Analysis of SURF and SIFT Representations to Recognize Food Objects

Mohd Norhisham bin Razali[1,2], Noridayu Manshor[1],Alfian Abdul Halin[1] ,Norwati Mustapha[1] and Razali Yaakob[1]
[1]*Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43300 Serdang, Selangor,Malaysia*
[2]*Faculty of Computing and Informatics, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia*
hishamrz@ums.edu.my

***Abstract*— The social media services such as Facebook, Instagram and Twitter has attracted millions of food photos to be uploaded every day since its inception. Automatic analysis on food images are beneficial from health, cultural and marketing aspects. Hence, recognizing food objects using image processing and machine learning techniques has become emerging research topic. However, to represent the key features of foods has become a hassle from the immaturity of current feature representation techniques in handling the complex appearances, high deformation and large variation of foods. To employ many kinds of feature types are also infeasible as it inquire much pre-processing and computational resources for segmentation, feature representation and classification. Motivated from these drawbacks, we proposed the integration on two kinds of local feature namely Speeded-Up Robust Feature (SURF) and Scale Invariant Feature Transform (SIFT) to represent the features large variation food objects. Local invariant features have shown to be successful in describing object appearances for image classification tasks. Such features are robust towards occlusion and clutter and are also invariant against scale and orientation changes. This makes them suitable for classification tasks with little inter-class similarity and large intra-class difference. The Bag of Features (BOF) approach is employed to enhance the discriminative ability of the local features. Experimental results demonstrate impressive overall recognition at 82.38% classification accuracy from the local feature integration based on the challenging UEC-Food100 dataset. Then, we provide depth analysis on SURF and SIFT implementation to highlight the problems towards recognizing foods that need to be rectified in the future research.**

***Index Terms*— Bag of Features; Food Recognition; Image Classification; Local Features.**

I. INTRODUCTION

Object Recognition research generally aims to solve the problems of classifying the objects into pre-defined category using image processing and machine learning techniques. Many recent works have been found to use object recognition techniques to recognize food objects as well which is important in developing an automatic dietary assessment system. However, food recognition is not a simple test case due lack of capability of current recognition approaches to handle to the complex appearances, high deformation as very large variation of foods[1]–[5]. In this context, feature representation to describe key features in an image is very crucial for reliable object recognition. A variety of low-level invariant features are available to describe the object appearances. Local features using Speeded-up Robust Feature Transform (SURF) is computationally efficient to detect and derive meaningful local descriptors. However, due to complex appearances of the real food images, using a single descriptor in isolation is not sufficient to effectively represent the large variation of foods. Therefore, using local features in combination is proven to be more beneficial. The contribution of this paper are four folds. First, we proposed the integration of Scale Invariant Feature Transform (SIFT) with SURF to capture denser key-points and more descriptive and discriminative features from large variation of foods. Bag-of-Feature(BOF)[6] approach is used to tokenize the local features key-points into two visual vocabularies. In many previous food recognition study, SIFT has used to represent food features as it provides a powerful descriptor due its stability under different scale and orientation changes as well as being robust to occlusion and clutter[7], [8]. In addition to that, the use of many feature types from both local and global will increase the computational cost during feature representation and classification as well as the pre-processing overhead from the segmentation process. Second, due to lack of evaluation of local feature in previous study, we provide few analysis on SURF and SIFT in term of feature representation efficiency, key-points detection, classification performance. Finally, we discussed the factors that contribute to the ineffectiveness to the key-points detection and feature representation methods classification performance, specifically in food objects. The feature representation method is evaluated using UEC-Food100 dataset, whose images have complex appearances, non-rigid deformation, fine-grained as well extremely huge in variations[1], [3], [9]–[11]. The remaining paper is organized as follows: In the second section, we provide the literature review on feature representation methods of food objects and the local features using SURF and SIFT. The third section described the experimental procedure undertaken to integrate the local features. In the fourth section, we present the results from the experiments and, the analysis of the key-point detections and overall classification performance. We draw the conclusions in the last section of this paper.

II. FEATURE REPRESENTATION METHODS

Food recognition is a specific topic in Object Category Recognition (OCR) which applied exiting methods in OCR to recognize food from images which mostly the prepared or cooked foods. The application of OCR to recognize foods are motivated from the popularity and advancement of mobile phone technology such as good imaging quality,

memory capacity, network connectivity and processers. Recognizing foods provide potentials to use the mobile phone in dietary assessment and other healthcare applications[1], [2], [9], [12]–[16] to combat obesity and overweight problems that lead to many serious diseases. In this paper, we are specifically look into the feature representation methods using local features and BoF to represent the features from food objects.

A. Food Object Recognition

The noteworthy performance of general object recognition methods are not a guarantee to be robust enough to recognize foods[1], [17]. Large variations in shape and deformation makes it difficult for recognition algorithms to distinguish among food categories[17]–[20]. Feature representation hence plays a vital role to map the low-level features to higher-level concepts. Recently, many works have sprung up for food recognition systems. Among the catalysts is the popularity of smartphones and social media services[4], [21], [22]. Numerous feature representation methods have been proposed to describe food images using both global and local features. Global features describe the entire image pixels meanwhile local features describe image patches based on detected key-points[23].

