling* (pp. 588–600).

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2014). Semantic image segmentation with deep convolutional nets and fully connected crfs. *arXiv preprint arXiv:1412.7062*.

Cordeiro, F., Epstein, D. A., Thomaz, E., Bales, E., Jagannathan, A. K., Abowd, G. D., & Fogarty, J. (2015). Barriers and negative nudges: Exploring challenges in food journaling. In *Proceedings of the 33rd annual acm conference on human factors in computing systems* (pp. 1159–1162).

Eigen, D., & Fergus, R. (2015). Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the ieee international conference on computer vision* (pp. 2650–2658).

Farinella, G. M., Allegra, D., & Stanco, F. (2014). A benchmark dataset to study the representation of food images. In *European conference on computer vision* (pp. 584–599).

Farinella, G. M., Allegra, D., Stanco, F., & Battiato, S. (2015). On the exploitation of one class classification to distinguish food vs non-food images. In *International conference on image analysis and processing* (pp. 375–383).

Forrestal, S. G. (2011). Energy intake misreporting among children and adolescents: a literature review. *Maternal & child nutrition*, *7*(2), 112–127.

Garriguet, D. (2008). Under-reporting of energy intake in the canadian community health survey. *Health reports*, *19*(4), 37.

Gold, B. C., Burke, S., Pintauro, S., Buzzell, P., & Harvey-Berino, J. (2007). Weight loss on the web: A pilot study comparing a structured behavioral intervention to a commercial program. *Obesity*, *15*(1), 155–155.

Helander, E., Kaipainen, K., Korhonen, I., & Wansink, B. (2014). Factors related to sustained use of a free mobile app for dietary self-monitoring with photography and

peer feedback: retrospective cohort study. *Journal of medical Internet research*, *16*(4), e109.

Horiguchi, S., Aizawa, K., & Ogawa, M. (2016). The log-normal distribution of the size of objects in daily meal images and its application to the efficient reduction of object proposals. In *Image processing (icip), 2016 ieee international conference on* (pp. 3668–3672).

Jia, W., Chen, H.-C., Yue, Y., Li, Z., Fernstrom, J., Bai, Y., ... Sun, M. (2014). Accuracy of food portion size estimation from digital pictures acquired by a chest-worn camera. *Public health nutrition*, *17*(08), 1671–1681.

Joutou, T., & Yanai, K. (2009). A food image recognition system with multiple kernel learning. In *Image processing (icip), 2009 16th ieee international conference on* (pp. 285–288).

Kagaya, H., & Aizawa, K. (2015). Highly accurate food/non-food image classification based on a deep convolutional neural network. In *International conference on image analysis and processing* (pp. 350–357).

Kagaya, H., Aizawa, K., & Ogawa, M. (2014). Food detection and recognition using convolutional neural network. In *Proceedings of the 22nd acm international conference on multimedia* (pp. 1085–1088).

Kawano, Y., & Yanai, K. (2014). Automatic expansion of a food image dataset leveraging existing categories with domain adaptation. In *European conference on computer vision* (pp. 3–17).

Kitamura, K., Yamasaki, T., & Aizawa, K. (2008). Food log by analyzing food images. In *Proceedings of the 16th acm international conference on multimedia* (pp. 999–1000).

Kroeze, W., Werkman, A., & Brug, J. (2006). A systematic review of randomized trials on the effectiveness of computer-tailored education on physical activity and dietary behaviors. *Annals of behavioral medicine*, *31*(3), 205–223.

Lin, M., Chen, Q., & Yan, S. (2013). Network in network. *arXiv preprint arXiv:1312.4400*.

Liu, J., Johns, E., Atallah, L., Pettitt, C., Lo, B., Frost, G., & Yang, G.-Z. (2012). An intelligent food-intake monitoring system using wearable sensors. In *Wearable and implantable body sensor networks (bsn), 2012 ninth international conference on* (pp. 154–160).

