ASM Sc. J., 15, 2021
https://doi.org/10.32802/asmscj.2020.485

Extremal Region Selection for MSER Detection in Food Recognition

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

[1]*Faculty of Computing and Informatics, Universiti Malaysia Sabah, Sabah, Malaysia*
[2]*Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Selangor, Malaysia*

The visual analysis of foods on social media by using food recognition algorithm provides valuable insight from the health, cultural and marketing. Food recognition offers a means to automatically recognise foods as well the useful information such as calories and nutritional estimation by using image processing and machine learning technique. The interest points in food image can be detected effectively by using Maximally Stable Extremal Region (MSER). As MSER used global segmentation and many food images have a complex background, there are numerous irrelevant interest points are detected. These interest points are considered as noises that lead to computation burden in the overall recognition process. Therefore, this research proposes an Extremal Region Selection (ERS) algorithm to improve MSER detection by reducing the number of irrelevant extremal regions by using unsupervised learning based on the k-means algorithm. The performance of ERS algorithm is evaluated based on the classification performance metrics by using classification rate (CR), error rate (ERT), precision (Prec.) and recall (rec.) as well as the number of extremal regions produced by ERS. UECFOOD-100 and UNICT-FD1200 are the two food datasets used to benchmark the proposed algorithm. The results of this research have found that the ERS algorithm by using optimum parameters and thresholds, be able to reduce the number of extremal regions with sustained classification performance.

Keywords: food recognition; object recognition; image processing; machine learning

I. INTRODUCTION

The advancement of mobile technology at a reasonable cost has indulged the people in photographing food and sharing their excitement when having a meal in the social media and it has become a worldwide phenomenon (Rich *et al.*, 2016). Food recognition has become an emerging research area in object recognition which has grown more substantially in the era of the smartphones and social media services revolutionary (Kagaya & Aizawa, 2015; R. Xu *et al.*, 2015). The revolution of big data and social media analytics technologies provides valuable encouragement that useful knowledge and information can be discovered from the massive volume of food images in social media, including trends of food consumption, eating habits and behaviour, and preferences for foods and restaurants (De Choudhury *et al.*, 2016; Fried *et al.*, 2015; Rich *et al.*, 2016). In previous research, the dense sampling and Different of Gaussian (DoG) are the two common interest points sampling used in earlier studies in food recognition (Kawano & Yanai, 2015; Martinel *et al.*, 2016; Sasano *et al.*, 2016). Inevitably, features will be extracted from irrelevant interest points (i.e. from the background, especially if it is complex) (Altintakan & Yazici, 2015) and will generate less informative descriptions regardless of the sampling techniques being used. Interest region-based detectors using Maximally Stable Extremal Region (MSER) that were used in the previous study (Razali *et al.*, 2017) use global segmentation and take into account regions from images with complex backgrounds as well. The configuration parameter of MSER based on Extremal Region Detection (ERD) technique as proposed by Razali *et al.*, 2019)

*Corresponding author's e-mail: ayu@upm.edu.my

have detected a massive amount of Extremal Regions especially from the food images with complex background. Detectors based on DoG also unavoidably detect interest points within complex and noisy backgrounds (Yu *et al.*, 2013). Furthermore, the number of interest points is still very high for real-time applications, and the irrelevant interest points increase the computational cost of the feature encoding process (Lin *et al.*, 2016; Mukherjee *et al.*, 2016; Xu *et al.*, 2015). Figure 1 shows examples of ER detection on a complex background.

The illustrations in Figure 1 (b), (e), and (h) demonstrate ER detection, while illustrations (c), (f), and (i) show interest points detection on the centroid of each ER. The examples feature a complex image background that has higher contrast and brightness density, and a pronounced texture, leading MSER also to detect ERs from these regions. Therefore, this research proposes the interest regions selection algorithm, or ER selection algorithm (ERS) to reduce the quantity of ERs, especially from the image background based on the spatial information of extremal regions. The evaluation and analysis have been conducted to determine the optimal threshold and parameters configuration in ERS.

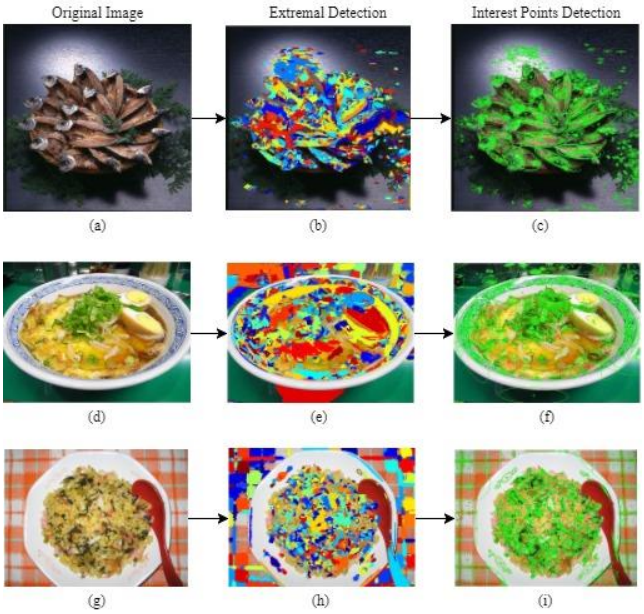


Figure 1. Examples of ER detection on a complex background

II. RELATED WORKS

The massive number of interest points produced by local feature extraction will provide a more discriminative visual dictionary but may also increase the computational effort for feature encoding (Lin *et al.*, 2016). It is good practice to reduce the number of interest points but at the same time retain or increase the recognition performance. A minimal amount of local (or useful) features may conserve memory and reduce the computational burden of feature encoding as well.