In the literature, local features are frequently used due to their capability to represent the unique properties of specific food types. SIFT falls within this category of features, and is frequently used along with Bag-of Feature encoding. Local features can be complemented with global features in order to provide better representation. Joutou and Yanai[24] implemented a Multiple Kernel Learning (MKL) technique to adaptively learn the diversity of foods using Bag of SIFT, Gabor filter and color histogram features, where they obtained a 61.34% classification accuracy on 50 food categories. Hoashi et. al[25] enhanced this work with an increased set of features, specifically by adding Histogram of Gradient(HOG) feature, and yielded slightly improved classification accuracy of 62.52% on 85 food categories. To cater for food images containing multiple food classes, Matsuda[26] proposed a Deformable Part Model (DPM), circle detector and JSEG segmentation. For each candidate region, Bag of SIFT and CSIFT, HOG and Gabor filter responses were extracted. Their method obtained a lower classification accuracy of 58% for multiple objects compared to 68.9% for single objects, where 100 food categories were considered. Later on, Kawano et. al[27], [28] used two separate feature combinations, where firstly they tried Bag of SURF and color patch, and then HOG patch and color patch. The latter combination yielded the best results with 79.2% recognition accuracy on 100 food categories as it used fisher vector which is known is effective in the recent image representation. However, feature representation using HOG is less compact compared to SIFT as SIFT compute more key-points during localization. In addition to that, HOG create highly sparse features from the local region around the corners which is less sufficient to describe the object[29].

In summary, it can be seen that the combination of local and global features can provide more discriminative prowess. However, one trade-off is the higher pre-processing overhead especially when global features are used for food region segmentation. Also, different kind features require different extraction techniques that could generate lengthy feature vectors. All these can potentially increase complexity as well as overall computational cost. In addition to that, comparison between feature types is difficult since many existing work are evaluated using different datasets[30].

B. Local Features

Local feature representation such as SURF and SIFT have been proven to be effective to represent images due to their capability to provide a high discriminability. Bag-of-Features (BOF) model is adopted to represent the local features using histogram have gained several popularity due to its simplicity and robustness[31].

1) Speeded-up Robust Feature (SURF)

This detector proposed a ‘fast-hessian’ detector. It is based on the basic approximation of Hessian matrix which relies on integral images to decrease computation time. Hessian matrix has a good performance in term of computation time and accuracy[32]. SURF describe the distribution of Haar-wavelet responses within the interest point neighbourhood and 64 dimensions are produced. Basically, there are four steps in SURF which are 1) to find image interest points by using Hessian matrix, 2) to find the major interest points in scale space by using non-maximal suppression, 3) to find feature direction to produce rotationally invariant features and finally 4) to produce feature vectors. The interest points are selected at different locations such as corners, blobs and T-junctions. A good detector should be repetitive where it can detect the same interest points in different views. The number of interest points can be controlled by setting up the threshold to select the major features. The neighbourhood of each interest point is represented by a feature vector before it finally matched between different images. The matching is based on a distance between the vectors, e.g. the Mahalanobis or Euclidean distance. The dimension of the descriptor has a direct impact on the time this takes, and less dimensions are desirable for fast interest point matching. However, lower dimensional feature vectors are in general less distinctive than their high-dimensional counterparts.

2) Scale Invariant Feature Transform (SIFT)

The Scale invariant feature transform(SIFT) generates a set of local descriptors that compute the interest points using DoG. It generate patch with size 16 X 16 and divided into 4 X 4 sub-regions. A 128-dimensional histogram will be generated after concatenating the histogram from these sub-regions. Various types of SIFT have been emerged such as PCA-SIFT, color SIFT and ASIFT. PCA-SIFT apply Principal Component Analysis(PCA) to reduce the patch dimensionality to become 20 instead of 128. Color SIFT processed the color value instead of grayscale value and the variants including HSV-SIFT, HueSIFT, OpponentSIFT, C-SIFT, rgSIFT, Transformed color SIFT and RGB-SIFT. The Affline SIFT (ASIFT) generates a set of patches by warping the original patch to handle the changes of viewpoints. SIFT is computer over the warped patch. Another variant of SIFT been proposed are GLOH which is more distinctive [33].

III. SURF-SIFT FEATURE REPRESENTATION

A. Bag of Features Model

We adopted BoF model to encode the low-level features produced by SURF and SIFT as shown in Figure 1. The proposed feature combination method is evaluated using the

UEC-FOOD100[16], [26] dataset, that consists of 100 food categories. In total, 14,467 JPEG food images were used (each picture having a different pixel dimensions). On average, there are around 150 images per category. However, it is worth noting that few categories contain up to 700 food images. This dataset is considered challenging as the images were collected from the World Wide Web from real world settings. There are multiple classes of food types, with great differences in image contrast, lighting and appearance. An adapted sample (images were slightly cropped to fit into this article) from the dataset is shown in Figure 2.

