

A Food Recognition System for Diabetic Patients Based on an Optimized Bag-of-Features Model

Marios M. Anthimopoulos, *Member, IEEE*, Lauro Gianola, Luca Scarnato, Peter Diem, and Stavroula G. Mougiakakou, *Member, IEEE*

Abstract—Computer vision-based food recognition could be used to estimate a meal’s carbohydrate content for diabetic patients. This study proposes a methodology for automatic food recognition, based on the bag-of-features (BoF) model. An extensive technical investigation was conducted for the identification and optimization of the best performing components involved in the BoF architecture, as well as the estimation of the corresponding parameters. For the design and evaluation of the prototype system, a visual dataset with nearly 5000 food images was created and organized into 11 classes. The optimized system computes dense local features, using the scale-invariant feature transform on the HSV color space, builds a visual dictionary of 10000 visual words by using the hierarchical k -means clustering and finally classifies the food images with a linear support vector machine classifier. The system achieved classification accuracy of the order of 78%, thus proving the feasibility of the proposed approach in a very challenging image dataset.

Index Terms—Bag of features (BoF), diabetes, feature extraction, food recognition, image classification.

I. INTRODUCTION

THE treatment of Type 1 diabetic (T1D) patients involves exogenous insulin administration on a daily basis. A prandial insulin dose is delivered in order to compensate for the effect of a meal [1]. The estimation of the prandial dose is a complex and time-consuming task, dependent on many factors, with carbohydrate (CHO) counting being a key element. Clinical studies have shown that, in children and adolescents on intensive insulin therapy, an inaccuracy of ± 10 g in CHO counting does not impair postprandial control [2], while a ± 20 g variation significantly impacts postprandial glycaemia [3]. There is also evidence that even well-trained T1D patients find it difficult to

estimate CHO precisely [4]–[6]. In [4], 184 adult patients on intensive insulin were surveyed with respect to the CHO content of their meals. On average, respondents overestimated the CHO contained in their breakfast by 8.5% and underestimated CHO for lunch by 28%, for dinner by 23%, and for snacks by 5%. In [5], only 23% of adolescent T1D patients estimated daily CHO within 10 g of the true amount, despite the selection of common meals. For children with T1D and their caregivers, a recent study has shown that 27% of meal estimations are inaccurate in ranges greater than ± 10 g [6].

The increased number of diabetic patients worldwide, together with their proven inability to assess their diet accurately raised the need to develop systems that will support T1D patients during CHO counting. So far, a broad spectrum of mobile phone applications have been proposed in the literature, ranging from interactive diaries [7] to dietary monitoring based on on-body sensors [8]. The increasing processing power of the mobile devices, as well as the recent advances made in computer vision, permitted the introduction of image/video analysis-based applications for diet management [9]–[14]. In a typical scenario, the user acquires an image of the upcoming meal using the camera of his phone. The image is processed—either locally or on the server side—in order to extract a series of features describing its visual properties. The extracted features are fed to a classifier to recognize the various food types of the acquired image, which will then be used for the CHO estimation.

A food recognition application was introduced by Shroff *et al.* [9] for the classification of fast-food images into four classes. For each segmented food item, a vector of color (normalized RGB values), size, texture (local entropy, standard deviation, range), shape, and context-based features is computed and fed to a feed-forward artificial neural network (ANN), resulting in recognition accuracy of the order of 95%, 80%, 90%, and 90% for hamburgers, fries, chicken nuggets, and apple pies, respectively. A set of color (pixel intensities and color components) and texture (Gabor filter responses) features was used by Zhu *et al.* [10], together with a support vector machine (SVM) classifier, for the recognition of 19 food classes, leading to a recognition rate of the order of 94% for food replicas and 58% for real food items. Kong and Tan [11] proposed the use of scale-invariant feature transform (SIFT) features clustered into visual words and fed to a simple Bayesian probabilistic classifier that matches the food items to a food database containing images of fast-food, homemade food, and fruits. A recognition performance of 92% was reported given that the number of references per food class in the database is larger than 50 and the number of food items to be recognized is less than six.

Manuscript received April 9, 2013; revised November 3, 2013 and February 11, 2014; accepted February 18, 2014. Date of publication March 11, 2014; date of current version June 30, 2014. This work was supported in part by the Bern University Hospital, “Inselhospital,” and by the European Union Seventh Framework Programme (FP7-PEOPLE-2011-IAPP) under Grant 286408.

M. M. Anthimopoulos, L. Scarnato, and S. G. Mougiakakou are with the Diabetes Technology Research Group, ARTORG Center for Biomedical Engineering Research, University of Bern, 3010 Bern, Switzerland (e-mail: marios.anthimopoulos@artorg.unibe.ch; luca.scarnato@artorg.unibe.ch; stavroula.mougiakakou@artorg.unibe.ch).

L. Gianola and P. Diem are with the Department of Endocrinology, Diabetes, and Clinical Nutrition, Bern University Hospital, 3010 Bern, Switzerland (e-mail: lauro.gianola@students.unibe.ch; peter.diem@insel.ch).

Color versions of one or more of the figures in this paper are available online at <http://ieeexplore.ieee.org>.

Digital Object Identifier 10.1109/JBHI.2014.2308928

Recently, the bag-of-features (BoF) model was introduced into the area of computer vision as a global image descriptor for difficult classification problems [12]. BoF was derived from the bag-of-words (BoW) model. BoW is a popular way of representing documents in natural language processing [13], which ignores the order of the words belonging to a previously defined word dictionary, and considers only how frequently they appear. Similarly, in the image analysis context, an image is represented by the histogram of visual words, which are defined as representative image patches of commonly occurring visual patterns. The concept of the BoF model adequately fits the food recognition problem, since a certain food type is usually perceived as an ensemble of different visual elements mixed with specific proportions, but without any typical spatial arrangement, a fact that encourages the use of a BoF approach, instead of any direct image-matching technique.

Puri *et al.* [14] proposed a pairwise classification framework that takes advantage of the user's speech input to enhance the food recognition process. Recognition is based on the combined use of color neighborhood and maximum response features in a texton histogram model, feature selection using Adaboost, and SVM classifiers. Texton histograms resemble BoF models, using though simpler descriptors, such that histograms of all possible feature vectors can be used. In this way, the feature vector clustering procedure can be omitted; however, less information is considered by the model which might not be able to deal with high visual variation. Moreover, the proposed system requires a colored checker-board captured within the image in order to deal with varying lighting conditions. In an independently collected dataset, the system achieved accuracies from 95% to 80%, as the number of food categories increases from 2 to 20.

A database of fast-food images and videos was created and used by Chen *et al.* [15] for benchmarking of the food recognition problem. Two image description methods were comparatively evaluated based on color histograms and bag of SIFT features for a seven fast-food classes problem. The mean classification accuracy using an SVM classifier was 47% for the color histogram based approach and 56% for the SIFT-based approach. However, the used patches are sampled with the SIFT detector which is generally not a good choice for image classification problems, and described by the standard grayscale SIFT that ignores any color information.

