sustainability

MDPI

Article

Food Image Recognition via Superpixel Based Low-Level and Mid-Level Distance Coding for Smart Home Applications

Jiannan Zheng *, Z. Jane Wang and Chunsheng Zhu

Electrical and Computer Engineering, the University of British Columbia, 5500-2332 Main Mall, Vancouver, BC V6T 1Z4, Canada; zjanew@ece.ubc.ca (Z.J.W.); cszhu@ece.ubc.ca (C.Z.)

* Correspondence: jiannanz@ece.ubc.ca

Academic Editor: Qingchen Zhang

Received: 28 February 2017; Accepted: 12 May 2017; Published: 19 May 2017

Abstract: Food image recognition is a key enabler for many smart home applications such as smart kitchen and smart personal nutrition log. In order to improve living experience and life quality, smart home systems collect valuable insights of users' preferences, nutrition intake and health conditions via accurate and robust food image recognition. In addition, efficiency is also a major concern since many smart home applications are deployed on mobile devices where high-end GPUs are not available. In this paper, we investigate compact and efficient food image recognition methods, namely low-level and mid-level approaches. Considering the real application scenario where only limited and noisy data are available, we first proposed a superpixel based Linear Distance Coding (LDC) framework where distinctive low-level food image features are extracted to improve performance. On a challenging small food image dataset where only 12 training images are available per category, our framework has shown superior performance in both accuracy and robustness. In addition, to better model deformable food part distribution, we extend LDC's feature-to-class distance idea and propose a mid-level superpixel food parts-to-class distance mining framework. The proposed framework show superior performance on a benchmark food image datasets compared to other low-level and mid-level approaches in the literature.

Keywords: food image recognition; smart home applications; low-level and mid-level approaches; superpixels segmentation

1. Introduction

Food image recognition has been attracting increasing attention as a key component in many smart home applications. For instance, in smart kitchen applications, food images are collected by a smart fridge to better plan sustainable grocery shopping; food images are used in smart oven to assist cooking. Compared with other image/object recognition problems, food image recognition for smart home applications have several challenges: (1) training data might be limited and with poor quality; (2) computation power might be limited; (3) food images are deformable with large in-class variance (e.g., different sandwiches from the same class) and small between-class variance (e.g., hamburgers and sandwiches from different classes). Methods providing the best performance on other image databases (e.g., ImageNet Large Scale Visual Recognition Challenge [1]) may not be the best choice for food image recognition in smart home applications. Therefore, more specific and efficient food image recognition methods are still needed to satisfy the special characteristics of food images. Recently, superpixels segmentation methods have been developed and applied to many computer vision problems [2]. Superpixels segmentation methods are suitable for food images since they can effectively segment a food image into food items and parts which are more meaningful in recognition

decision. Therefore, in this work, we adopt the superpixels segmentation idea to improve performance, efficiency and robustness of food image recognition.

During the past decade, image recognition has made great progress in various applications. Generally speaking, recent image recognition researches can be categorized into three major directions: low-level approaches [3–5], mid-level approaches [6–9] and deep learning approaches [10–12]. Despite the superior performance, the major drawbacks of deep learning approaches are their high computational complexity and large memory footprint caused by the huge number of variables and layers. High performance GPUs and large memory are desired for deep learning applications. As a consequence, current applications are mainly based on using remote servers. Thus reliable, fast transmission is required, which could be affected by bad wireless connections and limited server capacity during peak periods. Furthermore, such cloud based applications may raise privacy-related concerns. Since many food image recognition applications are based on mobile devices, which generally are not suitable for direct employment of deep learning approaches, in this work, we focus on cost efficient low-level and mid-level approaches.

There have been many researches tackling the problem of food image recognition via simple low-level feature extraction and coding methods. Low-level features are hand crafted feature vectors, such as Scale Invariant Feature Transform (SIFT) [13], Speeded Up Robust Features (SURF) [14], Histograms of Oriented Gradient (HoG) [15], and color descriptors [16]. For an input image, low-level features are usually densely sampled in a volume of thousands or even tens of thousands. Then the extracted low-level features will be encoded into image representation vectors via feature encoding methods such as Bag-of-Words (BoW), Locality-constrained Linear Coding (LLC) [4] or Fisher Vector (FV) [5,17]. Rahmana et al. proposed a novel food image texture feature with Gabor filter banks to capture rotation and scale invariant texture features of food images and reported good performances in texture rich data [18]. Wazumi et al. used circle segmentations to extract rotation invariant SPIN features suitable for circular Japanese dishes [19]. Zhu et al. employed color and texture features to classify segmented food items in a food database with 63 images and 19 food items [20,21]. He et al. first employed global color texture descriptors on segmented food items for classification, then incorporated two types of contextual information co-occurrence patterns and personalized learning models to improve the performance [22,23]. Aizawa et al. employed global color, circle, BoF and block features and proposed improving the performance with the personal likelihood [24,25]. Anthimopoulos et al. tested various types of features and coding methods and showed that the combination of the SIFT with the color feature yields better performance [26].

The above works mostly focused on less challenging food image datasets. Considering the real scenario in smart home applications where only limited training data is available, Chen et al. introduced a challenging Pittsburgh Fast-Food Image Dataset (PFID) database of popular fast food items which contains 1098 images in 61 categories [27]. Figure 1a shows image samples from the PFID fast food images database. Tests on the PFID database show poor performance when using the basic color and SIFT features with the BoW model. Yang et al. proposed a unique feature learning framework to learn pairwise features based on semantic clustering of image pixels into food items [28]. Qi et al. proposed a co-occurrence local binary pattern which shows significant improvements on the PFID database [29]. However, the usage of these pairwise feature approaches is limited by the dimensionality. For example, the dimension of the image representation of Orientation and Midpoint (OM) in [28] will increase dramatically with the increasing number of food items (e.g., n^3). Recently Wang et al. proposed Linear Distance Coding (LDC) to transform low-level features to a feature-to-class distance pattern [30]. The results on PFID yield state-of-the-art performance. However, the LDC feature transformation structure does not consider the food component information and color information that are crucial in food images. Also, since in LDC, for each image feature, the method calculates the distances with n_{class} codebooks, it requires n_{class} times more computational cost when compared with a regular coding scheme. In this work, we improve LDC for food images by incorporating superpixels segmentation. We extract color information within each superpixel and SIFT descriptors on the edge

of superpixels. In this way, we can better extract food component color information while retaining discriminative edge information between neighbouring superpixels. Furthermore, the proposed approach greatly reduces the computational cost for LDC since we only extract a small number of discriminative SIFT features and calculate distance features. We test the proposed framework on PFID and UEC Food 100 datasets.

(**a**) PFID database (**b**) UEC Food 100 database

Figure 1. Image samples from (**a**) the PFID fast food image database (PFID); (**b**) the UEC Food 100 image database. Each row contains images from the same class.