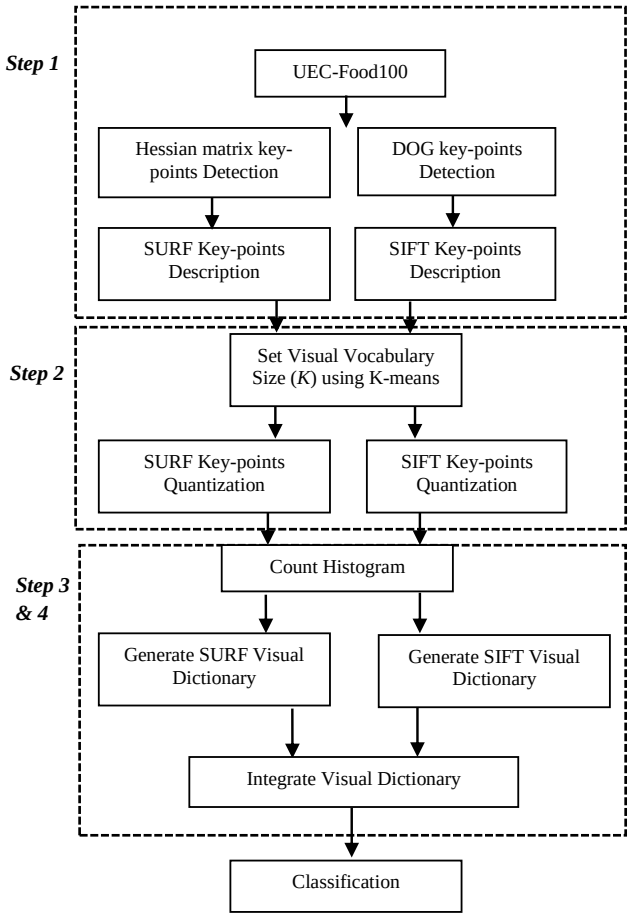

Figure 2: BoF Model

Figure 3: Samples adapted from the UEC-Food100 dataset

We used Matlab for feature extraction and to generate visual dictionary, and the Weka LibLINEAR classification package (L2-regularized L2-loss Support Vector Classification (dual - with default parameters)) for classification.

B. SURF and SIFT Integration

There are four steps to integrate SURF and SIFT as described as follows:

Step 1. Low-level Feature Extraction

The process begin with the individual low-level extraction of SURF and SIFT. There are two sub-processes within the extraction namely key-points detection and description. We use the key-point detector to find the salient regions of food. The SURF is using Hessian matrix while SIFT is using Different of Gaussian (DoG) detector. The Hessian matrix relies on integral images to decrease the computation time and find the major interest points in the scale space by using non-maximal suppression. Given an image with a point x= (x, y), the Hessian matrix of H(x, σ) can be defined as follows:

$$H(x,\sigma) = \begin{bmatrix} L_{xx}(x,\sigma) & L_{xy}(x,\sigma) \\ L_{xy}(x,\sigma) & L_{yy}(x,\sigma) \end{bmatrix} \quad (1)$$

In SIFT, series of DoG applied to detect the scale-space extrema and to localize the key-points. Given an image *I(x,y)*, the DoG convolve an image using the following formula:

$$L(x,y,\sigma) = G(x, y, \sigma) * I(x, y) \quad (2)$$

Both detectors used by SURF and SIFT selects key-points from corners, blobs and T-junction. However, there are differed in term of descriptor since the SURF sums up the Haar wavelength response and SIFT sums up the gradients. The Haar wavelet responses are built in x and y direction to produce 64 feature dimensions while the SIFT descriptor generate 128 feature dimensions from 4 x 4 image gradient and 16 x 16 sample arrays.

Step 2. Key-points Quantization

The patch-level features generated by SURF and SIFT is contained highly diverse and massive of key-points contributed by the image variations. Hence, the next process is to convert the patch-level representation into region-level representation to summarize the relevant cues at large scale. It is performed by grouping the key-points into pre-defined cluster by using clustering algorithm. This process is called key-points/feature quantization or feature encoding. Given a set of local features $\{x_1 \ldots\ldots x_m\}$ where $x_m < R_D$ and $d_k \in R_d$ is a prototype associated with *k*-th cluster. Then, there are partitioned set of K cluster $\{d_1 \ldots\ldots\ldots d_k\}$ where $d_k < R_D$. We used hard quantization as it is the simplest encoding technique to assign the key-points to the closest cluster $\hat{c}_i$ defined as,

$$\hat{c}_i = args_{ck}\ min\ \|f_i - c_k\|,\ k \in \{1,\ldots\ldots,N_c\} \quad (3)$$

The selection of vocabulary size is also contribute to the recognition performance since too small vocabulary size may weaken discriminability ability while large vocabulary size may generalize the key-points distribution as well as to increase the computation cost. In this paper, k-means clustering is adopted and we set the vocabulary size to be 500[16].

Step 3. Generate Visual Dictionaries.

