An Intelligent Nutritional Assessment System

Yulia Eskin and Alex Mihailidis
Intelligent Assistive Technology and Systems Lab
University of Toronto, Canada

Abstract

Higher life expectancies lead to an increased prevalence of dementia in older adults, which is projected to rise dramatically in the future. The link between malnutrition and dementia highlights the need to closely monitor nutrition as early as possible. However, current self-report assessment methods are labor-intensive, time-consuming and inaccurate. Technology has the potential of assisting in nutritional analysis by alleviating the cognitive load of recording food intake and lessening the burden of care for the elderly. Therefore, we propose an intelligent nutritional assessment system that will monitor the dietary patterns of older adults with dementia at their homes. Our computer vision-based system consists of food recognition and portion estimation algorithms that, together, provide nutritional analysis of an image of a meal. We create a novel food image dataset on which we achieve an 87.2% recognition accuracy. We apply several well-known segmentation and recognition algorithms and analyze their suitability to the food recognition problem.

Introduction

The World Health Organization (WHO) states that dementia is one of the most serious threats to aging populations (Organization 2002). Alzheimer’s Disease International (ADI) (International 2010) defines dementia as a syndrome that can be caused by a number of progressive disorders that affect memory, thinking, behavior and the ability to perform everyday activities. The 2010 World Alzheimer Report (International 2010) estimated that there were 35.6 million people suffering from dementia in the world, doubling by 2030 and more than tripling by 2050. Moreover, the total worldwide estimated costs of dementia were US$604 billion in 2010 and are set to increase by 85% by 2030.

The WHO states that good nutritional status is an important determinant of quality of life due to its effects on the nervous system and brain function (Organization 2002). Many of the diseases from which older adults suffer are associated with the natural aging process but are compounded by dietary factors (Organization 2011). Malnutrition and dehydration are major issues in elderly populations worldwide, particularly for those suffering from dementia due to loss of appetite and a decline in intake. Malnutrition, in turn, leads to a decreased immune function and magnifies risks of chronic diseases. The link between dementia and malnutrition is examined in studies (Organization 2002) that point to the correlation between concentrations of micronutrients and the Mini-Mental Status Examination (a cognitive function screening method for dementing disorders) (Folstein, Folstein, and McHugh 1975).

Fortunately, dietary factors that affect morbidity and mortality are modifiable. While proper nutrition appears to be an important factor in disease prevention or delay, the mechanisms are still unclear. To shed light on nutrition’s impact on disease, reliable and accurate nutritional data is essential, especially in community-dwelling older populations. Current studies on nutritional interventions use food-intake assessment methods which include 24-hour recalls, diet records and food-frequency questionnaires (Organization 2002). However, these rely on self-reporting and were shown to be unreliable, underestimating intake by as much as 37% or more (Martin, Kaya, and Gunturk 2009). Additionally, untrained people struggle to provide accurate portion data, accounting for the largest source of error in intake reporting (Martin, Kaya, and Gunturk 2009). Methods relying on physical measures, such as the body mass index, body composition and biomarkers, suffer from incompatibility to older populations due to the bias to data gathered from young adults (Organization 2002). Aside from the pen-and-paper methods being labor-intensive and time-consuming for both the individual and the nutritionist, a major concern is that no measure of ground truth exists to which one could compare the data provided (Zhu et al. 2010). Moreover, older adults who suffer from dementia are not able to reliably report their own food intake due to memory problems. Consequently, further burden falls on their caregivers.

Given the importance of understanding dietary patterns of older adults and the challenges with current food-intake methods, we propose an intelligent nutritional assessment system that would eventually be placed in the home. This system would monitor dietary-patterns and provide nutritional analysis through the application of computer vision and machine learning. The novelty of our work lies in the design of such a system for the challenging home setting, and the research of appropriate machine learning algorithms.
to ensure adaptability and continuous learning.

