A Survey on AI Nutrition Recommender Systems

Thomas Theodoridis
Centre for Research and Technology Hellas
Thessaloniki, Greece
tomastheod@iti.gr

Vassilios Solachidis
Centre for Research and Technology Hellas
Thessaloniki, Greece
vsol@iti.gr

Kosmas Dimitropoulos
Centre for Research and Technology Hellas
Thessaloniki, Greece
dimitrop@iti.gr

Lazaros Gymnopoulos
Centre for Research and Technology Hellas
Thessaloniki, Greece
lazg@iti.gr

Petros Daras
Centre for Research and Technology Hellas
Thessaloniki, Greece
daras@iti.gr

ABSTRACT

The goal of this work is to provide an overview of existing approaches regarding AI nutrition recommender systems. A breakdown of such systems into task-specific components is presented, as well as methodologies concerned with each individual component. The components of an idealized AI nutrition recommender system are presented and compared to state-of-the-art approaches in the corresponding area of research. Finally, identified issues in some of these areas are also discussed.

CCS CONCEPTS

• **General and reference → Surveys and overviews**; • **Information systems → Recommender systems**; • **Computing methodologies** → *Object recognition*; *Machine learning*.

KEYWORDS

survey, nutrition, food, recommender systems

ACM Reference Format:
Thomas Theodoridis, Vassilios Solachidis, Kosmas Dimitropoulos, Lazaros Gymnopoulos, and Petros Daras. 2019. A Survey on AI Nutrition Recommender Systems. In *The 12th PErvasive Technologies Related to Assistive Environments Conference (PETRA '19), June 5–7, 2019, Rhodes, Greece.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3316782.3322760

1 INTRODUCTION

Eating is for some people just a necessary everyday activity, while for others, a unique moment in their daily schedule that gives them great enjoyment. No matter the side that each person has chosen, it is becoming more and more evident that the role food plays in our overall health is of utmost importance. From a superficial point of view, our bodies need a specific amount of energy to function properly and food provides just this. However, in reality, not all calories are created equal; the accompanying nutrients play a vital role in the way food is processed by the human body, thus affecting our overall health. To this end, the consumption of a wide variety of food items is necessary in order for the human body to obtain the right amounts of nutrients. Failing to follow such a well-balanced diet, in combination with a generally unhealthy way of living, has been shown to increase the risk for cardiovascular disease, type II diabetes and some forms of cancer. Taking all these factors into consideration, food intake monitoring can provide substantial benefits in certain cases.

Traditional approaches towards food intake monitoring relied on 24-hour recalls and food frequency questionnaires in order to obtain relevant information. Although effective to some extent, the margin of error was high and the process was labour-intensive. In the last few years, the convergence of several technological advances, both from a hardware and software perspective, has made possible the existence of automated systems that can analyze users' eating habits and preferences, and provide recommendations in order to achieve specific goals (e.g., weight loss, muscle gain or eating healthy).

In the rest of this work, we provide an overview of such methods. These have been classified into task-specific categories, as shown in Figure 1. In section 2, methods concerned with food analysis are presented, including methods about food category recognition, food ingredient and cooking instructions recognition and food quantity estimation. In section 3, an overview of methods regarding eating behaviour analysis is presented. In section 4, an idealized AI nutrition recommender system is presented, and each of the needed components is compared to state-of-the-art methods in the corresponding area of research. Additionally, an overview of recent literature and EU-funded projects is provided regarding AI nutrition recommender systems. Finally, in section 5 conclusions are drawn.

2 FOOD ANALYSIS

Food analysis is a core component of AI nutrition recommender systems, as it provides the prerequisites for obtaining a high-level understanding of the type and the amount of food consumed by the user. This category can broadly be divided into methods related to food category recognition, food ingredient and cooking instructions recognition and food quantity estimation. In the next sections

PETRA '19, June 5–7, 2019, Rhodes, Greece

ACM ISBN 978-1-4503-6232-0/19/06…$15.00
https://doi.org/10.1145/3316782.3322760

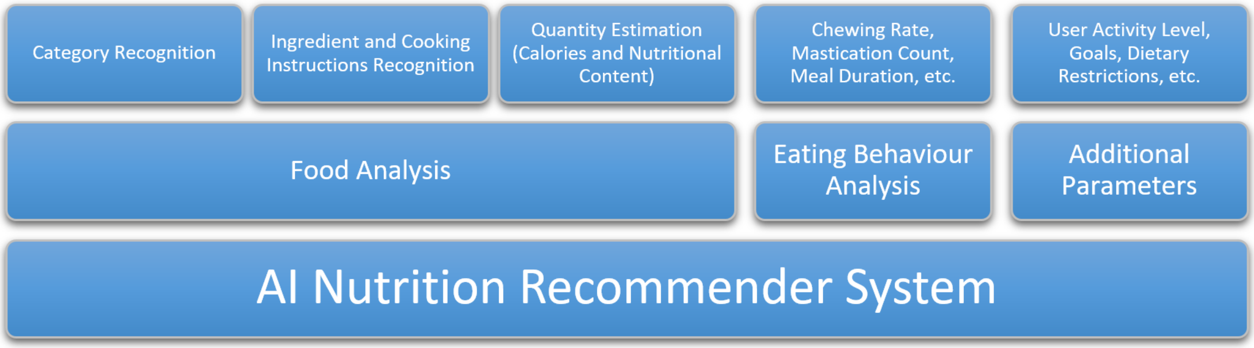
Category Recognition
Ingredient and Cooking Instructions Recognition
Quantity Estimation (Calories and Nutritional Content)
Chewing Rate, Mastication Count, Meal Duration, etc.
User Activity Level, Goals, Dietary Restrictions, etc.
Food Analysis
Eating Behaviour Analysis
Additional Parameters
AI Nutrition Recommender System