This process is also known as pooling which is to aggregate the encoded vector by using certain pooling techniques. For the coding coefficient of every local descriptor γ, this process convert the patch-level into region-level image representation $\rho \in R^M$ where M representing the visual vocabulary size. Basically, there are two pooling techniques which are sum and average pooling and max-pooling[34]. We apply sum-pooling to get the histogram of number of occurrences from cluster. By using sum pooling technique, the i^{th} component of T is $T_j = \sum_{i=1}^{1} \gamma_{ij}$ where the i is the total number of image key-points.

Step 4. Integration of Visual Dictionaries

As mentioned earlier, the SURF describes objects based on Haar wavelet responses while SIFT describe the gradient information around the detected key-points other than using different kind of detector. By integrate these features, it may increase the reliability and preciseness of recognition performance[35]. Basically, there are two ways of local feature integration which are patch-level and image level integration[36]. The patch-level integration is performed before key-points quantization stage while image-level integration is performed after the pooling stage. We use image-based SURF-SIFT feature integration as SIFT generate much of key-points compared to SURF and SIFT will become more dominant if patch-level integration is used[36]. Therefore, image-based integration will merge the SURF and SIFT visual dictionary and produce 1000 dimensions feature vector.

C. Classification

We investigate the effect of the local feature integration in image classification and we choose Linear SVM classifier as in [16] work. In addition to that, we are also make comparisons on other classifiers such as Naïve Bayes, k-nearest neighbor(KNN) and LIBSVM as often used in previous food recognition study. As for training and test procedure, we used 10-fold cross validation strategy. There are two stages during this exercise. In the first stage, we evaluate the SURF and SIFT individually and in the second stage, we evaluate the integration of SURF and SIFT.

IV. Results And Analysis

We divide this section into four parts. In the first part, we present the representation efficiency and the volume of key-points detection of SURF and SIFT. The second part provides the analysis on the foods with low volume of key-points. Then, the third part presents the classification performance using four kinds of classifiers. The last part provide the analysis on overall performance of SURF, SIFT and the effect from the integration.

A. Feature Representation Efficiency

Table 1 shows the processing time to represent the features and the total of key-points detected by both local features. Based on the results, SIFT was significantly time consuming for representing the features which is about 12 times higher than SURF. We are also specifically recorded the time taken by the Hessian Matrix and DoG detector as well as the descriptors and the hard quantization using k-means algorithm. It is found that lot of time is spent for feature quantization instead of description and detection for both local features as an image may generate up to thousands of key-points. In term of the amount of key-points, SIFT detects much denser key-points which are 13,912,613 and there are only 4,407,004 key-points detected by SURF. The amount of key-points have direct impact towards the feature representation processing time as dense key-points will require more processing time for extraction and quantization.

Table 1
Total of Key-points and Extraction Time

	SURF	SIFT
Overall Feature Representation	46.3	544.72
1. Key-points detection and description (min.)	12.8	176.7
2. Quantization(min)	33.5	368.02
Number of key-points	4,407,004	13,912,613

B. Key-points Detection Analysis

We examined the number of the key-points detected by SURF and SIFT for each categories as shown in Table 2. The average key-points detected for each category for SURF and SIFT are about 44 and 139 thousands key-points respectively. Based on these threshold, we list out all the food categories that yield key-points below than the average in Table 2. The pattern of categories for both local features are almost very similar except Croissant, Oden and Potato Salad. Figure 3 showed SURF key-point detection on few food samples from Table 2.

Table 2
Low Volume of Key-points

Food Categories	SURF	Food Categories	SIFT
Pilaf	29508	Pilaf	96018
Croissant	23418	Roll bread	43257
Roll bread	12052	Tensin noodle	80324
Tensin noodle	21500	Gratin	66412
Gratin	7235	Potage	31346
Potage	21500	***Oden***	96055
Ganmodoki	17699	Ganmodoki	91968
Stew	21945	Stew	56481
Steamed egg hotchpotch	22901	Steamed egg hotchpotch	77968
Seasoned beef with potatoes	26609	Seasoned beef with potatoes	74629
Beef steak	14093	Beef steak	55607
Cabbage roll	25362	Cabbage roll	46716
Rolled omelet	16838	Rolled omelet	82529
Egg roll	21405	Egg roll	55607
Simmered pork	26025	Simmered pork	70962
Boiled chicken and vegetables	22951	Boiled chicken and vegetables	76655
Fish-shaped pancake with bean jam	28234	Sushi bowl	55607
Shrimp with chill source	20058	Fish-shaped pancake with bean jam	98155
Steamed meat dumpling	22758	Shrimp with chill source	92669
Omelet with fried rice	24737	Steamed meat dumpling	64696
Pork miso soup	23128	Omelet with fried rice	75147
Hot dog		***Potato salad***	98888
		Pork miso soup	82329
		Hot dog	80625

Figure 4: Samples of Low Key-Points (SURF)