Although low-level features are simple, fast and effective, since their discriminative power is limited by their small patch size and low dimensionality, they have been shown to be less powerful in capturing image components with higher complexity in images (e.g., a wheel of a car) [6]. When more data is available (e.g., 10,000), researchers found that mid-level visual elements with larger sampling patches are more adaptable to the real-world objects and scene appearance distributions [7,8]. Mid-level visual elements are to capture more complex object parts when compared with edges and corners that can be captured by low-level local features. In order to describe mid-level features, many researches apply HoG patches (with the dimension around 2000) [6–9] and show good results in scene recognition. However, since food dishes are highly deformable, unlike scene parts (doors, wheels), food parts are mostly with arbitrary shapes and sizes, thus the square HoG patches have been shown to be less effective on food image recognition [31]. Another limitation of HoG is that it needs thousands of tests to match a HoG patch in an image, thus its potential in real time mobile based applications is limited. Recently, Bossard et al. proposed a superpixel based mid-level feature extraction approach [31]. Instead of creating thousands of square HoG patches, this approach creates only 30 meaningful patches with arbitrary shapes that are represented by Improved Fisher Vector (IFV).

Once mid-level food parts are extracted, relationships between parts and classes can be discovered. In LDC, the k-means clustering is applied to the feature space within each class to create a set of discriminative features as the class manifolds. Each local feature is then compared with class manifolds to model the feature-to-class distances. However, when applied to high dimensional mid-level parts, both k-means clustering and the l_2 distance metric perform poorly because of the "curse of dimensionality". In the literature, many researches sample parts and test their probabilities of appearances across all classes, then select discriminative parts that appear more frequently in each class. Doersch et al. proposed a mean shift estimation of density ratio to mine discriminative HoG patches [6]. However, this process requires a large amount of testing HoG patches to find the most discriminative ones. Bossard et al. proposed a Random Forest (RF) based method to find high probability parts for each class, and it was shown to be effective for food images [31]. However, this RF based approach also requires a large amount of training data and shows less competitive results on a smaller dataset (with 10,000 data points). In this work, we focus on the limited training data problem to simulate the real scenario in smart home applications. Therefore,

we expand the LDC's feature-to-class distance mining framework to a mid-level food parts-to-class distance mining framework by introducing the following aspects: (1) extracting and representing food parts using superpixels segmentation and IFV representation, in this way better food part selection and representations are formed; (2) constructing food parts density ratio maps across all classes since the number of parts within each class is greatly reduced; and (3) selecting discriminative parts by density ratio and cosine similarity, then an image-to-class distance representation can be formed for each image and further fed into classifiers. A short version of this mid-level approach can be found in a conference version of this paper [32].

To summarize, the main contributions of this work are stated as follows:

1. To tackle the limited data problem, we propose a efficient superpixel based LDC low-level feature extraction framework based on superpixels segmentation to extract discriminative food color and edge distance features. We test the proposed feature extraction framework on a small yet challenging food image dataset, PFID 61, and show improvements in recognition accuracy and robustness against noise and occlusions.
2. To tackle the problem when more training data is available, we expand the LDC's feature-to-class distance mining framework to a mid-level food parts-to-class distance mining framework to learn food parts based on superpixels segmentations. Our proposed framework achieves state-of-the-art performance in the single feature tests. When combining with multiple features, the proposed food image recognition framework can achieve significant performance improvement on the UEC Food 100 database with 10,000 images [33].

The rest of the paper is organized as follows: Section 2 will introduce the proposed superpixel based LDC low-level feature extraction framework and mid-level food parts mining framework, a short version of the presented mid-level food parts mining framework can be found in a conference version of this paper [32]; in Section 3, experiment results will be presented; and Section 4 will draw conclusions and overview future works.

2. Materials and Methods

2.1. Superpixel Based Linear Distance Coding

Here we present the proposed superpixel based LDC low-level feature extraction framework for food image classification. As stated in the previous section, LDC has shown impressive performance by introducing a feature-to-class distance feature. However, LDC brings considerable additional computational cost to the classification process. To improve LDC's efficiency and performance for food image classification, a powerful feature extraction framework for food images is needed to: (a) extract food image edge and color information as well as distance features; (b) reduce computational cost of LDC distance transformation and the following feature quantization problem. The recently developed Simple Linear Iterative Clustering (SLIC) Superpixels Segmentation algorithm [2] has the ability to segment images into meaningful image segments, which are small food pieces in this research, such as a blueberry on a blueberry muffin. Thus, we integrate SLIC Superpixels Segmentation to improve LDC's efficiency and performance for food images.

2.1.1. Introduction to LDC

Low-level local features such as SIFT with common coding frameworks, such as BoW, Soft Assignment BoW and LLC, will introduce information loss due to the feature encoding, which limits the image classification performance. To address this issue, Naive Bayes Nearest Neighbor (NBNN) is introduced to employ an image-to-class distance representation for image classification. LDC is recently introduced in [30] to incorporate NBNN's distance feature between local feature and class as a more discriminative feature to avoid local feature information loss in the feature encoding process and to improve image classification performance.

In NBNN, the distance between a local feature $\mathbf{x}_i$ and a class c is denoted as [30]:

$$d(\mathbf{x}_i, c) = \min_{x \in \mathbf{F}_c} ||\mathbf{x}_i - \mathbf{x}||^2 = ||\mathbf{x}_i - \mathbf{x}_i^c||^2 \quad (1)$$

where $\mathbf{F}_c$ is the local feature set of class c, $\mathbf{x}_i^c$ is the mapping point of $\mathbf{x}_i$ in class c. Since the NBNN method intends to find $\mathbf{x}_i^c$ from the whole set of $\mathbf{F}_c$, two problems arise: (1) distance measure $d(\mathbf{x}_i, c)$ is highly sensitive to noise and outliers in local feature sets $\{\mathbf{F}_c\}$; (2) computational cost is expensive. LDC tackles these problems by sampling discriminative local features within each class. LDC calculates class manifold M_c of class c by clustering the local feature set $\mathbf{F}_c = \{\mathbf{x}_j^c\}$ into n_c anchor points $\{\mathbf{m}_i^c\}_{i=1}^{n_c}$. Similar to the codebook in BoW coding, class manifold can be denoted as [30]:

$$M_c = [\mathbf{m}_1^c, \mathbf{m}_2^c, \ldots, \mathbf{m}_{n_c}^c] \quad (2)$$

Then LDC searches $\mathbf{x}_i$'s mapping point $\mathbf{x}_i^c$ in M_c. One can note that in this way, the complexity is reduced from $\mathcal{O}(NN_D log N_D)$ to $\mathcal{O}(NN_c n_c log(N_c n_c))$, where N_D is the number of all training local features, N_c is the number of classes and $N_c n_c << N_D$. Finally, LDC calculates the distance feature as [30]:

$$d(\mathbf{x}_i, c) = \min_{x \in M_c} ||\mathbf{x}_i - \mathbf{x}||^2 = ||\mathbf{x}_i - \mathbf{x}_i^c||^2 \quad (3)$$

$d(\mathbf{x}_i, c)$ is further normalized with minimum subtraction.