Figure 1: Categorization of AI nutrition recommender systems into task-specific components.

each category is further analyzed and the most important relevant literature is presented.

2.1 Food Category Recognition

This category of food analysis is concerned with recognizing the class that an item of interest belongs to. The majority of the existing literature regarding automated food analysis belongs to this category, since the first data-sets that were made publicly available only contained information regarding the class of the items. Due to the same reason, the dominant focus area of existing methods is image analysis, but approaches based on audio [30, 34, 45], motion [30], colour [19] and odour [35] also exist. Technically, methods concerned with multiple tasks at the same time (e.g., food recognition and calorie estimation) could also be included in this category, but in order to avoid duplication, they will be described in sections 2.2 and 2.3. Regarding food categories, they can range from very broad ones, like *rice*, to very specific, such as *chicken feet with black bean sauce*. Another distinguishing factor for the methods of this category is the setting under which the food categorization was performed. On one hand, there are methods that operate on a completely unconstrained setting (no information about the food source is known), while on the other hand, there are methods that have confined the recognition setting to menu items of specific restaurants. Further analysis of some of the aforementioned methods, as well as other important works is presented next. Additionally, some of the most commonly used data-sets for this task are presented in Table 1.

Regarding the image-based methods, there is a clear distinction between earlier methods, which used traditional image representations, and methods developed during the last few years, and are almost exclusively based on learned representations by convolutional neural networks (CNNs). For instance, Chen et al. [10] used representations based on color histograms and SIFT features, which were then processed by a support vector machine (SVM) classifier. Yang et al. [50] used an SVM classifier, trained on statistics of pairwise local features, such as the distance between 2 image pixels and the orientation of the line connecting 2 pixels. In [20] a 2-layer convolutional neural network (CNN) was used for the tasks of food detection (food / non-food classification) and food recognition. In both tasks the CNN outperformed an SVM classifier trained with traditional features. Kawano and Yanai [22] used a CNN architecture as a feature extraction mechanism. The features of the second to last layer were used in order to train an SVM classifier to distinguish among 100 food classes. Traditional features were also used for comparison. The authors noted that the learned CNN features were the best single representation, but the combination of learned and traditional features yielded the best overall result. Wu et al. [48] proposed a multi-level loss function which guides the underlying machine learning algorithm to classify the same object using multiple semantic levels of the provided hierarchy. For example, an image depicting *chicken wings* will simultaneously belong to the *chicken* and *meat or poultry* categories, which the algorithm must jointly optimize. The authors mentioned that this approach improves classification performance, because the algorithm tends to stay within the same semantic category, even if the fine-grained category was incorrectly predicted. This loss function, when utilized with the GoogLeNet CNN architecture, improved the baseline accuracy in both Food-101 and 5-Chain data-sets. Martinel et al. [25] adopted the wide residual networks architecture proposed in [51] and combined it with a slice convolution and pooling branch in order to arrive at their final architecture. Slice convolution is a regular convolution operation performed using filters that are as wide as the input image, instead of being square (e.g., 3×3). When tested on the UEC-Food100, UEC-Food256 and Food-101 data-sets, this method achieved an accuracy of 89.6%, 83.2% and 90.3% respectively.

Regarding food recognition from sources other than images, Shuzo et al. [45] used a bone conduction microphone in order to capture the sounds produced while eating. One of their test cases was to classify foods of different texture (hardness, springiness and brittleness). In [34] microphones were employed that simultaneously captured in-ear and environmental sounds. Using a finite-state grammar decoder based on the Viterbi algorithm for the recognition task, they were able to classify new samples into the categories: potato chips, peanut, walnut, carrot, apple, chocolate, pudding and drink with 66% accuracy. Kadomura et al. [19] used a sensing fork with an embedded RGB sensor in order to classify 17 food types from Japanese, Chinese and Western cuisines with an overall F-measure of 87.5%.

Table 1: Common data-sets used for the task of food category recognition.

Name	Classes	Images	Reference
Food-101	101	101,000	[4]
UEC-FOOD100	100	14,361	[26]
UEC-FOOD256	256	25,088	[21]

Table 2: Common data-sets used for food ingredient and cooking instructions recognition.