Related Work
Several research groups have explored the food recognition problem, however, none targeted the population of people with dementia to support better eating behavior. Most notably, Zhu et al. (Zhu et al. 2010) developed semi-automatic food recognition and portion estimation algorithms, and an intensive user-interface for cell-phones, targeted towards adolescents suffering from obesity. Similarly, Puri et al. (Puri et al. 2009) required a user to capture cell-phone images of their food and provide speech input. Others focused on creating a comprehensive dataset of images and videos of fast-foods (Chen et al. 2009). In (Yang et al. 2010) Yang et al. relied on the standardized making of fast-foods to develop a representation of a food item in terms of the spatial arrangement of its ingredients. Wu et al. (Wu and Yang 2009) analyzed eating videos but admitted their method was too slow for real-life applicability. Lastly, Veggie Vision (Bolle et al. 1996) was a system developed for recognizing produce items in stores and involved special hardware. The major drawbacks of these systems are that they require a high level of user-interaction, are not fully automated and often do not provide any nutritional analysis. Moreover, their focus is on universal design rather than on adaptability to an individual in a home setting.

System Overview
The system developed in this research consists of a webcam placed above a plate of food and a computer to run the developed algorithms. Starting with a food image database of commonly consumed foods, and an image of a meal taken by the user, the system would proceed to analyze it by: (1) segmenting and (2) recognizing the food objects, and (3) estimating the portion size of each one. Then, the nutritional intake of the meal will be displayed to the user by linking the food labels and portion data to a food composition table and extracting the relevant information. In this paper, we only focus on the segmentation and recognition components.

At this preliminary stage of research, no interaction with a user is required except for the meal images the user has to provide. However, there is non-visual information in an image that is essential for accurate nutritional analysis, e.g. distinguishing soy milk from regular milk. Therefore, we envision an adaptable system that would exploit user feedback in order to learn the user-specific dietary preferences to improve its performance (particularly during recognition) and expand its food vocabulary. Personalization is critical since even a large food database could not encompass the endless variety of foods and their preparations. Hence, rather than building a universal food recognition system, we envision one that could adapt to a particular user in the same way that humans become familiar with the cuisines that they consume regularly.

Data Collection
Due to the lack of available and relevant datasets we created our own image dataset - FOOD. FOOD consists of two sub-datasets, FRUITS and MEALS, with different levels of complexity in terms of appearance and portion sizes. In total, it consists of 676 images of 49 food classes, taken under variable lighting conditions with two kinds of backgrounds. The images were taken of food before consumption, and contain examples of realistic meals in terms of the food combinations and portion sizes. We focus on commonly consumed foods in general, and in particular by older adults.

The experimental setup consisted of a Logitech Webcam Pro 9000 webcam mounted at a height of 40 cm directly above a plate of food. The plate was white and 26 cm in diameter. Neither the position of the plate within the camera’s field of view nor the position of the food on the plate was fixed. Each image consisted of one plate of food and was recorded at 3264x2448 pixels with an eight megapixel resolution.

We collected several pieces of ground truth information for each food item in an image: a segmentation mask, class name, weight and nutrient content. The segmentation masks were obtained by manual annotation as polygons. The weight of the plate and each food item on the plate were recorded using a kitchen scale. To gather nutritional information, we utilized the UK Nutrient Databank (Agency 2002), which provides information on the nutrient content of foods commonly consumed in the UK. We consider each of its entries as a different food class and extract the relevant nutritional data.

FRUITS is a simplified dataset consisting of 254 images of nine fruit categories such as apples, bananas and strawberries. Only whole fruits were used with mostly one class present per image. In all images, a white foam plate was placed against a dark surface. Lighting conditions were diffused and uniform across this dataset. Different portion sizes were created by either using differently sized fruits or by placing several of them on the plate, non-touching. Each food class includes at least three instances and varieties, when applicable. On average, each class has about 29 images (i.e. 29 plates of food per class).

The MEALS dataset consists of 422 images of 42 food classes such as rice, steak and peas. Two kinds of white plates were used and images were taken under different lighting conditions. The background setting was either of a yellow wooden coffee table or a dark gray office desk. The number of food items per image ranged from one to four, which were touching but not overlapping. With respect to portion sizes, portions were made by the researcher using personal estimates of how much is commonly eaten. While most of the food was home-cooked, some of it was store bought or ordered from restaurants. Each food class includes at least three different instances of the food and has about 19 images per class.