2.1.2. Superpixel Based LDC

Although LDC shows significant performance improvements on the PFID dataset in the literature, it adds additional complexity $\mathcal{O}(NN_c n_c log(N_c n_c))$ to the overall classification cost. Furthermore, food color and structure information are critical for food image recognition. To improve LDC for food images feature extraction, SLIC Superpixels Segmentation is incorporated into the LDC framework.

Superpixels segmentation clusters pixels of an image into meaningful pixel groups which are perceptible by humans. Among various superpixels segmentation algorithms, SLIC Superpixels Segmentation algorithm has outstanding adherence to image boundaries [2]. SLIC is able to segment an input image into a given number of superpixels in similar size with low computational cost. For a given image, the algorithm first samples n_s superpixels centers $\mathbf{C} = [\mathbf{r}, \mathbf{g}, \mathbf{b}, \mathbf{x}, \mathbf{y}]^\top$ on a regular grid, where $[\mathbf{r}, \mathbf{g}, \mathbf{b}]$ are the 3 color components, $[\mathbf{x}, \mathbf{y}]$ are the pixels' position. Then the algorithm assigns each pixel to its nearest center with an adaptive distance measure D given by [2]:

$$D = \sqrt{(\frac{d_c}{m_c})^2 + (\frac{d_s}{m_s})^2} \quad (4)$$

where d_c and d_s are the Euclidean distances of color and spatial components between pixels and centers, m_c and m_s are normalization coefficients decided by the maximum observed color and spatial distances. The algorithm iteratively performs pixel assignments and cluster center updates until residual error is below a threshold. Figure 2 shows some sample food images from the PFID database. As can be observed from the images, food dishes are composed of small food pieces with different color and shape. For instance, a sandwich is composed of green lettuce, brown bread, scarlet beef and yellow cheese. Those food pieces can be easily segmented by SLIC Superpixels Segmentation (Figure 2). Inside each superpixel, color information is consistent and rich while between superpixels edge and corner information is rich. In this work we adopt SLIC Superpixels to extract 300 superpixels $\{S_j\}$ for an input image X_i.

Following superpixels segmentation, color information can be extracted within each superpixel. For each superpixel, a 56-bin histogram is formed which consists of a 24-bin histogram for transformed r, g and b distribution, and a 32-bin histogram for Hue component in HSV color space. Since food images

are taken under various lighting conditions, a scale-invariant and shift-invariant RGB transformation is further applied by normalizing the pixel value distributions according to the following equations [16]:

$$R' = \frac{R - \mu_R}{\sigma_R} \tag{5}$$

$$G' = \frac{G - \mu_G}{\sigma_G} \tag{6}$$

$$B' = \frac{B - \mu_B}{\sigma_B} \tag{7}$$

where μ and σ are the mean and standard deviation of each color component.

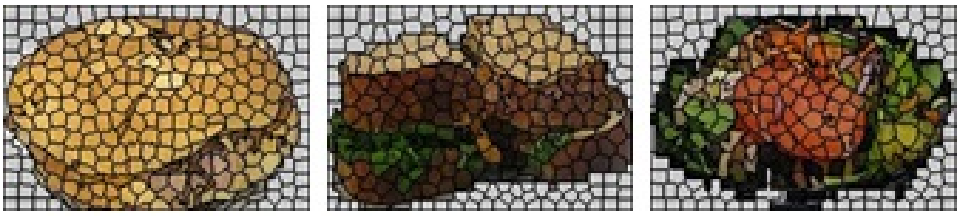

Figure 2. SLIC Superpixels Segmentation results of example images from the PFID fast food image database.

For SIFT features, other than densely sample SIFT patches and treat dense SIFT features equally in LDC process, we only sample SIFT patches at the edge of superpixels where edge information is more salient. In this work we select SIFT patches which cover the superpixel edges for each food image as salient SIFT features. The resulting number of SIFT features is around 15% to 20% of the dense sampled SIFT features. Our reported results show that this sampling scheme is sufficient to capture edge information while reduces the number of data points significantly.

2.1.3. Final Image Representation

In addition to two distance features calculated from superpixel-SIFT and superpixel-Colour features, we extract SIFT and color patch features following [30]. Then an encoding process is applied to each feature channel to generate 4 image representations: R_{SIFT}, R_{COLOR}, $R_{SIFT-LDC}$, $R_{COLOR-LDC}$. To combine different image representations, there are two existing approaches: (1) concatenate different image representations with pre-normalization or post-normalization and feed into one classifier; (2) feed different image representations into different classifiers with following late fusion. Regarding the food image recognition problem, [30] adopted the first approach with pre-normalization; [33–36] adopted the second approach with late fusion. In this work, we concatenate two local feature image representations (R_{SIFT}, R_{Color}) as R_{Low} and two distance feature image representations $(R_{SIFT-LDC}, R_{Color-LDC})$ as R_{LDC} separately with pre-normalization, then feed R_{Low} and R_{LDC} into two linear SVMs classifiers [37]. Finally, late fusion is applied to these two classifiers' outputs Y_{Low} and Y_{LDC} to generate the final label Y_i.

2.1.4. System Flowchart

Figure 3 shows the flowchart of the proposed superpixel based LDC framework, which consists of the following major steps:

1. For an given food image X_i, segment 300 superpixels $\{S_j\}$ by the SLIC Superpixels Segmentation.
2. Extract SIFT features from X_i, based on $\{S_j\}$, and extract color features and sample discriminative SIFT features. Then calculate LDC features from color features and Sampled SIFT features. Finally apply the feature encoding process to generate image representations: (R_{SIFT}, R_{Color}, $R_{SIFT-LDC}$, $R_{Color-LDC}$).

3. Concatenate two local feature representations (R_{SIFT}, R_{Color}) as R_{Low} and two distance feature representations ($R_{SIFT-LDC}$, $R_{Color-LDC}$) as R_{LDC} separately with pre-normalization. Then feed R_{Low} and R_{LDC} into two linear SVMs classifiers [37]. Finally, late fusion is applied to these two classifiers' outputs Y_{Low} and Y_{LDC} to predict the image label Y_i.

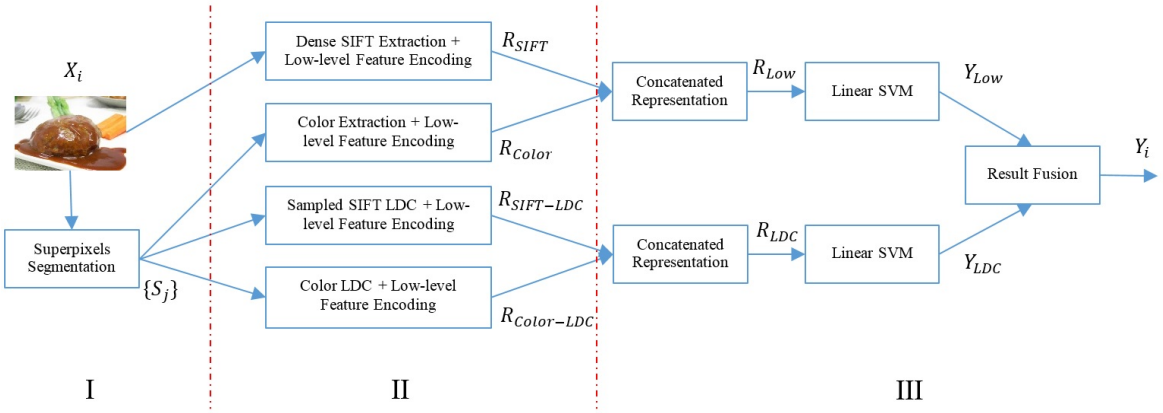

Figure 3. Flowchart of the proposed improved LDC low-level feature extraction framework for food image recognition.