Interest points can be sampled densely and may contain redundancies. Redundant interest points are defined as those with other interest points close to them and which may provide unnecessary information (Rudinac, Lenseigne, & Jonker, 2009). Using a certain neighbourhood threshold, redundant interest points can be filtered. Another filtering technique is based on information entropy, where a small subset of the most representative interest points with the highest information content is selected.

The distinctiveness of an image can also be captured with a visual saliency or attention analysis approach, which has become popular in recent years (Li, Wang, Tian, & Ding, 2015). Visual saliency refers to the rarity of the colour, gradient, edges, and boundary of an image in a way intended to reflect how humans gaze at certain regions that attract them. Saliency detection allows efficient resource allocation when exciting areas can be determined early on before further processing. The rarity of appearance is driven by low-level image features. For example, lightness, colour, contrast, intensity, edge, orientation, shape, gradient, coarseness, and sharpness (X. Xu *et al.*, 2015).

The study conducted by (Gao & Yang, 2011) identified two types of visual saliency techniques: local contrast-based and global contrast-based which are integrated to obtain a distinctive SIFT interest point. Local contrast-based techniques identify the rarity of regions in a small local neighbourhood where intensity, colour, and edges are calculated. The global contrast-based approach uses the entire image to evaluate saliency. In the research of (Buoncompagni, Maio, Maltoni, & Papi, 2015), the saliency of FAST interest points was measured based on an estimation of distinctiveness, repeatability, and detectability. High distinctiveness, repeatability, and detectability of interest points indicate a high level of saliency. Distinctiveness computes the difference between the different interest points detectors, reproducibility means the invariance of an indicator towards different conditions, and detectability calculates the suitability of interest points under various viewpoints and illuminations. A similar saliency measurement was adopted by (Mukherjee *et al.*, 2016), where the scores for distinctiveness, repeatability, and detectability were retrieved to detect salience and rank the KAZE and SIFT

interest points.

A different approach is used by (Lin *et al.*, 2016) in selecting interest points, where k-means is used in two iterative stages known as IKS1 and IKS2. In IKS1, interest points are sampled using SIFT, whereas in IKS2 k-means is used to filter the interest points if the distance from the identified representative interest points is less than a pre-defined threshold.

Interest points that are located nearest to each other are assumed to be unnecessary since they may contain redundancy. Selecting only a few interest points is considered enough to provide more representative interest points. However, this technique may remove too many interest points and require an exhaustive threshold evaluation to deal with images of a different nature and variability.

One of the challenges in identifying salient interest points is when an image has a distracting foreground and cluttered background. Most of the techniques that are based on the statistics of centre distance are limited to centralised objects. Hence, to alleviate these problems (Liu, Ling-Yu Duan, Jie Chen, & Huang, 2016) introduced depth cue information to interest point selection. Interest points were selected based on correlation analysis between depth cue and scale.

Inaccuracy in identifying interest points using Laplacian of Gaussian (LOG) in SIFT occurs during the convolution process. A modification in the stage of selecting extreme points can decrease the bias (Zuchun, 2013). Instead of identifying the salient regions of an image (Yoo & Kim, 2013) proposed a scheme to model the image backgrounds in the BoF model. The background can be well represented by visual words since its local structure is not too varied. Hence, the dynamic background modelling using a soft assignment approach can effectively subtract the background. However, the scope of this approach is limited only to dynamically textured scenes.

Feature or instance selection and interest points selection are distinct (Lin *et al.*, 2016). Feature selection aims to reduce feature dimensionality by removing redundant or irrelevant features, whereas interest point selection is intended to remove useless interest points. Technically, feature selection removes the columns of the vector, in contrast to interest points selection which removes the rows. However, the main limitation of feature selection is its computational expense, especially for high feature dimensions. On the other hand, interest points selection yields good speed performance.

In general object recognition, a procedure to reduce the interest points is known as interest point or region selection. Interest point selection should be not be confused with feature selection used in data mining. Feature selection normally works by pruning the attributes represented in columns from the extracted features. Still, interest point selection prunes whole interest points represented in rows and is performed before or after feature description.

Moreover, interest point selection is not necessarily intended to remove interest points solely from the background but rather to remove any irrelevant interest points and can be performed exclusively to reduce computational cost. In work conducted by (Lin *et al.*, 2016), interest point selection aims to reduce the number of interest points by identifying those that can be considered redundant and less representative. However, their method is less useful for detecting the interest points of scenic images. Food and scene recognition have highly similar characteristics, and therefore problems that exist in scene recognition may also appear in food recognition. Just as scene images consist of multiple entities with regions of arbitrary shape, food images are also composed of many multi-class foods with high deformation and significant variation in colour and texture. Additionally, the DoG and dense sampling used to detect interest points in previous studies provide a dense search that might cause redundancy. The MSER detector used in this study locates interest points in each region and generates a less dense distribution of interest points.

However, the interest region selection procedure in food recognition seems to be an uncommon practice. Feature selection to reduce the dimensionality of descriptors and feature vectors is still rare in food recognition. For instance, (Kawano & Yanai, 2015) used PCA as a form of feature selection to reduce the Histogram Of Gradients (HOG) dimensionality, and (Faria, Alex, Rocha, & Torres, 2012) has adopted a heuristic-based approach to feature selection. Whether or not feature selection is conducted in an online or offline process, this procedure will prolong the feature representation process since it involves crucial computation, not only in choosing a subset of features but also in transforming them into another level of representation.

Therefore, an exciting point selection approach is preferable as it involves a less complicated procedure (Ghosh, Dhamecha, Keshari, Singh, & Vatsa, 2015) and also has greater potential to eliminate interest points from the background. The elimination of features from the image background via segmentation techniques can be considered a significant task in recent food recognition studies. The main reason is the multi-class appearance of food images and the

current interest in measuring food portion size for caloric and nutritional estimation. As mentioned earlier in this section, food segmentation is never a straightforward process due to the complex nature of food appearance.