The combined use of bag of SIFT, Gabor filter responses, and color histograms features in a multiple kernel learning (MKL) approach was proposed by Joutou *et al.* [16] for the recognition of Japanese food images. However, the employed BoF model uses the conventional scheme of fixed-size SIFT features clustered with standard k -means, while the additional color and texture features are global and are not included into the BoF architecture. For the 50 food classes problem, a mean recognition rate of 61% was reported.

Although over the last few years food recognition has attracted a lot of attention for dietary assessment, most of the proposed systems fail to deal with the problem of the huge visual diversity of foods, so they limit the visual dataset considered to either too few or too narrow food classes, in order

to achieve satisfactory results. The BoF model appears to be an appropriate way for dealing with generic food description, but still there is no systematic investigation of the various technical aspects related to feature extraction and classification.

The present study makes several contributions to the field of food recognition. A visual dataset with nearly 5000 homemade food images was created, reflecting the nutritional habits in central Europe [17]. The foods appearing in the images have been organized into 11 classes of high intravariability. Based on the aforementioned dataset, we conducted an extensive investigation for the optimal components and parameters within the BoF architecture. Three key point extraction techniques, fourteen local image descriptors, two clustering methods for the creation of the visual dictionary, and six classifiers were tested and comparatively assessed. Descriptors' fusion and feature selection were also tested. Moreover, the effects of various parameters like the number of extracted key points, the descriptor size(s), and the number of visual words are illustrated after conducting extensive experiments. Finally, a system for the recognition of generic food is proposed based on an optimized BoF model.

The rest of the paper is organized as follows. Section II presents the different methods considered for image description and classification. In Section III, the dataset and the experimental setup are described, while in Section IV the results, as well as the final system configuration, are presented. Finally, Section V summarizes and provides discussion and future research directions.

II. METHODS DESCRIPTION

The proposed food recognition system consists of two stages: food image description and image classification (see Fig. 1). During food image description, a set of characteristics representing the visual content of the image is extracted and quantified. This set provides input to the second stage, where a classifier assigns to the image one class out of a predefined set of food classes. The design and development of both stages involves two phases: training and testing. During the training phase, the system learns from the acquired knowledge, while during the testing phase the system recognizes food types from new, unknown images.

A. Food Image Description

In order to describe the appearance of the different food classes, the BoF model was adopted, due to its proven ability to deal with high visual diversity and the absence of typical spatial arrangement within each class. BoF consists of four basic steps: 1) key point extraction, 2) local feature description, 3) learning the visual dictionary, and 4) descriptor quantization. All the steps, as presented in Fig. 1, are involved in both training and testing, except for the learning of the dictionary, which is performed only once, during the training phase.

1) *Key Point Extraction*: Key points are selected points on an image that define the centers of local patches where descriptors will be applied. In the current study, three different key point extraction methods were tested: interest point detectors, random sampling, and dense sampling.

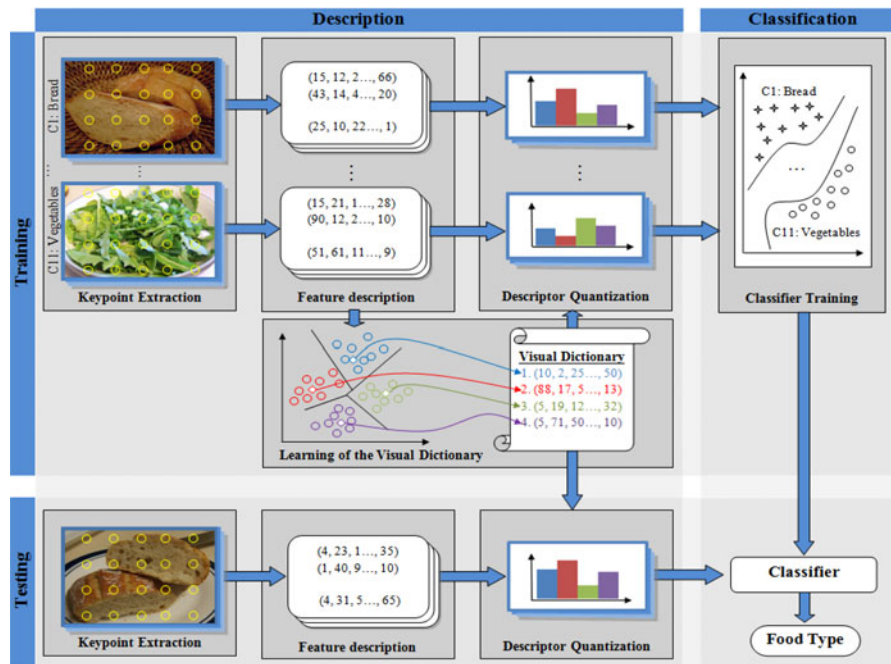


Fig. 1. Architecture of the proposed BoF-based food recognition system. The two major stages of food description and classification are illustrated within the training and testing phases. Food description includes key point extraction, feature description, descriptor quantization, and dictionary learning.



Fig. 2. Key point extraction techniques: (a) SIFT detector, (b) random sampling, and (c) dense sampling.

Interest point detectors, such as SIFT [18], are considered as the best choice for image matching problems where a small number of samples is required, as it provides stability under local and global image perturbations. SIFT estimates the key points by computing the maxima and minima of the difference of Gaussians (DoG), applied at different scales of the image. However, these techniques often fail to produce a sufficient number of image patches for classification problems [see Fig. 2(a)], where the number of sampled patches constitutes the most important factor. Random and dense sampling methods [see Fig. 2(b) and (c)] have been widely used in image classification with great success since they are able to provide a BoF-based system with the required number of image patches [19]. Random sampling is based on the random selection of point coordinates which usually follows a uniform distribution. Dense sampling is performed by extracting key points from a dense grid on the image. Any required number of points can be achieved by adjusting the spacing among them.

2) *Local Feature Description*: After the key point extraction, a local image descriptor is applied to a rectangular area around each key point to produce a feature vector. Identifying the appropriate descriptor size and type for a recognition problem is a challenging task that involves a number of experiments.

For the determination of the optimal descriptor size, the size of the object to be recognized should be considered. Although the SIFT interest point detector provides the position of the key points together with their scale, it is rarely used for image classification, as already explained. Hence, the size of the descriptor must be specified somehow after the dense or random key point sampling. A minimum size of 16×16 is often used as proposed in [18], since a smaller patch would not provide sufficient information for the description. However, the use of larger sizes or combination of sizes can often give better results by resembling the multiscale image description of the SIFT detector. It should be noted that food images are scaled to a standard size, so differences in food items scale should not be extreme.

To choose the best descriptor, the low-level visual features of the specific classes should be considered, so that each class can be adequately characterized and distinguished from the rest. The current literature proposes several feature descriptors, which can be roughly divided into two main categories: color and texture. Food description obviously depends on both color and texture information.