2.2. Expand Low-Level Linear Distance Coding to Mid-Level Food Parts-to-Class Similarity Mining

Although the proposed superpixel LDC framework has already shown improvements in food image recognition as well as computational efficiency on small food image dataset, its performance is limited by low-level local features' discriminative power. Since low-level features have a small patch size, they are capable of capturing only low-level edge and corner information. When more data is available (10,000 data points), more powerful mid-level based method shows superior performance in the literature. To better model deformable food part distribution, we expand LDC's feature-to-class distance coding for low-level features to a mid-level food parts-to-class distance mining, which has the ability to capture larger food parts with higher complexity and discriminative power. Our proposed framework has three main steps: (1) extracting and representing food parts using superpixels segmentation and IFV representation, in this way better food part selection and representations are formed; (2) constructing food parts density ratio maps across all classes since the number of parts within each class is greatly reduced; (3) selecting discriminative parts by density ratio and cosine similarity, then an image-to-class distance representation can be formed for each image and further fed into classifiers. Here we will present the proposed mid-level food parts-to-class distance mining framework in detail. A short version of this mid-level approach can be found in a conference version of this paper [32].

2.2.1. Candidate Food Parts Extraction and Representation

In order to extract image parts as meaningful mid-level features, some researches employed square HoG patches with large size (around 3000 pixels) as image or object parts [6–9]. This approach has two drawbacks: (1) the sliding window sampling and testing of HoG patches are computationally expensive; (2) square patches are weak in capturing deformable food items [31]. To address the first issue, Juneja et al. first introduced superpixels segmentation into mid-level feature mining as a cue to initialize parts set [8]. Bossard et al. employed graph based superpixels segmentation to detect around 30 meaningful food parts with arbitrary shapes [31]. Furthermore, for food parts representation, they adopted IFV to encode low-level features within each superpixel into a high dimensional discriminative representation. In this work, dense RootSIFT and color patches are extracted following Principal

component analysis (PCA) whitening. We adopt graph based superpixels segmentation with IFV as in [31] to represent image parts as candidate mid-level food parts.

2.2.2. Food Parts Density Ratio Map Construction

Following candidate food parts extraction, discriminative and representative food parts need to be selected. In LDC, discriminative local features are selected by k-means clustering to form class manifolds for each class. However, since the feature dimension increases dramatically from low-level features (around 100) to mid-level food parts (around 10,000), neither k-means clustering nor l_2 distance measure works as well as before. Many researches tend to evaluate mid-level features by their probability or coverage within each class. These approaches need to test a large number of candidate HoG patches [6–9] or superpixel patches [31] to achieve better performance. In this work, to tackle limited training data, a novel mid-level food parts selection scheme based on density ratio maps and cosine similarity measure is designed.

To select discriminative image parts, Doersch et al. proposed a mean-shift gradient ascent method to sample HoG patches by density ratios [6]. Given a class c, and its feature set $\mathbf{F}_c^+$, one can sample a negative set $\mathbf{F}_c^-$. The density ratio is defined as [6]:

$$r_i = \frac{\sum_{j=1}^{n^+} max(b - d_{ij}, 0)}{\sum_{j=1}^{n^-} max(b - d_{ij}, 0)} \tag{8}$$

where d_{ij} is the distance measure between two features, b is the bandwidth which defines how much the density is smoothed. To tackle the variance of density across negative feature space, [6] further introduced an adaptive bandwidth method to make the process more robust. That is, to set the denominator in (8) to a constant β and calculate adaptive bandwidth b_i which satisfies [6]:

$$\sum_{j=1}^{n^-} max(b_i - d_{ij}, 0) = \beta \tag{9}$$

r_i can be further defined as [6]:

$$r_i = \frac{\sum_{j=1}^{n^+} max(b_i - d_{ij}, 0)}{\beta} \tag{10}$$

In order to mine discriminative image parts, [6] samples many batches of parts and find local maxima of density ratio distributions with the time consuming gradient ascent method. In this research, since we have considerably small food parts set, we are able to construct a full density ratio map in each class instead of sampling feature batches and finding local maxima of density ratio. To better compare high dimension IFVs, we select cosine similarity which has been shown to be the most powerful distance metric in high-dimensional space:

$$similarity(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\| \|\mathbf{b}\|} \tag{11}$$

Since IFV is l_2 normalized as in [17], the similarity between two food parts $\mathbf{p}_i$ and $\mathbf{p}_j$ can be simplified as:

$$s(\mathbf{p}_i, \mathbf{p}_j) = \mathbf{p}_i \cdot \mathbf{p}_j \tag{12}$$

Then for each food part $\mathbf{p}_i$, the density ratio r_i can be defined as:

$$r_i = \frac{\sum_{j=1}^{n^+} max(s(\mathbf{p}_i, \mathbf{p}_j^+) - b_i, 0)}{\sum_{j=1}^{n^+} max(s(\mathbf{p}_i, \mathbf{p}_j^+) - b_i, 0) + \beta} = \frac{\sum_{j=1}^{n^+} max(\mathbf{p}_i \cdot \mathbf{p}_j - b_i, 0)}{\sum_{j=1}^{n^+} max(\mathbf{p}_i \cdot \mathbf{p}_j - b_i, 0) + \beta} \quad (13)$$

As we can see in (13), r_i is regularized to be in the range of $(0, 1)$. It is beneficial for the following food parts selection and image representation. Algorithm 1 summarizes the procedures of density ratio map ($\{Map_c\}$) construction.

Algorithm 1 Construct the Density Ratio Maps

Input: Extracted Food parts set **F**

Output: Density ratio maps $\{Map_c\}$

1: **for** *each class c* ($c = 1, ..., n_{class}$) **do**
2: construct $\mathbf{F}_c^+$ and sample $\mathbf{F}_c^-$ from **F**
3: **for** *each part* $\boldsymbol{p}_i$ *in* $\boldsymbol{F}_c^+$ **do**
4: find b_i as in Eqn. 9
5: calculate r_i as Eqn. 13
6: **end for**
7: construct the Density Ratio Map Map_c from $\{r_i\}$
8: **end for**
9: **return** $\{Map_c\}$