We realize that the meals we constructed and which foods were more commonly present in the dataset, were influenced by regional preferences of food in North America. Hence, the number of images taken for each class is not uniform, since some foods are consumed more than others due to either regional preferences or the role of a food in a meal, such as being an accompaniment.
Segmentation

Segmentation is an essential component of many computer vision systems and constitutes a critical first step in our recognition framework. We address segmentation as a two-step process. First, we extract the plate in an image and then propose a simple figure-ground segmentation algorithm, which retrieves a single mask for the food objects. Further research is required to separate the food objects from each other. All images were resized to 640x480 and color corrected (van de Weijer, Gevers, and Gijsenij 2007) prior to segmentation.

Plate Segmentation

Plate detection and segmentation are necessary in a real-world setting since they enforce the context in which food tends to appear and constrain the search space for relevant food objects. The key idea in our approach is to utilize a colour and shape prior, based on the assumptions in our data. Thus, we use the plate’s uniform white colour for detection and elliptical shape for segmentation. We start by transforming an image to the HSV (Hue, Saturation, Value) color map which is well suited for discriminating hues. We only retain the intensity channel, V. Next, we wish to threshold the V channel such that we obtain a white blob that corresponds roughly to the plate. Due to lighting variations and different shades of white, we choose a fixed image dependent threshold that is based on image percentiles. Then, a sequence of morphological operations (Dougherty 1992) is applied to the image to fill any gaps and discontinuities. Afterwards, we find connected components (Chang, Chen, and Lu 2004) and retain the blob with the largest area. Lastly, we apply an elliptical shape prior by fitting a RANSAC (Forsyth and Ponce 2003) ellipse to the contours of the detected blob. From the estimated ellipse we obtain a mask image of the plate.

Food Segmentation

Once the plate mask is obtained we can proceed to food segmentation. We observe that most foods are more colorful than the plate and that it is saturation rather than brightness that distinguishes foods from shadows or the plate. Saturation is a measure of the quantity of hue in a pixel, independent of the particular hue. Therefore, we proceed by thresholding the saturation channel of the plate segment image by an empirically determined constant. Then, we apply a series of morphological operations and find sufficiently large connected components. To deal with the issue of separating segmentation artifacts from real objects we design a simple measure of elongation defined as the perimeter of a blob over its area.

Essentially, we recognize that food objects are compact, i.e. they usually occupy most of the area of their convex hull; therefore, very elongated, curved or twisted shapes likely do not correspond to food objects. Hence, our perimeter over area (POA) metric would be very low for food objects and high for long and thin objects. Thus, we compute the POA measure for each blob and retain blobs that are below a certain threshold. To smooth the blobs and fill any gaps and discontinuities we make use of a series of morphological operations. Finally, we obtain the foreground mask that corresponds to the food objects.

Recognition

Food recognition is a core component of a nutritional assessment system. We formulate it as a supervised learning classification problem in which a segment can belong to a single class. In our approach, a segment is represented as a feature vector that is classified by a discriminative classifier. We explore several types of features and classifiers which are outlined below and examined in the Evaluation section.

Features

The features that we exploit fall under color, texture and shape descriptors. We explore four different color spaces, for which we create either a full or reduced normalized histogram. Typically, a full histogram would include all color components whereas a reduced one would only comprise component averages and standard deviations. The color spaces are RGB, YCbCr, HSV and LAB (Forsyth and Ponce 2003). RGB and YCbCr are linear and hardware-oriented color spaces whereas HSV and LAB are nonlinear and aim to approximate human vision. For texture, we use a well-known statistical method, the Gray Level Co-occurrence Matrix (GLCM) (Tuceryan and Jain 1993), whose second-order statistics are derived to estimate various texture properties. It represents texture as a two-dimensional spatial distribution of gray levels in an image. We compute the GLCMs for four orientations and three scales. Then, from each we derive the contrast, correlation, energy, homogeneity and entropy properties. In one feature variation, we use all of the properties from each of the orientations (ORIENTED), and in the other we average the first four over the orientations for each scale (MEANS).

