Indonesian Journal of Electrical Engineering and Computer Science
Vol. 21, No. 1, January 2021, pp. 179~195
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v21.i1.pp179-195

Fuzzy encoding with hybrid pooling for visual dictionary in food recognition

Mohd Norhisham Razali[1], Noridayu Manshor[2], Alfian Abdul Halin[3], Norwati Mustapha[4], Razali Yaakob[5]
[1]Faculty of Computing and Informatics, Universiti Malaysia Sabah, Malaysia
[2,3,4,5]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Malaysia

Article Info

Article history:

Received Feb 16, 2020
Revised Apr 5, 2020
Accepted Jul 21, 2020

Keywords:

Food recognition
Object recognition

ABSTRACT

Tremendous number of f food images in the social media services can be exploited by using food recognition for healthcare benefits and food industry marketing. The main challenges in food recognition are the large variability of food appearance that often generates a highly diverse and ambiguous descriptions of local feature. Ironically, the ambiguous descriptions of local feature have triggered information loss in visual dictionary constructions from the hard assignment practices. The current method based on hard assignment and Fisher vector approach to construct visual dictionary have unexpectedly cause errors from the uncertainty problem during visual word assignation. This research proposes a method of combination in soft assignment technique by using fuzzy encoding approach and maximum pooling technique to aggregate the features to produce a highly discriminative and robust visual dictionary across various local features and machine learning classifiers. The local features by using MSER detector with SURF descriptor was encoded by using fuzzy encoding approach. Support vector machine (SVM) with linear kernel was employed to evaluate the effect of fuzzy encoding. The results of the experiments have demonstrated a noteworthy classification performance of fuzzy encoding approach compared to the traditional approach based on hard assignment and Fisher vector technique. The effects of uncertainty and plausibility were minimized along with more discriminative and compact visual dictionary representation.

Corresponding Author:

Noridayu Manshor
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
43400 UPM Serdang Selangor Darul Ehsan, Malaysia
hishamrz@ums.edu.my

1. INTRODUCTION

The advancement of mobile devices technology that provides a better imaging quality has attracted an interest from researchers to adopt object recognition method to facilitate self-dietary assessment via an automatic food recognition [1, 2]. Dietary assessment is a treatment for diet-related chronic diseases such as diabetes, hypertensions, and heart diseases that been strongly linked with the obesity and being overweight that are caused by imbalanced nutrition intake and lack of physical activities. This health issue has seriously affected nations worldwide as 1.9 billion adults were categorized as overweight in 2018 and 650 million of them were obese [3]. In addition to that, the explosion of social media services have witnessed the popularity of food images which creates potential to food recognition algorithm to be used for analysing the eating habits and food preferences which are useful for food industry.

In general, there are two main steps to accomplish food recognition task which are image description and classification [4]. Image description is a process to extract the visual content of foods. In particular, local feature is more suitable to represent food features as the properties of local features that capture minuscule parts of the food beside its robustness towards illumination, scale, rotation, and orientation which made it capable to deal with the cluttered appearance of foods [1, 5, 6]. The interest points that were detected and described have produced a high volume and diverse features that require the features to be transformed into another more simplified representation by using certain feature encoding technique.

Feature encoding is a process in bag of feature (BoF) model to construct the visual dictionary in order to represent the characteristics of image features from the highly diverse and massive volume of interest points. BoF model has been employed to encode local features in many food recognition studies [1, 5, 7, 8]. Feature encoding is a crucial step in BoF as it has significant impact on the classification performance [9]. The most common feature encoding technique used in previous studies is by using hard assignment approach where k-means clustering algorithm was used to generate centroids or visual words [1, 7, 10]. The feature encoding by using hard assignment works by assigning each feature description from interest point to visual word solely based on the distance between interest point and visual word. The high variations of foods that produce highly diverse ambiguous feature descriptions [1, 11, 12] may strongly lead to information loss or error while assigning feature descriptions to visual words. The errors in feature encoding occurred due to uncertainty and plausibility problem. Uncertainty and plausibility problem had been earlier discovered in scene image classification [13, 14] which showed that any image with a large variety of appearance suffer from uncertainty and plausibility problem.

The uncertainty and plausibility problem are triggered when a feature description is assigned to only one visual word without consideration of other visual words that could be more relevant. In a visual word uncertainty, an interest point can have a similar or just a slight difference of distance with two visual words, especially for interest points that are located near to the boundary of clusters. These interest points represented by using hard assignment can be ambiguous as they are assigned to a visual word without further evaluation with other visual words and interest points. On the other hand, visual word plausibility can also occur when interest points populated far away from any visual word assigned to any nearest visual word and might be wrongly grouped. Figure 1 shows a demonstration of uncertainty and plausibility effect on the samples in a food category from UECFOOD-100 dataset.

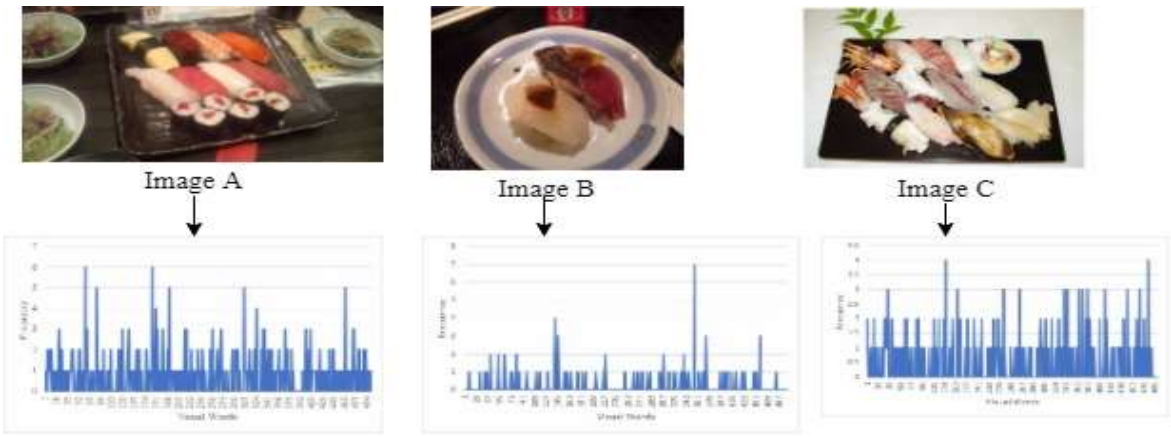

Figure 1. Uncertainty and plausibility effect on visual dictionary construction by using hard assignment

The samples of food image shown in Figure 1 demonstrated the variation of foods that lead to the confusion in generating visual dictionary. Obviously, the visual dictionary generated by the three samples have different pattern. This will definitely enlarge the intra-class variation in a food category. As mentioned in [2] and [12], the intra-class variations and deformable nature of food images have caused uncertainty and plausibility problem to become significant. Fisher vector (FV) technique is a more recent feature encoding than hard assignment introduced in image classification [15] that can generate a compressed representation via a small vocabulary size. FV has been introduced in food recognition in [16] which provides more advantage for mobile applications. However, the problem of uncertainty and plausibility still exist in FV representation as shown in Figure 2.

According to Figure 2, three different patterns of feature representation generated by using fisher vector on three food images A, B, and C from the sushi food category. In addition to that, the first-order statistic computation in Fisher vector end up with extremely longer feature vector that requires high computational cost for classification and less suitable for large-scale application [17, 18]. Therefore, this study is mainly proposing a soft assignment feature encoding technique to improve the visual dictionary construction in BoF. This study is motivated from the research conducted in [13, 19-21] that had suggested the soft assignment technique to deal with the uncertainty and plausibility in scene recognition. Specifically, the fuzzy feature encoding technique based on fuzzy set theory (FST) has been selected as it may lead to better construction of visual dictionary which in turn lead to better classification performance [22].