III. EXTREMAL REGION SELECTION IN MSER

MSER is an interest-region-based detector which, along with its variants, has proven effective in scene recognition as it yields the best score in term of effectiveness and efficiency (Lee & Park, 2017). MSER for scene classification can detect objects of arbitrary shape in scenes containing multiple entities, as well as small objects. Therefore, MSER is chosen here as it is expected to be able to handle the complex appearance of foods, especially small foods and mixed foods that have solid mixtures of ingredients. MSER works by identifying a set of connected candidate regions that are discovered by using a global segmentation technique, specifically the watershed algorithm. Based on an intensity threshold, pixels are grouped into two sets, namely black and white. The threshold value is changed at each iteration, which changes the cardinality of each set. Finally, extremal regions are generated as connected regions, and each region is represented by an interest point located at its centre.

MSER locates interest regions based on the existence of ridgelines and connected regions using intensity threshold adjustment. The use of global segmentation takes into account regions from the entire food image. Food images with high proportions of background (and backgrounds with a complex appearance) will increase the number of unnecessary regions detected. Thus, the complex appearance of food images which include a complex background will inevitably lead to the detection of many background regions. The features generated by the image background can be considered as noise and do not contribute to classification performance (Altintakan & Yazici, 2015; Zhang *et al.*, 2016) and occupy more time for interest detection, description, and feature encoding (Lin *et al.*, 2016).

The motivations for ERS algorithm are twofold. First, the capability of interest points detectors to find interest points most densely in the image foreground, meaning that fewer ERs are detected from the background. Second, in the study conducted by (Lin *et al.*, 2016), unsupervised learning based on k-means clustering was used to remove redundant interest points. Generally, k-means initialises its centroids randomly on denser regions. Figure 2 shows the centroids plotted by using k-means on the spatial information (i.e. location) of interest points, in a food image.

As depicted in Figure 2, the coloured points represent the location of interest points, and the black circles mark the distribution of the centroids. The centroids are placed on a dense area of points which are in the image foreground. This example uses a small cluster size, but the larger cluster sizes used in this study will also place the centroids in the surrounding area that might be the image background. Therefore, interest points from image background and foreground are differentiated by the density of interest points held by the centroids.

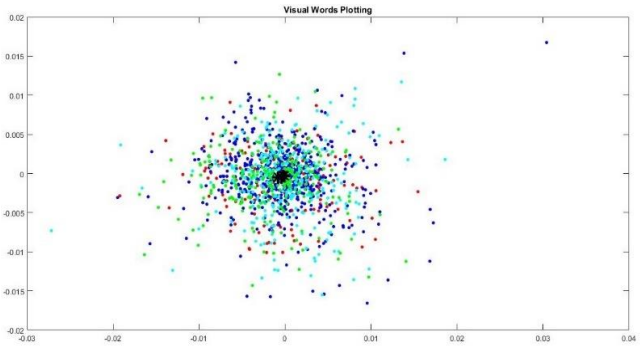


Figure 2. Centroid plotting in the k-means algorithm on a food image

As shown in Figure 3, ERS is performed after ERD. Based on the number of ERs detected by ERD on each image, it will determine whether that image should perform interest points selection. If the number of ERs is greater than the pre-defined threshold Z, ER selection will be applied. The preliminary analysis used to determine threshold Z will be explained in Section 4.

The data retrieved from each ER contains several items of information, including spatial information that indicates the coordinates of the x and y-axis of the located interest points in each ER. A clustering by k-means is performed to group the ERs based on their location on the x and y-axis. The centroids are placed randomly in the denser areas of data points. The clustering using k-means uses the distance function where points that are located within a certain neighbourhood will be grouped into a cluster. Therefore, all spatial information of ERs will be clustered. The centroid that distinguishes the ER from the image background is recognised by calculating the histogram, or occurrence frequency (OF), of the centroids.

As mentioned earlier, the use of a detector tends to identify denser ERs in the image foreground. Based on this

assumption, centroids with a small quantity of ERs, or low OF, are probably from the background. Additionally, spatial information indicates that interest points located too closely might be redundant (Rudinac *et al.*, 2009). Therefore, if the OF of the centroid is less than threshold *V*, the ERs that belong to the respective centroid will be regarded as noise and will be removed. The evaluation to determine threshold *V* (*OF*) is provided in Section 5.

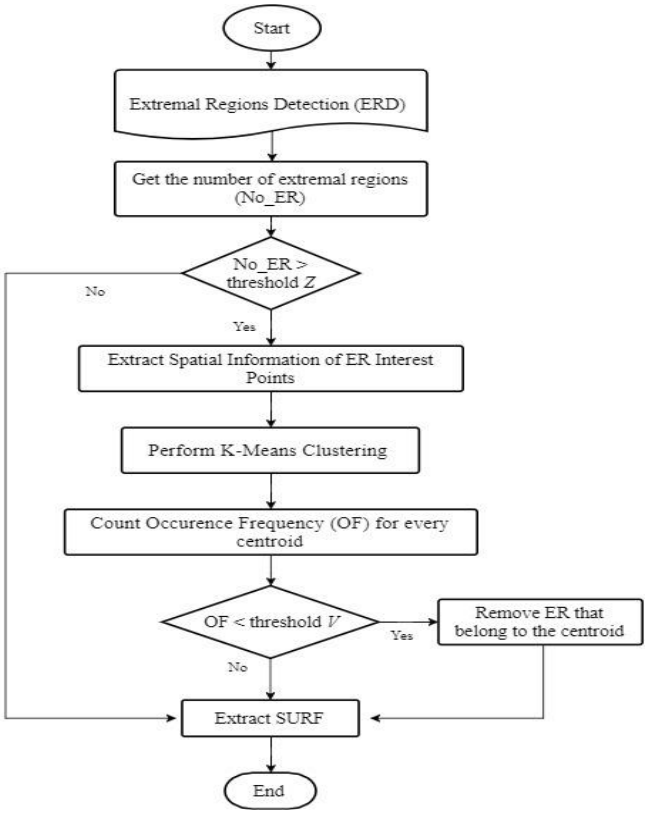
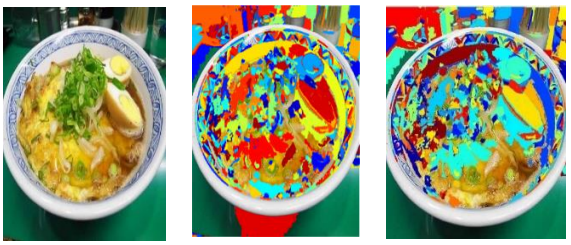