Lastly, for shape, we focus on region-based global shape features rather than on local contour-based models, since foods have irregular and easily deformable shapes without a very stable topology. We explore both size dependent and independent rotation-invariant features, which are: physical area, diameter, major axis length, minor axis length, eccentricity and extent (MathWorks Inc. 2011). The only feature we designed is the area of a segment, measured in cm$^2$. It is an absolute measure of size, independent of the camera position or the plate’s dimensions. Once the desired features are extracted from a given segment, they are concatenated into a single feature vector. To speed up the computation and eliminate redundant feature dimensions, we use a simple feature selection (FS) heuristic. Dimension variances are thresholded to retain the relatively informative ones. Afterwards, the remaining dimensions are normalized to have zero mean and unit variance.

Classifiers

Supervised learning is used to predict the class of a given food segment. We take a multi-class classification approach which means that a feature vector could belong to solely one class. The feature space is divided into decision regions that correspond to classes whose boundaries are called decision boundaries. We examine three well known discriminative classifiers: Logistic Regression (LR) (Bishop 2006), Artificial Neural Networks (ANN) (Bishop 2006) and Support Vector Machines (SVM) (Bishop 2006). In the learning phase of these algorithms, the decision boundaries are determined based on training data. At the classification stage, unlabeled food segment features are mapped onto the
learnt feature space. The category to which the feature point is closest to is the category to which the food segment belongs.

LR is a generalized linear classifier which means it assumes that the decision boundaries in the feature space are linear. Alternatively, ANNs represent the class labels as a series of nonlinear transformations of the inputs. We use a 3-layer network with ten hidden units. Neither method has a closed-form solution to its parameter optimization problem in which case energy minimization methods (Bishop 2006) are applied. SVM is a non-probabilistic classifier which takes as input a model for the decision boundaries and then finds the optimal slicing of the feature space. We use a formulation (Crammer et al. 2001) in which a single SVM classifier is trained on all classes with a Gaussian kernel function for the decision boundaries.

Evaluation

Segmentation
To evaluate the performance of our algorithm in a segmentation context, we wish to assess how closely the resulting segmentation mask corresponds to the ground truth mask. This is achieved by the normalized overlap metric that computes the intersection of two masks relative to their union. The plate is annotated in 190 FOOD images. The training set comprises 26 images and the test data consists of total of 164 images. The parameters for our algorithm are determined empirically during training. Test performance is 95.06% (SD = 4.75) for the FOOD dataset. The high accuracy and the stability of the parameters between training and test indicate that the algorithm is robust. Moreover, we found our method to be resistant to lighting variations which is critical in real-world settings. Its performance remains high when applied to the non-color corrected images while thresholding with a constant achieves about 65% accuracy. Qualitatively, finding the exact boundaries of the segmentation is not essential. Our technique tends to include more of the plate and the background than constant thresholding. However, in this context, undersegmentation is preferable to oversegmentation which may eliminate important visual information on food segments.

For the evaluation of the food segmentation method, the training set consists of 20 images, 10 from each of FRUITS and MEALS while the rest of the images are reserved for testing. We use again the normalized overlap metric for the evaluation of the performance. Ground truth segmentation annotations are made for each object and are combined per image to obtain a single foreground mask. On the test datasets the performance is 81.64% (SD = 20.27), 62.72% (SD = 25.5) and 69.76% (SD = 25.38) for FRUITS, MEALS and FOOD, respectively. Figure 1 demonstrates successful results in the FOOD dataset. As evident from Figure 1 (a) the algorithm produces separate masks for non-touching objects and captures quite well their irregular shape. However, a single foreground mask is produced for objects that are touching or in close proximity, as in Figure 1 (b-c). Plate artifacts are not produced for uniformly white plates in contrast to the plate in Figure 1 (b). Overall, the performance is very good on the FRUITS dataset since fruits tend to have uniform coloring, relatively regular shapes and are typically eaten individually, so the objects are often non-touching and straightforward to segment.

Our algorithm has several failure modes. They all occur as a result of not satisfying one or more of the thresholds in the algorithm. Foods that have low saturation values due to being too dark, such as plum or egg whites, do not produce a mask. In cases when partial masks are produced, they are rejected either because they are too small or too elongated as a result of their unnatural shape. This tends to happen to foods that have low saturation, high texture, or are unevenly coloured, such as pasta, rice and coleslaw. Another failure mode emerges from the segmentation artifacts of the plate pattern leading to undersegmentation. In conclusion, the performance of the FRUITS dataset mostly suffers from objects that are too dark, whereas the MEALS dataset suffers from objects that are too light or unevenly coloured. Consequently, partial segmentations are generated that either do not produce any masks or oversegment the object. Patterned backgrounds can also hinder the segmentation quality.