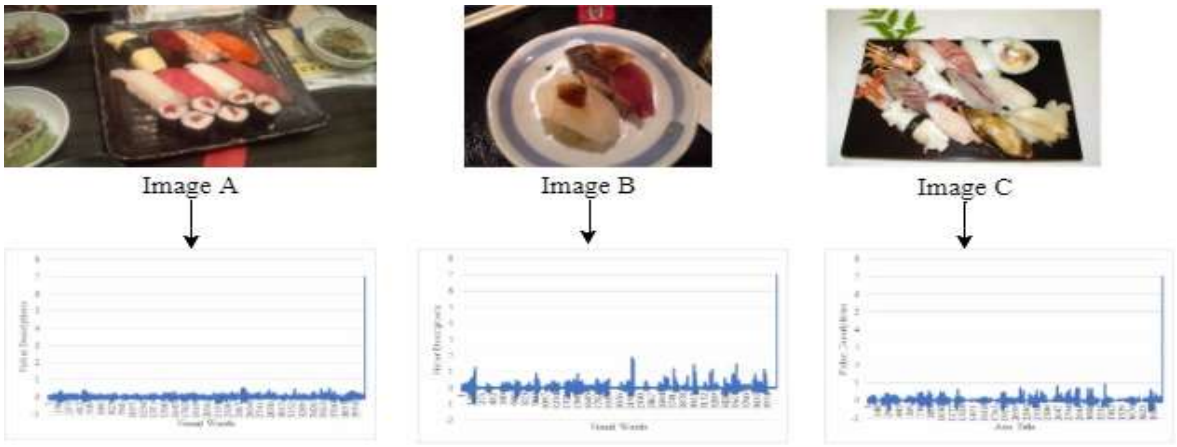

Figure 2. Uncertainty and plausibility effect on visual dictionary construction by using fisher vector

In summary, the contributions of this study are four folds:

a) We improved the BoF model in food recognition by incorporating FST in feature encoding process which reduced the uncertainty and plausibility problem.
b) We proposed the combination of maximum pooling and sum pooling techniques to produce a highly discriminative visual dictionary to summarize the encoded features.
c) We provide evaluation of feature encoding techniques towards different features.
d) We provide evaluation of feature encoding techniques towards different machine learning classifiers.
e) We provide evaluation on using different vocabulary sizes in fuzzy encoding technique.

The rest of the paper is organized as follows. The following sections discuss related works regarding feature encoding techniques, the fuzzy encoding technique, the experimental design, experimental results and findings are finally concluded in the last section.

2. RELATED WORKS

The bag of feature (BoF) model has been widely used in previous research to represent features from food images [5, 7, 8]. Feature encoding is a process within the BoF model that constructs the visual dictionary in order to represent the characteristics of image features from the highly diverse and massive volume of local feature interest points. The most common feature encoding technique is through hard assignment where k-means clustering and hierarchical k-means are used to construct the visual words. In these techniques, the centroids or visual words are placed near the most occurring interest points and each will be assigned to a single nearest centroid.

In general, the hard assignment approach to encode the local feature was criticised due to the errors created while building the visual dictionary [9, 20, 23]. Hard assignment approach is less suitable for images that contain large appearance variability [21]. The reason is that a single visual word feature using hard assignment may cause its relevance to other possible visual words to be lost. This can potentially lead to visual word uncertainty and plausibility. Visual word uncertainty occurs when a group of nearest interest points are assigned to different visual words. These interest points are usually populated near to the border in between the cluster regions. On the other hand, visual word plausibility is a case where the interest points are populated far from any visual word where they are assigned to any nearest visual word. Consequently, these interest points are wrongly grouped with other dissimilar interest points.

The problems of uncertainty and plausibility have been long identified in object recognition datasets dealing with natural scenes, such as Caltech and Pascal [9, 20, 21, 23, 24]. These datasets exhibit large variabilities in image appearances with high intra-class difference and inter-class similarities. As summarized by [25], many encoding approaches have been proposed such as Fisher vector, sparse coding, local tangent coding, and saliency coding to replace the traditional hard assignment technique.

However, the soft assignment based encoding using fuzzy logic has been found to be more representative at modelling uncertainty and plausibility by allowing multiple degrees of membership assignments to each visual word [22, 26, 27]. However, despite the success of fuzzy encoding approaches, to the best of our knowledge, the problem of uncertainty and plausibility in food recognition domain have yet to be explored. Since food images have very diverse appearances in terms of colour and texture, as well as being highly deformable in nature, they are not exempt from uncertainty and plausibility [1, 13]. Moreover, the uncertainty problem in food recognition contributes to high intra-class variations where the foods in the same category can vary depending on ingredients, location, and individual preferences [11]. In many of the previous works pertaining to food recognition, hard assignment by using k-means was adapted in the BoF model [1, 7, 28, 29].

Some of the works have used the fisher representation [30] and sparse coding [31] as alternatives to using hard assignment. However, the concentration of fisher representation is more to provide richer gradient representations with respect to the mean and co-variance from a gaussian mixture model (GMM), which results in a lengthy feature vector. Sparse coding on the other hand is more to capture the salient properties of local features. Both methods are not designed for eliminating the uncertainty and plausibility issues in food images.

3. FUZZY ENCODING IN BOF

The fuzzy technique to encode the features in BoF is based on the fuzzy set theory (FST) in [32]. FST allows soft assignment to be performed where initially each interest point is assigned to multiple visual words with different degrees of membership values. The membership value is determined by using a Gaussian Probability Density Function based on the distance between interest points and the visual words. The closer visual words are assigned higher membership values.

Fuzzy c-means (FCM) [33] and possibilistic c-means (PCM) [34] are among two established fuzzy clustering techniques that perform soft assignment using FST. FCM is known to perform very well on noise free data but tends to be sensitive to outliers. PCM on the other hand is able to deal with noisy data. Early work by [13] that dealt with the uncertainty problem in visual word had suggested that the higher dimensional feature space will create visual word uncertainty due to more interest points lying close to the clustering boundary. The uncertainty modelling were used to alleviate the problems and showed improvement on classification accuracy on general object recognition datasets. Then subsequent works are found in [21] and [19] that adopted uncertainty modelling using FCM for an automatic scene recognition, object tracking [23] and many other works [22, 24, 35, 36] have demonstrated the effectiveness of fuzzy encoding approach over the traditional approach using hard assignment technique.

4. SOFT ASSIGNMENT VS. HARD ASSIGNMENT

In this section, we illustrate the comparison between hard and soft assignment when constructing the visual dictionary. The illustration is adopted from the research conducted in [23]. Figure 3 shows the feature encoding technique by using hard assignment. The yellow square denote the visual words and are labelled as A, B, C and D. The round shape with the colours red, grey, purple, and green denoted the interest points or feature descriptions. The same colour of feature descriptions indicate the high similarity between these features.

Based on Figure 4, the hard assignment by using k-means would generate the centroids or visual words and place randomly at high density points. Then, each feature description is assigned to the nearest visual words which is shown by the arrow symbol. Table 1 shows the histogram calculation to generate the visual dictionary based on Figure 4.

As shown in Table 1, the histogram of each visual words A, B, C and D is calculated or sum pooling from all the feature descriptions that have been assigned to them. Figure 4 shows the histogram of visual word A, B, C and D.

