2010

On the combination of local texture and global structure for food classification

Zhimin Zong
University of Wollongong

Duc Thanh Nguyen
University of Wollongong, dtn156@uow.edu.au

Philip O. Ogunbona
University of Wollongong, philipo@uow.edu.au

Wanqing Li
University of Wollongong, wanqing@uow.edu.au

Publication Details

Zong, z., Nguyen, D., Ogunbona, P. & Li, W. (2010). On the combination of local texture and global structure for food classification. IEEE International Symposium on Multimedia, ISM 2010 (pp. 204-211). Piscataway, New Jersey, USA: IEEE.

On the combination of local texture and global structure for food classification

Abstract
This paper proposes a food image classification method using local textural patterns and their global structure to describe the food image. In this paper, a visual codebook of local textural patterns is created by employing Scale Invariant Feature Transformation (SIFT) interest point detector with the Local Binary Pattern (LBP) feature. In addition to describing the food image using local texture, the global structure of the food object is represented as the spatial distribution of the local textural structures and encoded using shape context. We evaluated the proposed method on the Pittsburgh Fast-Food Image (PFI) dataset. Experimental results showed that the proposed method could obtain better performance than the baseline experiment on the PFI dataset.

Keywords
combination, local, texture, global, structure, for, food, classification

Disciplines
Physical Sciences and Mathematics

Publication Details
Zong, z., Nguyen, D., Ogunbona, P. & Li, W. (2010). On the combination of local texture and global structure for food classification. IEEE International Symposium on Multimedia, ISM 2010 (pp. 204-211). Piscataway, New Jersey, USA: IEEE.

2010 IEEE International Symposium on Multimedia

On the Combination of Local Texture and Global Structure for Food Classification

Zhimin Zong, Duc Thanh Nguyen, Philip Ogunbona, and Wanqing Li
Advanced Multimedia Research Lab, ICT Research Institute
School of Computer Science and Software Engineering
University of Wollongong, Australia
Email: {zmz225, dtn156, philipo, wanqing}@uow.edu.au

***Abstract*—This paper proposes a food image classification method using local textural patterns and their global structure to describe the food image. In this paper, a visual codebook of local textural patterns is created by employing Scale Invariant Feature Transformation (SIFT) interest point detector with the Local Binary Pattern (LBP) feature. In addition to describing the food image using local texture, the global structure of the food object is represented as the spatial distribution of the local textural structures and encoded using shape context. We evaluated the proposed method on the Pittsburgh Fast-Food Image (PFI) dataset. Experimental results showed that the proposed method could obtain better performance than the baseline experiment on the PFI dataset.**

***Keywords*-Local binary pattern; shape context; food classification**

I. Introduction

There is increased attention being paid to the nutritional value of the food intake of human populations around the world. This is largely due to the high cost of providing health care and the drain on national economy resulting from diseases caused by unhealthy nutrition. Furthermore, there is a need for data collection in order to measure nutritional value of dietary and supplement intake in health studies and treatment. Accurate and passive acquisition of dietary data from human populations is essential for a better understanding of the etiology and the development of effective health management programs.

Conventionally, this task has been conducted manually through self-reports. Despite widespread use of questionnaires and structured interviews, numerous studies have revealed that data obtained by self reporting seriously underestimates food intake, and thus do not accurately reflect the habitual behaviour of individuals in real life [1], [2]. In addition, the practicality of manually acquiring large datasets is diminished because of the labour intensity of the exercise and further compounded by the inaccuracy generally associated with this mode of data collection.

The above-mentioned considerations have motivated the use of computer vision and image processing approaches to improve the accuracy of food intake reporting by developing automated food recognition systems [3]. Such systems are beneficial to the researchers but also provide a convenient way for people to estimate the nutritional and calorific value of their food intake. The main aim of this paper is to develop a robust food image classification method. The proposed classification algorithm is inspired by the Bag-of-Word (BoW) approach originated for text classification [4]. However, in this paper, the spatial information of the codewords is considered and represented using shape context [5]. In addition, we employ the local binary pattern (LBP) originally developed for texture classification [6] to encode the local textural structures (codewords) of the food images.

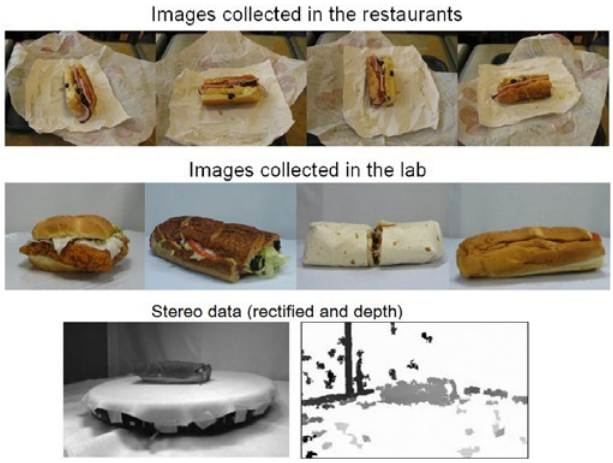

Figure 1. Some examples of the PFI dataset [7].