Figure 3. Extremal region selection (ERS) technique

The k-means clustering used in ERS requires the cluster size *k* to be set before performing the clustering. However, the high diversity and variability of food appearance make determining *k* an exhaustive process that may subsequently affect the performance of the ERS algorithm. If the value of *k* is too small, there will be too many ERs grouped into a cluster, and this will possibly mix ERs from foreground and background. In contrast, a larger *k* may remove too many ERs, including those from the front, as the OF holds a small number of ERs. An example of ER removal on a sample image is shown in Figure 4.



(a) Original (b) ER Detectio (c) ER Selection-

Figure 4. Extremal region selection

As shown in Figure 4 (b), 554 ERs have been detected in total, with many detected from the foreground. Then, the ERS is performed in (c), and 69 ERs have been removed. In this example, the value of *K* and *OF* used is 10 and 50 respectively. The *OF* for each cluster *K* is shown in Table 1.

Table 1. The Value of K and OF in ERS

Cluster (*K*)	**Occurrence Frequency (OF)**
1	50
2	31
3	57
4	55
5	64
6	64
7	66
8	38
9	52
10	77

As shown in Table 1, the number of ERs (*OF*) that belong to each cluster (*K*) is computed. As this example using *OF*=50, the cluster with OF less than 50 will be removed. Thus, the ERs in cluster 2 and 8 will be removed as it consists of 31 and 32 ERs, respectively. This example shows that the variable of *K* and *OF* is crucial in ERS algorithm as it is affecting the number and location of ERs to be eliminated. Therefore, empirical evaluation to determine the optimal value of *K* and *OF* is performed as presented in Section 5. Table 2 shows the algorithm for ER selection. The algorithm is developed based on the flowchart shown in Figure 3.

Table 2. Extremal Region Selection Algorithm

Extremal Region Selection (ERS)
Input: Extremal Regions (ER) detected by using ERD $\mathbf{ER} = \{\mathbf{ER}_1, \mathbf{ER}_2, \mathbf{ER}_3 \ldots . \mathbf{ER}_a\}$ **Output: The selected and extracted ER**
1. **for** all images i = $\{i_1, i_2, i_3 \ldots i_n\}$ **do**
2. extremal_region{i} ← Detected ER using ERD for each i

3. No_ER{i}← Get the number of ER for each i
4. **if** No_ER{i} > threshold Z **then**
5. extremal_region. Location{i} ← Access spatial information of coordinate (x, y) for each ER
6. mserpointslocation{i}← store coordinate (x, y) for each ER
7. k-means(mserpointslocation{i}, *K*) ← Perform clustering using k-means on the coordinate (x, y) for each ER.
8. cluster_location{ER} ← return centroid for each ER.
9. frequency{i} ← Calculate frequency for each centroid
   if frequency{i} < *OF* **then**
   Remove ER_i
   end if
   end if
10. all_regions{i} ← keep all updated ER
11. extractfeatures{i} ← Describe by using SURF
12. **end for**

As shown in Table 2, the input to the ERS algorithm is a list of ERs detected by ERD, and the output is a set of selected ERs, which are kept in a cell array of *all_regions{i}* before their features are described using a SURF descriptor.

IV. PRELIMINARY EXTREMAL REGION QUANTITY ANALYSIS

The purpose of ER quantity analysis is to provide a basic evaluation to determine a threshold value *Z* for which images with fewer than *Z* ERs will use the ERS technique. A preliminary experiment is conducted to reveal the classification performance for each food category using the traditional MSER detector and SURF descriptor.

The results of the experiments are used to form two groups. Group A retains all food categories that yield classification rate below to 80%, and Group B keeps food categories that yield classification rate above or equal to 80%. This group is specified to observe the difference in the number of interest points detected between these groups. There are 78 food categories with a classification rate below 80%, representing the majority of the food categories. A total of 22 of food categories obtained classification rate above 80%. The number of ERs per category in both groups is shown in Figure 5. As seen in Figure 5, the numbers of ERs for group B are almost all higher than those in group B. The average of ER quantity in group A is 25,446 and 50,130 in group B. This finding indicates that food categories with a higher number of ERs have better classification performance.

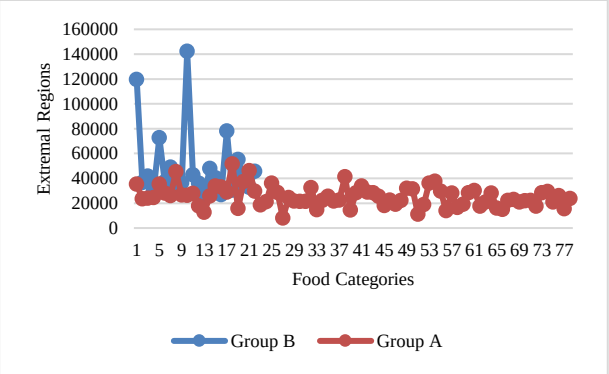


Figure 5. Comparison of ER quantity between group A and group B

The maximum, minimum, mean, median, and mode of the ER number for all images are calculated for both groups A and B. The maximum and minimum numbers of ERs guide how to group these images into a few ranges with an interval of 100 to identify the distribution of ERs. An interval of 100 is fixed as many texture-less foods images have the number of interest points within the range of 0-100.