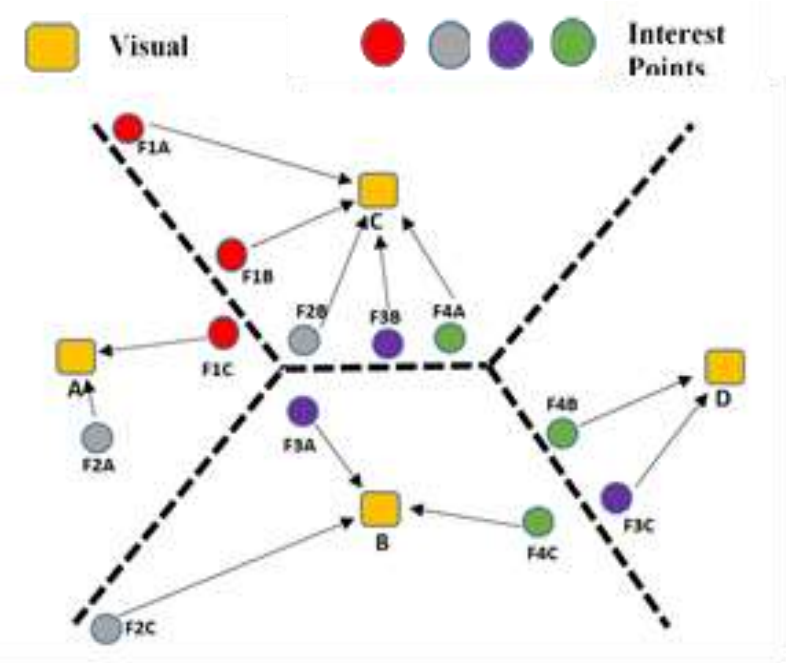

Figure 3. Feature encoding by using hard assignment

Table 1. Visual dictionary construction by using hard assignment

	A	B	C	D
F1A	0	0	1	0
F1B	0	0	1	0
F1C	1	0	0	0
F2A	1	0	0	0
F2B	0	1	0	0
F2C	0	1	0	0
F3A	0	1	0	0
F3B	0	0	1	0
F3C	0	0	0	1
F4A	0	0	1	0
F4B	0	0	0	1
F4C	0	1	0	0
Histogram	2	5	4	2

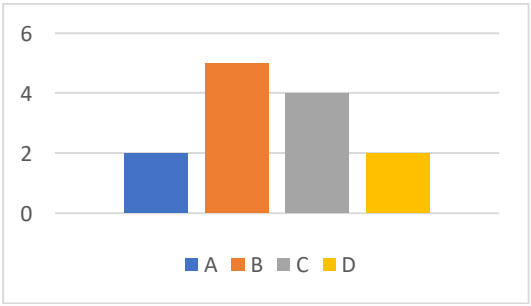

Figure 4. Visual word histogram by using hard assignment

In the graph shown in Figure 5, all visual words have been assigned with feature descriptions. Based on Figure 5, the uncertainty situations occur among the interest points located in the boundary of cluster. The hard assignment technique has assigned similar feature descriptions into different visual words.

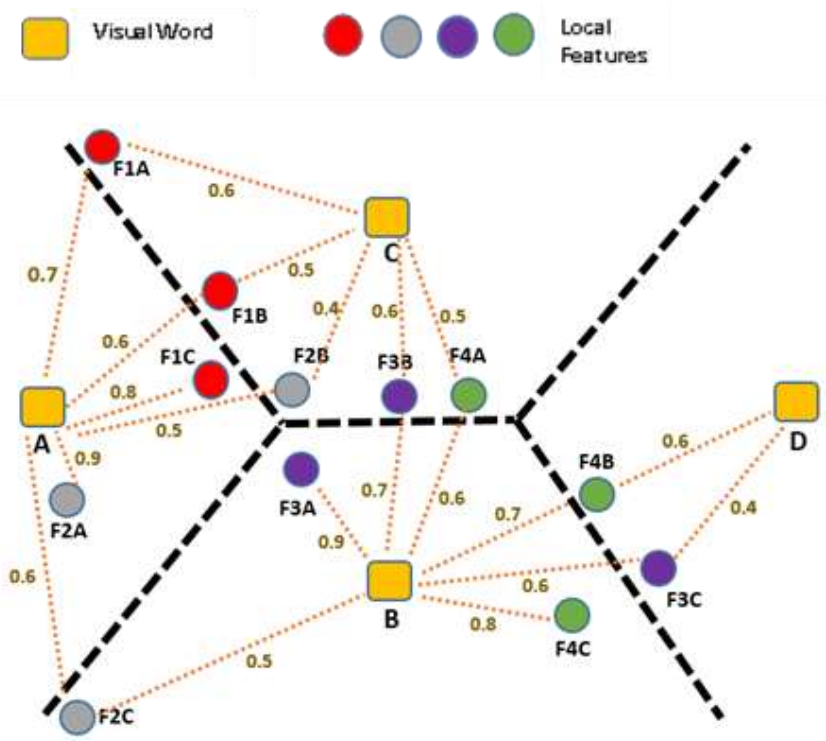

Figure 5. Feature encoding by using soft assignment

Based on Figure 5, every feature description is supposed to be assigned to all visual words. However, for illustration clarity, the demonstration in Figure 6 shows the assignation with only two visual words. Every assignation will return membership value calculated by using FST. As mentioned earlier, the uncertainty problem occurs when interest points are located at the boundary of a visual word. For example, F1A and F1B are assigned to visual words A and C, both having membership values for both clusters. However, careful examination of the weights reveal that both points have heavier weights for visual word A and are hence assigned to that cluster. The interest point F2C demonstrates the plausibility problem since it is located far from the visual words and the membership value contribute more on visual word A. Table 2 shows the histogram calculation to generate visual dictionary based on Figure 6.

Based on Table 2, maximum pooling is initially performed to choose one highest membership value for each feature descriptions as highlighted. This is followed by performing sum pooling to finalize the histogram. Figure 6 shows the visual word histogram.

Table 2. Visual dictionary construction by using soft assignment

	A	B	C	D
F1A	0.7	0.3	0.6	0.1
F1B	0.6	0.4	0.5	0.2
F1C	0.8	0.6	0.5	0.3
F2A	0.9	0.5	0.4	0.1
F2B	0.5	0.3	0.4	0.2
F2C	0.6	0.5	0.2	0.1
F3A	0.7	0.9	0.6	0.4
F3B	0.4	0.7	0.6	0.5
F3C	0.2	0.6	0.3	0.4
F4A	0.2	0.6	0.5	0.4
F4B	0.1	0.7	0.5	0.6
F4C	0.5	0.8	0.4	0.3
Histogram	6	6	0	0

Figure 6. Visual word histogram by using soft assignment

According to the histogram shown in Figure 6, the feature descriptions have been grouped into two visual words only. However, the uncertainty and plausibility problem has caused the histogram distribution of hard assignment in Figure 6 more sparse. Hence, the soft assignment has demonstrated more discriminative histogram.

5. EXPERIMENTAL DESIGN

In this section, the overall recognition process to classify food images by using fuzzy encoding approach is explained as shown in Figure 8. Experiments using the UECFOOD-100 dataset [37], which contains 100 food categories as shown in Figure 7 were conducted. The images are the real setting of food images as it was collected from the Internet consisting of multiple classes of food categories whose image contrast, lighting, and appearances differ greatly.