Color definitely constitutes a valuable source of information for image description. Most descriptors in this category capture specific color attributes and ignore information related to change and/or shift of intensity and/or color gaining corresponding invariances. In this study, the following types of color descriptors were tested:

Color histograms: Color histograms are probably the most common color descriptors. They represent the color distribution of an image in a certain color space and—despite their simplicity—they have been successfully used in various object recognition applications [20]. For the proposed system, five color histograms were considered covering different

combinations of invariants: $HistRGB$, $HistOp$, $HistRG_{norm}$, $HistHue$, and $HistRGB_{trans}$ calculated in the RGB color space (1), the opponent color space (2), the RG normalized channels (3), the Hue channel (4), and the transformed RGB color space (5), respectively:

$$RGB = \begin{pmatrix} R \\ G \\ B \end{pmatrix} \quad (1)$$

$$Op = \begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{2}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix} \quad (2)$$

$$RG_{norm} = \begin{pmatrix} R_{norm} \\ G_{norm} \end{pmatrix} = \begin{pmatrix} \frac{R}{R+G+B} \\ \frac{G}{R+G+B} \end{pmatrix} \quad (3)$$

$$Hue = \text{atan2} \left(\sqrt{3} * (G - B), 2 * R - G - B \right) \quad (4)$$

$$RGB_{trans} = \begin{pmatrix} R_{trans} \\ G_{trans} \\ B_{trans} \end{pmatrix} = \begin{pmatrix} \frac{R - \mu_R}{\sigma_R} \\ \frac{G - \mu_G}{\sigma_G} \\ \frac{B - \mu_B}{\sigma_B} \end{pmatrix} \quad (5)$$

where R , G , and B are the Red, Green, and Blue channels of the RGB color space, and μ_i, σ_i ($i = R, G, B$) are the means and the standard deviations of the distributions in channels R , G , and B , respectively. The histograms of multichannel color spaces are simple concatenations of the single channels' histogram values.










Generalized color moments: Color moments have been successfully used for recognition purposes [21]. They are based on the fact that any probability distribution is uniquely characterized by its moments. Generalized color moments have been proposed in order to overcome the inability of histogram based features to capture spatial information [22]. They combine powers of the pixel coordinates and their intensities in different color channels within the same integral. Thus, some information regarding the spatial color distribution is also considered. The generalized color moments are estimated by the following equation:

$$M_{pq}^{abc} = \sum_i^W \sum_j^H x^p y^q R(x, y)^a G(x, y)^b B(x, y)^c \quad (6)$$

where $R(x, y)$, $G(x, y)$, and $B(x, y)$ are the RGB values of the pixel at position (x, y) in an image of size $W \times H$. M_{pq}^{abc} is defined to be a generalized color moment of order $p + q$ and degree $a + b + c$. In this study, only generalized moments of order 0 or 1 and degree 1 or 2 were used, leading to 27 possible combinations.

Color moment invariants: Generalized color moments can be combined to construct color moment invariants which achieve

TABLE I
EXAMPLES OF CLUSTERED IMAGE PATCHES AND VISUAL WORDS

	Image patches				Visual word (cluster center)
Cluster 1			...		(35, 2, ..., 12)
Cluster 2			...		(3, 82, ..., 72)
Cluster 3			...		(16, 9, ..., 42)

viewpoint and illumination invariance, by dealing with different combinations of affine geometric and linear photometric changes. In this study, all three color band invariants from Min-drú *et al.* [22] are considered, producing a total of 24 features.

All of the aforementioned descriptors focus on the color distribution, without providing any texture information. SIFT and its color variants constitute popular local texture descriptors and have been successfully used within the BoF framework for detection and classification problems [23].

SIFT: SIFT considers a region of 16×16 pixels around a given key point. Then, the region is divided into 4×4 sub-regions and for each of them an eight-bin histogram of the intensity gradient orientation is computed, leading to a 128-dimensional feature vector. SIFT features provide remarkably informative texture description, and are invariant to light intensity changes. Despite the superior texture description capabilities of the original SIFT, its inability to capture any color information constitutes a problem for the description of many objects, including foods.

Color SIFT variants: Color SIFT is computed similarly to SIFT, but with one major difference. Instead of using just the intensity image, the histograms of gradient orientations are computed in various color channels, so the resulting descriptor constitutes the concatenation of the individual descriptors. Thus, the size of the created feature vector is $128 \times NC$, where NC is the number of color channels used. The variants used for experimentation in the framework of this study operate in the RGB, HSV, Hue, Opponent, and C-Invariant color spaces [24], producing the rgbSIFT, hsvSIFT, hueSIFT, opponentSIFT, and cSIFT, respectively. In addition, rgSIFT has also been used; this applies SIFT only in the R_{norm} and G_{norm} components of the normalized RGB color model.

3) *Learning the Visual Dictionary:* Once the descriptors of each training image patch have been computed, the most representative patches need to be identified which will constitute the system's visual words. To this end, the feature vectors that correspond to the various patches are clustered in a predefined number of clusters. The centers of the created clusters constitute the visual words of the dictionary, while the entire clustering procedure is known as the dictionary learning. Table I provides some examples of clustered image patches and visual words. The most common clustering technique used for the creation of visual dictionaries is the k -means [25] clustering algorithm.

In this study, the classic k -means with its hierarchical version (hk -means) algorithm [26] were comparatively evaluated.

Although k -means is considered to be one of the best clustering techniques, its high computational complexity significantly increases the processing time of the dictionary learning, while the large number of clusters produced slows down the descriptors' quantization. Hierarchical k -means (hk -means) has been used in order to reduce the time needed for both training and testing. hk -means performs clustering in a hierarchical way by building a tree of clusters. Initially, k -means clustering is performed to partition the N observations into K clusters of descriptor vectors. Then, each new cluster is iteratively divided by subsequent k -means runs until a predefined number of clusters is reached. Thus, in this case, k defines the branch factor of the tree while the final number of clusters is K^l , where l is the number of tree levels.

4) *Descriptors Quantization*: Descriptor quantization is the procedure of assigning a feature vector to the closest visual word of a predefined visual vocabulary. Once the visual dictionary is learnt, each descriptor of an image is quantized and the histogram of visual word occurrences serves as a global description of the image. Then, the histogram values are usually scaled to $[0\ 1]$ and fed to the classifier either for training or testing. The efficiency of this part of the system is crucial, since it affects processing times for both training and testing. The complexity of the descriptor quantization mainly depends on the dimensions of the descriptor and the number of visual words.

B. Food Image Classification

The image classification stage is involved in both training and testing phases. In order to identify the appropriate classifier for the specific problem, several experiments with three supervised classification methods were conducted: SVM, ANN, and Random Forests (RF).

The SVM [27] with linear or nonlinear kernels constitutes the most common classifier among the BoF approaches. The following three kernels were used in the experiments of the current study:

$$\text{Linear: } k_{\text{linear}}(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{x}_1^T * \mathbf{x}_2 \quad (7)$$

$$\text{RBF: } k_{\text{RBF}}(\mathbf{x}_1, \mathbf{x}_2) = \exp(-\gamma * \|\mathbf{x}_1 - \mathbf{x}_2\|^2) \quad (8)$$

$$\text{exp } X^2 : k_{x^2}(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\frac{\gamma}{2} * \sum_i \frac{(\mathbf{x}_{1i} - \mathbf{x}_{2i})^2}{\mathbf{x}_{1i} + \mathbf{x}_{2i}}\right) \quad (9)$$

where \mathbf{x}_1 and \mathbf{x}_2 are feature vectors, and γ is a scaling parameter which needs to be tuned. Thus, the following SVMs were tested: $\text{SVM}_{\text{linear}}$, SVM_{RBF} and SVM_{x^2} .