2.2.3. Food Parts Selection and Image Representation

In order to select discriminative image parts, the intuitive way is to select the top ranked image parts for each class c from Map_c. To make the selection more representative and diverse, [6] proposed a purity-coverage plot constructed by counting image parts' pixel coverage in a test set. [31] ignores models that have more than half of the same image parts with higher ranking models after sorting. These approaches bring additional computational cost and implementation complexity. In this work, we propose a more directive and less complex approach. For a given class c and its density ratio map Map_c constructed in the previous paragraph, we iteratively sort and sample image parts based on r_i and b_i. In every iteration, we select the image part $\mathbf{p}_i$ with the highest density ratio r_i, then we remove parts $\{\mathbf{p}_j\}$ from the density ratio map which satisfy the following condition:

$$s(\mathbf{p}_i, \mathbf{p}_j) - b_i * (r_i * \eta) > 0 \quad (14)$$

where η is a parameter to control the mining rate. With increasing r_i, the number of removed parts $\{\mathbf{p}_j\}$ decreases. Thus, the proposed process tends to select more discriminative parts while also taking less discriminative parts into account to increase diversity of selection. The selection will stop as either there are less than a certain number of parts remaining in the density ratio map, or the number of selected parts reaches the give number. The algorithm of food part selection is summarized in Algorithm 2. In experiments, considering the trade off between performance and computational cost, we select around 200 food parts per class. Figure 4 shows selected food parts with their regularized density ratio from 3 different food categories: sushi, hamburger steak and meat sauce spaghetti. In the example of sushi, parts of salmon, different sushi and laver are selected. In the example of hamburger steak, different beef components and vegetables are selected. In the example of meat sauce spaghetti, meat sauce, spaghetti, green beans and part of the plate are selected as food parts. Here, since spaghetti is more likely to be served on a round white plate, it can be used to distinguish spaghetti from other food

classes such as "sushi" or "eel on rice". Therefore, spaghetti plate can be considered as a representative food part as well.

After selection, codebook $Codebook_{d\times n_p}$ containing n_p selected parts from all classes is built. Given an input food part $\mathbf{p}_{1\times d}$ represented as a d dimension IFV, the cosine similarity score vector can be calculated via cosine similarity weighted by density ratio:

$$Score_{1\times n_p} = \mathbf{p}_{1\times d} * Codebook_{d\times n_p} \cdot \{r_i\}_{1\times n_p} \quad (15)$$

For the an given input food image, we apply max pooling method over cosine similarity score vectors $\{Score\}$ to form a $1 \times n_p$ final image representation.

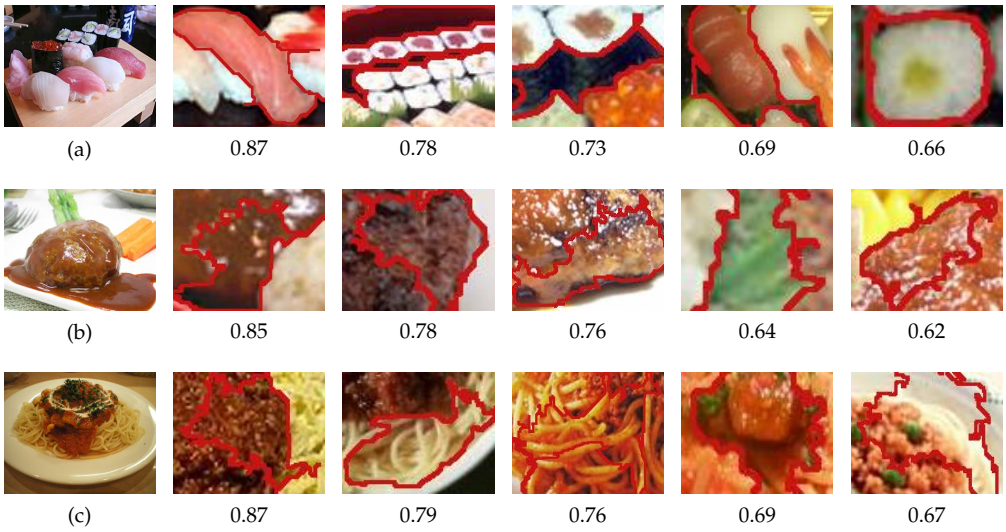

Figure 4. Typical examples of selected mid-level food parts and their corresponding regularized density ratio results, (**a**) sushi; (**b**) hamburger steak; (**c**) meat sauce spaghetti.

Algorithm 2 Food Parts Selection

Input: Density ratio maps $\{Map_c\}$

Output: $Codebook_{d\times n_p}$

1: **for** *each class c* ($c = 1, ..., n_{class}$) **do**
2: **while** stopping criteria not satisfied **do**
3: select $\mathbf{p}_i$ with highest r_i
4: remove $\{\mathbf{p}_j\}$ satisfies Equationn (14)
5: **end while**
6: **end for**
7: **return** $Codebook_{d\times n_p}$

2.2.4. System Flowchart

Figure 5 shows the flowchart of the proposed mid-level food parts-to-class distance mining framework, which consists of the following major steps:

1. For an given food image X_i, segment 30 superpixels $\{S_j\}$ by graph based superpixels segmentation as in [31]. Then extract SIFT and color features following IFV feature encoding process to generate food parts set **F** [31].
2. Construct density ratio maps $\{Map_c\}$ as described in Algorithm 1.
3. Select discriminative food parts and construct $Codebook_{d\times n_p}$ as described in Algorithm 2.

4. For each given image and its superpixels set $\{S_j\}$, cosine similarity score vectors $\{Score\}$ will be calculated by (15). Then max pooling method will be applied to form a $1 \times n_p$ final image representation. A linear SVM will be applied to predict the image label Y_i.

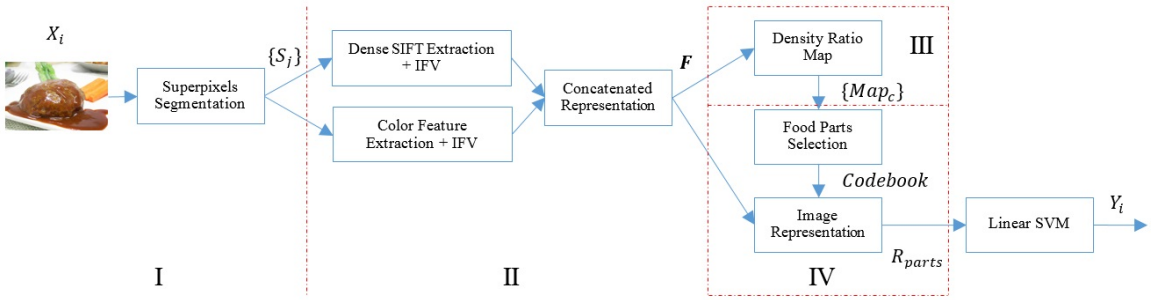

Figure 5. Flowchart of the proposed Mid-Level Food Parts-to-Class Similarity Mining Framework.

3. Results and Discussion

In this section, we evaluate the performances of the proposed superpixel based LDC for low-level food image feature extraction framework and mid-level food parts-to-class distance mining framework. We tested the proposed frameworks on two food image databases: the PFID food image database and the UEC Food 100 database.

3.1. Experiment on PFID Database

We first tested the proposed low-level approach on PFID to evaluate the performance as well as robustness through two tests with very limited training data. The PFID dataset is a small and challenging dataset with fast food images from 13 chain restaurants acquired under both lab and realistic settings. Each food category only contains three different instances of food (bought on different days from different branches of the restaurant chain). Each food instance has six images from six viewpoints (60 degrees apart), 1098 images in total.

