International Journal of Engineering & Technology, 7 (4.31) (2018) 204-208

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Recognition of Food with Monotonous Appearance using Speeded-Up Robust Feature (SURF)

Mohd Norhisham Razali[1], Noridayu Manshor[2*], Alfian Abdul Halin[3], Razali Yaakob[4], Norwati Mustapha[5]

[1]*Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, 43300 Serdang, Malaysia*
[2]*Faculty of Computing and Informatics, Universiti Malaysia Sabah, 88400 Kota Kinabalu, Sabah, Malaysia*
**Corresponding author E-mail: ayu@upm.edu.my*

Abstract

Food has become one of the most photographed objects since the inceptions of smart phones and social media services. Recently, the analysis of food images using object recognition techniques have been investigated to recognize food categories. It is a part of a framework to accomplish the tasks of estimating food nutrition and calories for health-care purposes. The initial stage of food recognition pipeline is to extract the features in order to capture the food characteristics. A local feature by using SURF is among the efficient image detector and descriptor. It is using fast hessian detector to locate interest points and haar wavelet for descriptions. Despite the fast computation of SURF extraction, the detector seems ineffective as it obviously detects quite a small volume of interest points on the food objects with monotonous appearance. It occurs due to 1) food has texture-less surface 2) image has small pixel dimensions, and 3) image has low contrast and brightness. As a result, the characteristics of these images that were captured are clueless and lead to low classification performance. This problem has been manifested through low production of interest points. In this paper, we propose a technique to detect denser interest points on monotonous food by increasing the density of blobs in fast hessian detector in SURF. We measured the effect of this technique by performing a comparison on SURF interest points detection by using different density of blobs detection. SURF is encoded by using Bag of Features (BoF) model and Support Vector Machine (SVM) with linear kernel adopted for classification. The findings has shown the density of interest point detection has prominent effect on the interest points detection and classification performance on the respective food categories with 86% classification accuracy on UEC100-Food dataset.

Keywords: *bag of features; food recognition; local features; object recognition; SURF*

1. Introduction

Food recognition deals with image processing and machine learning techniques to extract the features from food images and classify to the respective food categories. The patch-level feature description using local features such as Scale Invariant Feature Transform (SIFT) has been widely adopted to represent the visual properties of image in object recognition, image retrieval and image classification. This is due to its distinctiveness and stability towards different image orientation and occlusions [1], [2] in which the global feature is unable to provide. Global feature is a pixel level feature description. It generates too large number of features and noises [3] which may increase the computational cost for feature extraction and classification.

In general, food images are characterized by its complex appearances and variations which is a hindrance to effectively represent the features as well as to recognize the food category [4]–[6]. This condition has also shrink-up inter-class similarities which affect classification performance [7]. SURF is a popular gradient-type local features after SIFT as it has been reported more computationally efficient compared to SIFT and also provide more distinctive features [8]. However, SURF produces extremely minimum amount of interest points on food categories that consists of many texture-less food images, images with small pixel dimension and images with low contrast and brightness. In this context, appropriate volume of interest points should be taken into account as too small interest points volume may pull out the informative features that are crucial for effective feature quantization and classification [9]. The main objective of this paper is to provide an informative and discriminative Bag of Feature (BoF) model for food recognition. It is conducted by proposing a technique to increase the density of SURF interest points detection via the manipulation of blobs density in SURF.

The rest of the paper is organized as follows. In section II, we provide related works on the local feature representation methods to optimize the overall BoF model performance. Section III describes the steps conducted to perform the experiments, section IV presents the experimental results and the last section summarizes the overall and future works.

2. Local Feature Representation

This section will provide the related works on local feature representation for object recognition and food object recognition domain.