ANNs also constitute popular machine learning models for solving complex computer vision problems [28]. Two different feed-forward ANN models were used during the present study: a linear one without hidden layers (ANN_{nh}) and a nonlinear with one hidden layer (ANN_{wh}). ANN_{nh} was trained using the simple gradient-descent algorithm, while ANN_{wh} used the scaled conjugate gradient back-propagation algorithm [28].

Generally, the conjugate gradient back-propagation algorithm leads to faster convergence to better minima than standard steepest descent methods [29]. Both ANNs are fully connected, with initial weights randomly selected in the range $[-1.0, 1.0]$. As an activation function, the saturated linear was used for the output layer and the hyperbolic tangent sigmoid for the hidden layer of the second network. The ANN_{wh} topology and the internal parameters were determined using a trial-and-error process.

RFs have become popular because of their extreme efficiency, the simplicity of training, and their ability to give estimates of the variables importance. They are an ensemble of decision trees such that each tree depends on the values of a random vector sampled independently with the same distribution for all trees in the forest [30]. The forest chooses the classification having the majority of votes over all the trees in the forest. In this study, one RF was used for the experiments consisting of 31 trees, with each split randomly choosing a number of features equal to the squared root of the total number of features.

III. DATA AND EXPERIMENTAL SETUP

A. Food Image Dataset

For the experimental needs of the system developed a dataset of 4868 color images was created by collecting images from the web. The food types and their categorization were identified in collaboration with the Department for Endocrinology, Diabetology and Clinical Nutrition of Bern University Hospital, Inselspital. A total of eleven (11) classes were considered (the number in the parenthesis denotes the number of images per class): $C1$: Bread (506), $C2$: Breaded food (310), $C3$: Cheese (412), $C4$: Egg products (398), $C5$: Legumes (257), $C6$: Meat (174), $C7$: Pasta (564), $C8$: Pizza (731), $C9$: Potatoes (440), $C10$: Rice (399), and $C11$: Vegetables (677).

The downloaded images were filtered in order to keep only those images with at least one side larger than 500 pixels and then all images were rescaled so that their greatest side becomes equal to 500. Downloading images from the web enabled the generation of a real-world visual dataset with very high intra-class variability. Images present a wide range of lighting conditions, arbitrary viewing angle, and different food servings within each food class. An overview of the visual dataset is presented in Fig. 3.

B. Evaluation

In order to evaluate the performance of the food recognition system, the overall recognition accuracy (ORA) was considered, which represents the percentage of the correctly classified test images. ORA is defined as

$$\text{Overall Recognition Accuracy} = \frac{\sum_i^N \text{CM}_{ii}}{\sum_i^N \sum_j^N \text{CM}_{ij}} \quad (10)$$

where CM_{ij} is the number of images that belong to class i and were classified in class j and N is the number of classes. The performance is visualized using the confusion matrix.

Applying the aforementioned evaluation metric required splitting the dataset into training and test sets. Hence, 60% of



Fig. 3. Sample images of the developed visual dataset. The dataset contains nearly 5000 food images organized into 11 classes.

the images were randomly chosen for training and the remaining 40% for testing. Finally, for testing the optimized BoF system, a cross-validation approach with fivefolds was adopted. In this case, the dataset is randomly split into five sets and each of these is tested by using the rest of the images for training. The obtained result is generally more representative but the method is too time expensive to be used for extensive experiments.

C. Implementation

All the experiments were carried out within the MATLAB environment on a machine with an Intel Q8300 CPU and 8 GB of RAM. The VLFEAT library [31] was used for the SIFT detector/descriptor and the k -means/ hk -means clustering techniques. Color descriptors were computed using the implementation provided by Van de Sande *et al.* [32]. For classification, LIBLINEAR [33] and LIBSVM [34] libraries provided the linear SVM and its nonlinear variants, respectively. Furthermore, the MATLAB ANN toolbox was utilized for the ANN experiments while for the RF, Google's open source code was used [35].

IV. EXPERIMENTAL RESULTS

This section provides an extensive report of the experimental results, together with the corresponding discussion as well as the specification of the proposed optimized food recognition system.

A. Experiment 1: Evaluation of Key Point Extraction Techniques

In this experiment, three key point extraction strategies were comparatively evaluated: the SIFT interest point detector, random sampling, and dense sampling. For dense sampling, three different patch sizes were used, namely 16, 24, and 32, with spacing among the patches equal to 8, 12, and 16, respectively.

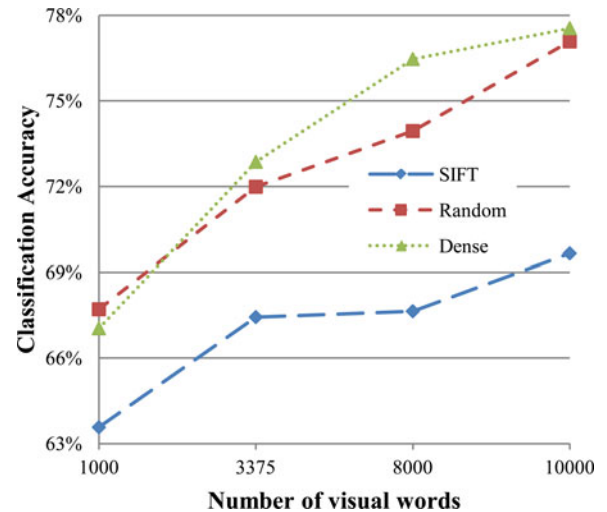


Fig. 4. Comparison of sampling strategies in terms of overall classification accuracy. SIFT key point detector, random, and dense sampling have been used.

This setup resulted in approximately 6000 densely sampled patches for an image of 500×500 pixels, while the random sampler was also configured in order to produce the same number of patches. For the SIFT key point detector, the corresponding parameters were adjusted so the maximum number of points was reached; however, the resulting key points were still less than 1000, too few when compared to the other sampling techniques.

Fig. 4 shows the results of the comparison between the three sampling techniques for different dictionary sizes. The obviously poor performance of the SIFT point detector is due to the inability of SIFT to produce enough key points for capturing the image characteristics. The performances of dense and random sampling methods are comparable, with the former having slightly better results probably because of the uniform spatial distribution of the extracted points that produced less correlated features.

B. Experiment 2: Descriptor Size Configuration

The scope of this experiment is to identify the proper descriptor size or combination of sizes that should be used together with the best performing key point extraction technique. To this end, different sizes were evaluated and then combined into a multi-scale scheme using a dense sampler. The used descriptor sizes were 16, 24, 32, and 56 all their combinations with a spacing among them equal to 1/2 of each size in order to guarantee a sufficient number of patches.