Table shows the statistics for ER numbers, comprising the maximum, minimum, mean, median, and mode for groups A and B.

Table 3. Analysis of ER quantity for groups A and B

Analysis	Group A (CR < 80%)	Group B (CR> 80%)
Maximum	1548	2106
Minimum	0	5
Mean	195	250
Median	155	214
Mode	105	118

Based on Table 3, the maximum and minimum number of ERs are used to construct the ranges to identify the distribution of ERs in that respective range. As mentioned earlier, the range is specified based on the observation that revealed many texture-less images have interest points below 100. For example, the food category potage consists of 79.65% of images with ERs below 100. The maximum is from the image that generates the highest number of ERs, and the minimum is from the image that generates the lowest number. The mean, median, and mode of ERs are used to support the selection of thresholds *Q* and *Z*. Since the number of ERs of the images in group B is greater than group A, group B has greater values for all statistics.

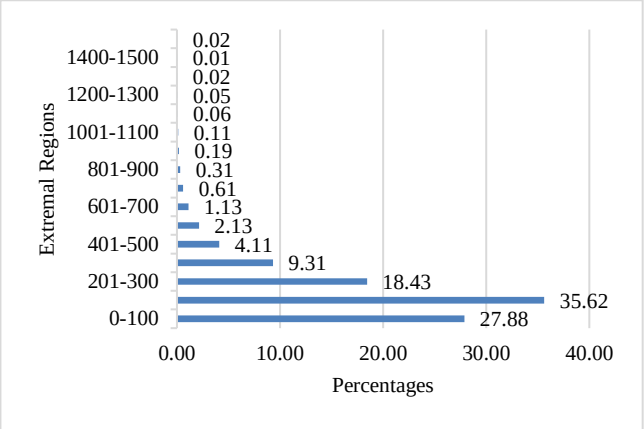


Figure 6. Distribution of ERs in group A (CR<80%)

Based on Figure 6, most images contain ERs in the range of (0, 100) and (101, 200) making up 63.5% of the total. The value of the mean, median, and mode are within the range of (0, 200), as shown in Table 3. Figure 7 shows the histogram of ERs for group B.

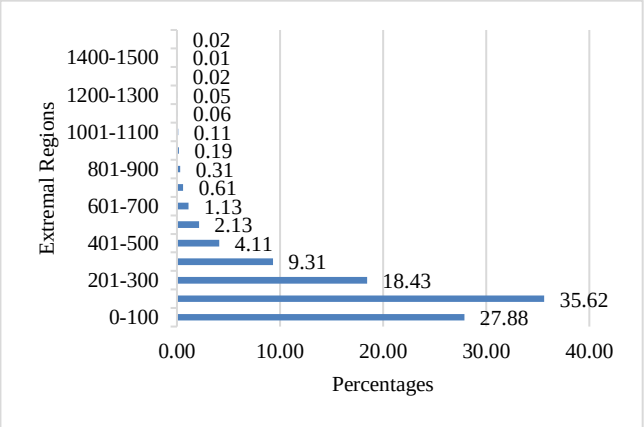


Figure 7. Distribution of ERs in group B (CR ≥ 80%)

Based on Figure 7, most images are found in the ranges (101, 200), (201, 300), and (301, 400), accounting for 68.61% of the total. The most remarkable difference between group A and group B shown in this analysis is the number of images belonging in the range (0, 100). In group B, only 17.28% of the images have less than 100 ERs. In contrast, the percentage of images with ER quantity below 100 in group A is much larger, namely 27.88%. In the other hand, a significant number of images in group A have the number of interest points in the range (100,200). Based on the graph in Figure 7, most of the food images with better classification performance have between 100 and 500 ERs. Therefore, any image with more than 500 ERs is too dense and might contain too many ERs from the background. Figure 8 shows three examples from three food categories with more than 500 ERs.

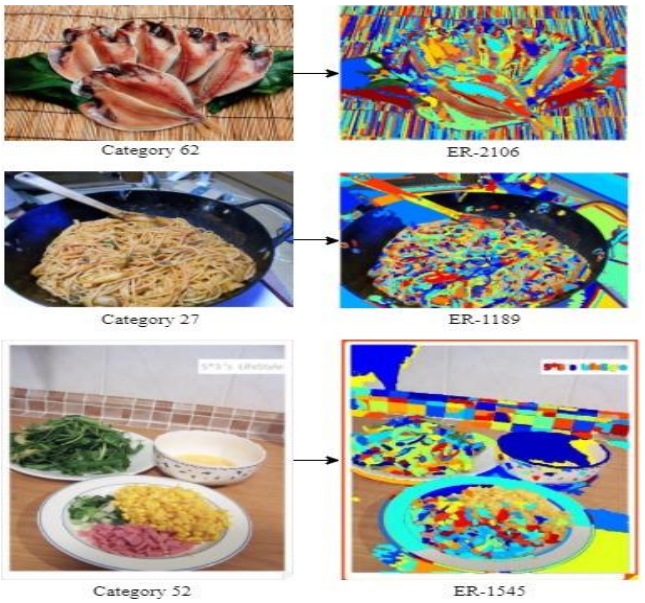


Figure 8. Examples of foods with ERs above 500

Figure 8 shows dense detection with many ERs from the image background. The image background is considered complex as it has a pronounced texture (as illustrated by category 62), contains miscellaneous objects (as per category 27) and contains visible ridgelines and high contrast (as seen in category 52). As a result, the image background and foreground are difficult to distinguish by MSER. In this context, the image with ERs more than 500 can be considered for ERS. However, an empirical evaluation to find an optimal value of *Z* will be performed as provided in Section 5.