Lustria, M. L. A., Cortese, J., Noar, S. M., & Glueckauf, R. L. (2009). Computer-tailored health interventions delivered over the web: review and analysis of key components. *Patient education and counseling*, *74*(2), 156–173.

Lustria, M. L. A., Noar, S. M., Cortese, J., Van Stee, S. K., Glueckauf, R. L., & Lee, J. (2013). A meta-analysis of web-delivered tailored health behavior change interventions. *Journal of health communication*, *18*(9), 1039–1069.

Matsuda, Y., Hoashi, H., & Yanai, K. (2012). Recognition of multiple-food images by detecting candidate regions. In *Multimedia and expo (icme), 2012 ieee international conference on* (pp. 25–30).

Meyers, A., Johnston, N., Rathod, V., Korattikara, A., Gorban, A., Silberman, N., ... Murphy, K. P. (2015). Im2calories: towards an automated mobile vision food

diary. In *Proceedings of the ieee international conference on computer vision* (pp. 1233–1241).

Noh, H., Hong, S., & Han, B. (2015). Learning deconvolution network for semantic segmentation. In *Proceedings of the ieee international conference on computer vision* (pp. 1520–1528).

Noronha, J., Hysen, E., Zhang, H., & Gajos, K. Z. (2011). Platemate: crowdsourcing nutritional analysis from food photographs. In *Proceedings of the 24th annual acm symposium on user interface software and technology* (pp. 1–12).

O'Loughlin, G., Cullen, S. J., McGoldrick, A., O'Connor, S., Blain, R., O'Malley, S., & Warrington, G. D. (2013). Using a wearable camera to increase the accuracy of dietary analysis. *American journal of preventive medicine*, *44*(3), 297–301.

Pikholz, C., Swinburn, B., & Metcalf, P. (2004). Under-reporting of energy intake in the 1997 national nutrition survey. *New Zealand medical journal*, *117*(1202), 1–11.

Ragusa, F., Tomaselli, V., Furnari, A., Battiato, S., & Farinella, G. M. (2016). Food vs non-food classification. In *Proceedings of the 2nd international workshop on multimedia assisted dietary management* (pp. 77–81).

Schoeller, D. A., Bandini, L. G., & Dietz, W. H. (1990). Inaccuracies in self-reported intake identified by comparison with the doubly labelled water method. *Canadian journal of physiology and pharmacology*, *68*(7), 941–949.

Singla, A., Yuan, L., & Ebrahimi, T. (2016). Food/non-food image classification and food categorization using pre-trained googlenet model. In *Proceedings of the 2nd international workshop on multimedia assisted dietary management* (pp. 3–11).

Turner-McGrievy, G. M., Beets, M. W., Moore, J. B., Kaczynski, A. T., Barr-Anderson, D. J., & Tate, D. F. (2013). Comparison of traditional versus mobile app self-monitoring of physical activity and dietary intake among overweight adults participating in an mhealth weight loss program. *Journal of the American Medical Informatics Association*, *20*(3), 513–518.

Webb, T., Joseph, J., Yardley, L., & Michie, S. (2010). Using the internet to promote health behavior change: a systematic review and meta-analysis of the impact of theoretical basis, use of behavior change techniques, and mode of delivery on efficacy. *Journal of medical Internet research*, *12*(1), e4.

Wharton, C. M., Johnston, C. S., Cunningham, B. K., & Sterner, D. (2014). Dietary self-monitoring, but not dietary quality, improves with use of smartphone app technology in an 8-week weight loss trial. *Journal of nutrition education and behavior*, *46*(5), 440–444.

Wu, H., Merler, M., Uceda-Sosa, R., & Smith, J. R. (2016). Learning to make better mistakes: Semantics-aware visual food recognition. In *Proceedings of the 2016 acm on multimedia conference* (pp. 172–176).

Yang, Z., He, X., Gao, J., Deng, L., & Smola, A. (2016). Stacked attention networks for image question answering. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 21–29).