C. *Classification Performance*

We showed the results of classification using individual and combinational local features over four types of machine learning classifiers as depicted in Table 3. The performance between SURF and SIFT are comparable with SIFT slight higher in overall classification accuracy. SIFT is also to be found compatible with KNN while SURF perform better than SIFT when using non-linear SVM. Remarkably, both features are performed worst when using Naïve Bayes while Linear SVM is the best classifier for them. The results from the local feature integration shown in both ways positive and negative impact depending on the types of classifier being used. The best classification performance from the integration of SURF and SIFT are 82.38% using Linear SVM classifier. Then, we compared the results that we obtained from the local feature integration with other two previous research that using the same dataset and classifier as shown in Table 4. The previous research employed multiple kind of feature combining both local and global features. For instance, [26] using three types of local feature combining SIFT, CSIFT and HOG as well as a global feature using gabor filter. In the other hand, the work in [16] combining Histogram of Gradient(HOG) feature and color histogram and both methods have obtained 68.9% and 79.2% classification accuracy.

D. *Analysis on Overall Performance*

We look further into specific food categories that yield low classification accuracy that using individual and combinational feature as depicted in Table 5. It can be summarized based on these figures, the amounts of key-points is one of the factor that give impact to the classification performance as there are many foods with low key-points volume are suffered from low recognition accuracy. The results have shown despite of very slight different on overall performance between SURF and SIFT, SURF is to be found performed poorly in many food categories. The integration between SURF and SIFT have improved the overall recognition performance as well on certain food categories. However, despite of these improvements, still there are certain food categories that are consistently with low performance using two or all kind of feature representations as highlighted in red and blue font such as *pizza, takoyaki, cabbage roll, boiled chicken and vegetables, sashimi bowl, gratin, jiaozi and sushi bowl.* In addition to that, there are still lot of food categories with low recognition rate. Although overall SURF has very competitive performance and efficient, it is less robust towards food categories as SIFT recognize better in many food categories.

Table 3
Classification Performance

Classifiers	Performance Rate	SURF	SIFT	SIFT +SURF
LIBSVM	Training (Sec.)	779.58	1234.04	2007.22
	Tp Rate (%)	**54.71**	28.87	41.87%
Linear SVM	Training (Sec.)	30.24	96.23	111.83
	Tp Rate (%)	62.08	**64.65**	**82.38%**
KNN	Training (Sec.)	0.01	0	0
	Tp Rate (%)	33.94	**50.67%**	45.22%
Naïve Bayes	Training (Sec.)	0.95	0.86	3.53
	Tp Rate (%)	**33.53**	32.24%	39.04%

Table 4
Comparison of Local Feature Performance

Feature Representation Method	Classification Accuracy (%)
SIFT + CSIFT + HOG + Gabor[26] (4 feat.)	68.90 %
HOG + Color[16] (2 feat.)	79.20 %
SURF + SIFT (Proposed Method)	82.38 %

Table 5
Foods with Low TP Rate

SURF	TP Rate	SIFT	TP Rate	SIFT + SURF	TP Rate
Roll bread	0.458	*Pizza*	0.381	Croissant	0.75
Raisin bread	0.485	*Takoyaki*	0.485	*Roll bread*	0.692
Pizza	0.388	*Gratin*	0.417	*Pizza*	0.694
Takoyaki	0.433	*Croquette*	0.508	*Takoyaki*	0.672
Sausage	0.5	*Pilaf*	0.53	*Gratin*	0.687
Ganmodoki	0.451	*Jiaozi*	0.503	*Croquette*	0.746
Sirloin cutlet	0.493	*Beef steak*	0.537	Grilled eggplant	0.755
Seasoned beef with potatoes	0.483	*Cabbage roll*	0.533	*Sausage*	0.771
Hambarg steak	0.504	*Boiled chicken and vegetables*	0.552	*Sushi Bowl*	0.766
Beef steak	0.491	*Sashimi bowl*	0.524	*Jiaozi*	0.766
Yakitori	0.495	*Sushi bowl*	0.514	*Stew*	0.745
Cabbage roll	0.486	Shrimp with chill source	0.441	Fried chicken	0.773
Rolled omelet	0.458	Tempura bowl	0.551	*Sirloin cutlet*	0.764
Boiled chicken and vegetables	0.457	Tensin noodle	0.545	*Nanbanzuke*	0.755
Sashimi bowl	0.497	*Nanbanzuke*	0.529	*Hambarg steak*	0.741

Steamed meat dumpling	0.487	Spaghetti meat sauce	0.552	*Ginger pork saute*	0.726
Fried shrimp	0.504	Mixed rice	0.565	***Cabbage roll***	0.71
Fried noodle	0.504			*Rolled omelet*	0.771
Gratin	0.522			Egg roll	0.743
Pilaf	0.53			***Boiled chicken and vegetables***	0.714
Omelet	0.505			***Sashimi bowl***	0.741
Jiaozi	0.533			*Fish-shaped pancake with bean jam*	0.754
Stew	0.538			*Steamed meat dumpling*	0.722
Grilled salmon	0.534			Omelet with fried rice	0.778
Chicken n egg on rice	0.521			*Fried shrimp*	0.765
Ginger pork saute	0.513			*Kinpira-style sauteed burdock*	0.766
Chilled noodle	0.453			Rice ball	0.778
Sushi bowl	0.532				
Potato salad	0.531				
Kinpira-style sauteed burdock	0.532				
Steamed egg hotchpotch	0.546				
Fish-shaped pancake with bean jam	0.541				