V. RESULTS AND DISCUSSION

This section provides the experimental results of the proposed method: the ERS algorithm designed to reduce the quantity of ERs, especially from the background. ERS is performed after ER detection and initially is used on any food images with ER quantity greater than *Z*. The determination of threshold *Z* was initially explained in Section 4 and, as mentioned then, an ER range of 100 to 500 is sufficient to provide good classification performance. An excessive number of ERs could be overwhelming since too many are detected from the image background, especially in images with a complex background. However, an empirical evaluation threshold *Z* is conducted by using 500 and 1000 to determine the best performance of ERS. There are only two values are evaluated to determine threshold *Z* as only 10 food images have ERs greater than 1500.

A. Evaluation of ERS Parameters on UECFOOD-100 Dataset

Two parameters govern ERS: the so-called cluster size *k* and the occurrence frequency (OF). OF computes the histogram of the centroid on each cluster *k*. K-means is used to partition the ERs based on the location of the interest points within them. Table 4 presents the overall performance of ERS parameters and threshold *Z*.

As shown in Table 4, the evaluation can be grouped based on the two threshold *Z:* 500 and 1000. The *K* and *OF* evaluation consist of two stages, namely, stage 1 and stage 2. Stage 1 evaluates the effect of *K* while stage 2 evaluates the effect *OF* towards the ERS algorithm. The findings showed the ERS variants have successfully reduced some amounts of ERs and sustained the ERT performance obtained previously in ERD. In addition to that, the threshold *Z* required different parameter configuration for ERS to performed in an optimum manner.

Table 1. Evaluation of ERS parameters and threshold Z

Threshold *Z*	**Stages**	**ERS Variants**	**Performance Measurement**				
			CR%	**ERT%**	**Prec.%**	**Rec.%**	**Extremal Regions**
500 (5062 images)	Stage 1 (*OF*=10)	**ERS1 (*k* =20)**	**90.31**	**0.10**	**90.30**	**90.30**	**6577569**
		ERS2 (*k* =40)	89.81	0.10	90.00	89.80	6324237
	Stage 2 (*K*=20)	ERS3 (OF =20)	89.59	0.10	89.70	89.60	6338721
		ERS4 (OF =30)	88.87	0.10	88.90	88.90	5705136
1000 (292 images)	Stage 1 (*OF*=10)	ERS5 (*k* =20)	90.47	0.10	90.50	90.50	6608064
		ERS6 (*k* =40)	**90.59**	**0.10**	**90.60**	**90.60**	**6604616**
		ERS7 (*k* =60)	90.33	0.10	90.40	90.30	6592816
	Stage 1 (*K*=40)	ERS8 (OF =20)	90.13	0.10	90.20	90.10	6570537
		ERS9 (OF =30)	90.31	0.10	90.40	90.30	6472556
		ERS10 (OF =40)	90.39	0.10	90.40	90.40	6382108
		ERS11 (OF =50)	88.74	0.10	88.90	88.70	6310263
ERD			90.93	0.10	91.10	90.90	6608363

Figure 9 depicted a combo graph of classification rate and number of ERs in ERS variants by using threshold *Z*=500.

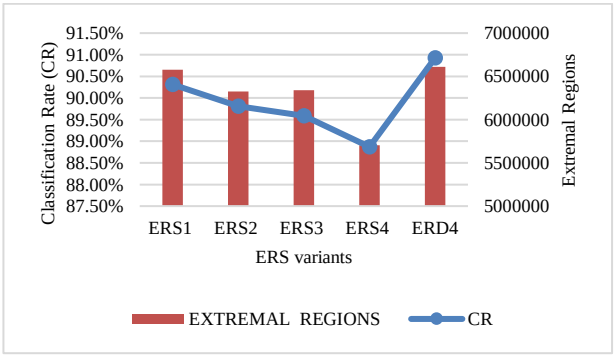


Figure 9. Classification rate and ERs of ERS variants for threshold Z=500

Based on Figure 9, it is apparent that the classification rate dropped very slightly from 90.93% in ERD to 90.31% in ERS1. As the parameters value changed in ERS2, ERS3 and ERS4, the classification rate has minorly affected but has reduced the number of ERs significantly. In this case, ERS1 yielded the best classification performance overall. Although the ERS algorithm decreased the ERD4 classification rate by 0.62% in ERS1, the error rate is still low as in ERD4.

Figure 10 showed a graph of classification rate and number of ERs in ERS variants by using threshold *Z*=1000.

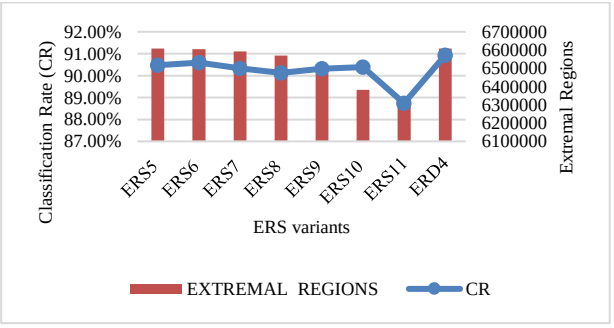


Figure 10. Classification rate and ERs of ERS variants for threshold Z=1000

Based on the graph presented in Figure 10, the classification rate pattern over ERS5, ERS6, ERS7, ERS8, ERS9 and ERS10 does not severely affected and can be sustained above 90%. However, the number of extremal regions demonstrated a dramatic reduction. For instance, the number of ERs in ERS10 is only 6382108 to achieved 90.39% classification rate.

In conclusion, based on results showed in Table 4, Figure 9 and Figure 10, food images with the number of ERs above 1000 (threshold *Z*) obtained better classification performance from using ERS algorithm. Specifically, ERS6 has yielded the best classification performance among ERS variants.

B. Visual Effect of the ERS Variants

This section will provide the illustrations to show the effect of ERS algorithm by using different parameter configuration on two samples of food image. The two samples are taken from food image with the number of ERs more than 500 and 1000.