Fig. 5 presents the recognition rates of the different scale setups. Small scales proved to be more informative than larger ones. This is probably related to the fact that small scales provide larger number of descriptors which give statistically sufficient samples, in order to estimate the actual distribution of visual words in large dictionaries. In view of the poor performances obtained by size 56, larger sizes were not evaluated. Moreover, the results presented prove the initial assumption that a combination of different scales will benefit the image description. The best recognition rates were obtained with the combination

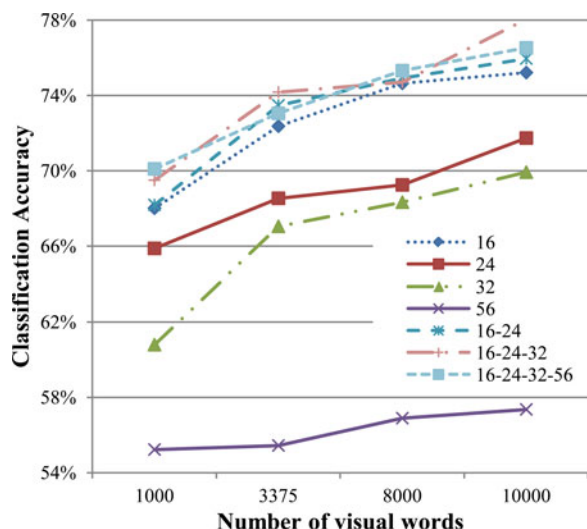


Fig. 5. Comparison of descriptor size configurations in terms of overall classification accuracy. Seven combinations of sizes have been reported. The rest achieved worst performance and therefore are omitted for clarity.

of sizes 16, 24, and 32, while adding 56 did not add more information. The rest of the combinations that are not presented in Fig. 5 produced far worse results, so they were excluded for clarity.

C. Experiment 3: Comparing Local Feature Descriptors

In order to identify the best descriptor for the specific problem, 14 popular color and texture descriptors were compared. Half of these constitute pure color descriptors, while the rest include SIFT and some of its color variants.

Fig. 6 shows the classification rates of the proposed system for each of the 14 different descriptors. The results show that the SIFT-based descriptors perform significantly better than color descriptors, probably because they are less sensitive to intensity changes and color shifts. Although the color-SIFT variants are able to describe the color texture of images, they still constitute differential descriptors so that they do not rely on the actual color values. Among the SIFT-based descriptors, hsvSIFT achieved the best performance followed by rgSIFT, rgbSIFT, and opponentSIFT, thus proving the superiority and intuitiveness of the HSV color representation. With respect to color, the two best performing descriptors were the color moment invariants and the opponent color histogram.

To investigate the complementarity of the color/texture features, the two best performing descriptors per type were fused. The fusion of the descriptors was carried out in two different ways. First, we tried concatenating the descriptor values before learning the visual dictionary, in order to achieve a better description of local patches. As an alternative, we created two different visual dictionaries, one for each descriptor, and then concatenated the generated histograms before feeding them to the classifier. Fig. 7 presents the corresponding results. The best result (77.8%) was achieved by concatenating hsvSIFT and color moment invariant descriptors before the creation of the visual dictionary. However, the performance achieved was

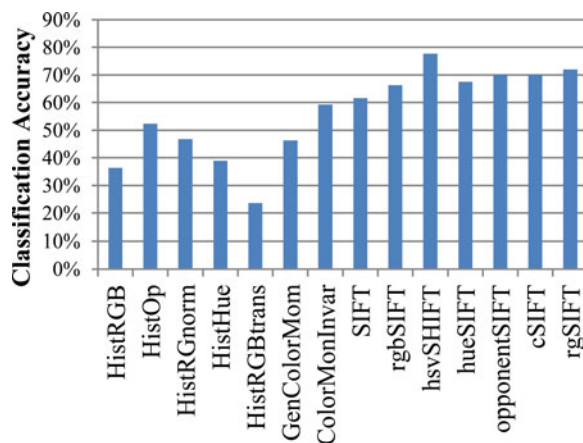


Fig. 6. Comparison of local descriptors in terms of overall classification accuracy. Seven color and seven texture descriptors have been used.

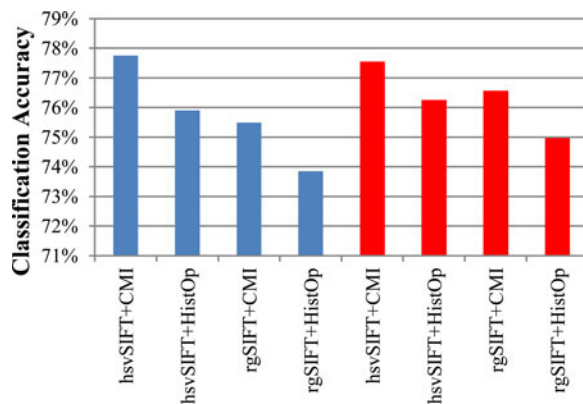


Fig. 7. Comparison of local descriptors' combinations in terms of overall classification accuracy. Blue denotes descriptor concatenation, while red corresponds to histogram concatenation.

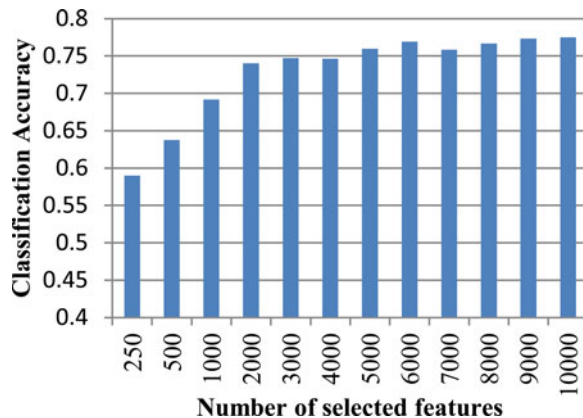


Fig. 8. Effect of feature selection in the overall classification accuracy.

equivalent to hsvSIFT alone (77.6%). This result proved the ability of the hsvSIFT additionally to capture color information apart from texture.

As an additional experiment, we applied a feature selection procedure based on the random forest feature ranking [30]. Fig. 8 presents the obtained results. As it can be seen, after ignoring more than half of the 10000 features, the system is still able to

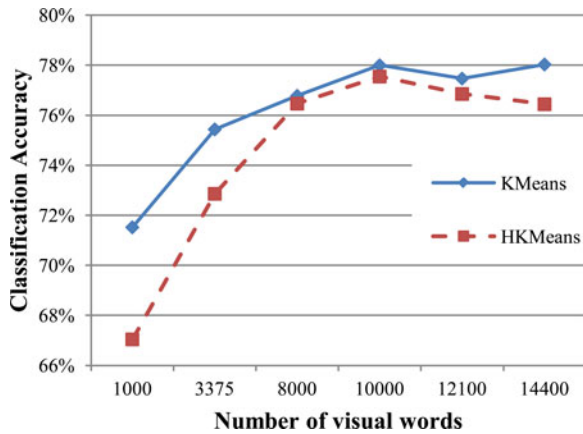


Fig. 9. Comparison of clustering techniques in terms of overall classification accuracy. The classic k -means and its hierarchical variant were used.

achieve almost the same overall accuracy. However, the result was not improved in any of the experiments and the gained computational efficiency is negligible since the prediction time for most classifiers is on the order of milliseconds.