Figure 7. Samples of UECFOOD-100 dataset

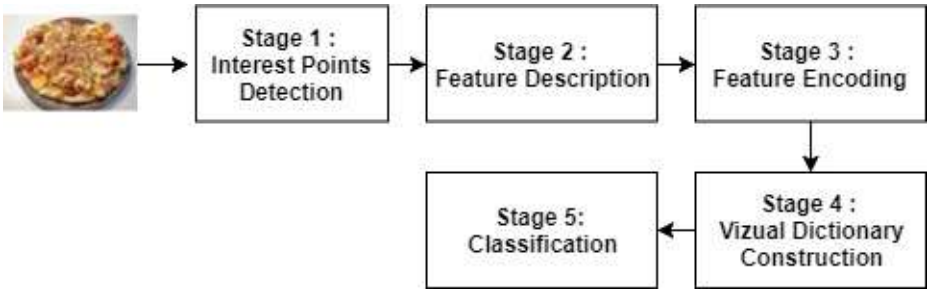

Figure 8. Food recognition based on fuzzy encoding approach

5.1. Stage 1: interest points detection

The interest points detector identifies a set of salient regions from an image. It provides stable and discriminative interest points that are robust to illumination variation [38]. The interest points detector, besides providing a distinctive set of interest points, is also more computationally efficient [39]. Specifically, maximally stable extremal region (MSER) detector is used to sample the interest points. MSER is an interest region based detector which has proven to be effective among its variant as it yields the best score in terms of effectiveness and efficiency in recent study [40]. MSER may provide a more discriminative interest points on food categories that have very strong mixture of ingredients as it may represent the irregular shape of foods, typically in parallelograms. Indeed, the larger size of patch detection in interest region detectors are more suitable to detect food interest points [41]. MSER works by identifying a set of connected candidate regions that are discovered by using a global segmentation by using watershed algorithm. Based on an intensity threshold, pixels are grouped into two sets which are black and white. The threshold value is changed at each iteration, which changes the cardinality of each set. Finally, the extremal regions are generated as connected regions and each region will be represented by an interest point that is located in the centre of extremal region.

5.2. Stage 2: feature descriptions

Feature description is a process to generate feature vector from each detected interest point. The gradient-based features are more effective to deal with the problem of various object deformation, viewpoints, illumination, occlusion, and blur resolution [42]. MSER detector however was not built with its own descriptor. The empirical study conducted in [40] to evaluate the descriptors for MSER have suggested that the speeded-up robust feature (SURF) descriptor is very close to real-time applications. This is because SURF used integral image and Haar wavelet to approximate gradient information and very minimum of noises were generated. SURF has also scored very well in terms of repeatability, distinctiveness, robustness, detection error as well as geometric and photometric deformation [43]. Basically, the SURF describes the intensity content surrounding the interest point neighborhood. In the first place, the orientation of each feature is identified via pixel convolution in its neighborhood together with the horizontal and vertical Haar wavelet filter. The Haar wavelet filters can be illustrated as a block to calculate the directional derivatives of the image's intensity. The features can be described regardless of their orientation via the intensity changes to characterize the orientation.

5.3. Fuzzy feature encoding

A huge and diverse feature have been generated in feature description stage. At this point, the feature descriptions can be represented as from N dimensional features from an image. For instance, hundreds or even thousands of interest points were generated per image and the amount of interest points for all images may reach up to hundreds of thousands of interest points. Soft assignment technique encodes the feature descriptions by assigning them into several visual words and the response on each visual word is calculated by using kernel function of the distance between feature descriptions with visual word. Initially, visual words *v* are generated to define the feature descriptions *X*:

$$v(i) = \frac{esp(\|x-b_i\|_2^2/\sigma)}{\sum_{k=1}^{K} esp(\|x-b_i\|_2^2/\sigma)} = 1,2,\ldots\ldots,M$$

Where $\sum_{k=1}^{K} \mathrm{esp}(\|x - b_i\|_2^2/\sigma)$ is the normalization factor, σ is smooth parameter and M is the vocabulary size. Specifically, fuzzy c-means (FCM) is used to encode the features. FCM is an extension of k-means where objective function is modified by incorporating fuzzier parameter. FCM assigned each SURF descriptions into all visual words where different degrees of membership values are computed for each assignation. The membership value is determined using a Gaussian Probability Density Function based on the distance between interest points and the visual words. Let $S = \{ s_1, s_2, s_3, s_4 \ldots s_n\}$ be the set of SURF descriptors, where $\widetilde{\boldsymbol{A}}$ represents the fuzzy set in *S* and $\mu_{\tilde{A}}(s_i)$ is the fuzzy membership function. The $\widetilde{\boldsymbol{A}}$ can be computed as follows:

$$\tilde{A} = \mu_{\tilde{A}}(s_1)/ s_1 + \mu_{\tilde{A}}(s_2)/s_2 + \ldots . \mu_{\tilde{A}}(s_n)/s_n = \sum_{i=1}^{n} \mu_{\tilde{A}}(s_i)/s_i$$

Where *n* is the number of SURF descriptors in *S*, $\mu_{\tilde{A}}(s_i)/s_i$ represents the fuzzy membership value of s_i to $\tilde{A}$ and $\sum_{i=1}^{n}$ represents the relationship between SURF descriptor and the membership function. The membership function is initially started with the visual words construction by using k-means algorithm to produce list of visual words $W = \{w_1, w_2, .., w_{j,..} w_{m,}\}$. The distance $D_{i,m}$ between s_i and w_m can be represented as:

$$D = \{D_1, D_2, \dots D_n\}' = \begin{Bmatrix} d_{1,1} & d_{1,2} & d_{1,m} \\ d_{2,1} & d_{2,2} & d_{2,m} \\ d_{n,1} & d_{n,2} & d_{n,m} \end{Bmatrix}$$

Where n is the number of SURF descriptors extracted from an image g, and m is the number of visual words in W. The fuzzy set can be expressed as:

$$\tilde{A}_i = \sum_{j=1}^{n} \frac{\mu(d_{i,j})}{d_{i,j}}$$

The value $\mu(d_{i,j})$ determine the similarity between s_i and w_j based on the Gaussian membership function which transforms the distance set to fuzzy set which is calculated as:

$$\mu(d_{i,j}) = 1 - \frac{1}{\sqrt{2\pi\sigma_j}} exp\left(-\frac{(d_{i,j}-\theta_j)^2}{2\sigma_j^2}\right)$$

Where θ_j is the expected value and σ_j is the membership function for w_j. Both are parameters that are derived by using maximum likelihood estimation. Table 3 shows the algorithm for feature encoding by using FCM.

Table 3. Algorithm for fuzzy feature encoding

Feature encoding using fuzzy technique and feature summarization using maximum pooling
Input: $S_g = \{S_1, S_2, S_3 \dots S_n\} \leftarrow$ Set of SURF feature from an image, g.
Output: $F_g = \{F_1, F_2, F_3 \dots F_b\} \leftarrow$ Fuzzy visual dictionary
1. 1. Perform k-means by partitioning n features in S_g into k cluster represented by $= \{w_1, w_2, \dots, w_k\}$.
2. Generate fuzzy set $\tilde{A}_i$ by using equation 4.4 to measure the similarity between the features in S_g and visual words W.
3. $U_{g,w} \leftarrow$ Get membership value for each w in g from $\tilde{A}_i$.
4. $P_g = max(U_{g,1}, U_{g,2}, U_{g,3} \dots \dots U_{g,k}) \leftarrow$ Perform maximum pooling on $U_{g,w}$.
5. $F_g = \sum_{w=1}^{k} P_{kg} \leftarrow$ count occurrence frequency for each W

The fuzzy feature encoding algorithm shown in Table 3 has used k-means to generate visual words. Afterwards, the fuzzy set is calculated to measure the similarity between SURF descriptions and visual words represented by the membership value in each visual word. Finally, maximum pooling is applied where the highest membership value of visual word is assigned to the SURF description. Table 4 shows the algorithm for FCM.

Table 1. Algorithm for fuzzy c-means (FCM)

Fuzzy C-Means
Input: $s_g = \{s_1, s_2, s_3 \dots s_n\} \leftarrow$ Set of SURF feature Fuzzification parameter *m*. Maximum number of iterations *max_it.* Tolerance criterion ε
Output: Center matrix *C*; Membership matrix *U*.
1. $\tau \leftarrow 0$
2. Initialize randomly $U^{(\tau)}$
3. while $((\lvert U^{(\tau)} - U^{(\tau-1)} \rvert < \varepsilon \; V \; \tau)$<*max_it* update $C^{(\tau)}$ update $U^{(\tau)}$ $\tau \leftarrow \tau + 1$
4. end while
5. $c_j = (c_{j1}, c_2, \dots, c_{jd}) \leftarrow$ cluster centers
6. $U=[u_{ij}]_{i=1..n}^{j=1..k} \leftarrow$ Fuzzy assignment of feature vector x_i to clusters.