Figure 11 shows the effect ERS parameters on a sample image by using threshold *Z*=500. In Figure 11, the *OF* value was set to 10 in ERS1 and ERS2, and cluster size was set to 20 and 40, respectively. While the OF in ERS3 and ERS4 were set to 20 and 30 respectively. Any cluster containing fewer ERs than the OF value was removed. Initially, ERD in (a) detects 525 ERs in a sample image belonging to this image. Both parameter *K* and *OF* has played an important role to remove the ERs, especially from the image background as can be seen in (d), (e) and (f). Figure 12 shows the effect of ERS parameter configuration on a sample image by using *Z*=1000.

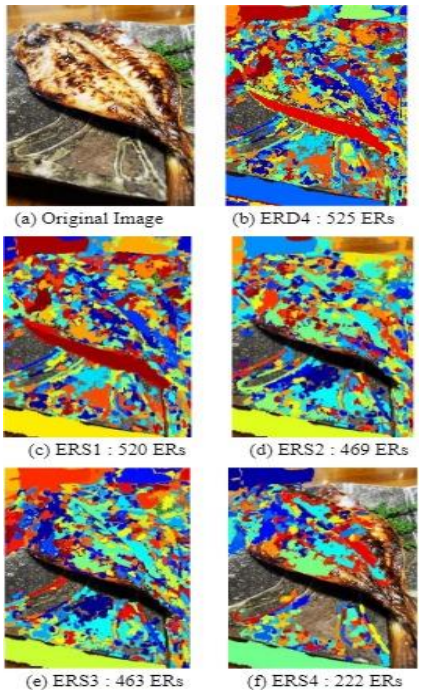


Figure 11. Effect ERS variants by using Z=500

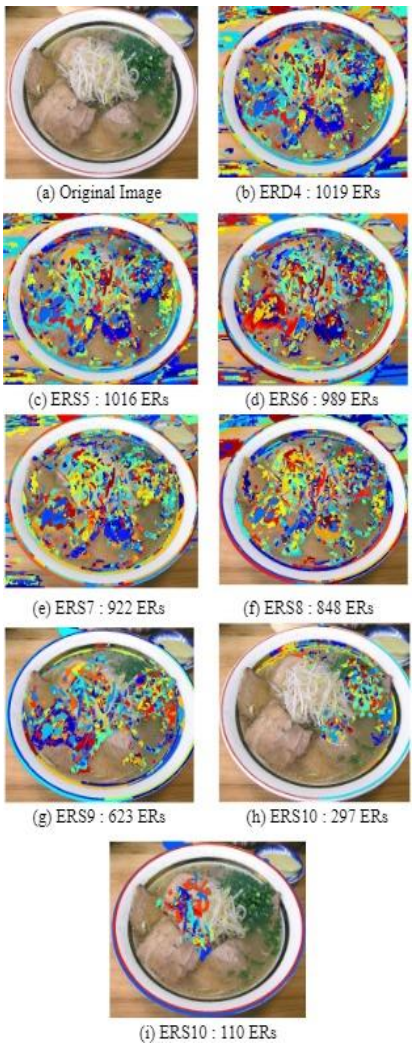


Figure 12. Examples of foods with ERs above 500

Based on Figure 12, the number of ERs, especially from the background, has been removed gradually in ERS6, ERS7, ERS8 and ERS9. However, the ERS10 and ERS11 have removed ERs dramatically, including many ERs from the foreground image.

Conclusively, the ERS variants have successfully reduced the number of ERs, including many ERs from the background. In general, the ERS algorithm is applicable for any image with more than 500 ERs, provided a suitable ERS parameter configuration needs to be used for according ERs density to images based on the empirical evaluation conducted in Table 4.

Denser ERs are usually detected from food images with complex background. Unavoidably, ERs from the foreground are eliminated as well due to the heterogeneous colour and texture of foods. The global segmentation in MSER samples food regions into granular parts. However, aside from visual inspection, the primary evaluation metrics of the proposed ERS algorithm are still its classification performance and the quantity of ERs it produces (Lin *et al.*, 2016).

VI. CONCLUSIONS AND FUTURE WORKS

This research has proposed the extremal region selection (ERS) algorithm in the MSER detector to reduce the number of extremal regions. Specifically, ERS recognised the extremal regions detected from the complex background of food image by calculating the occurrence frequency of the k-means centroids. The clustering is performed on the spatial information of the extremal region. The results demonstrated the reduction number of extremal regions by 37 ERs using ERS variants with ERS6 yields the most optimal performance with 0.1 error rate. This research can be improved in the future by incorporating visual saliency evaluation as a criterion in performing ER selection. Saliency values for each interest point might provide a clear and more accurate separation between relevant interest regions and outliers because, in some cases, the image background is not an outlier and may be a useful clue for recognition.

VII. ACKNOWLEDGEMENT

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

VIII. REFERENCES

Altintakan, U & Yazici, A 2015, 'Towards effective image classification using class-specific codebooks and distinctive local features', *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp. 323–332. https://doi.org/10.1109/TMM.2014.2388312

Buoncompagni, S, Maio, D, Maltoni, D & Papi, S 2015, 'Saliency-based keypoint selection for fast object detection and matching', *Pattern Recognition Letters*, vol. *62*, pp. 32–40. https://doi.org/10.1016/j.patrec.2015.04.019

De Choudhury, M, Sharma, S & Kiciman, E 2016, 'Characterising Dietary Choices, Nutrition, and Language in Food Deserts via Social Media', in *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing*, pp. 1157–1170. https://doi.org/10.1145/2818048.2819956

Faria, FA, Alex, J, Rocha, A & Torres, RS 2012, 'Automatic Classifier Fusion for Produce Recognition', in *25th SIBGRAPI Conference on Graphics, Patterns and Images*. https://doi.org/10.1109/SIBGRAPI.2012.42