The classification accuracy of the proposed method was evaluated on the Pittsburgh Fast-Food Image (PFI) dataset [7]. The generation of this dataset was a collaborative effort of Intel Labs Pittsburgh, Carnegie Mellon University and Columbia University and was published in late 2009. Indeed, it is considered as the first mature dataset in food domain with a huge collection of visual data to facilitate research in automated food recognition. The dataset contains three instances of 101 foods from 11 popular fast food chains; with images and videos captured in both restaurant conditions and controlled lab setting. Some examples of the PFI dataset are shown in Fig. 1.

The rest of the paper is organized as follows. Section II reviews the related work in this area. Sections III and IV respectively present the basic concepts of LBP and shape context descriptor, which are the key techniques we used

978-0-7695-4217-1/10 $26.00 © 2010 IEEE
DOI 10.1109/ISM.2010.37

to analyse local texture feature and spatial information. Section V presents the proposed food image classification method. Experimental results along with comparative analysis are presented in Section VI. Section VII concludes the paper and discusses further work.

II. Related Work

As presented in the introductory section, automatic food recognition is a contemporary research topic. It would appear that available object recognition and classification techniques with minor modifications can be applied to food recognition and classification. However, it does not seem that the problem of food recognition is simply a test case of object recognition. The results obtained in a number of food recognition and classification methods proposed in the literature provide evidence. Food objects with possible variations in appearance (color, texture, shape) and viewpoints make the problem of food recognition and classification challenging in practice. Therefore, in this section, rather than review object recognition and classification methods, we limit the literature review to food recognition and classification techniques.

Generally speaking, food recognition and classification methods can be categorized based on food image representation using color, texture, or shape features. For color information, color histogram is commonly used to describe the food image. For example, Chen et al. [7] employed a 64-bin RGB color histogram (i.e. $4 \times 4 \times 4$ bins for three components Red, Greed, Blue) to represent the food image. Each pixel in the food image was then mapped to its closest bin in the histogram to generate a 64 dimensional image representation for that food image. The 64-dimensional feature vectors of all training food images were finally used to train a SVM classifier for food recognition.

For texture information, Gabor texture features of local regions of 3×3 and 4×4 with several scales and orientations were employed in the work of Joutou and Yanai [8]. In this paper, 24 Gabor filters with four scales and six orientations were used. The 24 Gabor filters were applied on each local region and the average filter responses within the block were computed to obtain a 24-dimensional texture feature vector for each block. Similar to the color histogram, the 24-dimensional vectors of all blocks were concatenated to create a richer and higher dimensional feature vector for each food image.

An example of the the application of shape information is the work of Pishva et al. [9] in which, based on the observation of the bread image, they analyzed the size (counted as the area) and shape (represented by the ratio of the difference between the major and minor axes to their sum) of the bread objects for bread classification. However, this method requires the knowledge of the size and shape of the food object and may not be generally applicable. For example in a multi-food classification and recognition task the food objects may occlude each other and in general their shape may not be persistent.

To exploit the advantages of each type of features, some methods have combined different types of image features in defining the feature vectors. For example, in [8] color histogram, Gabor texture features, and bag of SIFT features (which will be presented later) were employed to train a multiple kernel learning (MKL) SVM classifier (a sub-kernel was assigned to each type of image features). The optimal combined kernel was obtained by estimating the weights by the MKL method. We note that in [9], in addition to the use of bread's shape, color and texture were also applied to enhance the recognition performance.

The methods we have described so far represent the entire food image by feature vectors and would be referred to as global approach. On the other hand, local methods describe the food image by its components. Compared with the global approach, the local approach could provide a more compact and discriminative description and representation of food objects.

One issue with local methods is how to identify the components of a food object image. Another concern is that these components need to be invariant under the transformations and various viewpoints that are normally encountered in the practice of food recognition and classification. SIFT (Scale Invariant Feature Transform) interest point detector introduced by David Lowe [10], [11] has been considered as an appropriate solution for this purpose due to the following reasons. First, the interest points detected using SIFT are local features with high informative content. Second, they are stable under local and global perturbations in the image domain. In particular, SIFT feature is invariant to image scale and rotation, and has been shown to provide robust matching across a substantial range of affine distortion, change in 3D viewpoint, addition of noise, and change in illumination.