Name	Recipes	Images	Reference
Recipe1M	1,029,720	887,706	[43]
VIREO Food-172	65,284	110,241	[8]
Yummly-28K	27,638		[28]
Yummly-66K	66,615		[29]

2.2 Food Ingredient and Cooking Instructions Recognition

In the last few years, the emergence of data-sets with accompanying information regarding the ingredients used in the recipe and the necessary cooking instructions to reproduce it, has brought about interesting new research ideas and directions. Methods belonging to this category are mainly concerned with creating appropriate representations for ingredients, cooking instructions and images and combining them in such a way so as to be possible to transition from one representation to another. This is most commonly employed for ingredient and cooking instructions recognition and retrieval from food images. The distinction between methods that perform recognition (classification) and retrieval is important, as the end-goal in each case is different. In the case of classification, a method predicts ingredients and instructions from a given image. In the case of retrieval, on the other hand, a method searches for and retrieves the closest ingredients and instructions from a given data-set for an input image. It is worth mentioning that in the case of retrieval, the opposite problem has also been studied; to retrieve images based on instructions and ingredients. Also, since this research direction is relatively recent, neural networks have dominated the field as the method of choice. Below we present some notable works in this field and list the most common data-sets for this task in Table 2.

In [9] a single CNN architecture was proposed that takes as input an image of a food item, extracts features at different resolution levels in order to obtain fine-grained information, and predicts the ingredients, cutting method and cooking method. This triplet is then used for text-based recipe retrieval. The evaluation results showed that the additional tasks of cutting and cooking method prediction were beneficial to the ingredient prediction as well. This method was also able to find the location of each detected ingredient in the input image. Min et al. [29] proposed a deep belief network (DBN) that is able to learn a joint representation for images and ingredients. An interesting detail of this work is that visible ingredients in the image were modelled differently than non-visible ingredients, and this proved to be beneficial for the performance of the network. This method was then tested on cuisine classification (e.g., Mediterranean), image retrieval (given the ingredients, course and cuisine information, find the most relevant image) and ingredient, course and cuisine prediction (given a food image). In the work of Salvador et al. [43], a joint network architecture that embeds images, ingredients, cooking instructions into a common space was presented. Image representations were obtained using a traditional convolutional architecture (ResNet-50). Ingredient representations resulted from a bi-directional recurrent neural network (RNN) processing word2vec encodings [27] of the ingredients. Finally, instruction representations were the result of skip-thought encodings [23] being processed by another RNN. This architecture was then used for image to recipe and recipe to image retrieval, where recipe is considered the pair of ingredients and cooking instructions. When compared to human performance on the image to recipe task, their evaluation showed comparable results. Carvalho et al. [6] adopted the network architecture presented in [43] and proposed a new optimization objective in order to improve retrieval performance. The new objective function combines the classification and retrieval tasks at the same level, therefore eliminating the need for an additional classification layer in the model architecture. In a continuation of their previous work, Salvador et al. [42] proposed a network architecture that predicts a list of ingredients from a given image. Then, both the image and the predicted ingredients are used in order to predict the cooking instructions for the recipe. Compared to their previous work, where the network architecture was used for retrieval, they showed that this method improves the accuracy of the results.

2.3 Food Quantity Estimation

The purpose of methods included in this category is to obtain an estimation of the quantity of consumed food or of its nutritional content and calories. The problem of calorie estimation has been approached in two different ways. One group of methods works by first obtaining an estimation of the food volume, or of its ingredients, and then translating this information into calories. The second group performs a direct estimation of the calories without any intermediate representation. Methods belonging to both groups are presented next, while the available data-sets for this task are presented in Table 3.

Puri et al. [39] employed a combination of color and texture features at different scales, which were used for training an SVM-based AdaBoost classifier for segmenting and classifying the food items on a given image. To obtain an estimation of the volume, three pictures of each dish were used (where a checker-board was also visible for scale), for which key-points were detected and matched using Harris corners and RANSAC. Finally, a dense 3D reconstruction was computed. Using a small data-set for evaluation, their method achieved an average error of 6% in volume estimation. In [36], an SVM classifier trained with color, texture, size, and shape features is used for the task of food recognition. Then, the area that each food item occupies is measured and a side photo of the same dish is used in order to calculate depth and volume information. It should be noted that users’ thumb is also visible in the images for scale assessment. Finally, the calories for each item in the image are calculated using predefined tables for food density and caloric content. A similar architecture is presented in [11], with the difference, however, that instead of calculating calories based on the whole meal, the system estimates the leftover quantity and subtracts the corresponding calories. In the work of Myers et al. [31], a CNN architecture (GoogLeNet) was used for food recognition, both in an unconstrained setting as well as a setting of menu items from 23 restaurants. Another network was used for segmenting the image into individual items and a third network was employed for the task of depth estimation. The last network was trained with pairs of RGB and depth images in order to learn depth estimation. The final step of translating the volume of each food item into calories was performed using predefined tables from the USDA National Nutrient Database [12]. Ege and Yanai [15] presented a multi-task CNN architecture for simultaneous food recognition and calorie estimation. The authors employed a VGG-16 network architecture, which was shared between the two tasks up to the last two layers. These last layers were independent and produced the class probabilities on the one side, and calories on the other. Their evaluation results showed a 28% relative error in the calorie estimation task. Fang et al. [16] employed a Conditional Generative Adversarial Network (CGAN) in order to learn an energy mapping (calories) from a single input image. Given a noise vector, and conditioned on the input RGB image, the network learns to produce an energy distribution image, where the value of each pixel corresponds to the caloric content of the displayed food item. The authors tested two architectures for the network and concluded that the U-Net produced the best results, with a relative energy estimation error of 10.9%. In contrast to the image-based works presented so far, Mirtchouk et al. [30] presented a method based on audio, motion and high-level features for food category recognition and quantity estimation. Six participants were instructed to eat foods of their choice in any quantity they wanted, while wearing an earbud with internal and external microphones, a smart-watch in each hand and Google Glass smart-glasses. Random forest models were trained using features extracted from these sensors, as well as high-level annotation features such as the number of chews after an intake and the duration of the intake window, in order to predict food category and estimate weight for each intake.