There are numerous researches that have been working on improving the BoF model in order to produce a compact and discriminative local feature representation. Specifically, it concentrates on improving the method in quantizing the interest points as this is considered as a crucial step to produce a good visual dictionary [1]. Several techniques have been proposed to improve the local feature representation using the BoF model. Li et.al [2] improved

the BoF algorithm by addressing the problem of SIFT that is unable to handle image with complex background. Instead of using SIFT, SURF with Spatial Pyramid Matching (SPM) is used to generate multiple frames of BoF. The method is evaluated on Graz, Caltech-256, and Pascal VOC 2012 datasets and achieved 85%, 89.43% and 73.75 respectively using LIBSVM classification package. SPM has the capability to minimize the interference and noises from the image background. To capture the invariance of an image, Xie et. al [3] improved the visual dictionary of the BoF by using feature summarization technique. Feature summarization involved two processes namely feature pooling and normalization which are executed using the Generalized Regular Spatial Pooling (GRSP) and Hierarchal Feature Normalization (HFN) algorithm. The method is evaluated by using Support Vector Machine (SVM) on three types of datasets for scene recognition, generic object recognition and fine-grained object recognition.

SIFT and SURF are the well-known local-patch descriptor features as both have proven to provide the distinctive features. SIFT employed Different of Gaussians (DoG) to locate the local extrema in the scale space. The interest points are determined based on the high local contrast in non-edge extrema. Then, the descriptor computes the image gradients in a 16 X 16 window and grouped into 4X4 sub-regions. The eight-bin histogram is obtained by quantizing the direction of gradients and finally getting 128 dimensions feature vector by combining all the bins from 16 histograms. While SURF local interest points using determinant of hessian (DoH). DoH is based on 'fast-hessian' detector which relies on integral images to decrease computation time. The SURF descriptor used the distribution of Haar-wavelet responses within interest point neighborhood to produce 64 dimensions feature vector. More recent local features are Oriented Fast and Rotated Brief (ORB) [4] and A-KAZE [5]. The ORB integrates the FAST detector and BRIEF descriptor to produce a more computationally efficient local features. The FAST is repeated at multiple times on different scales and exclude the non-corner key-points as non-corner key-points may include many key-points from the background. The KAZE detect and describe features in non-linear scale space using additive operator Splitting techniques and variable conductance diffusion.

In food object recognition domain, several research use local feature to describe the image [6]–[10] because the gradients based features are the best representatives to describe food image [11]–[13]. Recent work has used Deep convolutional neural network based approach to deal with the problem of detecting multiple food objects via semantic localization, segmentation and recognition [14]. However, the properties of local feature are capable enough to deal with the complexity of food such as irregular and deformation shape. The problem of recognizing texture-less of food have been discussed by [15], [16] where the local features based on Local Binary Pattern (LBP) and shape context were reported less effective for meat and donut category due to their surface being less diverse. Both SIFT and SURF are also less to be found for irregular type and small in size of food due to a few number and incorrect interest points detected [16]. To the best of our knowledge, there is no research that have been conducted to address these problems.

3. Experimental Setup

The steps taken to accomplish the experiments are shown in Figure 1. We evaluate the proposed techniques using UEC100-Food dataset [17]. This dataset consists of 100 food categories with 14,467 food images and vary pixels dimensions. Each category consists of roughly about 150 images taken from the World Wide Web. These images are characterized by the real world settings as it contains multiple classes of food types, image variations such as occlusions, contrast and illumination. There are two main stages of the experiments. The first stage extract different configurations of SURF and the second stage build the visual dictionary and classification.