An example of using SIFT feature for food recognition in video is the work of Wu and Yang [1]. In this work, SIFT was used to extract the interest points and their local regions which were considered as the components of the food object. The recognition proceeded frame-by-frame by matching individual SIFT features from a newly acquired food image to a database of pre-trained features similar to matching keypoints SIFT descriptors [11]. The performance was then improved by employing different matching schemes.

Since food object can appear in images with complicated and cluttered backgrounds, some SIFT features will represent the background regions and thus are less discriminative. To reduce those irrelevant SIFT features while maintaining important features, the Bag-of-Features (BoF) approach has been used in [8], [7]. This approach is inspired by the Bag-of-Words (BoW) approach originated for text classification [4]. However, in the BoF approach the codewords are represented by the image features (SIFT feature in this case).

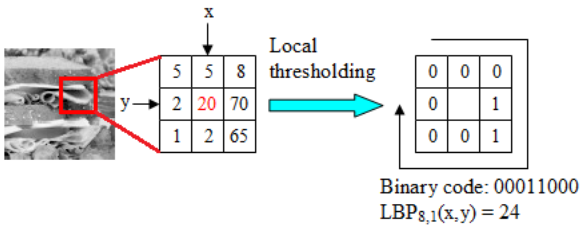

Figure 2. An illustration of the LBP descriptor.

Similar to the BoW, the basic idea of the BoF approach is to represent each image as a histogram of occurrence frequencies defined over a discrete vocabulary of features and then to use the histogram as a high-dimensional vector in a conventional discriminative framework, e.g. classifiers.

III. Local Binary Pattern (LBP)

Local binary pattern (LBP) is an effective technique to describe local texture feature [6] and has been successfully applied in a wide range of applications including texture classification [12], object recognition [13] and detection [14], [15], [16]. Fig. 2 represents an example of the LBP in which the LBP code of the center pixel (in red color and value 20) is obtained by comparing its intensity with neighbouring pixel intensities. The neighbour pixels whose intensities are equal or higher than the center pixel's are labeled as "1"; otherwise as "0".

We adopt the following notation. Given a pixel $c = (x_c, y_c)$, the value of the LBP code of c is defined as:

$$LBP_{P,R}(x_c, y_c) = \sum_{p=0}^{P-1} s(g_p - g_c)2^p \qquad (1)$$

where p is a neighbour pixel of c and the distance from p to c does not exceed R. g_p and g_c are the gray values (intensities) of p and c respectively. Furthermore,

$$s(x) = \begin{cases} 1, & \text{if } x \geq 0 \\ 0, & \text{otherwise.} \end{cases}$$

In (1), R is the radius of a circle centered at c and P is the number of sampled points. For example, in Fig. 2, R and P are 1 and 8 respectively. The following are important properties of the LBP descriptor.

Uniform and non-uniform LBP. Uniform LBP is defined as the LBP that has at most two bitwise transitions from 0 to 1 and vice versa in its circular binary representation. LBPs which are not uniform are called non-uniform LBPs. As indicated in [6], an important property of uniform LBPs is the fact that they often represent primitive structures of the texture while non-uniform LBPs usually correspond to unexpected noise structures and hence are less discriminative.

Scanning a given image in pixel-wise fashion, LBP codes are accumulated into a discrete histogram called LBP histogram. It is easy to see that the number of $LBP_{P,R}$ histogram bins is 2^P. However, the dimensionality of the LBP histogram can be reduced by casting all the non-uniform LBPs into one bin. Two given images can be compared by measuring the similarity between their two LBP histograms. A number of measures can be used for histogram comparison: χ^2-distance, $\mathcal{L}_p$ distance, Bhattacharyya distance, etc.

IV. Shape Context

Shape context as a descriptor, was proposed by Belongie et al. [5] and has been successfully used in various shape matching and object recognition tasks. Shape matching using shape context is a variant of the conventional Hausdorff matching which employs the Hausdorff distance to measure the similarity between two shapes.

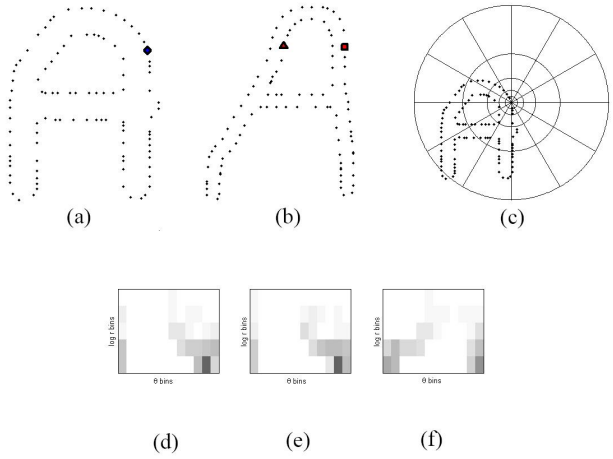

Figure 3. (a) and (b) are the sampled edge points of the two shapes. (c) is the diagram of the log-polar bins used to compute the shape context, 5 bins for $\log r$ and 12 bins for θ. (d),(e), and (f) are example shape contexts for reference samples marked by different symbols in (a) and (b). (d) is the shape context for the circle, (e) is that for the diamond, and (f) is that for the triangle. Notice that dark intensity represents high value of the histogram [5].