By reexamining the images of the respective foods appearances and the amount of detected key-points, we made a few conclusions on why the results were poorer compared to other food types. These conclusions were also based on claims supported by the literature. As mentioned earlier, there is a correlation between the amount of key-points and classification accuracy as lower number of key-points contributed to the overall poorer performance. Low key-points detection have been linked to the inability of local features to handle certain image and object characteristics such as very little image contrast difference between foreground and background[5], small food regions in multi-class objects[5], [37], small image dimensions, arbitrary food appearances[19] and the mixed kinds of foods[11], [38], [39] that have variety of shape and color. We described below the factors of these problems.

1) Low Dimensions and Contrast

The small image dimensions limit the capability of the local feature detector to provide enough samples of key-points. For instance, the average number of key-points detected by SURF for roll breads are about 113, cabbage roll are about 130 key-points and egg roll around 154 key-points. In addition to that, there are lot of irrelevant or noises key-points are included. Also, it can be observed that SURF and SIFT perform poorly on the low contrast type of image to distinguish foreground from background.

Figure 5: Low Contrast

2) Food Appearances

As shown in Figure 5, foods have large variability in term of appearance. It may contain multi-class objects, arbitrary shape, variety of colors, very high deformation as well as very smooth texture. Hence, representing them in feature space become very complex and difficult. For instance, the multi-class appearance makes the region of interest become too small which limit the interest points and create a massive of noises from the unnecessary object classes as well as image background. The deformation of food objects are also makes the food regions become tiny and create arbitrary shape and the local feature like SIFT has lack capability in describing these kind of images[1], [36], [40]. There are also found not working well with the smooth texture kind of images[29].

Figure 6: Appearance Variability

3) The Mixed kind of Foods

The foods appearance become more complex when different kind foods are mix together which finally produce many variety shape and colours as shown in Figure 6. When SURF and SIFT are used, even there are lot of key-points were detected, but they became too sparse and less unique as it will be generalized into too many clusters during quantization stage. This problem is consistent with the findings in [39] when they yield low recognition accuracy in mixed kind of foods.

Figure 7: Mixed Foods

Next we identify set of food categories that has obtained a good classification rate as shown in Table 6. We mark using

blue and green font on the food categories that get a good classification rate on all and any of two feature representation method respectively. Based on the observation in respective food categories, it is contain lot of high dimensional size of images and the region of food is closed-up dominating almost the entire image. The level of contrast of food regions are also high which easily to compare with the background images. There are also have more consistent colour around the region. High classification rate also caused by lot of instances as mentioned in previous. For instance the total instances for rice is 620 and Miso Soup is 728, All these factors contribute to better quality of key-points produced and good enough to represent the uniqueness of foods. We show some food samples in Figure 7.

Table 6
Foods with High TP Rate

SURF	TP Rate	SIFT	TP Rate	SIFT + SURF	TP RATE
Rice	0.792	Hamburger	0.76	Goya chanpuru	0.856
Goya chanpuru	0.712	Rice	0.795	Sandwiches	0.853
Hamburger	0.721	Ramen noodle	0.762	Rice	0.894
Vegetable tempura	0.748	Beef noodle	0.777	Soba noodle	0.853
Miso soup	0.795	Potage	0.761	Ramen noodle	0.907
Sashimi	0.758	Sashimi	0.758	Beef noodle	0.871
Lightly roasted fish	0.755	Sukiyaki	0.705	Japanese-style pancake	0.869
Tempura	0.771	Lightly roasted fish	0.706	Potage	0.858
Beef bowl	0.79	Boiled fish	0.725	Sashimi	0.927
Dipping noodles	0.73	Dried fish	0.727	Sukiyaki	0.885
		Yakitori	0.703	Lightly roasted fish	0.922
		Beef curry	0.715	Dried fish	0.864
		Beef bowl	0.749	Yakitori	0.865
		Dipping noodles	0.714	Egg sunny-side up	0.871
		Sauteed spinach	0.706	Fermented soybeans	0.891
				Beef curry	0.858
				Roast chicken	0.882
				Cutlet curry	0.859
				Spaghetti meat sauce	0.856
				Green salad	0.906
				Beef bowl	0.922
				Dipping noodles	0.889
				Fried rice	0.87