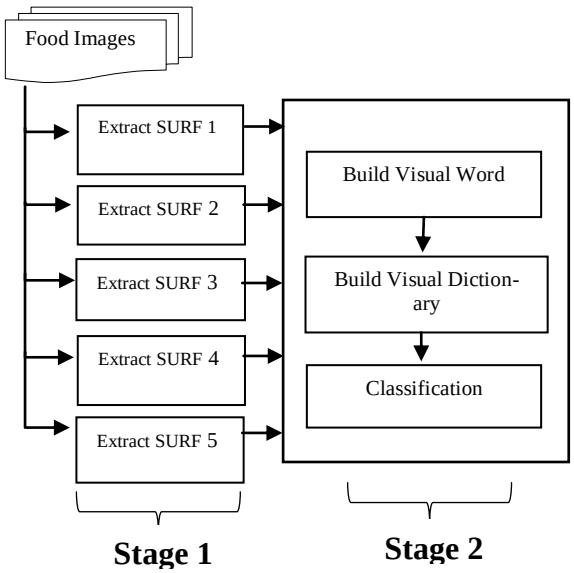


Fig. 1: Research Framework

3.1. Stage 1- SURF Extraction

The Hessian matrix detector used the integral images to speed-up the box convolution for interest points detection. The value of the integral image at the point of (x, y) is the total of all the intensities in the rectangle box. Basically, there are three components that affect the interest points density which are *Hessian Threshold*, *number of octaves* and *number of scale levels*. However, the scope of this research is limited to observe the effect of hessian threshold only. Hessian threshold controls the amount of blobs and by reducing this value will return more amount of blobs. After the Hessian detects the interest points, the next stage is to find the major interest points in scale space using non-maximal suppression. The scale space is formed from a number of octaves that contain a series of filters obtained from the convolution. By increasing the number of octaves, it will increase the size of blobs as well. Lastly, every octave will be divided into a few scale levels and increasing the number of scale will detect more blobs at the finer scale increments. Based on the range of threshold value suggested in [18], we have performed an analysis on SURF on five different values of Metric threshold configuration as shown in Table 1. Each configuration is run separately.

Table 1: Hessian Threshold Configuration

Components	SURF 1	SURF 2	SURF 2	SURF 2	SURF 2
HessianThreshold	1000	800	600	400	200
Number of Octaves	3	3	3	3	3
Number of Scale Levels	4	4	4	4	4

3.2. Stage 2- Build Visual Dictionary and Classification

We implement the Bag-of Feature (BoF) approach to represent the huge diversity of local features [19] as our general framework. We extract SURF individually and by category basis. The SURF descriptor generates 64-feature dimensions which describes the Haar wavelet responses. Clustering is performed to group each interest points into visual words. We used k-means clustering algorithm to create the visual dictionary of the local features as suggested in many previous works [19]. The vocabulary size depends on the size of cluster k. An optimal value of k is usually determined via trial and error process. This is because too small of k may cause a different pattern of interest points assigned to the similar cluster

while a large of k may lead to the abundant of noises, lack or generalizability and processing overhead [20]–[22]. However, we set the vocabulary size to 500 [9] since it is optimal. Then, the interest points will be quantized by assigning to the nearest centroid of cluster followed by counting the total of interest points for each cluster. Finally, the SURF visual dictionary is generated.

3.3. Performance Evaluation

We have recorded the processing time for the local feature detection, description and quantization as well as the amount of interest points to measure the quality of the local feature. We have also evaluated the effectiveness of the features by measuring classification rate using Linear SVM classifier which is following the research conducted in [24] and [25]. The training and testing strategy is based on k-fold cross-validation and we used 10 folds. The classification is ran 10 times with a random of different set of training and testing.

4. Experimental Results

This section provides the experimental results which can be divided into three parts. The first part will present the performance comparisons between local features. The second part provides an analysis using different SURF interest points density configuration and the final part provides the samples of food categories that suffered from low interest points detection.

Table 2 shows the performance comparisons between FAST, HOG, HARRIS, SIFT and SURF local features in term of the amount of interest points, the extraction time and classification rate. The amount of interest points indicate how sufficient the information of foods can be provided and affect the whole extraction time. The classification rate measures how accurate the instances are recognized. If possible, a good local feature performance should generate an optimal amount of interest points, reduce the extraction time and to get a high classification rate.

Table 2: Performance Comparisons of Local Features

Local Features	Total Interest Points	Detection and Description (Min.)	Quantization (Min.)	Classification Rate
FAST	4,801,093	68.06	51.13	0.59
HOG	4,801,093	95.48	46.89	0.43
HARRIS	4,948,907	102.00	47.54	0.55
SIFT	13,912,613	176.74	368.02	0.65
SURF	4,407,004	12.80	33.50	0.62

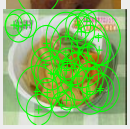
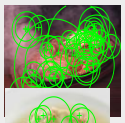
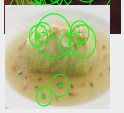
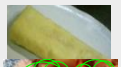
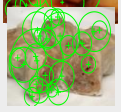
The results show that the SIFT yield the best classification rate of 0.65 followed by SURF, FAST, HARRIS and Histogram Of Gradient (HOG). In flipside, SIFT produced the most dense interest points which require much processing time for both interest points detection and quantization. On the other hand, SURF produces the smallest amount of interest points and the shortest time for extraction. In addition to that, the classification rate between SURF and SIFT were just ~~a~~ slightly different. Table 3 shows the comparisons in terms of interest point's volume, interest point's extraction time and classification rate in five types of SURF density configurations.