To extract the shape context at a point p, the vectors connecting p to all the other points are computed. For each point p, a histogram of the relative coordinates determined by the length r and orientation θ of the remaining points is created. To make the histogram more sensitive to nearby points (rather than farther away points), the histogram is represented in log-polar space. The shape context of each point is then created by flattening and concatenating all bins of the histogram. Fig. 3 shows an example of the shape contexts of two different variants of the letter 'A'. As can be seen in Fig. 3, the shape contexts of (d) and (e) are similar since they represent the two corresponding points.

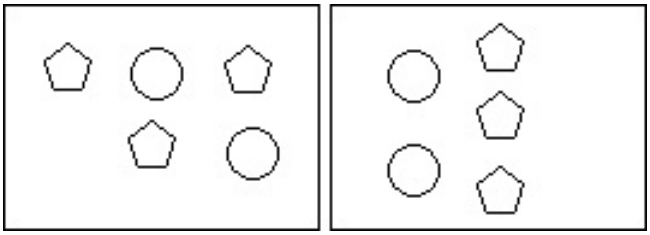

Figure 4. Left and right image represent two different objects with the same BoW histogram. Notice that the pentagons and circles represent the codewords.

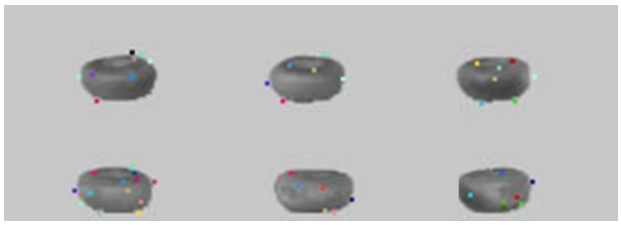

Figure 5. Some examples of using SIFT to locate the codewords (best viewed in color).

The matching of the shape contexts of two shapes is conducted by determining the correspondences between pairs of points on the two shapes so as to minimize the total cost of matching. A typical matching cost at a pair of points is computed using the χ^2-distance and the optimization can be performed by using the shortest augmenting path algorithm.

V. Proposed Method

The proposed method is motivated by the Bag-of-Words (BoW) approach [4]. In particular, we used the SIFT detector to locate the local structures and created a visual codebook of these local structures to describe the object of interest (food in this case). However, in contrast to the approach adopted in [7], [8], [1], (i.e. the use of SIFT descriptor), we employed the LBP to encode the local textural structures of the food images. The LBP has advantages in describing the objects for recognition and classification. It is robust under illumination changes, simple to compute, and discriminative in classification.

Another novelty introduced by the proposed method is the exploitation of the spatial relationship between the codewords. In [7], [8], the spatial information of the codewords, which may have special meaning in practice, were totally ignored. As shown in Fig. 4, the left and right images represent two different structures but the frequency of codewords (icons) in both structures are similar. In this paper, the spatial relationship between codewords represents the global structure of the food object described using shape context [5]. As indicated earlier, shape context is robust under deformations and transformations. The proposed method comprises two procedures, namely, training and classification and they will be presented in the following sections.

A. Training

The main aim of the training procedure is to create the visual codebooks of local textural structures and learn their spatial relationship using shape context. Assume that we have a number of images containing the food objects. For each training image, the SIFT detector [11] is invoked to detect the interest points. For each interest point, the LBP histogram with $P = 8$ and $R = 1$ of a local image patch centered at that interest point is then computed and normalized using the L_1 norm. Subsequently, the LBP histograms of all interest points of the training images are clustered using a K-means algorithm in which the similarity between two histograms is measured using the Bhattacharyya distance (see Eq. (6)). This step results in a codebook in which codewords are the image patches nearest to the cluster means. Fig. 5 shows some examples of using the SIFT to detect interest points; different codewords are represented by different colours.