6. EXPERIMENTAL RESULTS

The performance of feature encoding techniques are measured based on classification rate. The training and testing strategy is based on five-folds where the dataset is divided into five training and testing set following the procedure in [30]. The final classification rate is taken based on its average.

6.1. Visual words and fuzzy membership value generation

We illustrate a demonstration of FCM in a sample image. The visualization of FCM mechanism is shown in Figure 9.

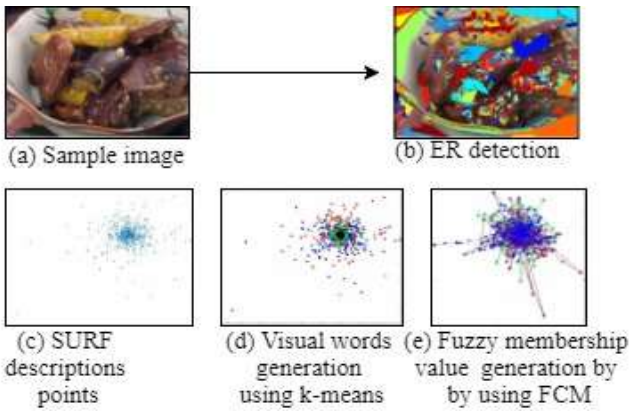

Figure 9. Fuzzy feature encoding by using FCM

As shown in Figure 9, initially the extremal regions (ER) of sample food image in (a) are detected by using MSER in (b) and the features are described by using SURF as plotted in (c). Then, the features are fed into k-means algorithm to obtain the visual words as shown in (d). This example sets the cluster size k as 3. In hard assignment, the process of assigning visual words to the SURF descriptions will proceed in (d) where the SURF descriptions are just assigned to the three visual words which have been marked in dotted green, blue and red color. However, the FCM has extended the evaluation by assigning each SURF descriptions into several visual words and perform the similarity checking between SURF descriptions as shown in (e). When a set of SURF descriptions are regarded as similar, they will be assigned to the nearest visual word.

6.2. Classification performance

The given chart in Figure 10 depicts the percentages of classification between hard assignment, soft assignment, and fisher vector on different features. In general, the soft assignment technique has obtained the best as well as the most consistent classification rate over different features. It is also apparent from the chart the inconsistencies of fisher vector and hard assignment technique classification rate across multiple features. It is found that the Fisher vector only performed well only with Dog-SIFT. The reason is perhaps SIFT descriptor generates 128 feature dimensions compared to SURF and HOG descriptor that produce shorter feature dimensions which are 64 and 32 respectively.

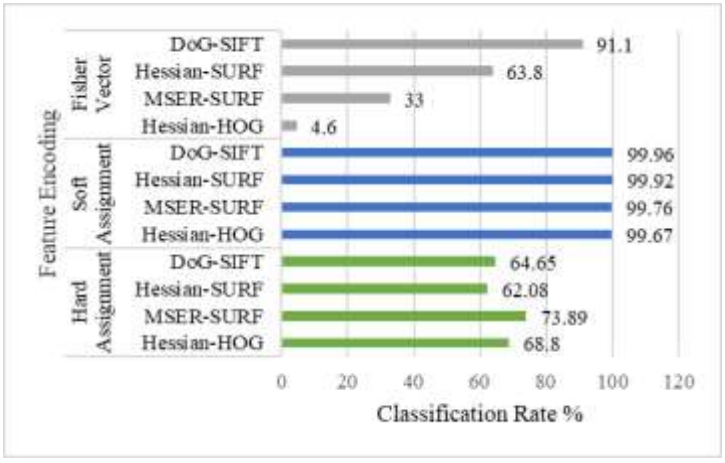

Figure 10. Performance of feature encoding techniques across different features

The superiority of Fisher vector that was reported in object recognition study was only evaluated by using SIFT descriptor [25]. Even the hard assignment encoding technique was better for certain features compared to Fisher vector.

6.3. Visual dictionary representation

The visual dictionary is a set of visual words containing the occurrence frequency or histogram of the visual words generated after pooling. Visual dictionary is the final representation of local feature Ixk before it can be fed into the classifier. The I denotes set of training and testing images, while k denotes vocabulary size respectively. This section provides the comparisons in terms of the visual dictionary representation pattern through the plotting of the graphs between soft assignment, hard assignment, and Fisher vector. To get a clear overview of the visual dictionary patterns, only ten visual dictionary of food categories were randomly picked for this study. Figure 11 provides the histogram of occurrence frequency of 500 visual words on ten food categories that were generated by using soft assignment.

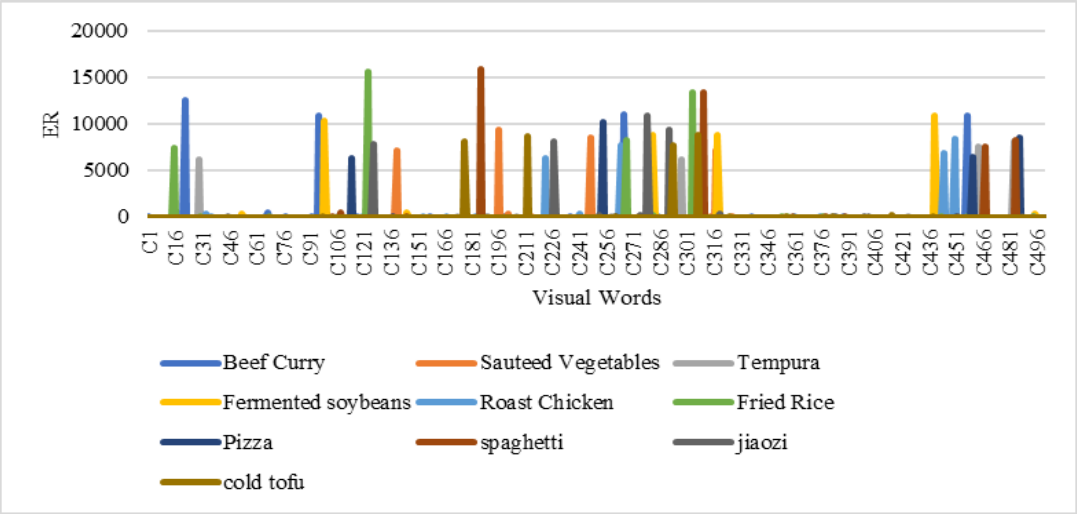

Figure 11. Visual words occurrence frequencies on the samples of food categories by using soft assignment

Based on Figure 11, it is worth noticing the high density of frequency in certain visual words. In soft assignment, the feature descriptions in a food category were obviously grouped into a certain number of visual words. A discriminative visual dictionary has been produced as the FCM algorithm has reduced the effect of uncertainty and plausibility problem in the visual words assignation performed using k-means. The FCM performed the evaluation on every interest points throughout the visual words by calculating the membership function for every single visual word assignation. The membership function value was calculated using a Gaussian Probability Density Function based on the distance between interest points and the visual words. Then, the maximum pooling chose the highest membership value as the most relevant visual word to the respective interest point. The results showed that the uncertainty and plausibility problems can be overcome where a consensus can be achieved when choosing the best visual word to be assigned to the interest points. Figure 12 shows the histogram of occurrence frequency of 500 visual words on ten food categories that were generated by using hard assignment.