3 EATING BEHAVIOUR ANALYSIS

This section is concerned with methods that analyze human eating behaviour, and more specifically with chewing rate, mastication count, overall meal duration estimation, as well as distinguish between eating events and non-eating-related events. Unlike food category recognition, where most methods are based on image analysis, a number of different approaches have been used in the literature regarding eating behaviour analysis. These are based on weight [7], audio [2, 45], hand motion [14], image [5] and jaw motion [44] analysis, to name a few. Below we describe some of the methods in more detail. Available data-sets for this task are presented in Table 4.

Chang et al. [7] employed embedded RFID and weighing sensors underneath a dining table. The RFID sensor is used for determining the presence of specific tabletop objects, while the weighting sensors can detect changes to the food containers with half a gram resolution. An eating event is recognized when there is a decrease in the weight of the container corresponding to each individual. The audio-based method proposed by Shuzo et al. [45], described previously, was also able to estimate mastication count and meal duration. Bi et al. [2] used a neck-worn microphone for recording audio signals during eating. Initially, hidden Markov models (HMMs) were used for identifying chewing and swallowing events based on Mel frequency cepstrum coefficients (MFCCs). Then, features such as the zero crossing rate, the energy of different frequency bands and the fractal dimension were extracted and processed by a decision tree for classifying food types. In [14] a wrist-worn device with an embedded gyroscope sensor was used in order to estimate the number of bites. The method is based on the observation that a specific wrist motion takes place before each bite, as part of the movement of bringing food to the mouth. Therefore, an event was detected if the roll velocity surpassed a first threshold, then went below another threshold and also satisfied two timing requirements, put in place to reduce false detections. Cadavid et al. [5] used a video-based approach for detecting chewing events. An active appearance model (AAM) was used in order to capture shape and appearance information regarding a person's face. Frequency analysis of the active appearance model parameters and dimensionality reduction were used for obtaining the final representations, which were then processed by an SVM classifier. The sensing fork developed by [19], besides food recognition, was also able to detect biting events through the use of a conductive probe. Sazonov and Fontana [44] investigated the use of a jaw motion sensor, placed below a person's ear, for the detection of chewing events. The authors extracted a set of 25 time and frequency features, such as the number of zero crossings and the peak frequency. Classification was handled by an SVM model. This methodology is different from [45] for example, since it uses voltage signals generated by jaw motion through a piezoelectric film element sensitive to stress, not audio. Finally, the recent works of Prioleau et al. [37] and Vu et al. [47] provided a comprehensive review of the relevant literature.

4 AI NUTRITION RECOMMENDER SYSTEMS

This section initially provides a description of the components that an idealized AI nutrition recommender system would have. Each component is then compared to state-of-the-art methods and an assessment of its feasibility with current technology is provided. Finally, recent literature and EU-funded projects relevant to this task are presented, including the approach followed by the PROTEIN project, in which the authors of this work participate.

To begin with, an ideal AI nutrition recommender system would be able to identify the type of food consumed by the user, providing as detailed a description as possible. For example, identifying a dish as *Chicken Salad with Wild Rice* instead of *Salad*. As described previously in Section 2.1, this field of study has received the most attention from the research community and is in a mature state, with standardized large-scale data-sets being available for evaluation purposes. Although recent approaches in food category recognition have reported results above the 90% mark in Food-101, good evaluation results on a data-set equivalent in scale to ImageNet [13], such as Recipe1M, would be needed in order to get closer in fulfilling this requirement.

Table 3: Data-sets regarding food quantity estimation.

Name	Images	Calories	Macro-nutrients	Micro-nutrients	Reference
Food-pics	896	✓	✓	-	[3]
Menu-Match	646	✓	-	-	[1]
Yummly-28K *	27,638	✓	✓	✓	[28]
Recipe1M **	887,706	✓	✓	✓	[43]

* Yummly-28K contains nutritional information for about 90% of the recipes.
** Recipe1M contains nutritional information for about 50k recipes [40].