Some codewords might not be sufficiently discriminative to represent a certain category of the food. Thus a codeword filtering step is initiated so that we can keep typical and important codewords to characterize all food categories. Specifically, for each category C_i and codeword w_{ij}, we compute the relative frequency of assignments of w_{ij} to C_i, i.e. $f(w_{ij}|C_i)$, and not to C_i, i.e. $f(w_{ij}|NonC_i)$. The histograms of frequencies, (respectively, for the assignments inside and outside the category), can be considered as the conditional distributions of the codeword given category label. A codeword w_{ij} is selected if the following condition is satisfied:

$$\log\left[\frac{f(w_{ij}|C_i)}{f(w_{ij}|NonC_i)}\right] \geq \phi \tag{2}$$

where ϕ is a predefined threshold and C_i is some food category. The code word filtering step yields a general codebook $\mathcal{W} = \{W_1, W_2, ..., W_N\}$ including N category codebooks corresponding to N food categories. Each category codebook $W_i, i \in \{1, .., N\}$ can contain different number of codewords $W_i = \{w_{i1}, w_{i2}, ...\}$.

In addition to generating the codebooks, the spatial relationship between codewords needs to be learned. This process can be performed as follows. For each category codebook W_i of the category C_i, we re-examine all food images belonging to that category (assume that all food images are labeled with the corresponding categories). Given a training image of the category C_i and its interest points generated using SIFT detector, considering every codeword $w_{ij} \in W_i$ as a point on a shape, the best matching interest point of w_{ij} and its location on that training image can be determined as in Eq. (4).

Similar to [5], the shape context at a codeword is created as the histogram of the relative coordinates from that code-

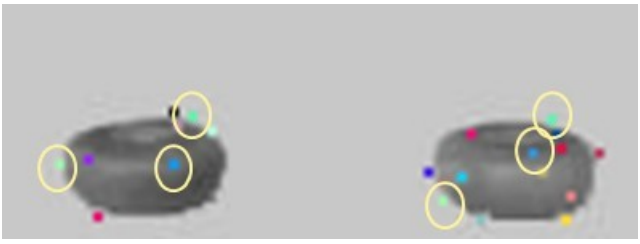

Figure 6. Different pictures of a donut, each has three visual codewords marked by a circle (best viewed in color).

word to the remaining codewords. Notice that these shape contexts are defined for one training image. Training the whole dataset of the images of the category C_i, the shape context at a codeword w_{ij} is then considered as the mean (histogram) of all shape contexts at that codeword calculated on all images of the category C_i. Since each codeword is represented by a histogram of the LBPs, the value of each bin of the mean histogram can be obtained by averaging the values of all LBP histograms at that bin. Note that not all codewords can be found in every training image; the shape contexts at those missing codewords can be simply filled with 0s for all bins.

Fig. 6 shows two different pictures of a donut. Each codeword is represented as a coloured point. As shown in Fig. 6, the relative spatial information of the codewords is useful for determining the correspondences between codewords. In order to make the spatial distribution of the codewords invariant under the relative changes of location of codewords, instead of computing the absolute value of the orientation θ between two codewords (compared with the x-axis), we calculate the relative orientation of the vector connecting a codeword to the center location of all codewords belonging to the same object category.

B. Food Classification

Let I denote the image containing an instance of the food object and $\mathcal{F} = \{C_1, C_2, \ldots, C_N\}$ be the set of food categories. The matching cost between I and food category C_i is denoted as $\mathcal{C}(I, C_i)$. The problem of food classification then becomes finding the best matching C_i^* so that,

$$C_i^* = \arg\min_{C_i \in \mathcal{F}} \mathcal{C}(I, C_i) \tag{3}$$

We now present how to compute $\mathcal{C}(I, C_i)$. Let W_i denote the category codebook of a food category C_i and $E = \{e_1, e_2, \ldots\}$ denote a set of the image patches located at interest points in I obtained using SIFT detector [11]. For each codeword $w_{ij} \in W_i$, its best matching image patch $e(w_{ij})$ can be determined as,

$$e(w_{ij}) = \arg\max_{e \in \hat{E}_{w_{ij}}} \rho(w_{ij}, e) \tag{4}$$

where $\hat{E}_{w_{ij}}$ is a set of image patches of I which have the best matching codeword w_{ij} (i.e. $\hat{E}_{w_{ij}} \subset E$). This is computed as,

$$\hat{E}_{w_{ij}} = \{e \in E | w_{ij} = \arg\max_{w_{il} \in W_i} \rho(w_{il}, e)\} \tag{5}$$

where $\rho(w_{il}, e)$ represents the similarity between the codeword w_{il} and image patch e and can be defined using the Bhattacharyya distance as,

$$\rho(w_{il}, e) = \sum_{b=1}^{B} \sqrt{w_{il}(b) e(b)}. \tag{6}$$

In Eq. (6), B is the number of LBP histogram bins; $w_{il}(b)$ and $e(b)$ are the values of the LBP histograms of w_{il} and e at the b-th bin.