We followed the experimental protocol proposed by Chen et al. and performed 3-fold cross-validation for the experiments, only using the 12 images from two instances for training and the other 6 images from the third for testing [27]. We repeated this procedure three times, with a different instance serving as the test set and averaged the results. The protocol ensures that no image of any given food item ever appears in both the training and testing sets and guarantees that food images are selected from different restaurants on different days. In our experiment, to have a fair comparison, we employed common settings in feature extraction and coding as in the literature [28,30,36]. For SLIC Superpixels Segmentation, we set n_s to 300 for all images. For SIFT feature extraction, we extracted features based on image patches of 16-by-16 pixels in the scale of every 4 pixels. For color feature, we extracted 56 bins histogram of each superpixel as stated in Section 2.1.2. For LDC, we set the size of class manifolds to 1024 as in [30]. For coding process, in LLC, we set the codebook size to 2048. In IFV, we set the number of Gaussian mixture model (GMM) to 256. We only applied a 3 level spatial pyramid for LLC.

3.1.1. PFID Clean Data

In this experiment we tested the performance of the proposed superpixel based LDC framework with three different coding methods: Orientation Midpoint category (OM) [28], LLC and IFV and compared with state-of-the-art results. In [28] the authors randomly sample 1000 pixels from each training food image, then cluster pixels into 8 classes and construct pairwise OM features for classification. In the experiment we extracted 300 color descriptors per image by superpixels segmentation and clustered

them into food elements. Our superpixel based LDC approach improves OM by 4%, yet has much less computational cost in coding process (300 * 300 compared to 1000 * 1000 searching for each food image).

Then we tested the proposed framework with LLC and compared with the state-of-the-art performance, which is 48.45% achieved by LDC in [30]. First, we implemented SIFT + LLC/IFV with/without LDC to reproduce performance in [30]. As in Table 1, the performances of our implementation of SIFT + LLC and SIFT + LDC + LLC are slightly worse than [30] due to implementation details. For better comparison of the proposed feature extraction framework, we mainly compare the results from our own implementation. Compared with SIFT + LDC with LLC, the proposed approach boosts the performance to 50.45%, which is an improvement of about 3%. The reason for this improvement is that the proposed low-level feature extraction framework captures both discriminative edge information between food components and color information inside food components. Together a more structured food image representation is formed.

Table 1. Recognition Accuracy Results on the PFID 61 database.

Methods	Accuracy
ColourH [28]	11.3%
OM [28]	28.2%
BoW [27]	30.6%
PRICoLBP [29]	45.4%
SIFT + LLC [30]	44.63%
SIFT + LDC + LLC [30]	48.45%
Colour Descriptor + OM [28] [1]	32.31%
SIFT + LLC [30] [2]	43.25%
SIFT + LDC + LLC [30] [2]	47.42%
Superpixel LDC + LLC	**50.45%**
SIFT + IFV	45.72%
SIFT + LDC + IFV	46.63%
Superpixel LDC + IFV	48.63%

[1] The performance of our improved implementation of OM in [28]; [2] The performance of our implementation of the algorithms in [30].

We also tested the proposed framework with IFV and compared performance with LLC intuitively. From results, we can see SIFT + IFV shows 2% higher accuracy than SIFT + LLC. However, improvement of LDC with IFV is lower than LDC with LLC. When comparing LLC and IFV, LLC does not capture distance features, while IFV captures first and second order distances between local features and GMM centers. In addition, LDC captures the distances between SIFT and class manifolds, which is complement of basic IFV's distance measure. As a consequence, LDC boosts IFV for around 1% in classification accuracy. Results of the proposed superpixel based LDC approach has 2% improvement to SIFT + LDC with IFV.

Considering the computational cost, LDC produces $\mathcal{O}(NN_cn_clog(N_cn_c))$ complexity, while the proposed superpixel based LDC approach extracts 80% fewer numbers of local descriptors (N) than dense SIFT sampling, which reduces LDC's computational cost and memory footprint around 5 times. In summary, the proposed superpixel based LDC food image feature extraction framework is effective and efficient on food image data even with very limited training data.

3.1.2. PFID Noise Data

Since a large proportion of food image data is captured by mobile devices, noise and occlusions may occur. Thus, robustness is also a crucial concern in food image classification. In this experiment we evaluated robustness of the proposed improved low-level food image feature extraction framework with PFID noise data.

We followed experiment settings in [38] and compared performance with their approach. To tackle corrupted image data, Bao et. al. proposed a Corruptions Tolerant Discriminant Analysis algorithm which learns three underlying subspaces from training data to separate desired properties from undesired properties and corruptions. To investigate their algorithm, contiguous occlusions and random pixel corruptions are randomly added to the PFID food image dataset. For each class in PFID, 50% of training images and 50% of testing images are corrupted randomly. For contiguous occlusions, black blocks with sizes of 80 × 80 (7%), 100 × 100 (11%), 120 × 120 (16%) and 140 × 140 (22%) are added to different locations in the images. For random pixel corruption, 5%, 10% and 15% of image pixels are corrupted randomly.

Figure 6 shows the superpixels segmentation result of pixel corruption and block occlusion image data. In Figure 7, compared with baseline SIFT + LLC and SIFT + LDC + LLC, the proposed superpixel based LDC+LLC is less sensitive to change of corruption rate and block size. Especially with increasing block size of occlusions, the proposed approach shows more robustness improvement than the baseline approaches. This robustness improvement can be explained in Figure 6, the blocks in the images are successfully segmented by superpixels segmentation. We can explore that pixel noise affect superpixels segmentation, but food image items are still extracted successfully, which guarantees the robustness of the proposed improved LDC approach against pixel noise.

(**a**) PFID noisy data

(**b**) SLIC Superpixels Segmentation results

Figure 6. PFID noisy data and corresponding SLIC Superpixels Segmentation results. (**a**) Clean image, 5% pixel corruption, 15% pixel corruption, 80 * 80 block occlusion and 140 * 140 block occlusion. (**b**) Corresponding SLIC Superpixels Segmentation results.

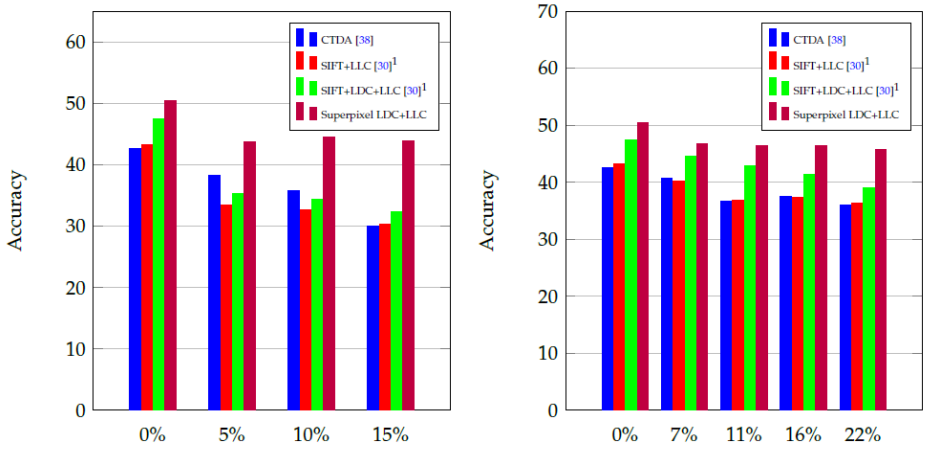

Figure 7. Classification accuracy results on the PFID noisy dataset.