Table 4: Available data-sets for eating behaviour analysis.

Name	Description	Reference
Food Intake Cycle	Accelerometer and gyroscope data from wrist-worn smartwatches. Hand micromovement annotations during eating.	[24]
SPLENDID chewing detection	PPG, audio and acceleration data from body-worn sensors. Eating event annotations.	[33]

Next, an ideal recommender system would have to provide an accurate estimation of the ingredients, calories and nutrients present in the food. These three attributes constitute the building blocks of the recommender system, as any attempt at creating personalized nutrition plans depends on at least one of them. This requirement is considerably harder to satisfy compared to the previous one, because of the inherent difficulties of the task at hand. For example, there is the issue of visually occluded ingredients. There are recipes where some of the ingredients are not visible in the final form of the dish, e.g., olive oil in *Moussaka*, or almost every ingredient in *Soup*. As discussed in Section 2.2, an interesting approach towards this problem was presented by Min et al. [29], where visible and non-visible ingredients were modeled separately within their architecture, providing improved performance. Another issue is calorie estimation based on predefined tables. As presented in Section 2.3, most methods first identify the displayed food, then obtain a volume estimate and finally translate this information into calories using predefined tables. Although this may be a good strategy in the absence of any dish-specific information, such tables provide food densities and calories for a 'typical' interpretation of the dish, which may be quite different from the one at hand. Another issue related to this requirement is nutrient estimation without knowledge of the cooking method. Since nutrients are affected by the way the food is cooked, estimations that disregard this information, and are based on raw ingredients instead, could still be far off the actual value. In Section 2.2, the method of Chen et al. [9] headed towards this direction by incorporating ingredient cooking method estimation into their architecture. Regarding the state of each aforementioned area, ingredient recognition is currently under active development, after the recent introduction of large-scale data-sets for this purpose (see Table 2). The methodologies presented in Section 2.2 show promising results, but further improvements are needed if such systems are to be deployed for everyday use. On the other hand, it seems that the area of calorie and nutrient estimation has not reached a mature state yet, as literature approaches are often reporting evaluation results on small-scale private data-sets. Although not advertised as such, the Yummly-28K and Recipe1M data-sets could also be used for this task (see Table 3), providing up to two orders of magnitude more samples than the traditional data-sets.

Such a system would also be able to analyze the way food is consumed by the user, extracting information regarding the time of day the user eats, the duration of each meal, the chewing rate and the response of the user to the specific food items (e.g., blood glucose levels). As discussed previously in Section 3, approaches targeting some of these areas already exist, providing promising evaluation results as well. However, the lack of standardized data-sets for this task makes it difficult to assess progress, as the reported evaluation results are mostly in private data-sets. Public data-sets that could be used for this task are presented in Table 4. Moreover, as the challenges in these areas are much easier to deal with than those in ingredient, calorie and nutrient estimation, standardization would help these areas progress rapidly.

Finally, an ideal AI recommender system would be able to modify its recommendations based on the behaviour, preferences and needs of the user. This includes the omission or substitution of specific ingredients from recipes and the re-calculation of nutrition and activity plans based on the (changing) goals of the user (weight loss, muscle gain or eating healthy), or on the deviation of the user from the provided plans. As we will see next, current literature approaches are already concerned with some of these requirements, but there is still a lot of margin for improvement.

Regarding the relevant literature on this topic, Ge et al. [18] presented a recommender system based on user preferences towards recipes and ingredients. The system uses this information to generate candidate recipes, which the user can select for cooking or request that they be adjusted, such as make a recipe spicier. When changes are requested, the system then provides updated recommendations. The system statically estimates caloric needs based on user profile data, such as weight, height and activities. Finally, users have the ability to influence the recommendations towards tasty or healthy foods. The recommender system proposed in [49], initially asks the users about any dietary restrictions they may have, such as Kosher, and about their nutritional expectations towards calories, protein and fat (increase, decrease or maintain). Then the system presents to the user a list of food images and the user must select the most preferred ones. Based on the information gathered so far, an iterative process is then initiated with as many as 13 repetitions, where the system presents food images to the user, gathers the provided preferences and updates its state regarding which items are more likely to appeal to the user. An online learning framework is proposed for this task, with the food (image) similarity sub-component being handled by a multi-task siamese convolutional neural network. The provided results of a user study showed a 73% acceptance rate for the proposed recommender, compared to 51% for the baseline. Ribeiro et al. [41] proposed a meal recommender that first estimates the nutritional requirements of the user, then filters the available food items based on several rules and finally scales the recipe ingredients to match the caloric needs of the user. Nutritional requirements are calculated through user-provided information, such as age, sex, weight, height and activity level. The last one is measured using a Fitbit activity tracker. Food items are selected for each meal based on criteria such as the food and dietary preferences of the user, the avoidance of recipe repetition within the same week and promoting the consumption of meat dishes for lunch and fish for dinner. The system also has the ability to create shopping lists with the ingredients of the recommended food items. Based on the results of a user study, 70% of participants replied that they would follow the recommended weekly food plan.