The above step is done for all codewords w_{ij} of the food category C_i to result in a set of image patches $e(w_{ij})$. Considering the set of image patches $e(w_{ij})$ as the first shape and all codewords w_{ij} of the food category C_i as the second shape, the shape contexts of $e(w_{ij})$ and w_{ij}, denoted as $s_{e(w_{ij})}$ and $s_{w_{ij}}$, can be obtained as described in section V-A. Notice that these shape contexts are related only to the spatial information of $e(w_{ij})$ and w_{ij}. The cost of matching $e(w_{ij})$ and w_{ij} is defined as,

$$\Delta_{ij} = \delta(s_{e(w_{ij})}, s_{w_{ij}})[1 - \rho(w_{ij}, e(w_{ij}))] \tag{7}$$

where $\delta(s_{e(w_{ij})}, s_{w_{ij}})$ denotes the matching cost between the two shape contexts $s_{e(w_{ij})}$ and $s_{w_{ij}}$. Similar to [5], this matching cost can be computed as,

$$\delta(s_{e(w_{ij})}, s_{w_{ij}}) = \frac{1}{2} \sum_{k=1}^{|W_i|} \frac{[s_{e(w_{ij})}(k) - s_{w_{ij}}(k)]^2}{s_{e(w_{ij})}(k) + s_{w_{ij}}(k)} \tag{8}$$

where $|W_i|$ represents the number of codewords of the food category C_i.

Given the set of costs Δ_{ij} between all pairs of $e(w_{ij})$ and w_{ij}, the matching cost $\mathcal{C}(I, C_i)$ in Eq. (3) is defined as,

$$\mathcal{C}(I, C_i) = \sum_{j=1}^{|W_i|} \Delta_{ij} \tag{9}$$

By using shape context to represent the relative spatial relationship between codewords, the proposed method can accommodate deformations and transformations in the shape of food objects. In addition, the proposed method does not require all codewords to appear in a given food object image to be classified. Thus, it is robust even when the SIFT interest point detector does not work well, as may be the case in the challenging conditions such low illumination. Finally, by using multiple codebooks for different types of food, the proposed method can provide a compact and discriminative description for food objects.

VI. Experimental Results

The proposed method was evaluated on the PFI dataset created by Chen et al. in [7]. This dataset contains a total of 4,545 still images, 606 stereo pairs, 303 360-degree videos for structure from motion, and 27 privacy-preserving videos of eating events of volunteers. Some examples are shown in Fig. 1. In addition to providing a huge database of food images with labeled categories, two food classification baseline methods are also provided in [7]. The two baseline experiments can be summarized as follows.

Baseline 1: Colour Histogram + SVM Classifier. Colour histogram is employed in which a standard RGB 3-dimensional histogram is quantized with four levels per channel. Each pixel in the image is mapped to its closest cell in the histogram to generate a 64-dimensional representation for each image.

Baseline 2: Bag of SIFT Features + SVM Classifier. SIFT interest point detector [11] is employed to identify a sparse set of 'key points' or locations at which descriptors should be computed. These key points are localized both in scale and space. At each of the key points identified by SIFT, an image patch is extracted and its 128-dimensional SIFT feature is computed (similar to the work in [11]). These image patches are quantized into a 1000-word vocabulary using a K-means clustering algorithm. For an input image containing a food object, a 1000-dimensional feature vector is obtained by counting the frequency of all 1000 codewords. The feature vector is called 'bag of SIFT features' and then used to train a SVM classifier.

In the result of the baseline experiments (shown in Fig. 9), one can see that SIFT baseline often outperforms the color histogram (4 of 6 cases). Notice that SIFT baseline also gives the higher accuracy compared with the color histogram baseline even in the case of *salads*, where one would expect the colors to have significant advantage. It is probably due to how SIFT descriptor somehow represents the object's shape. However, it is interesting to see that even some categories that seem visually distinctive are recognized with lower accuracy than expected.

Based on the given labeled food categories, we separate the training and test data samples so that no instance of a food item appears in both the training and test set as in [7]. The performance of the proposed method is shown in Table. I. It can be seen that the *sandwich* category achieves the highest accuracy, while the *meat* category is of lowest accuracy. One possible reason might be that the *meat* category usually has smaller food items and their surfaces are less diverse, which leads to fewer interest points. Therefore, the shape contexts become less stable and discriminative.