Table 3: Performance Comparisons between different SURF Density Configuration

SURF Config.	Tot. IP	Detection and Description (Min.)	Quant. (Min.)	Classification Rate
SURF 1	4,407,004	12.80	33.50	0.62
SURF 2	7,633,950	14.7	58.2	0.67
SURF 3	6,216,641	14.1	56.1	0.71
SURF 4	7,794,832	12.2	70.3	0.71
SURF 5	10,610,965	16.6	141.0	0.86

SURF 1 has been configured using the maximum value of Hessian Threshold (1000) while SURF 5 using the smallest Hessian Threshold value (200). As we decrease the Hessian Threshold, more blobs can be detected which in turn increase the volume of interest points as well as the time for extraction and quantization. The effect from the configuration has resulted with an increase of classification rate as more information were gained. This has indicated that the interest point volume is important to capture the features from the large variations of food and to provide more informative feature representation. The volume of interest points have increased about 58% using SURF 5 to become 10,610,965 interest points which in turn increased the overall extraction time. However, the classification rate has improved significantly compared to SIFT and can also sustain the extraction time. Table 4 shows the list of food categories that obtained low classification performance due to low interest points detection.

Table 4: Foods with Low Interest Points Detection

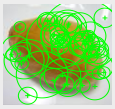
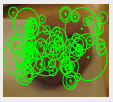
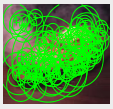
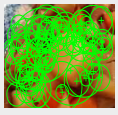
Category Name	Total Points	Classification Rate	Detection Sample	Total Points
Roll Bread	12052	0.46		7
Gratin	21500	0.52		21
Ganmodoki	27590	0.45		6
Seasoned beef with potatoes	22901	0.48		31
Beef Steak	26609	0.49		39
Cabbage Rolls	14093	0.49		8
Roll Omelet	25362	0.46		0
Boiled Chicken and Vegetables	26025	0.46		35
Steamed meat dumpling	20058	0.49		23

Low interest points detection were caused by two main factors which are the appearance of food and the quality of images. SURF is less effective in detecting and describing foods with low texture variability and monotonous appearance such as roll bread, ganmodoki, cabbage rolls and roll omelet. In term of image quality, there were many food images with small pixel dimensions and low contrast and brightness such as gratin, ganmodoki, beef steak and steamed meat dumpling which affect the volume and quality of interest points.

In Table 5, we provide the improvement in terms of interest points detection and classification rate on food categories listed in Table 4. The table shows that by detecting more amount of blobs through reducing the metric threshold value in SURF, it may significantly cater the problem of detecting features in food catego-

ries that consists of many foods with low texture variability and monotonous appearance.

Table 5: The Effect of SURF Configuration

Cat. Name	Total Points	Classification Rate	Detection Sample	Total Points
Roll Bread	35457	0.70		51
Gratin	47039	0.76		116
Ganmodoki	70785	0.72		82
Seasoned beef with potatoes	52888	0.71		89
Beef Steak	55161	0.81		100
Cabbage Rolls	37908	0.65		50
Roll Omelette	61649	0.75		35
Boiled Chicken and Vegetables	56350	0.76		75
Steamed meat dumpling	47093	0.78		92

5. Conclusion

In this paper, we have evaluated different types of local features to recognize food objects. SURF is found to be the most optimal local features as it detects small volume of interest points with decent classification rate compared to the other local features. Then, we have performed an analysis on using different value of hessian threshold in SURF to increase the density sampling of SURF interest points. Specifically, this technique is proposed to deal with problem of SURF in detecting and describing interest points on food images with minimum texture appearance and also the small image pixel dimensions with low contrast and brightness. In this analysis, five types of SURF using different value of Hessian threshold were evaluated. The experimental results have demonstrated the improvement on classification accuracy on the respective food categories perpendicular with an increase of local feature interest points. As a consequence, a high volume of interest points were produced. Future research should investigate a mechanism in producing an automatic and optimal volume of interest points of local without sacrificing the recognition accuracy. In addition to that, due to lack of research conducted in dealing multi-class food detection, a technique such as multi-label classification can be considered which is effective in tackling the high diversity of physical human activity [24].