3.2. Experiment on UEC Food 100 Database

In this experiment we tested both the proposed superpixel based LDC (low-level) and mid-level approaches on the UEC Food 100 image database. UEC Food 100 contains 100 food categories with more than 100 images for each category. Some images contain multiple classes of food. Bounding box information is provided in the database. The total number of food images in UEC Food 100 is 12,905. We followed the experiment protocol as stated in [36]. For each category we selected 20 images as testing data, and the rest are training data. We evaluated classification accuracy in 5 trials in the five-fold cross validation manner.

For fair comparison, we employed the same settings for feature extraction and coding as in [36]. We extracted SIFT at the scale of every 6 pixels and the same color descriptors as in [36]. For the proposed superpixel based LDC low-level feature extraction, we segmented 300 superpixels per image and set the class manifold size to 1024. For low-level feature encoding, we employed IFV with 64 GMMs and a 3 level spatial pyramid. For the proposed mid-level food parts approach, we segmented 30 superpixels per image and extracted dense SIFT and color within each superpixel's bounding box as in [31]. Following PCA-whitening and dimension reduction, an IFV with 64 GMMs is constructed for each superpixel. We selected 200 discriminative and representative food parts per category and a 200×100 image representation is formed for classification.

Figure 4 shows some selected food parts from three different food categories with their density ratio r_i. In order to make selected food parts more diverse, the proposed food parts selection scheme selects not only discriminative parts with a high density ratio, but also less discriminative parts with a low density ratio. We further weighted select parts with regularized density ratio and formed final image representations based on cosine similarity.

Table 2. Recognition Accuracy Results on UEC Food 100.

	Methods	**Accuracy**
Single feature	HoG + IFV [36]	50.14%
	Colour + IFV [36]	53.04%
	DCNN [36]	57.87%
	SIFT + IFV	48.25%
	Colour + IFV [36] [1]	52.80%
	Mid-level approach	**60.50%**
Combined feature	HoG + Colour + IFV [36]	65.32%
	HoG + Colour + IFV + DCNN [36]	**72.26%**
	SIFT + Colour + IFV	66.12%
	SIFT + Colour + IFV + Superpixel LDC	67.65%
	SIFT + Colour + IFV + Mid-level approach	70.84%

[1] The performance of our implementation of algorithm in [36].

Table 2 shows the classification accuracy of the proposed approaches and state-of-the-art results. For single feature performance comparison, the proposed mid-level food parts-to-class approach achieves the best performance with 60.50% and outperforms low-level feature with IFV and Deep Convolution Neural Networks (DCNN) feature's 57.87% in the literature. When combined with low-level feature based methods, our implementation of RootSIFT + Colour is slightly better than RootHoG + Colour since SIFT is considered to be more powerful than HoG. Then we tested the proposed superpixel based LDC with RootSIFT + Colour and showed a performance improvement of 1.53%. Finally, we combined the proposed mid-level food parts approach with the superpixel based LDC approach. Our proposed food parts approach significantly improves the classification accuracy by 4.7%, achieving 70.84% which is only beaten by DCNN combined approach's 72.26% in [36]. We would like to mention that DCNN in [36] have large memory footprint and requires large amount of training data and computation resources (e.g., one million data). In comparison, the proposed method is more efficient with competitive performance and

thus it is more suitable for potential smart home applications. As a result, the proposed mid-level food parts-to-class mining framework yields the best performance in single feature tests, and comparable performance with DCNN trained with large scale data in combined feature tests.

4. Conclusions

In this paper, we tackled the challenging smart home food image recognition problem with different setups. For the limited data problem, we proposed a superpixel based LDC low-level feature based approach which is suitable for very limited training data. We improved the LDC method by extracting discriminative food item information based on the superpixels segmentation. In the experiments on challenging PFID food image database, our proposed superpixel based LDC approach shows promising performance improvement and robustness against noise and occlusions. In addition, the proposed superpixel based LDC approach significantly reduces the computational cost when compared with the original LDC approach. When more data points are available, we proposed a mid-level food image parts based method by expanding the LDC's local feature-to-class distance to a mid-level food parts-to-class distance mining approach and designed a simple and effective food parts selection scheme. In the experiments on middle-size database UEC Food 100, the proposed mid-level approach significantly outperforms other single feature based approaches. When combined with low-level feature based approach, the proposed mid-level approach improves classification accuracy from 66.12% to 70.84%, only beaten by DCNN approach in [36] which was trained with one million images.

Acknowledgments: The work described in this paper has been supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) (with both a Discovery grant and a Strategic Partnership grant.

Author Contributions: Jiannan Zheng and Z. Jane Wang conceived and designed the methods; Jiannan Zheng performed the experiments; Jiannan Zheng , Z. Jane Wang and Chunsheng Zhu analyzed the data; Jiannan Zheng wrote the paper.

Conflicts of Interest: The authors declare no conflict of interest. The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

LDC	Linear Distance Coding
SIFT	Scale Invariant Feature Transform
SURF	Speeded Up Robust Features
HoG	Histograms of Oriented Gradient
BoW	Bag-of-Words
LLC	Locality-constrained Linear Coding
FV	Fisher Vector
IFV	Improved Fisher Vector
PFID	Pittsburgh Fast-Food Image Dataset
OM	Orientation and Midpoint
RF	Random Forest
SLIC	Simple Linear Iterative Clustering
NBNN	Naive Bayes Nearest Neighbor
PCA	Principal component analysis
GMM	Gaussian mixture model
DCNN	Deep Convolutional Neural Networks