Recognition
For the evaluation of the different features and classifiers described above, we performed 10-fold cross validation twice and averaged the results. The classification accuracy of a dataset is defined as an average of the individual classification performances of each class, which is the percentage of a class’s correctly labeled segments. The standard deviations reported are the averages of the standard deviations of class performances in the two runs. We trained the classifiers on all of the data in a dataset, tuned parameters on a validation set and report results on a test set. The validation set for FRUITS consists of three classes and the test set of six. In MEALS, the validation set consists of seven classes with the other 35 classes making the test set. Since FOOD is a combination of FRUITS and MEALS, its validation and test sets are simply the union of the respective sets for FRUITS.
and MEALS. Manually annotated segmentation masks were used in the experiments.

First we evaluate the different variations for each feature type. When only color features are used, the LAB and HSV spaces achieve the best performance for both datasets. On the MEALS dataset, our feature selection heuristic improves the performances for all color spaces using the SVM classifier. Moreover, the worst color features tend to be the reduced histogram versions as they contain less information. For texture, we are comparing directional (ORIENTED) and averaged (MEANS) texture features. For the FRUITS dataset, MEANS slightly outperform ORIENTED while for the MEALS dataset, ORIENTED are significantly superior. This result can be explained by the uniform texture fruits have, reducing the need for oriented features. Nonetheless, complex objects tend to have dominant orientations which can be leveraged by directional texture features, whereas that knowledge is lost by averaging over orientations. When all of the feature types are combined, the best combination for FRUITS is when the HSV or LAB space is used together with oriented texture, shape and feature selection. For MEALS and FOOD, the LAB space with the oriented texture, shape and feature selection is best with an 86% (SD = 12.7) and 87.2% (SD = 12.56) performance with the LR classifier, respectively. Although, LAB and HSV seem equivalent for the FRUITS dataset, for MEALS, LAB performs nearly perfectly on two challenging classes, *steak* and *bread* while HSV confuses them with *ribs* and *toast*, respectively.

In order to discuss the contribution of each feature type to the test recognition performance, we fix the classifier as LR and compare the performance on the best variation of each feature type, see Figure 2. On FRUITS, color alone performs nearly perfectly. Combining colour and texture helps disambiguate classes, which were confused when only one feature was used. For instance, the class *bananas* was mistaken as an *apple* when colour alone was used, and mistaken as *strawberries* by texture alone. Shape alone achieves an overall performance of 73.13%. Joining all of the feature types results in an increase to 99.87% (SD = 1.1), a 3% boost over using colour and texture and a 4.4% decrease in the SD. From the performances of the features on the FRUITS dataset, it is clear that colour plays the biggest role in the classification, with shape adding a boost of 3% to the accuracy and texture contributing only about 1.5% overall. Hence, when combined, colour and shape are the most significant features for FRUITS; however, individually, colour and texture are the most discriminative.

On the MEALS dataset, colour alone achieves a performance of 68.9% (SD = 22.13) while texture alone achieves a performance of 54.02% (SD = 26.34). When combined, they gain a 13.6% and a 28.5% advantage over colour or texture alone, respectively. Shape alone achieves quite a low overall performance of 17.38% (SD = 22.08) on MEALS. We notice two interesting behaviors in feature fusion. On some categories the stronger feature helps to disambiguate a category while other times, the classifier seems to fuse the feature information to gain a boost in performance, even though, individually, both features exhibit a lot of confusion over a category. When all of the features are combined, an additional 4% boost is achieved compared to colour and texture. More importantly, the SD drops significantly which demonstrates higher stability in the class performances. The addition of shape hinders the accuracy of some classes, highlighting that adding features is not always helpful. Nevertheless, classification benefits from all feature types, while color is again the most discriminative on its own. Texture is quite crucial in the MEALS dataset, since cooked foods tend to be similar in terms of color leading to texture playing a bigger role in their identification. Individually, none of the features have an acceptable performance for recognition but together they achieve high accuracy on a challenging dataset. Relative to FRUITS, shape is much less important in MEALS with color and texture taking the lead.