Acknowledgement

The authors acknowledge the financial supported by the Putra Grant (Cost Center: 9569000) funded by the Universiti Putra Malaysia (UPM).

References

[1] Y. Kuang, K. Åström, L. Kopp, M. Oskarsson, and M. Byröd, "Optimizing visual vocabularies using soft assignment entropies," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6495 LNCS, no. PART 4, pp. 255–268, 2011.

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

[3] L. Xie, Q. Tian, and B. Zhang, "Simple Techniques Make Sense: Feature Pooling and Normalization for Image Classification," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 7, pp. 1251–1264, 2016.

[4] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski, "ORB: An efficient alternative to SIFT or SURF," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2564–2571, 2011.

[5] A. J. D. Pablo F. Alcantarilla Adrien Bartoli, "KAZE features," *Proc. ECCV2012*, 2012.

[6] S. Giovany, "ScienceDirect ScienceDirect ScienceDirect ScienceDirect ScienceDirect ScienceDirect Machine Learning and SIFT Approach Indonesian Food Image Machine Learning Approach for Indonesian Food Image Recognition Machine Learning Approach for Indonesian Food Imag," *Procedia Comput. Sci.*, vol. 116, pp. 612–620, 2017.

[7] M. B. Fengqing Zhu Nitin Khanna, Carol J. Boushey, Edward J. Delp, "Multiple Hypotheses Image Segmentation and ClassificationWith Application to Dietary Assessment," *IEEE J. Biomed. Heal. INFORMATICS*, vol. 19, no. 1, pp. 377–388, 2015.

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

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

[10] 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.

[11] 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 J. Sel. Top. Signal Process.*, vol. 4, no. 4, pp. 756–766, 2010.

[12] Y. M. Kiyoharu Aizawa He Li, Chamin Morikawa, "Food Balance Estimation by Using Personal Dietary Tendencies in a Multimedia Food Log," *IEEE Trans. Multimed.*, vol. 15, no. 8, pp. 2176–2185, 2013.

[13] M. N. Razali, N. Manshor, and A. A. Halin, *Food Category Recognition using SURF and MSER Local Feature Representation*. Bangi,Malaysia: Springer, Cham, 2017.

[14] E. Aguilar, B. Remeseiro, M. Bolaños, and P. Radeva, "Grab, Pay and Eat: Semantic Food Detection for Smart Restaurants," pp. 1–10, 2017.

[15] Z. Zong, D. T. Nguyen, P. Ogunbona, and W. Li, "On the combination of local texture and global structure for food classification," *Proc. - 2010 IEEE Int. Symp. Multimedia, ISM 2010*, pp. 204–211, 2010.

[16] 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.

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

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

[19] 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 J. Biomed. Heal. INFORMATICS, vol. 18, no. 4, pp. 1261–1271, 2014.

[20] G. Csurka, C. Dance, L. Fan, J. Willamowski, and Cedric Bray, "Visual categorization with bag of keypoints," *Int. Work. Stat. Learn. Comput. Vis.*, pp. 1–22, 2004.

[21] Y. Jiang, J. Yang, C. Ngo, and A. G. Hauptmann, "Representations of Keypoint-Based Semantic Concept Detection : A Comprehensive Study Representations of Keypoint-Based Semantic Concept Detection : A Comprehensive Study," *IEEE Trans. Multimed.*, vol. 12, no. 1, pp. 42–53, 2010.

[22] T. Kirishanthy and A. Ramanan, "Creating Compact and Discriminative Visual Vocabularies using Visual Bits," 2015.

[23] Y. Matsuda, H. Hoashi, and K. Yanai, "Multiple-Food Recognition Considering Co-occurrence Employing Manifold Ranking," *2012 21st Int. Conf. Pattern Recognit.*, no. Icpr, pp. 2017 – 2020, 2012.

[24] R. Mohamed, M. N. S. Zainudin, N. Sulaiman, and T. Perumal, "Multi-label Classification for Recognition of Physical Activity from Various Accelerometer Sensor Positions," *J. ICT*, no. 2, pp. 209–231, 2018.