References

1. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet Large Scale Visual Recognition Challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252, doi:10.1007/s11263-015-0816-y.
2. Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Susstrunk, S. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *34*, 2274–2282, doi:10.1109/TPAMI.2012.120.
3. Fei-Fei, L.; Perona, P. A Bayesian hierarchical model for learning natural scene categories. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005 (CVPR 2005), San Diego, CA, USA, 20–25 June 2005; Volume 2, pp. 524–531.
4. Wang, J.; Yang, J.; Yu, K.; Lv, F.; Huang, T.; Gong, Y. Locality-constrained linear coding for image classification. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 3360–3367.
5. Perronnin, F.; Dance, C. Fisher kernels on visual vocabularies for image categorization. In Proceedings of the 2007 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07), Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
6. Doersch, C.; Gupta, A.; Efros, A.A. Mid-level visual element discovery as discriminative mode seeking. *Adv. Neural Inf. Process. Syst.* **2013**, 494–502.
7. Singh, S.; Gupta, A.; Efros, A.A. Unsupervised discovery of mid-level discriminative patches. In *Computer Vision—ECCV 2012*; Springer: New York, NY, USA, 2012; pp. 73–86.
8. Juneja, M.; Vedaldi, A.; Jawahar, C.; Zisserman, A. Blocks that shout: Distinctive parts for scene classification. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Portland, OR, USA, 23–28 June 2013; pp. 923–930.
9. Doersch, C.; Singh, S.; Gupta, A.; Sivic, J.; Efros, A. What makes Paris look like Paris? *ACM Trans. Graph.* **2012**, *31*, doi:10.1145/2185520.2185597.
10. Lawrence, S.; Giles, C.; Tsoi, A.C.; Back, A. Face recognition: A convolutional neural-network approach. *IEEE Trans. Neural Netw.* **1997**, *8*, 98–113.
11. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* **2012**, 1097–1105.
12. Karpathy, A.; Toderici, G.; Shetty, S.; Leung, T.; Sukthankar, R.; Fei-Fei, L. Large-scale video classification with convolutional neural networks. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 1725–1732.
13. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110.
14. Bay, H.; Tuytelaars, T.; Van Gool, L. Surf: Speeded up robust features. In *Computer Vision—ECCV 2006*; Springer: New York, NY, USA, 2006; pp. 404–417.
15. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
16. Van de Sande, K.E.A.; Gevers, T.; Snoek, C.G.M. Evaluating Color Descriptors for Object and Scene Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2010**, *32*, 1582–1596.
17. Perronnin, F.; Sánchez, J.; Mensink, T. Improving the fisher kernel for large-scale image classification. In *Computer Vision–ECCV 2010*; Springer: New York, NY, USA, 2010; pp. 143–156.
18. Rahmana, M.H.; Pickering, M.R.; Kerr, D.; Boushey, C.J.; Delp, E.J. A new texture feature for improved food recognition accuracy in a mobile phone based dietary assessment system. In Proceedings of the 2012 IEEE International Conference on Multimedia and Expo Workshops (ICMEW), Melbourne, Australia, 9–13 July 2012; pp. 418–423.
19. Wazumi, M.; Han, X.H.; Ai, D.; Chen, Y.W. Auto-recognition of food images using SPIN feature for Food-Log system. In Proceedings of the 2011 6th International Conference on Computer Sciences and Convergence Information Technology (ICCIT), Seogwipo, Korea, 29 November–1 December 2011; pp. 874–877.
20. Zhu, F.; Bosch, M.; Woo, I.; Kim, S.; Boushey, C.J.; Ebert, D.S.; Delp, E.J. The use of mobile devices in aiding dietary assessment and evaluation. *IEEE J. Sel. Top. Signal Process.* **2010**, *4*, 756–766.
21. Zhu, F.; Bosch, M.; Boushey, C.J.; Delp, E.J. An image analysis system for dietary assessment and evaluation. *ICIP* **2010**, 1853–1856.

22. He, Y.; Xu, C.; Khanna, N.; Boushey, C.J.; Delp, E.J. Food image analysis: Segmentation, identification and weight estimation. In Proceedings of the 2013 IEEE International Conference on Multimedia and Expo (ICME), San Jose, CA, USA, 15–9 July 2013; pp. 1–6.
23. He, Y.; Xu, C.; Khanna, N.; Boushey, C.J.; Delp, E.J. Context based food image analysis. In Proceedings of the 20th IEEE International Conference on Image Processing (ICIP), Melbourne, Australia, 15–18 September 2013; pp. 2748–2752.
24. Aizawa, K.; Maruyama, Y.; Li, H.; Morikawa, C. Food balance estimation by using personal dietary tendencies in a multimedia food log. *IEEE Trans. Multimedia* **2013**, *15*, 2176–2185.
25. Kitamura, K.; de Silva, C.; Yamasaki, T.; Aizawa, K. Image processing based approach to food balance analysis for personal food logging. In Proceedings of the 2010 IEEE International Conference on Multimedia and Expo (ICME), Singapore, 19–23 July 2010; pp. 625–630.
26. Anthimopoulos, M.; Gianola, L.; Scarnato, L.; Diem, P.; Mougiakakou, S. A Food Recognition System for Diabetic Patients Based on an Optimized Bag-of-Features Model. *IEEE J. Biomed. Health Inform.* **2014**, *18*, 1261–1271.
27. Chen, M.; Dhingra, K.; Wu, W.; Yang, L.; Sukthankar, R.; Yang, J. PFID: Pittsburgh fast-food image dataset. In Proceedings of the 2009 16th IEEE International Conference on Image Processing (ICIP), Cairo, Egypt, 7–10 November 2009; , pp. 289–292.
28. Yang, S.; Chen, M.; Pomerleau, D.; Sukthankar, R. Food recognition using statistics of pairwise local features. In Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), San Francisco, CA, USA, 13–18 June 2010; pp. 2249–2256.
29. Qi, X.; Xiao, R.; Guo, J.; Zhang, L. Pairwise rotation invariant co-occurrence local binary pattern. In *Computer Vision–ECCV 2012*; Springer: New York, NY, USA, 2012; pp. 158–171.
30. Wang, Z.; Feng, J.; Yan, S.; Xi, H. Linear distance coding for image classification. *IEEE Trans. Image Process.* **2013**, *22*, 537–548.
31. Bossard, L.; Guillaumin, M.; Van Gool, L. Food-101–Mining Discriminative Components with Random Forests. In *Computer Vision–ECCV 2014*; Springer: New York, NY, USA, 2014; pp. 446–461.
32. Zheng, J.; Wang, J. Superpixel-based Image Recognition for Food Images. In Proceedings of the 29th Annual IEEE Canadian Conference on Electrical and Computer Engineering (CCECE), Vancouver, BC, Canada, 5–18 May 2016.
33. Kawano, Y.; Yanai, K. Real-time mobile food recognition system. In Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Portland, OR, USA, 23–28 June 2013; pp. 1–7.
34. Matsuda, Y.; Hoashi, H.; Yanai, K. Recognition of multiple-food images by detecting candidate regions. In Proceedings of the 2012 IEEE International Conference on Multimedia and Expo (ICME), Melbourne, Australia, 9–13 July 2012; pp. 25–30.
35. Kawano, Y.; Yanai, K. FoodCam: A real-time food recognition system on a smartphone. *Multimedia Tools Appl.* **2015**, 74, 5263–5287.
36. Kawano, Y.; Yanai, K. Food image recognition with deep convolutional features. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*; ACM: New York, NY, USA, 2014; pp. 589–593.
37. Fan, R.E.; Chang, K.W.; Hsieh, C.J.; Wang, X.R.; Lin, C.J. LIBLINEAR: A library for large linear classification. *J. Mach. Learn. Res.* **2008**, *9*, 1871–1874.
38. Bao, B.K.; Liu, G.; Hong, R.; Yan, S.; Xu, C. General subspace learning with corrupted training data via graph embedding. *IEEE Trans. Image Process.* **2013**, *22*, 4380–4393.