As shown in Figure 14, the dots in the graph represent the ER of particular food images that have been assigned into the visual words ranging from 1 to 500. For example, the yellow dots in the graph represent the 394 feature descriptions produced by image '12929.jpg' that have been grouped into 185 visual words. Figure 15 shows the visual words distribution food samples by using fisher vector.

Based on Figure 15, the fisher vector has enriched the feature descriptions which then generates very long visual dictionary representation. The size of the visual dictionary generated by Fisher vector is highly influenced by the size of feature description. The size is calculated based on formula 2(D X K) where D represents the size of feature descriptions and K represents the cluster size. For instance, the SIFT descriptions with 128 D and 32 K will generate 8192 feature vector. The sparse visual dictionary representation with very long feature vector requires a lot of computational effort for classification. Figure 16 represents the visual word assignation distribution on five samples of food images in a beef curry food category by using soft assignment.

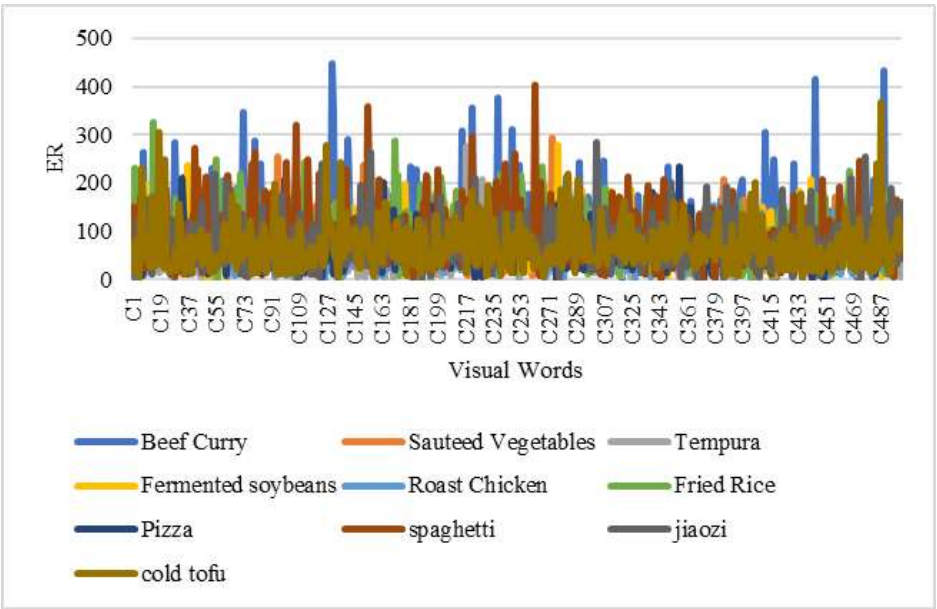

Figure 12. Visual words occurrence frequencies on the samples of food categories using hard assignment

Figure 13. Samples of beef curry food category

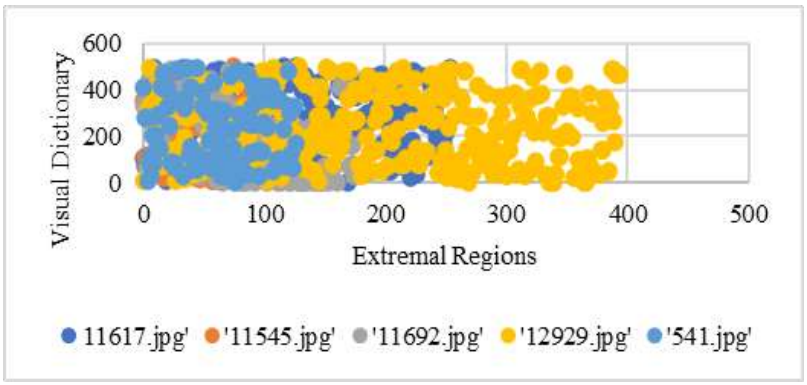

Figure 14. Visual words distribution of feature descriptions on food samples by using hard assignment

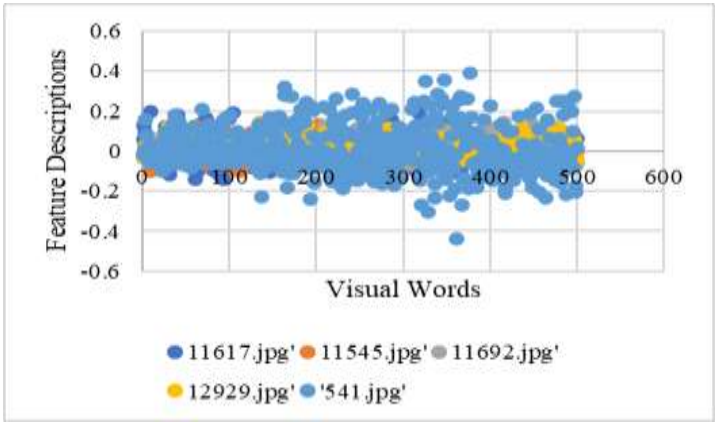

Figure 15. Visual words distribution of feature descriptions on food samples by using fisher vector

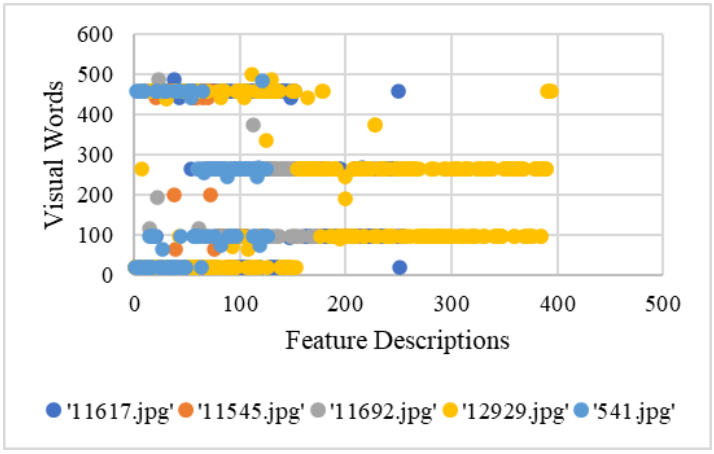

Figure 16. Distribution of the visual word of extremal regions on food samples in by using soft assignment

From the graph in Figure 16, a large proportions of feature descriptions were assigned to visual words 457, 266, 96, and 21. The yellow dots that represent the feature descriptions from image 12929.jpg only occupied 15 visual words compared to 185 when using hard assignment. Therefore, a discriminative visual dictionary can be produced by using soft assignment where the images in a food category are only assigned to few visual words. The classification can also be performed faster.

6.4. Evaluation of machine learning classifier and vocabulary size evaluation

Figure 17 shows the classification rate of Fisher vector, soft assignment, and hard assignment on different machine learning classifiers.

Based on Figure 17, the soft assignment technique was not only providing a good classification performance on Support Vector Machine that is synonym in image classification, but was also robust towards the other classifiers such as Naïve Bayes, K-Nearest Neighbor, and decision tree, while both fisher vector and hard assignment only performed well on Support Vector Machine. The soft assignment technique by using FCM required the cluster size or vocabulary size of K to be pre-defined. In previous experiments, the K is initially set to 500 to evaluate the performance of feature encoding technique. Therefore, the evaluation using different vocabulary sizes of 100, 1000, and 1500 on soft assignment were conducted and the result is shown in Figure 18.

From the results presented in Figure 18, the soft assignment may still sustain a good classification performance even by using even small vocabulary size. As mentioned previously, even the Fisher vector encoding technique uses small vocabulary size, but it generates enormous feature dimensions. This is different with soft assignment where the vocabulary size is equivalent to the feature dimensions. As larger vocabulary size is used, the classification rate of soft assignment has become flawless as more discriminative visual dictionary is produced.