EU-funded projects have also been concerned with the topic of personalized nutrition recommendations. The FOOD4ME project [17], which was completed in 2015, investigated the impact of various levels of personalized dietary recommendations on health markers and weight. Personalization levels varied by the amount and type of data taken into account by the system. The first level included the current diet, weight, BMI and physical activity of the person, the second level included the first level information plus blood markers (e.g., glucose and total cholesterol), while the third level included information from the previous levels plus genetic markers. The results showed that personalized nutrition is effective, but level two and three provided no additional benefit.

Regarding ongoing research projects, NUTRISHIELD [32] aims at providing personalized nutrition by taking into account factors such as phenotype, genome expression, microbiome composition and health condition of the person. Stance4Health [46] aims to develop personalized nutrition recommendations based on users' health status, microbiota composition, food preferences, lifestyle and budget. A wearable device that tracks body composition, physical activity and sleeping hours also provides information to the platform. Among the target groups of the platform are people with celiac disease, food allergies and people that are overweight. Finally, the PROTEIN project [38] intends to deliver a personalized nutrition and activity recommender system, with early warning functionality in case of suboptimal eating patterns, which will also help users make healthier choices in supermarkets and restaurants. A variety of sensors will be employed in order to fulfill these goals, including smart-watches (physical activity and eating rate analysis), smart food scales (eating rate, meal size and duration analysis), volatile organic compounds sensors (health and nutritional status analysis) and continuous glucose monitoring sensors.

5 CONCLUSION

This work provided an overview of existing AI nutrition recommender systems, a field that has experienced substantial growth in the last few years. A categorization of such systems into task-specific components was presented, along with approaches concerned with each component and relevant data-sets. An assessment of the feasibility of implementing an ideal AI nutrition recommender system using current methods was also provided, with the general conclusion being that some of the required components have not reached a mature state yet.

ACKNOWLEDGMENTS

This work was supported by the European Project: PROTEIN Grant no. 817732 within the H2020 Research and Innovation Programme.

REFERENCES

[1] Oscar Beijbom, Neel Joshi, Dan Morris, Scott Saponas, and Siddharth Khullar. 2015. Menu-match: Restaurant-specific Food Logging from Images. In *Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision*. IEEE, 844–851.

[2] Yin Bi, Mingsong Lv, Chen Song, Wenyao Xu, Nan Guan, and Wang Yi. 2016. Autodietary: A Wearable Acoustic Sensor System for Food Intake Recognition in Daily Life. *IEEE Sensors Journal* 16, 3 (2016), 806–816.

[3] Jens Blechert, Adrian Meule, Niko A Busch, and Kathrin Ohla. 2014. Food-pics: An Image Database for Experimental Research on Eating and Appetite. *Frontiers in Psychology* 5 (2014), 617.

[4] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–Mining Discriminative Components with Random Forests. In *Proceedings of the 2014 European Conference on Computer Vision*. Springer, 446–461.

[5] Steven Cadavid, Mohamed Abdel-Mottaleb, and Abdelsalam Helal. 2012. Exploiting Visual Quasi-periodicity for Real-time Chewing Event Detection Using Active Appearance Models and Support Vector Machines. *Personal and Ubiquitous Computing* 16, 6 (2012), 729–739.

[6] Micael Carvalho, Rémi Cadène, David Picard, Laure Soulier, Nicolas Thome, and Matthieu Cord. 2018. Cross-modal Retrieval in the Cooking Context: Learning Semantic Text-image Embeddings. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*. ACM, 35–44.

[7] Keng-hao Chang, Shih-yen Liu, Hao-hua Chu, Jane Yung-jen Hsu, Cheryl Chen, Tung-yun Lin, Chieh-yu Chen, and Polly Huang. 2006. The Diet-aware Dining Table: Observing Dietary Behaviors Over a Tabletop Surface. In *Proceedings of the 2006 International Conference on Pervasive Computing*. Springer, 366–382.

[8] Jingjing Chen and Chong-Wah Ngo. 2016. Deep-based Ingredient Recognition for Cooking Recipe Retrieval. In *Proceedings of the 24th ACM International Conference on Multimedia*. ACM, 32–41.

[9] Jing-jing Chen, Chong-Wah Ngo, and Tat-Seng Chua. 2017. Cross-modal Recipe Retrieval with Rich Food Attributes. In *Proceedings of the 25th ACM International Conference on Multimedia*. ACM, 1771–1779.

[10] Mei Chen, Kapil Dhingra, Wen Wu, Lei Yang, Rahul Sukthankar, and Jie Yang. 2009. PFID: Pittsburgh Fast-food Image Dataset. In *Proceedings of the 2009 IEEE International Conference on Image Processing (ICIP)*. IEEE, 289–292.