D. Experiment 4: Comparing Clustering Algorithms for the Visual Dictionary Creation

The visual dictionary is created by clustering a number of training descriptor vectors into a predefined number of clusters. Thus, the factors that affect the dictionary creation are three: the number of selected training vectors, the chosen clustering technique, and the number of clusters.

After experimenting, it was shown that the most crucial parameters were the last two, since the only requirement for the first parameter was to have a reasonably large number of descriptor vectors from each food type. To this end, 100000 patches were randomly chosen from each food class and described by hsvSIFT to produce a total of 1100000 descriptor vectors as input to the clustering procedure. Apparently, decreasing the number of considered vectors accelerates the clustering procedure. However, the learning of the dictionary is performed offline during the training phase of the system so its efficiency can be regarded as a minor issue compared to the recognition performance of the system. On the other hand, increasing the training vectors had no influence on the overall classification performance.

In order to clarify the influence of the clustering algorithm used and dictionary size on the system's performance, an experiment comparing k -means with hk -means was conducted by using different numbers of visual words (clusters). Since the number of words produced by hk -means is restricted to the powers of the branch factor, while k -means is able to generate any number of clusters, we defined the dictionary sizes on the former basis. Hence, by specifying branch factors of 10, 15, 20, 100, 110, and 120, hk -means produced 1000, 3375, 8000, 10000, 12100, and 14400 visual words, respectively, which resulted in three hierarchical levels for the first three cases and two levels for the rest of them. From the results of Fig. 9, it is observed that recognition rate is directly related to the dictionary dimension,

TABLE II
COMPARISON BETWEEN K -MEANS AND HK -MEANS FOR 10 000 WORDS

Attribute	k -means	hk -means
Dictionary learning time (training)	22 hours	16 minutes
Histogram computation time (testing)	1.6 seconds	0.03 seconds
Recognition accuracy	78%	77.6%

with a gain in performance as the dictionary size increases. By increasing the number of visual words, the dimensionality of the final feature space is also increased providing information even for the rarest visual words. However, after a threshold, the system's performance stops growing, since a further increase in the dictionary dimensionality will just increase the correlation among the feature vectors without adding any information. Specifically, the k -means curve shows a clear tendency to converge, while hk -Means even declines, probably due to its hierarchical nature, that produces more unbalanced clusters. These results are easily justifiable and consistent with the findings of other BoF-based studies in the relevant literature [36]. The optimal threshold was determined to be approximately 10000 for the specific problem, thus proving the great visual variability of the selected food images.

In addition, Fig. 9 provides a comparison between the two different clustering techniques; k -means and hk -means. It is easily observed that k -means produces more representative small dictionaries. However, as the number of visual words considered increases, hk -means provides equivalent results while at the same time greatly reducing the computational cost in both training and testing. Table II presents the corresponding comparison. This remarkable improvement is due to the tree structure of the hk -means dictionary which results in more efficient vector quantization having logarithmic complexity with respect to the number of clusters.

E. Experiment 5: Comparing Machine Learning-Based Classifiers

Based on the optimum BoF configuration as defined in Experiments 1–4, the effect of a variety of machine learning-based classifiers in the system performance was investigated. To this end, three different types of classification algorithms (SVM, ANN, RF) were comparatively assessed.

Fig. 10 compares the results with linear SVM (SVM_{linear}), two nonlinear kernel-based SVMs (SVM_{RBF} , SVM_{x^2}), the two ANNs (ANN_{nh} , ANN_{wh}) and the RF, which also constitutes a nonlinear classification approach. For the rest of the system, the optimal components were used as identified by the aforementioned experiments.

From the comparison it can be seen that SVM_{linear} and ANN_{nh} outperform the more complex classifiers (SVM_{RBF} , SVM_{x^2} , ANN_{wh} , and RF). The low performance of the RF is probably due to overfitting of the classifier to the training data that caused limited generalization capability. The nonlinear kernels (k_{RBF} , k_{x^2}) failed to reach the recognition rates of the linear SVM, despite their complex mapping to higher dimensional spaces. Similarly, the ANN_{nh} outperformed

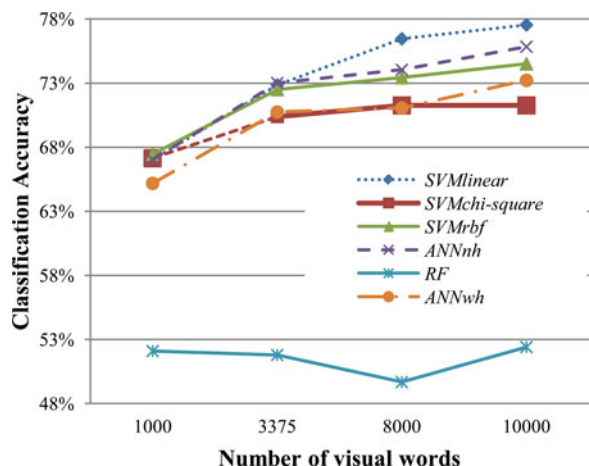


Fig. 10. Comparison of classifiers in terms of overall classification accuracy. Two linear (*SVMlinear*, *ANNnh*) and four nonlinear classifiers (*SVMchi-square*, *SVMrbf*, *ANNwh*, *RF*) were used.

the *ANNwh*. Although the performances of the *SVMlinear* and *ANNnh* are comparable for small visual dictionaries, the former seems more capable of exploiting the information provided by additional dimensions. Furthermore, *SVMlinear* produced a more balanced confusion matrix with a simpler training procedure. After the experimentation with the different classifiers, the general conclusion can be drawn that the proposed BoF approach not only gives an informative description of the food appearance but also creates a sparse high-dimensional space which proves to be linearly separated to some extent. These findings are in line with the recent literature regarding large-scale learning and high-dimensional classification [37].

F. Optimized BoF-based Food Recognition System

After experimenting with the various components and the corresponding parameters involved in the proposed food recognition architecture (see Fig. 1), the final optimized BoF-based system was determined. For each image, a set of hsvSIFT descriptors is computed in densely extracted rectangular patches with sizes 16, 24, and 32 and spacing 8, 12, and 16, respectively. The visual dictionary was created by clustering over 1 million randomly selected descriptors into 10000 clusters using the *hk*-means algorithm with a branch factor of 100 and 2 hierarchical levels. The created histograms of visual words are normalized by setting their L2-norm equal to 1, subtracting the minimum value and then divide by the maximum. Finally, *SVMlinear* is used for the classification of each image. The optimal C parameter of the *SVMlinear* was determined experimentally equal to 1. Moreover, class weights are provided to the SVM which are equal to the inverse of the occurrence frequency, for each class in the dataset.