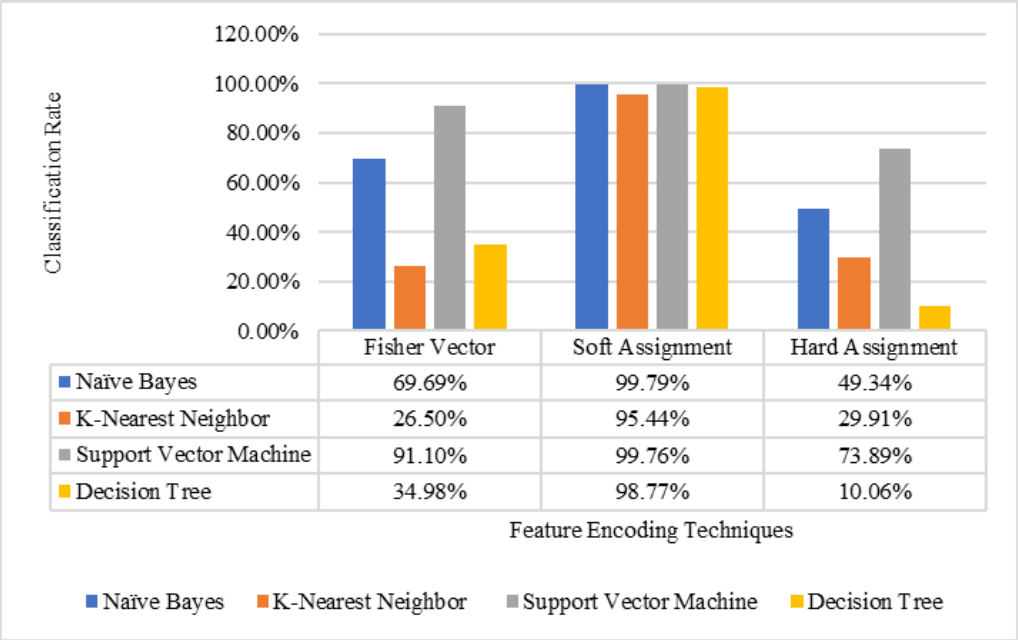

Figure 17. Classification rate of feature encoding techniques on various classifier

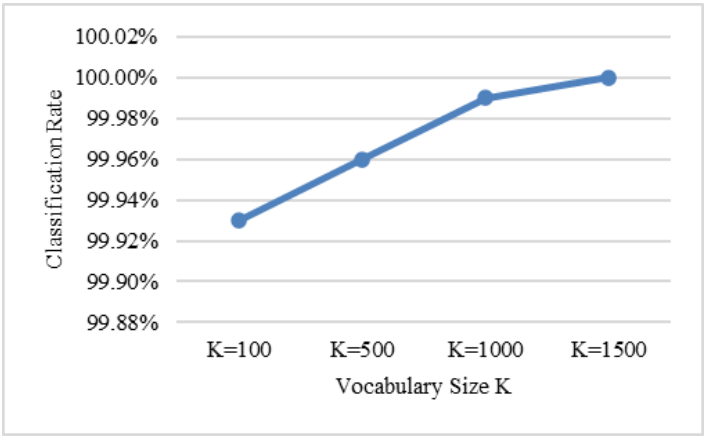

Figure 18. Evaluation of vocabulary sizes in soft assignment

7. CONCLUSION

This study has improved the performance of food recognition model by proposing soft assignment based on fuzzy encoding technique with maximum and sum pooling in building the visual dictionary. Specifically, the fuzzification in building the visual dictionary has reduced the errors while assigning feature descriptions to visual words by eliminating the problem of uncertainty and plausibility that exist in the large variations of colour and texture of foods. Fuzzy encoding compute fuzzy membership function to measure the similarity between the feature descriptions and visual words to assign them into relevant visual words. Another remarkable finding in this study is the level of robustness of the fuzzy feature encoding toward variety types of local features and machine learning classifier. A compact visual dictionary produced as a small size of visual vocabulary is good enough to provide informative and discriminative feature representation. Future research should investigate a mechanism in feature description-visual words assignation to avoid one-to-all assignation to improve the time efficiency in soft assignment. A selection procedure may be conducted to rank the visual words to decrease the number of visual words to be assigned to feature descriptions.

REFERENCES

[1] F. Kong, H. He, H. A. Raynor, and J. Tan, "DietCam: Multi-view regular shape food recognition with a camera phone," *Pervasive Mob. Comput.*, vol. 19, no. c, pp. 108-121, 2015.

[2] Z. Jie, W. F. Lu, S. Sakhavi, Y. Wei, E. H. F. Tay, and S. Yan, "Object Proposal Generation with Fully Convolutional Networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 8215, no. c, pp. 1-1, 2016.

[3] WHO, "Obesity and Overweight," 2018. [Online]. Available: http://www.who.int/en/news-room/fact-sheets/detail/obesity-and-overweight.

[4] M. Anthimopoulos, J. Dehais, P. Diem, and S. Mougiakakou, "Segmentation and recognition of multi-food meal images for carbohydrate counting," *13th IEEE Int. Conf. Bioinforma. Bioeng. IEEE BIBE 2013*, pp. 1-4, 2013.

[5] F. Zhu, M. Bosch, N. Khanna, C. J. Boushey, and E. 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.

[6] J. O. Pinzón-Arenas, R. Jiménez-Moreno, and C. G. Pachón-Suescún, "ResSeg: Residual encoder-decoder convolutional neural network for food segmentation," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 1, pp. 1017-1026, 2020.

[7] 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.

[8] H. Pooja and P. S. A. Madival, "Food Recognition and Calorie Extraction using Bag-of- SURF and Spatial Pyramid Matching Methods," *Int. J. Comput. Sci. Mob. Comput.*, vol. 5, no. 5, pp. 387-393, 2016.

[9] H. Chougrad, H. Zouaki, and O. Alheyane, "Soft assignment vs hard assignment coding for bag of visual words," *2015 10th Int. Conf. Intell. Syst. Theor. Appl. SITA 2015*, 2015.

[10] H. Hassannejad, G. Matrella, P. Ciampolini, I. De Munari, M. Mordonini, and S. Cagnoni, "A new approach to image-based estimation of food volume," *Algorithms*, vol. 10, no. 2, 2017.

[11] N. Martinel, C. Piciarelli, C. Micheloni, and G. L. Foresti, "A Structured Committee for Food Recognition," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2016-Febru, pp. 484-492, 2016.

[12] P. Pouladzadeh, S. Shirmohammadi, A. Bakirov, A. Bulut, and A. Yassine, "Cloud-based SVM for food categorization," *Multimed. Tools Appl.*, pp. 5243-5260, 2015.

[13] J. C. Van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271-1283, 2010.

[14] Q. Qiu, Q. Cao, and M. Adachi, "Filtering out background features from BoF representation by generating fuzzy signatures," in *International Conference on Fuzzy Theory and Its Applications (iFUZZY2014)*, pp. 14-18, 2014.

[15] G. Csurka and F. Perronnin, "Fisher Vectors : Beyond Bag-of-Visual-Words Image Representations," pp. 28-42, 2011.

[16] 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.

[17] H. Wang and W. Deng, "Face Recognition via Compact Fisher Vector," in Chinese Conference on Biometric *Recognition*, 2015, no. 10, pp. 68-7.

[18] 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.

[19] V. Garg, S. Vempati, and C. V. Jawahar, "Bag of visual words: A soft clustering based exposition," *Proc.-2011 3rd Natl. Conf. Comput. Vision, Pattern Recognition, Image Process. Graph. NCVPRIPG 2011*, pp. 37-40, 2011.

[20] L. Liu, L. Wang, and X. Liu, "In defense of soft-assignment coding," *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 2486-2493, 2011.