Since this paper proposes the use of LBP to describe the local textural structures and shape context to represent the relative spatial distribution of those local structures, we verified the importance of both of these cues. In particular,

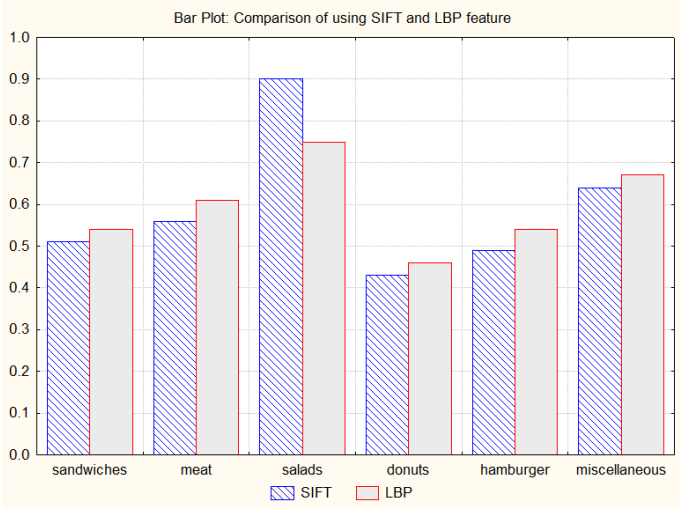

Figure 7. Comparison of using SIFT and LBP feature (best viewed in color).

to evaluate the role of LBP, we re-implemented the method proposed in the baseline 2 except that we replaced the SIFT feature by LBP feature. The results of this experiment are shown in Fig. 7. Generally speaking, LBP could slightly outperform SIFT when the overall performance is considered.

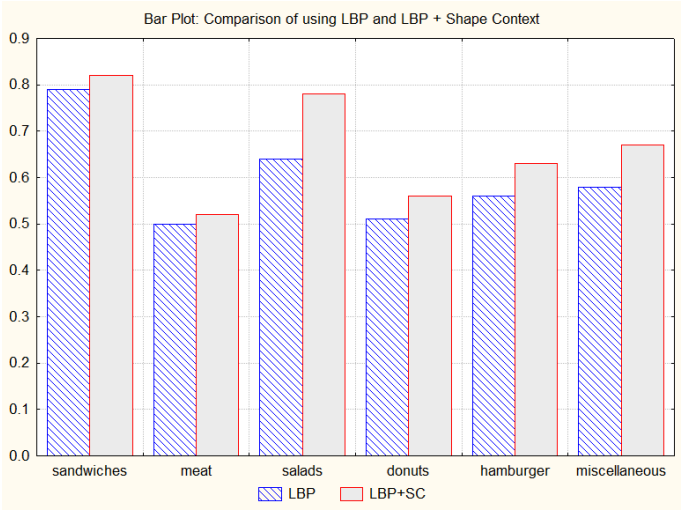

Figure 8. Comparison of using LBP and LBP + Shape Context (best viewed in color).

We also verified the importance of the spatial information of local structures represented using shape context. This experiment was conducted by employing only the LBP for generating the visual codebooks of all food categories and combining with shape context to describe the food object structure. The results are shown in Fig. 8. As can be seen in Fig. 8, the use of shape context could improve the classification performance.

Table I
CLASSIFICATION PERFORMANCE OF THE PROPOSED METHOD

Category	Accuracy
sandwiches, subs and wraps	0.82
meat	0.52
salads	0.78
donuts	0.56
hamburger	0.63
miscellaneous	0.67

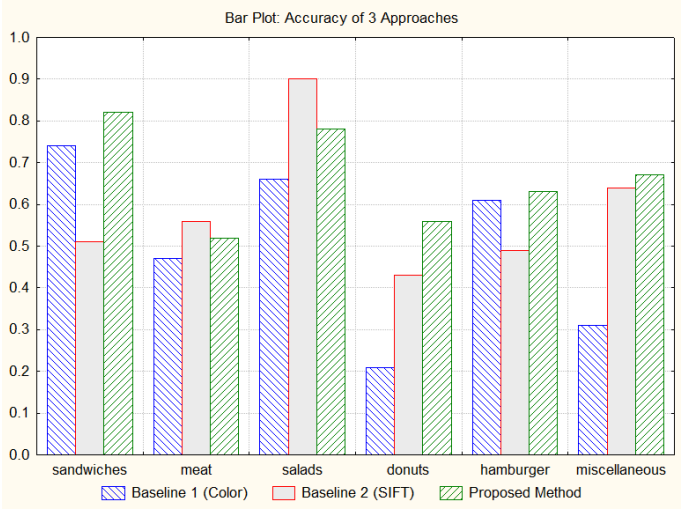

Figure 9. Comparison of the proposed method with the two baselines in [7] (best viewed in color).