[11] Gianluigi Ciocca, Paolo Napoletano, and Raimondo Schettini. 2015. Food Recognition and Leftover Estimation for Daily Diet Monitoring. In *Proceedings of the 2015 International Conference on Image Analysis and Processing*. Springer, 334–341.

[12] USDA National Nutrient Database. 2014. USDA National Nutrient Database for Standard Reference, Release 27 (revised). http://www.ars.usda.gov/ba/bhnrc/ndl

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale Hierarchical Image Database. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 248–255.

[14] Yujie Dong, Adam Hoover, Jenna Scisco, and Eric Muth. 2012. A New Method for Measuring Meal Intake in Humans via Automated Wrist Motion Tracking. *Applied Psychophysiology and Biofeedback* 37, 3 (2012), 205–215.

[15] Takumi Ege and Keiji Yanai. 2017. Simultaneous Estimation of Food Categories and Calories with Multi-task CNN. In *Proceedings of the 15th IAPR International Conference on Machine Vision Applications (MVA)*. IEEE, 198–201.

[16] Shaobo Fang, Zeman Shao, Runyu Mao, Chichen Fu, Edward J Delp, Fengqing Zhu, Deborah A Kerr, and Carol J Boushey. 2018. Single-view Food Portion Estimation: Learning Image-to-energy Mappings using Generative Adversarial Networks. In *Proceedings of the 25th IEEE International Conference on Image Processing (ICIP)*. IEEE, 251–255.

[17] FOOD4ME. 2011. Personalised Nutrition: An Integrated Analysis of Opportunities and Challenges. https://cordis.europa.eu/project/rcn/98657/factsheet/en

[18] Mouzhi Ge, Francesco Ricci, and David Massimo. 2015. Health-aware Food Recommender System. In *Proceedings of the 9th ACM Conference on Recommender Systems*. ACM, 333–334.

[19] Azusa Kadomura, Cheng-Yuan Li, Koji Tsukada, Hao-Hua Chu, and Itiro Siio. 2014. Persuasive Technology to Improve Eating Behavior Using a Sensor-embedded Fork. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 319–329.

[20] Hokuto Kagaya, Kiyoharu Aizawa, and Makoto Ogawa. 2014. Food Detection and Recognition Using Convolutional Neural Network. In *Proceedings of the 22nd ACM International Conference on Multimedia*. ACM, 1085–1088.

[21] Yoshiyuki Kawano and Keiji Yanai. 2014. Automatic Expansion of a Food Image Dataset Leveraging Existing Categories with Domain Adaptation. In *Proceedings of the 2014 European Conference on Computer Vision*. Springer, 3–17.

[22] Yoshiyuki Kawano and Keiji Yanai. 2014. Food Image Recognition with Deep Convolutional Features. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. ACM, 589–593.

[23] Ryan Kiros, Yukun Zhu, Ruslan R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought Vectors. In *Advances in Neural Information Processing Systems (NIPS)*. 3294–3302.

[24] Konstantinos Kyritsis, Christos Diou, and Anastasios Delopoulos. 2017. Food Intake Detection from Inertial Sensors using LSTM Networks. In *Proceedings of the 2017 International Conference on Image Analysis and Processing*. Springer, 411–418.

[25] Niki Martinel, Gian Luca Foresti, and Christian Micheloni. 2018. Wide-slice Residual Networks for Food Recognition. In *Proceedings of the 2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 567–576.

[26] Yuji Matsuda, Hajime Hoashi, and Keiji Yanai. 2012. Recognition of Multiple-Food Images by Detecting Candidate Regions. In *Proceedings of the 2012 IEEE International Conference on Multimedia and Expo*. IEEE, 25–30.

[27] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781* (2013).

[28] Weiqing Min, Shuqiang Jiang, Jitao Sang, Huayang Wang, Xinda Liu, and Luis Herranz. 2017. Being a Supercook: Joint Food Attributes and Multimodal Content Modeling for Recipe Retrieval and Exploration. *IEEE Transactions on Multimedia* 19, 5 (2017), 1100–1113.

[29] Weiqing Min, Shuqiang Jiang, Shuhui Wang, Jitao Sang, and Shuhuan Mei. 2017. A Delicious Recipe Analysis Framework for Exploring Multi-modal Recipes with Various Attributes. In *Proceedings of the 25th ACM International Conference on Multimedia*. ACM, 402–410.

[30] Mark Mirtchouk, Christopher Merck, and Samantha Kleinberg. 2016. Automated Estimation of Food Type and Amount Consumed from Body-worn Audio and Motion Sensors. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 451–462.

[31] Austin Myers, Nick Johnston, Vivek Rathod, Anoop Korattikara, Alex Gorban, Nathan Silberman, Sergio Guadarrama, George Papandreou, Jonathan Huang, and Kevin Murphy. 2015. Im2Calories: Towards an Automated Mobile Vision Food Diary. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 1233–1241.