Fried, D, Surdeanu, M, Kobourov, S, Hingle, M & Bell, D 2015, 'Analysing the language of food on social media', in *Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014*, (Section II), pp. 778–783. https://doi.org/10.1109/BigData.2014.7004305

Gao, H & Yang, Z 2011, 'Integrated Visual Saliency Based Local Feature Selection for Image Retrieval', in *Intelligence Information Processing and Trusted Computing (IPTC), 2011 2nd International Symposium On*, pp. 47–50. https://doi.org/10.1109/IPTC.2011.19

Ghosh, S, Dhamecha, TI, Keshari, R, Singh, R & Vatsa, M 2015, 'Feature and keypoint selection for visible to near-infrared face matching', in *2015 IEEE 7th International*

Conference on Biometrics Theory, Applications and Systems, BTAS 2015.
https://doi.org/10.1109/BTAS.2015.7358760

Kagaya, H & Aizawa, K 2015, 'Highly Accurate Food/Non-Food Image Classification Based on a Deep Convolutional Neural Network', in *International Conference on Image Analysis and Processing*, *9281*, pp. 350–357. https://doi.org/10.1007/978-3-319-23222-5

Kawano, Y & Yanai, K 2015, 'FoodCam: A real-time food recognition system on a smartphone', *Multimedia Tools and Applications*, vol. 74, no. 14, pp. 5263–5287. https://doi.org/10.1007/s11042-014-2000-8

Lee, MH & Park, IK 2017, 'Performance evaluation of local descriptors for maximally stable extremal regions', *Journal of Visual Communication and Image Representation*, vol. 47, pp. 62–72.
https://doi.org/10.1016/j.jvcir.2017.05.008

Li, Y, Wang, S, Tian, Q & Ding, X 2015, 'A survey of recent advances in visual feature detection', *Neurocomputing*, vol. 149(PB), pp. 736–751.
https://doi.org/10.1016/j.neucom.2014.08.003

Lin, W, Tsai, C, Chen, Z & Ke, S 2016, 'Keypoint selection for efficient bag-of-words feature generation and effective image classification', *Information Sciences*, vol. 329, pp. 33–51.
https://doi.org/10.1016/j.ins.2015.08.021

Liu, Z, Ling-Yu Duan, Jie Chen & Huang, T 2016, 'Depth-based Local Feature Selection for Mobile Visual Search', in *International Conference on Image Processing (ICIP)*, IEEE.

Martinel, N, Piciarelli, C & Micheloni, C 2016, 'A supervised extreme learning committee for food recognition', *Computer Vision and Image Understanding*, vol. 148, pp. 67–86.
https://doi.org/10.1016/j.cviu.2016.01.012

Mukherjee, P, Srivastava, S & Lall, B 2016, 'Salient keypoint selection for object representation', in *2016 22nd National Conference on Communication, NCC 2016*. https://doi.org/10.1109/NCC.2016.7561176

Razali, MN, Manshor, N, Mustapha, N, Yaakob, R & Zainudin, MNS 2019, 'Improving invisible food texture detection by using adaptive extremal region detector in food recognition', in *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 8(1.4 S1), pp. 68–74.
https://doi.org/10.30534/ijatcse/2019/1181.42019

Razali, MN, Manshor, N, Halin, AA, Yaakob, R & Mustapha, N 2017, 'Food Category Recognition using SURF and MSER Local Feature Representation', in *International Visual Informatics Conference*, pp. 212–223.
https://doi.org/978-3-319-70010-6_20

Rich, J, Haddadi, H & Hospedales, TM 2016, 'Towards bottom-up analysis of social food', in *DH 2016 - Proceedings of the 2016 Digital Health Conference*, pp. 111–120. https://doi.org/10.1145/2896338.2897734

Rudinac, M, Lenseigne, B & Jonker, P 2009, 'Keypoint extraction and selection for object recognition', in *MVA2009 IAPR Conference on Machine Vision Applications*, pp. 191–194.

Sasano, S, Han, X-H & Chen, Y 2016 'Food Recognition by Combined Bags of Color Features and Texture Features', in *International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, pp. 815–819.

Xu, R, Herranz, L, Jiang, S, Wang, S, Song, X & Jain, R 2015, 'Geolocalized Modeling for Dish Recognition', *IEEE Transactions on Multimedia*, vol. 17, no. 8, pp. 1187–1199. https://doi.org/10.1109/TMM.2015.2438717

Xu, X, Mu, N & Zhang, H 2015, 'Inferring Visual Perceptual Object by Adaptive Fusion of Image Salient Features', *Mathematical Problems in Engineering*, *2015*. https://doi.org/10.1155/2015/973241

Yoo, S & Kim, C 2013, 'Background subtraction using hybrid feature coding in the bag-of-features framework', *Pattern Recognition Letters*, vol. 34, no. 16, pp. 2086–2093. https://doi.org/10.1016/j.patrec.2013.07.008

Yu, J, Qin, Z, Wan, T & Zhang, X 2013, 'Feature integration analysis of bag-of-features model for image retrieval', *Neurocomputing*, vol. 120, pp. 355–364.
https://doi.org/10.1016/j.neucom.2012.08.061

Zhang, C, Wen, G, Lin, Z, Yao, N, Shang, Z & Zhong, C 2016, 'An Effective Bag-of-visual-word Scheme for Object Recognition', in *International Congress on Image and Signal Processing, BioMedical Engineering and Informatics*, pp. 417–421.
https://doi.org/10.1109/CISP-BMEI.2016.7852747

Zuchun, D 2013, 'An Effective Keypoint Selection Algorithm in SIFT', *International Journal of Signal Processing, Image Processing and Pattern Recognition*, vol. 6, no. 2, pp. 155–164.
https://doi.org/10.1.1.641.1181