[21] N. I. W. Warfield, S. K, N. I. Weisenfeld, and S. K. Warfield, "Kernel Codebooks for Scene Categorization," *Eur. Conf. Comput. Vis.*, pp. 696-709, 2008.

[22] D. Dell'Agnello, G. Carneiro, T.-J. Chin, G. Castellano, and A. M. Fanelli, "Fuzzy clustering based encoding for visual object classification," *Proc. 2013 IFSA World Congr.-NAFIPS Annu. Meet.*, pp. 1439-1444, 2013.

[23] T. Ren, Z. Qiu, Y. Liu, T. Yu, and J. Bei, "Soft-assigned bag of features for object tracking," *Multimed. Syst.*, vol. 21, no. 2, pp. 189-205, 2014.

[24] U. L. Altintakan and A. Yazici, "A novel fuzzy feature encoding approach for image classification," *2016 IEEE Int. Conf. Fuzzy Syst.*, pp. 1134-1139, 2016.

[25] Y. Huang, Z. Wu, L. Wang, and T. Tan, "Feature coding in image classification: A comprehensive study," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 3, pp. 493-506, 2014.

[26] S. Ghosh and S. K. S. Dubey, "Comparative analysis of k-means and fuzzy c-means algorithms," *Int. J. Adv. Comput. Sci. Appl.*, vol. 4, no. 4, pp. 35-38, 2013.

[27] U. L. Altintakan and A. Yazici, "An improved BOW approach using fuzzy feature encoding and visual-word weighting," *IEEE Int. Conf. Fuzzy Syst.*, vol. 2015-Novem, no. 114, 2015.

[28] D. Thanh Nguyen, Z. Zong, P. O.Ogunbona, Y. Probst, and W. Li, "Food image classification using local appearance and global structural information," *Neurocomputing*, vol. 140, pp. 242-251, 2014.

[29] R. S. Abbirami, A. Abhinaya, P. Kavivarthini, and T. Rupika, "Large Scale Learning for Food Image Classification," *J. Recent Innov. Trends Comput. Commun.*, vol. 3, no. 3, pp. 973-978, 2015.

[30] 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.

[31] M. Wazumi, X.-H. Han, and Y.-W. Chen, "Food recognition using Codebook-based model with sparse-coding," *Proc. 2013 IEEE/SICE Int. Symp. Syst. Integr.*, pp. 482-485, 2013.

[32] L. a. Zadeh, "Fuzzy sets," *Inf. Control*, vol. 8, no. 3, pp. 338-353, 1965.

[33] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Comput. Geosci.*, vol. 10, no. 2-3, pp. 191-203, 1984.
[34] R. Krishnapuram and J. M. Keller, "A Possibilistic Approach to Clustering," *IEEE Trans. Fuzzy Syst.*, vol. 1, no. 2, pp. 98–110, 1993.
[35] D. Park, "Image Data Classification Using Fuzzy c-Means Algorithm with Different Distance Measures," in *International Symposium on Neural Networks*, pp. 489–496, 2013.
[36] Y. Shinomiya and Y. Hoshino, "Bag of Features Based on Feature Distribution Using Fuzzy C-Means," International Conference on Human-Computer Interaction, HCI International 2014-Posters' Extended, pp. 546-550, 2014.
[37] Y. Matsuda, H. Hoashi and K. Yanai, "Recognition of Multiple-Food Images by Detecting Candidate Regions," *2012 IEEE International Conference on Multimedia and Expo*, Melbourne, VIC, pp. 25-30, 2012. doi: 10.1109/ICME.2012.157.
[38] W. X. Liu, J. Hou, and H. R. Karimi, "Research on vocabulary sizes and codebook universality," *Abstr. Appl. Anal.*, vol. 2014, 2014.
[39] E. Salahat and M. Qasaimeh, "Recent Advances in Features Extraction and Description Algorithms : A Comprehensive Survey," in *IEEE International Conference on Industrial Technology (ICIT)*, 2017.
[40] M. H. Lee and I. K. Park, "Performance evaluation of local descriptors for maximally stable extremal regions," *J. Vis. Commun. Image Represent.*, vol. 47, pp. 62-72, 2017.
[41] J. Zheng, Z. J. Wang, and C. Zhu, "Food Image Recognition via Superpixel Based Low-Level and Mid-Level Distance Coding for Smart Home Applications," *Sustainability*, vol. 9, no. 5, 2017.
[42] Y. Li, S. Wang, Q. Tian, and X. Ding, "Feature representation for statistical-learning-based object detection: A review," *Pattern Recognit.*, vol. 48, no. 11, pp. 3542-3559, 2015.
[43] S. Jabeen, Z. Mehmood, T. Mahmood, T. Saba, A. Rehman, and M. T. Mahmood, "An effective content-based image retrieval technique for image visuals representation based on the bag-of-visual-words model," *PLoS One*, pp. 1–24, 2018.

BIOGRAPHIES OF AUTHORS

Mohd Norhisham Razali received his Bachelor and Master's degree in Information Technology from Universiti Teknologi MARA. He is currently working as lecturer at Faculty of Computing and Informatics, Universiti Malaysia Sabah since 2009. In 2019, he has obtained PHD degree in Intelligent System at Universiti Putra Malaysia. He is doing research in image processing and machine learning that focuses on improving a recognition algorithm to detect the category of foods. His work addresses the challenges related to food images feature representation to bridge the low-level features into higher level features. He is also has investigated intelligent techniques to optimize the feature representation process for more effective image classification. Currently, he is doing research in vision-based social robotics under Intelligent Robotics Research Group in Faculty of Computing and Informatics UMS. He has authored and co-authored a number of journals, conference paper and attended numerous national and international conferences and workshops.

Noridayu Manshor received the Ph.D degree in 2014 from Universiti Sains Malaysia (USM).Currently, she is a lecturer at the Faculty of Computer Science and Information Technology, Universiti Putra Malaysia (UPM). Her main research includes image processing, computer vision and pattern recognition.

Alfian Abdul Halin received the B.Sc. degree in information technology from Universiti Teknologi MARA, Malaysia, in 2000, the master of multimedia computing degree from Monash University, Melbourne, VIC, Australia, in 2004, and the Ph.D. degree in computer science from Universiti Sains Malaysia, Malaysia, in 2011. He joined the Faculty of Computer Science and Information Technology (FSKTM), Universiti Putra Malaysia (UPM) in 2001 as a tutor. Upon completion of his PhD, he was promoted to Senior Lecturer in 2011. Dr. Alfian's research interests include the application of image/audio processing and machine learning techniques to solve real world problems. He holds memberships in local entities such as the Digital Information and Computation Research Group, the Road Safety Research Center (UPM) and the Intelligent Computing Research Group (FSKTM).

Norwati Mustapha is an Associate Professor at the Department of Computer Science. Currently, she is the Deputy Dean of Academic, Student Affairs and Alumni at the Faculty of Computer Science and Information Technology, University Putra Malaysia (UPM). She received the BSc in Computer Science from University Putra Malaysia in 1991 and MSc degree in Information Systems from University of Leeds UK in 1995. She obtained her PhD in Artificial Intelligence from University Putra Malaysia in 2005. She is an active researcher in the area of Data Mining, Web Mining, Social Networks and Intelligent Computing.

Assoc. Prof. Razali Yaakob received the Bachelor Degree in Computer Science in 1996 and Master in Computer Science from in 1999, from Universiti Putra Malaysia, and PhD from University of Nottingham, United Kingdom in 2008. Currently, he is a lecturer at the Faculty of Computer Science and IT, Universiti Putra Malaysia. His research areas include artificial neural network, pattern recognition, and evolutionary computation in game playing. He is a member of the Intelligent Computing Group at the faculty.
Orcid id: 0000-0002-8228-5753