In addition to evaluating the performance of the proposed method and its variants, we also compared the proposed algorithm with other state-of-the-art food classification methods. Particularly, the two baselines with the benchmarks in [7]: color histograms and bag of SIFT features were employed for this purpose. The comparison between the proposed method and these two baselines are shown in Fig. 9. Through experiment, we have found that, for the *salads* category, the accuracy of the proposed method was 78% which was 12% lower than that of the Bag of SIFT feature baseline (baseline 2), but 12% better than using color histogram (baseline 1). One possible reason is that *salads* category contains different salads made from different food; thus the texture feature may not outperform SIFT feature as the texture of each food component may vary. But it outperforms color-based method from the observed results. Similar situation is found in the *meat* category, i.e. the proposed method keeps the middle position compared with baseline 1 and 2. However, the difference between the performance figures was slight. In particular, for the *meat* category, the proposed method achieved 52% accuracy while the baseline 1 and 2 achieved 56% and 47% accuracies respectively. As described before, this might result from the stability of shape contexts of *meat* category. With the remaining food categories, the proposed method outperformed both of the baseline 1 and 2.

VII. CONCLUSION

This paper presents a food image classification method combining both local texture and global structure of the food object. In this paper, we employed SIFT interest point detector and local binary pattern (LBP) to locate and encode the local textural structures of the food object while shape context is used to represent the spatial constraints on the distribution of these local structures. The robustness of the proposed method was verified for food classification on the Pittsburgh Fast-Food Image (PFI) dataset. The proposed method was also compared with its variants and other state-of-the-art food classification methods. Since the food images can be captured from various viewpoints, developing view invariant texture feature would be our future work. The possibility of combining different type of features to improve the recognition performance is also being pursued.

REFERENCES

[1] W. Wu and J. Yang, "Fast food recognition from videos of eating for calorie estimation," in *Proc. IEEE International Conference on Multimedia and Expo (ICME'09)*, New York, USA, May 2009, pp. 1210–1213.

[2] N. Yao, R. J. Sclabassi, Q. Liu, J. Yang, J. D. Fernstrom, M. H. Fernstrom, and M. Sun, "A video processing approach to the study of obesity," in *Proc. IEEE International Conference on Multimedia and Expo (ICME'07)*, Beijing, China, Jul. 2007, pp. 1727–1730.

[3] L. Yang, J. Yang, N. Zheng, and H. Cheng, "Layered object categorization," in *Proc. IEEE International Conference on Pattern Recognition (ICPR'08)*, Florida, USA, Dec. 2008, pp. 545–576.

[4] K. Nigam, J. Lafferty, and A. Mccallum, "Using maximum entropy for text classification," in *Proc. Intl. Joint. Conf. on Aritificial Intelligence Workshop on Machine Learning for Information Filtering (IJCAI'99)*, Stockholm, Sweden, Jul. 1999, pp. 61–67.

[5] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002.

[6] T. Ojala, M. Pietikăinen, and D. Harwood, "Distinctive image features from scale-invariant keypoints," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, 1996.

[7] M. Chen, K. Dhingra, W. Wu, L. Yang, R. Sukthankar, and J. Yang, "PFID: Pittsburgh fast-food image dataset," in *Proc. IEEE International Conference on Image Processing (ICIP'09)*, Cairo, Egypt, Nov. 2009, pp. 289–292.

[8] T. Joutou and K. Yanai, "A food image recognition system with multiple kernel learning," in *Proc. IEEE International Conference on Image Processing (ICIP'09)*, Cairo, Egypt, Nov. 2009, pp. 285–288.

[9] D. Pishva, A. Kawai, and T. Shiino, "Shape based segmentation and color distribution analysis with application to bread recognition," in *Proc. IAPR Workshop on Machine Vision Applications (MVA'00)*, Tokyo, Japan, Nov. 2000, pp. 193–196.

[10] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE International Conference on Computer Vision (ICCV'99)*, Corfu, Greece, Sep. 1999, pp. 1150–1157.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[12] T. Ojala, M. Pietikăinen, and D. Harwood, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[13] T. Ojala, M. Pietikăinen, and D. Harwood, "View-based recognition of real-world textures," *Pattern Recognition*, vol. 37, no. 2, pp. 313–323, 2004.

[14] T. Ojala, M. Pietikăinen, and D. Harwood, "A texture-based method for modeling the background and detecting moving objects," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 657–662, 2006.

[15] Y. Mu, S. Yan, Y. Liu, T. Huang, and B. Zhou, "Discriminative local binary patterns for human detection in personal album," in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR'08)*, Anchorage, Alaska, Jun. 2008, pp. 1–8.

[16] D. T. Nguyen, Z. Zong, P. Ogunbona, and W. Li, "Object detection using non-redundant local binary patterns," in *Proc. IEEE International Conference on Image Processing (ICIP'10)*, Hong Kong, Sep. 2010.