Lastly, we discuss the performance of the classifiers we used, comparing them on the FOOD dataset using the best performing feature combination. LR and SVM achieve similar performance and are always superior to ANN. ANN achieves a performance of 49.81% (SD = 31.23), failing on classes with complex coloring and texture. ANNs require a lot of diverse training data as the variability and complexity of a class increases. Therefore, ANNs are less suitable for our application. LR achieves the highest accuracy of 87.2% and the smallest SD of 12.56, while SVM performs 3% worse in terms of both accuracy and SD. Both of the classifiers excel on categories of fruits and vegetables. LR seems to do better than SVM on categories with a salient colour distribution, such as *fried eggs*. Nonetheless, both struggle on the classes: *roasted chicken breast*, *drumsticks*, *baked potato* and *shepherd’s pie*. Also, LR has a higher sensitivity to texture and seems less affected by imbalance in the amount of training data per class. For the categories with fewer images, confusion tends to happen with categories sharing similar appearance that have more training data. For example, *drumsticks* is confused with *chicken leg* which has twice as many images. From a computational perspective, on the FOOD dataset the running time of LR is quite fast (a few minutes) while SVM with the Gaussian kernel takes about four hours to run. Hence, we conclude that logistic regression is the most suitable classifier in the context in which we posed the food recognition problem. In terms of generalizability, we tuned the classifiers’ parameters mostly on the FRUITS dataset and still achieved quite high accuracy on the MEALS dataset, which is encouraging.

**Discussion**

Comparison with other methods in the literature is difficult since a standard food dataset does not exist yet. Methods (Yang et al. 2010; Chen et al. 2009) that have used the large PFID dataset (Chen et al. 2009) achieved between 10% - 28% precision when classifying very similar categories *e.g.* different types of burgers. When re-organized into semantically meaningful groups, results improved to 78% which indicates high within-group confusion. Moreover, in the former, a food object is represented as a set of ingredients with fixed spatial relationships which would rarely apply to home-cooked food. Others (Puri et al. 2009) who used multisensory input achieved 80% precision on 20 food classes.
There are several limitations with the current implementation of our system. Our segmentation approach is knowledge-driven and is therefore very dependent on the assumptions we make. The complexity of scenes in free-living conditions demands the use of more generalizable approaches such as bottom-up and learning methods with a minimal level of expert knowledge. For classification, we have demonstrated that appearance-based features are needed but play different roles in the recognition of different categories. Fusing them boosts performance at times but may also be detrimental, since irrelevant information leads to confusion. Therefore, feature selection is critical to determining the most discriminative representation of each category. Boosting techniques are often utilized for that purpose. Also, foods that are frequently confused among themselves, such as chicken dishes, raise the question of taxonomy i.e. the optimal grouping of classes in terms of both nutrition and computer vision. A taxonomy could aid algorithms by eliminating irrelevant candidates, reducing computation and improving recognition. One option is exploring the natural co-occurrence relationships among food categories to determine their characteristics and types. Probabilistically modeling the co-occurrence could resolve ambiguity in segmentation and recognition. However, to learn these relationships without overfitting or introducing noise requires a lot more data.

The FOOD dataset we created is comprehensive and was built to evaluate both the recognition and portion estimation performance. However, to satisfy the variability in the portion sizes, we sacrificed some variability in the appearances. Therefore, further tests are needed to ensure generalizability and the dataset has to be significantly expanded. There is a tradeoff between accounting for all of the theoretically possible scenarios and what is realistic and relevant for the application. In fact, imbalance in the number of images of food classes in the dataset is a cue to the amount of true variability in appearance, portion size and consumption. Therefore, it is crucial to decrease any bias in our assumptions by collecting real data on people’s dietary patterns and their free-living environments. Discovering how many different foods and meals people eat on a daily basis and the amount of time that it will take to adapt to a user would be very valuable.

In conclusion, our preliminary work has allowed us to understand the important challenges in both food and general object recognition. There are still many limitations and further research is required for the real-life applicability of the system. However, we were already able to demonstrate the plausibility of an automatic nutritional assessment system which will drastically impact the health of individuals, societies and the world. We are hopeful that our new dataset and research is a contribution to this important goal.

References