The ORA of the final system on the testing dataset was calculated equal to 77.6%. In addition, a cross-validation scheme with fivefolds was used for further testing of the system which resulted in 78% accuracy. The difference is probably due to the bigger training sets used inside the cross-validation procedure which contained 80% of the images instead of 60%. Fig. 11

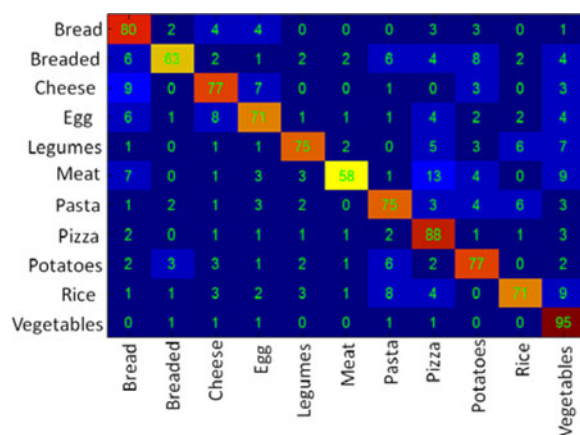


Fig. 11. Confusion matrix of the proposed optimized system. The entry in the *i*th row and *j*th column corresponds to the percentage of images from class *i* that was classified as class *j*.

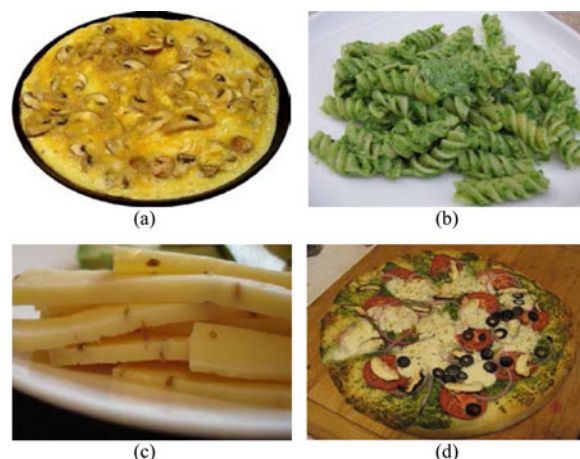


Fig. 12. Examples of incorrectly classified images: (a) eggs classified as pizza, (b) pasta classified as vegetable, (c) cheese classified as potatoes, and (d) pizza classified as vegetables.

shows the confusion matrix of the proposed optimized system. The highest recognition rate is achieved by the class C11 (Vegetables) due to their distinctive colors while the class C8 (Pizza) follows, as it has quite representative shape and texture. On the other hand, the class C6 (Meat) presents the lowest accuracy, since it is often misclassified as pizza or vegetables. This is mainly because of the different meat products (e.g., ham, bacon) used for pizza and because of the small parts of vegetables which often decorate meat plates. Moreover, the number of images used for training the system clearly affects the classification performance for each class. Class C6 (Meat), with the lowest rates, contains substantially fewer images than the rest of the classes, while classes C8 (Pizza) and C11 (Vegetables) have the best performance and are the most common classes of the image dataset.

Fig. 12 presents some examples of misclassified images. Fig. 12(a) shows an omelet which was classified as C8 (Pizza) instead of C4 (Egg products). The reason for this misclassification obviously lies in the color and shape of the specific food item that commonly appear in category C8 (Pizza). Shape

is described indirectly within the BoF model by the frequency of appearance of specific visual words with strong edges that correspond to the contour of an object. The second image [see Fig. 12(b)] was classified into $C11$ (Vegetables), despite belonging to $C7$ (Pasta). This failure of the classification system was apparently caused by the green color of pasta, which is almost exclusively associated with salads. The cheese slices of Fig. 12(c) were wrongly classified as $C9$ (Potatoes) due to their color, texture, and shape that resembles french fries, while the last example [see Fig. 12(d)] shows a vegetarian's pizza misclassified as $C11$ (Vegetables). This error can again be easily justified, since the pizza is mainly covered by vegetables.

V. DISCUSSION AND CONCLUSION

In this paper, we propose a BoF-based system for food image classification, as a first step toward the development of a portable application, providing dietary advice to diabetic patients through automatic CHO counting. A series of five major experiments was carried out for choosing and optimizing the involved components and parameters of the system. The experiments were conducted on a newly constructed food image dataset with 4868 images of central European food belonging to 11 different food classes.

The first experiment proved the superiority of dense key point extraction which was able to produce the required large number of patches with minimum overlap between them. The second experiment investigated the effect of the descriptor's size on the final performance. The best results were obtained by the combination of descriptors with sizes 16, 24, and 32. By using descriptors with different sizes, the BoF system gained multiresolution properties that increased the final performance, since the food scale may vary among the images. Then, the hsvSIFT was chosen among 14 different color and texture descriptors as giving the best results. hsvSIFT constitutes a differential descriptor that describes the local texture in all three different color channels of the HSV color space. This fact enables it to include color information, apart from texture, but also keep some invariance in intensity and color changes. The color capturing ability of hsvSIFT was also proved by the descriptors' fusion experiment that failed to increase the performance after combining it with the best color descriptors. As regards the learning of the visual dictionary, k -means was compared to its hierarchical version hk -means. The latter managed to produce almost equivalent results with k -means, for the optimal number of visual words, while being extremely computationally efficient. The optimal number of words was determined to be approximately 10000, since fewer words resulted in clearly worse results and more words did not improve the performance. For the final classification, two linear and four nonlinear machine-learning techniques were employed, with the linear giving the best results, especially for large number of features. This is probably caused by the high dimensionality of the feature space, as this makes the problem linearly separable, at least to some extent.

The final, optimized system achieved ORA in the order of 78%, proving the feasibility of a BoF-based system for the food recognition problem. For future work, a hierarchical classification approach will be investigated by merging visually similar

classes for the first levels of the hierarchical model, which can then be distinguished in a latter level by exploiting appropriate discriminative features. Moreover, the enhancement of the visual dataset with more images will improve the classification rates, especially for the classes with high diversity. The final system will additionally include a food segmentation stage before applying the proposed recognition module, so that images with multiple food types can also be addressed. As a final stage, food volume will be estimated by using multi-view reconstruction and CHO content will be calculated based on the computer vision results and nutritional tables.

ACKNOWLEDGMENT

The authors would like to thank their colleagues J. Dehais, S. Shevchik, P. Agbesi of the ARTORG Center at University of Bern, A. Greenburg, A. Soni, and D. Duke of the Roche Diagnostics Inc., USA, and C. Wenger and J.-P. Schnyder of the Department for Clinical Nutrition at Bern University Hospital "Inselspital" for their support in the creation of the visual dataset and the definition of the various food types. More information about the project can be found at www.gocarb.eu.