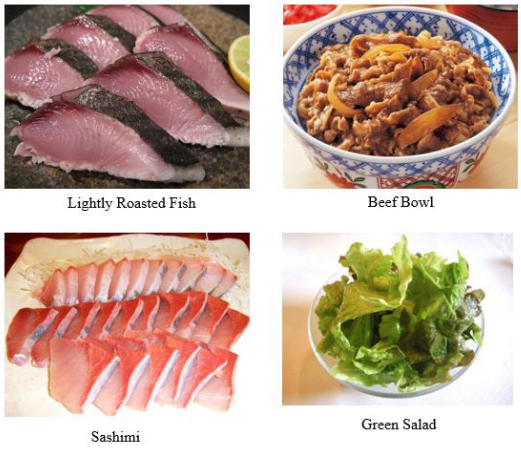

Figure 8: Food Categories with High TP Rate

V. CONCLUSIONS

An evaluation of the local features of SURF and SIFT towards recognizing food objects has been provided. The overall performance on individual feature showed SIFT outperformed the SURF in term of classification rate with little difference. SURF is to be found more efficient as it detects much fewer key-points compared to SIFT and outperformed SIFT in term of feature representation processing time. The SIFT detect denser key-points and longer descriptor dimensions. The integration of SURF and SIFT however, showed to be superior in term of classification performance with a slight increase in training time. Both features seem a good complement for each other, possible due to SIFT being robust towards scale and rotation and SURF is robust towards illumination changes[35]. Based on the findings obtained during feature extraction and classification, we provide an analysis and discussed several factors/problems lead to the lower key-points detection as well as classification accuracy which needs to be rectified in future endeavor.

REFERENCES

[1] H. H. Fanyu Kong Hollie A. Raynor, Jindong Tan, "DietCam: Multi-view regular shape food recognition with a camera phone," *Pervasive and Mobile Computing*, vol. 19, no. C, pp. 108–121, 2015.

[2] J. T. Fanyu Kong, "DietCam: Automatic dietary assessment with mobile camera phones," *Pervasive and Mobile Computing*, vol. 8, pp. 147–163, 2012.

[3] M. Bosch, F. Zhu, N. Khanna, C. J. Boushey, and E. J. Delp, "Combining Global And Local Features For Food Identification In Dietary Assessment Video and Image Processing Lab (VIPER) School of Electrical and Computer Engineering Department of Foods and Nutrition," pp. 1789–1792, 2011.

[4] H. Kagaya and K. Aizawa, "New Trends in Image Analysis and Processing -- ICIAP 2015 Workshops," vol. 9281, pp. 350–357, 2015.

[5] Z. Z. Duc Thanh Nguyen Philip O. Ogunbona, Yasmine Probst ,Wanqing Li, "Food image classification using local appearance and global structural information," *Neurocomputing*, vol. 140, pp. 242–251, 2014.

[6] G. Csurka, C. Dance, L. Fan, J. Willamowski, and Cedric Bray, "Visual categorization with bag of keypoints," *International Workhop on Statistical Learning in Computer Vision*, pp. 1–22, 2004.

[7] S. Keypoints and D. G. Lowe, "Distinctive Image Features from," *International Journal in Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[8] U. L. Altintakan and A. Yazici, "Towards effective image classification using class-specific codebooks and distinctive local features," *IEEE Transaction on Multimedia*, vol. 17, no. 3, pp. 323–332, 2015.

[9] V. C. Luciano Oliveira Gustavo Neves, Talmai Oliveira, Eduardo

Jorge, Miguel Lizarraga c, "A mobile, lightweight, poll-based food identification system," *Pattern Recognition*, vol. 47, pp. 1941–1952, 2014.

[10] L. G. MariosM. Anthimopoulos Luca Scarnato, Peter Diem, Stavroula G.Mougiakakou, "A Food Recognition System for Diabetic Patients Based on an Optimized Bag-of-Features Model," *IEEE Journal on Biomedical and Health Informatics*, vol. 18, no. 4, pp. 1261–1271, 2014.

[11] P. Pouladzadeh, S. Shirmohammadi, S. Member, and R. Al-maghrabi, "Measuring Calorie and Nutrition From Food Image," vol. 63, no. 8, pp. 1947–1956, 2014.

[12] F. Kong, "Automatic Food Intake Assessment Using Camera Phones," 2012.

[13] M. B. Fengqing Zhu InsooWoo, Sung Ye Kim, Carol J. Boushey, David S. Ebert, Edward J. Delp, "The Use of Mobile Devices in Aiding Dietary Assessment and Evaluation," *IEEE Journal of Selected Topics in Signal Processing*, vol. 4, no. 4, pp. 756–766, 2010.

[14] Y. K. Keiji Yanai, "Twitter Food PhotoMining and Analysis for One Hundred Kinds of Foods," 2014.

[15] S. V. B. P. Parisa Pouladzadeh Pallavi Kuhad, Abdulsalam Yassine, Shervin Shirmohammadi, "A virtualization mechanism for real-time multimedia-assisted mobile food recognition application in cloud computing," *Cluster Computing*, vol. 18, no. 3, pp. 1099–1110, 2015.

[16] Y. Kawano and K. Yanai, "FoodCam: A real-time food recognition system on a smartphone," *Multimedia Tools Application*, vol. 74, no. 14, pp. 5263–5287, 2015.

[17] G. M. Farinella, D. Allegra, M. Moltisanti, F. Stanco, and S. Battiato, "Retrieval and classification of food images," *Comput. Biol. Med.*, vol. 77, pp. 23–39, 2016.