[32] NUTRISHIELD. 2018. Fact-based Personalised Nutrition for the Young. https://cordis.europa.eu/project/rcn/218675/factsheet/en

[33] Vasileios Papapanagiotou, Christos Diou, Lingchuan Zhou, Janet van den Boer, Monica Mars, and Anastasios Delopoulos. 2017. A Novel Chewing Detection System Based on PPG, Audio, and Accelerometry. *IEEE Journal of Biomedical and Health Informatics* 21, 3 (2017), 607–618.

[34] Sebastian Päßler, Matthias Wolff, and Wolf-Joachim Fischer. 2012. Food Intake Monitoring: An Acoustical Approach to Automated Food Intake Activity Detection and Classification of Consumed Food. *Physiological Measurement* 33, 6 (2012), 1073.

[35] Michele Penza, Gennaro Cassano, Fiorentino Tortorella, and Giuseppe Zaccaria. 2001. Classification of Food, Beverages and Perfumes by WO3 Thin-film Sensors Array and Pattern Recognition Techniques. *Sensors and Actuators B: Chemical* 73, 1 (2001), 76–87.

[36] Parisa Pouladzadeh, Shervin Shirmohammadi, and Rana Al-Maghrabi. 2014. Measuring Calorie and Nutrition from Food Image. *IEEE Transactions on Instrumentation and Measurement* 63, 8 (2014), 1947–1956.

[37] Temiloluwa Prioleau, Elliot Moore II, and Maysam Ghovanloo. 2017. Unobtrusive and Wearable Systems for Automatic Dietary Monitoring. *IEEE Transactions on Biomedical Engineering* 64, 9 (2017), 2075–2089.

[38] PROTEIN. 2018. Personalized Nutrition for Healthy Living. https://cordis.europa.eu/project/rcn/218540/factsheet/en

[39] Manika Puri, Zhiwei Zhu, Qian Yu, Ajay Divakaran, and Harpreet Sawhney. 2009. Recognition and Volume Estimation of Food Intake using a Mobile Device. In *Proceedings of the 2009 Workshop on Applications of Computer Vision (WACV)*. IEEE, 1–8.

[40] Recipe1M. Accessed: April 1st 2019. A Dataset for Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. http://pic2recipe.csail.mit.edu/

[41] David Ribeiro, João Machado, Jorge Ribeiro, Maria João M Vasconcelos, Elsa F Vieira, and Ana Correia de Barros. 2017. SousChef: Mobile Meal Recommender System for Older Adults. In *ICT4AgeingWell*. 36–45.

[42] Amaia Salvador, Michal Drozdzal, Xavier Giro-i Nieto, and Adriana Romero. 2018. Inverse Cooking: Recipe Generation from Food Images. *arXiv preprint arXiv:1812.06164* (2018).

[43] Amaia Salvador, Nicholas Hynes, Yusuf Aytar, Javier Marin, Ferda Ofli, Ingmar Weber, and Antonio Torralba. 2017. Learning Cross-Modal Embeddings for Cooking Recipes and Food Images. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3020–3028.

[44] Edward S Sazonov and Juan M Fontana. 2012. A Sensor System for Automatic Detection of Food Intake Through Non-invasive Monitoring of Chewing. *IEEE Sensors Journal* 12, 5 (2012), 1340–1348.

[45] Masaki Shuzo, Shintaro Komori, Tomoko Takashima, Guillaume Lopez, Seiji Tatsuta, Shintaro Yanagimoto, Shin'ichi Warisawa, Jean-Jacques Delaunay, and Ichiro Yamada. 2010. Wearable Eating Habit Sensing System Using Internal Body Sound. *Journal of Advanced Mechanical Design, Systems and Manufacturing* 4, 1 (2010), 158–166.

[46] Stance4Health. 2018. Smart Technologies for Personalised Nutrition and Consumer Engagement. https://cordis.europa.eu/project/rcn/218510/factsheet/en

[47] Tri Vu, Feng Lin, Nabil Alshurafa, and Wenyao Xu. 2017. Wearable Food Intake Monitoring Technologies: A Comprehensive Review. *Computers* 6, 1 (2017), 4.

[48] Hui Wu, Michele Merler, Rosario Uceda-Sosa, and John R Smith. 2016. Learning to Make Better Mistakes: Semantics-aware Visual Food Recognition. In *Proceedings of the 24th ACM International Conference on Multimedia*. ACM, 172–176.

[49] Longqi Yang, Cheng-Kang Hsieh, Hongjian Yang, John P Pollak, Nicola Dell, Serge Belongie, Curtis Cole, and Deborah Estrin. 2017. Yum-me: A Personalized Nutrient-based Meal Recommender System. *ACM Transactions on Information Systems (TOIS)* 36, 1 (2017), 7.

[50] Shulin Yang, Mei Chen, Dean Pomerleau, and Rahul Sukthankar. 2010. Food Recognition using Statistics of Pairwise Local Features. In *Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2249–2256.

[51] Sergey Zagoruyko and Nikos Komodakis. 2016. Wide Residual Networks. In *Proceedings of the 27th British Machine Vision Conference*. British Machine Vision Association, 87.1–87.12.