REFERENCES

- [1] American Diabetes Association, "Standards of medical care in diabetes-2010," *Diabetes Care*, vol. 33, no. 1, pp. S11–S61, 2010.
- [2] C. E. Smart, K. Ross, J. A. Edge, C. E. Collins, K. Colyvas, and B. R. King, "Children and adolescents on intensive insulin therapy maintain postprandial glycaemic control without precise carbohydrate counting," *Diabetic Med.*, vol. 26, no. 3, pp. 279–285, 2009.
- [3] C. E. Smart, B. R. King, P. McElduff, and C. E. Collins, "In children using intensive insulin therapy, a 20-g variation in carbohydrate amount significantly impacts on postprandial glycaemia," *Diabetic Med.*, vol. 29, no. 7, pp. e21–e24, Jul. 2012.
- [4] M. Graff, T. Gross, S. Juth, and J. Charlson, "How well are individuals on intensive insulin therapy counting carbohydrates?" *Diabetes Res. Clinical Practice*, vol. 50, suppl. 1, pp. 238–239, 2000.
- [5] F. K. Bishop, D. M. Maahs, G. Spiegel, D. Owen, G. J. Klingensmith, A. Bortsov, J. Thomas, and E. J. Mayer-Davis, "The carbohydrate counting in adolescents with type 1 diabetes (CCAT) study," *Diabetes Spectr.*, vol. 22, no. 1, pp. 56–62, 2009.
- [6] C. E. Smart, K. Ross, J. A. Edge, B. R. King, P. McElduff, and C. E. Collins, "Can children with type 1 diabetes and their caregivers estimate the carbohydrate content of meals and snacks?" *Diabetic Med.*, vol. 27, pp. 348–353, 2010.
- [7] M. C. Rossi, A. Nicolucci, P. D. Bartolo, D. Bruttomesso, A. Girelli, F. Ampudia, D. Kerr, A. Ceriello, L. Mayor, F. Pellegrini, D. Horwitz, and G. , "Diabetes interactive diary: A new telemedicine system enabling flexible diet and insulin therapy while improving quality of life: An open-label, international, multicenter, randomized study," *Diabetes Care*, vol. 33, no. 1, pp. 109–115, 2010.
- [8] O. Amft and G. Tröster, "Recognition of dietary activity events using on-body sensors," *Artif. Intell. Med.*, vol. 42, no. 2, pp. 121–136, 2008.
- [9] G. Shroff, A. Smailagic, and D. P. Siewiorek, "Wearable context-aware food recognition for calorie monitoring," in *Proc. 12th IEEE Int. Symp. Wearable Comput.*, 2008, pp. 119–120.
- [10] F. Zhu, M. Bosch, I. Woo, S. Y. Kim, C. J. Boushey, D. S. Ebert, and E. J. Delp, "The use of mobile devices in aiding dietary assessment and evaluation," *IEEE J. Sel. Topics Signal Process.*, vol. 4, no. 4, pp. 756–766, Aug. 2010.
- [11] F. Kong and J. Tan, "DietCam: Automatic dietary assessment with mobile camera phones," *Pervasive Mobile Comput.*, vol. 8, pp. 147–163, Feb. 2012.
- [12] L. Fei-Fei and P. Perona, "A bayesian hierarchical model for learning natural scene categories," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2005, vol. 2, pp. 524–531.
- [13] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *Proc. 10th Eur. Conf. Mach. Learning*, 1998, pp. 137–142.

- [14] M. Puri, Z. Zhu, Q. Yu, A. Divakaran, and H. Sawhney, "Recognition and volume estimation of food intake using a mobile device," in *Proc. Workshop Appl. Comput. Vis.*, 2009, pp. 1–8.
- [15] M. Chen, K. Dhirra, W. Wu, L. Yang, R. Sukthankar, and J. Yang, "PFID: Pittsburgh fast-food image dataset," in *Proc. 16th IEEE Int. Conf. Image Process.*, 2009, pp. 289–292.
- [16] T. Joutou and K. Yanai, "A food image recognition system with multiple kernel learning," in *Proc. 16th IEEE Int. Conf. Image Process.*, 2009, pp. 285–288.
- [17] L. Scarnato, L. Gianola, C. Wenger, P. Diem, and S. Mougiakakou, "A visual dataset for food recognition and carbohydrates estimation for diabetic patients," in *Proc. 5th Int. Conf. Adv. Technol. Treatments Diabetes (ATTD2012)*, Barcelona, Spain, 2012.
- [18] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.
- [19] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 490–503.
- [20] M. Szummer and R. W. Picard, "Indoor-outdoor image classification," in *Proc. IEEE Int. Workshop Content-Based Access Image Video Database*, 1998, pp. 42–51.
- [21] A. Vailaya, M. Figueiredo, A. Jain, and H. J. Zhang, "Content-based hierarchical classification of vacation images," in *Proc. IEEE Int. Conf. Multimedia Comput. Syst.*, 1999, vol. 1, pp. 518–523.
- [22] F. Mindru, T. Tuytelaars, L. Van Gool, and T. Moons, "Moment invariants for recognition under changing viewpoint and illumination," *Comput. Vis. Image Understanding*, vol. 94, pp. 3–27, 2004.
- [23] T. Tamaki, J. Yoshimuta, M. Kawakami, B. Raychev, K. Kaneda, S. Yoshida, Y. Takemura, K. Onji, R. Miyki, and S. Tanaka, "Computer-aided colorectal tumor classification in NBI endoscopy using local features," *Med. Image Anal.*, vol. 17, pp. 78–100, 2013.
- [24] J. M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts, "Color invariance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 12, pp. 1338–1350, Dec. 2001.
- [25] S. Lloyd, "Least squares quantization in PCM," *IEEE Trans. Inf. Theory*, vol. 28, no. 2, pp. 129–137, 1982.
- [26] Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. 2006 IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recog.*, 2006, vol. 2, pp. 2161–2168.
- [27] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learning*, vol. 20, no. 3, pp. 273–297, 1995.
- [28] S. Haykin, *Neural Networks: A Comprehensive Foundation*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1999.
- [29] A. Gonzalez and J. R. Dorrnsoro, "Natural conjugate gradient training of multilayer perceptrons," *Neurocomputing*, vol. 71, no. 13, pp. 2499–2506, 2008.
- [30] L. Breiman, "Random forests," *Mach. Learning*, vol. 45, no. 1, pp. 5–32, 2001.
- [31] A. Vedaldi and B. Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *Proc. ACM Multimedia*, 2010, pp. 1469–1472.
- [32] K. E. A. Van de Sande, T. Gevers, and C. G. M. Snoek, "Evaluating color descriptors for object and scene recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1582–1596, Sep. 2010.
- [33] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *J. Mach. Learning Res.*, vol. 9, pp. 1871–1874, 2008.
- [34] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, pp. 27:1–27:27, 2011. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- [35] "Random forest (regression, classification and clustering) implementation for MATLAB," (2013, April 9). [Online]. Available: <https://code.google.com/p/randomforest-matlab/>
- [36] B. Fulkerson, A. Vedaldi, and S. Soatto, "Localizing objects with smart dictionaries," in *Proc. 10th Eur. Conf. Comput. Vis.*, 2008, pp. 179–192.
- [37] F. Perronnin, Z. Akata, Z. Harchaoui, and C. Schmid, "Towards good practice in large-scale learning for image classification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 3482–3489.

Authors' photographs and biographies not available at the time of publication.