[18] L. Bossard, M. Guillaumin, and L. Van Gool, "Food-101 - Mining discriminative components with random forests," *Lecture Notes Computing Science (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 8694 LNCS, no. PART 6, pp. 446–461, 2014.

[19] J. T. Fanyu Kong, "DietCam: Regular Shape Food Recognition with a Camera Phone," in *International Conference on Body Sensor Networks*, 2011.

[20] H. Kagaya, "Food Detection and Recognition Using Convolutional Neural Network," no. 3, pp. 1085–1088, 2014.

[21] L. H. Ruihan Xu Shuqiang Jiang, ShuangWang, Xinhang Song, Ramesh Jain, "Geolocalized Modeling for Dish Recognition," *IEEE Transaction on Multimedia*, vol. 17, no. 8, pp. 1187–1199, 2015.

[22] M. Giovanni Maria Farinella and S. Battiato, "Classifying Food Images Represented As Bag Of Textons Giovanni Maria Farinella Department of Mathematics and Computer Science Image Processing Laboratory - University of Catania," pp. 5212–5216, 2014.

[23] D. A. Lisin, M. A. Mattar, M. B. Blaschko, E. G. Learned-Miller, and M. C. Benfield, "Combining Local and Global Image Features for Object Class Recognition," *2005 IEEE Computer Society Conference on Computer Vision Pattern Recognition Workshop*, vol. 3, pp. 47–47, 2005.

[24] T. Joutou and K. Yanai, "A food image recognition system with Multiple Kernel Learning," in *2009 16th IEEE International Conference on Image Processing (ICIP)*, 2009, pp. 285–288.

[25] H. Hoashi, T. Joutou, and K. Yanai, "Image Recognition of 85 Food Categories by Feature Fusion," 2010.

[26] Y. Matsuda, H. Hoashi, and K. Yanai, "Recognition of multiple-food images by detecting candidate regions," in *Proceedings - IEEE International Conference on Multimedia and Exposition*, 2012, pp. 25–30.

[27] K. Y. Yoshiyuki Kawano, "FoodCam: A real-time food recognition system on a smartphone," *Multimedia Tools Application*, vol. 74, no. 14, pp. 5263–5287, 2015.

[28] Y. Kawano and K. Yanai, "Rapid Mobile Object Recognition Using Fisher Vector," *2013 2nd IAPR Asian Conference Pattern Recognition*, pp. 476–480, 2013.

[29] Y. Li, S. Wang, Q. Tian, and X. Ding, "Feature representation for statistical-learning-based object detection: A review," *Pattern Recognition*, vol. 48, no. 11, pp. 3542–3559, 2015.

[30] G. M. Farinella, D. Allegra, and F. Stanco, "A Benchmark Dataset to Study the Representation of Food Images."

[31] J. Cui, M. Cui, B. Xiao, and G. Li, "Compact and discriminative representation of Bag-of-Features," *Neurocomputing*, vol. 169, pp. 55–67, 2015.

[32] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," *Lect. Notes Comput. Sci. (including Subser. Lecture Notes Artificial Intelligent and Lecture Notes Bioinformatics)*, vol. 3951 LNCS, pp. 404–417, 2006.

[33] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool, "Speeded-Up Robust Features (SURF)," *Computer Vision and Image Understanding*, vol. 110, no. 3, pp. 346–359, 2008.

[34] C. Hiba, Z. Hamid, and A. Omar, "Bag of Features Model Using the New Approaches: A Comprehensive Study," *International Journal Advances Computer Science and Applications*, vol. 1, no. 7, pp. 226–234, 2016.

[35] N. Ali, K. B. Bajwa, R. Sablatnig, S. A. Chatzichristofis, Z. Iqbal, M. Rashid, and H. A. Habib, "A novel image retrieval based on visual words integration of SIFT and SURF," *PLoS One*, vol. 11, no. 6, pp. 1–20, 2016.

[36] J. Yu, Z. Qin, T. Wan, and X. Zhang, "Feature integration analysis of bag-of-features model for image retrieval," *Neurocomputing*, vol. 120, pp. 355–364, 2013.

[37] C. Morikawa, H. Sugiyama, and K. Aizawa, "Food region segmentation in meal images using touch points," pp. 7–12, 2012.

[38] P. Pouladzadeh, A. Yassine, and S. Shirmohammadi, "New Trends in Image Analysis and Processing -- ICIAP 2015 Workshops," vol. 9281, pp. 441–448, 2015.

[39] P. Pouladzadeh, S. Shirmohammadi, A. Bakirov, A. Bulut, and A. Yassine, "Cloud-based SVM for food categorization," *Multimedia Tools and Applications*, pp. 5243–5260, 2014.

[40] K. Li, F. Wang, and L. Zhang, "A new algorithm for image recognition and classification based on improved Bag of Features algorithm," *Opt. - International Journal on Light Electron Optics*, vol. 127, no. 11, pp. 4736–4740, 2016.