Oceanic Scene Recognition Using Graph-of-Words (GoW)

Xinghui Dong*
The University of Manchester
Oxford Road, Manchester, M13 9PT, UK
xinghui.dong@manchester.ac.uk

Junyu Dong
Ocean University of China
Songling Road, Qingdao, 266071, China
dongjunyu@ouc.edu.cn

Abstract

In this paper, we focus on recognition of oceanic scene images. This task is particularly important for monitoring the oceanic environment with the cameras mounted at different locations. For this purpose, a new image dataset, namely, Flickr Oceanic Scene Dataset *(FOSD)* [1] *, is collected. Although it is intuitive to use this dataset to train a Convolutional Neural Network (CNN) from scratch, the relatively limited size of this dataset prevents us from doing so. One option is to encode the visual words learnt from deep convolutional features and it has been shown that these visual words outperform the fully-connected (FC) features extracted using a pre-trained CNN. However, it is commonly known that these word encoders generally do not utilise the spatial layout of words, whereas the spatial information is important to representation of long-range image characteristics. Thus, we propose a new image descriptor: Graph-of-Words (GoW), to capture the higher order spatial relationship between words, simply because graphs are able to encode the complicated spatial layout of node. This descriptor is also fused with three state-of-the-art word encoders to exploit richer characteristics. The GoW descriptor and the fused variants produce promising results in the oceanic and aerial scene recognition tasks. We attribute these results to the fact that the GoW descriptor encodes both the short-range and long-range higher-order spatial relationships between words.*

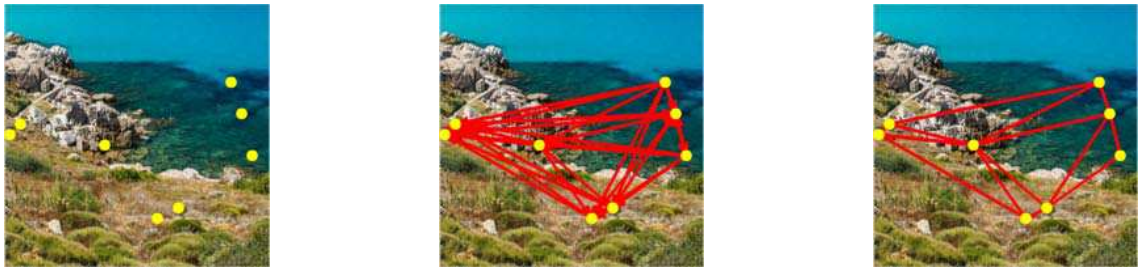
1. Introduction

Indoor and outdoor scene recognition [17], [23], [29], [32] has received much attention in the computer vision community as it involves the contexts for object recognition. However, none of previous studies focus on recognition of oceanic scenes. This task is key to monitoring the oceanic environment by analysing the images captured using a network of cameras that are fixed on autonomous underwater vehicles (AUV), docks, oil rigs, patrol aircrafts or ships, islands, etc. Although several scene datasets have been published such as 15 Scenes [17], Indoor Scenes [23], SUN [29] and Places205 [32], to our knowledge, there is no specific oceanic scene dataset available. Recently, online image albums have been used to source images [23], [29]. We are therefore inspired to collect 45 classes of oceanic scene images from Flickr [2] in a fashion of "in the wild". This dataset contains 4500 images and is titled "Flickr Oceanic Scene Dataset" or "FOSD". The overlapping between FOSD and other datasets is small even if some FOSD classes are shared by these datasets.

In recent years, CNN methods [16], [26], [32] have been extensively used in computer vision. As well-known, however, it is not feasible to train a CNN from scratch when only a relatively small dataset is available. Although the fully-connected features extracted from a pre-trained CNN are considered as generic, they can be outperformed by encoding the visual words [25] learnt from the convolutional features [5], [20], [30] computed using the same CNN. In general, visual word encoders are more generic than the features extracted by fine-tuning a CNN on a specific dataset, and these word encoders can further boost the performance of the fine-tuned model. Thus, in this study, we exploit the words learnt from the output of the final convolutional layer of a pre-trained CNN model: Places-CNN [32] which was trained using over seven million labelled scene images.

Yet it is also known that Bag-of-Words (BoW) [25] does not exploit the spatial relationship between words [11], [17] (see Figure 1 (a)). This is also the case for Fisher Vector (FV) [21] and Vector of Locally Aggregated Descriptors (VLAD) [13]. However, both local characteristics and global spatial layouts are important to image representation [11]. The former can be used to characterise the scenes with objects (e.g. offshore oil rigs in Figure 2) while the latter is more useful for representing the scenes containing long-range patterns (e.g. bays). In addition, Spatial Pyramid Matching (SPM) [17] and Pairs of Identical Words Angle Histogram (PIWAH) [14] were designed to utilise the spatial layout of words by exploiting the spatial partitioning of images and the 2nd-order orientation relationship between the locations of the same word (see Figure 1 (b)), respectively. In this context, none of the

[1]Available at: http://pan.baidu.com/s/1c2vn7JM



(a) (b) (c)

Figure 1. Three word encoders (only eight locations of the same word are shown). (a) BoW [25]: the number of the locations is collapsed into a bin of the BoW histogram; (b) PIWAH [14]: the angle computed between each pair of locations is quantised into an angle histogram; and (c) GoW: the distance computed along the shortest path (≥ two locations) connecting each pair of locations in a DT [8] graph is quantised into a distance histogram.

aforementioned descriptors use the higher order spatial relationship between words.

Therefore, we propose a new image descriptor based on Delaunay triangulation (DT) [8] in order to exploit the higher order spatial relationship between words. The characteristics of DT include that: (1) it is unique given a set of non-degenerate points; (2) it is locally stable; and (3) it can be efficiently built with the time complexity of $O(N\log N)$. The first two characteristics make DT be stable for encoding the spatial layout of points; while the third one guarantees the efficiency of the DT construction. To accelerate the computational speed, only the spatial layout of the locations of the same word is modelled. In this case, we build an individual DT graph for each word. The shortest path is used to represent the higher order spatial relationship between two nodes (see Figure 1 (c)). For each graph, the distances accumulated along each shortest path are quantised into a histogram, to form a sparse description of the graph. All histograms are concatenated into a single feature vector, which is referred to as "Graph-of-Words" or "GoW". The GoW descriptor is also fused with SPM [17], FV [21] and VLAD [13] in order to exploit richer image characteristics. These descriptors are tested in the scenarios of oceanic and aerial scene recognitions.

The contributions of this study are threefold. First, we propose a novel image recognition topic that focuses on oceanic scene images and accordingly collect a new dataset: FOSD. Second, we develop a new image descriptor, i.e. GoW, in order to encode the higher order spatial relationship between words. Third, we test this descriptor and its three fused variants against five baselines for oceanic scene recognition using three types of local features, which provides benchmarks for further research. The rest of this paper is organised as follows. Section 2 reviews the related work. The FOSD dataset is introduced in Section 3. The GoW descriptor and the fused variants are described in Section 4. The oceanic and aerial scene recognition experiments are reported in Sections 5 and 6 respectively. Finally, we draw our conclusions in Section 7.

2. Related work

2.1. Bag-of-Words (BoW) image descriptors

The BoW descriptors [25] can be calculated in four phases. First, local features, such as HoG (Histogram of Oriented Gradients) [7], SIFT (Scale-Invariant Feature Transform) [18] and deep convolutional features [5], [20], [30], are computed. Then, a word dictionary is learnt from all or a subset of these features. Next, soft assignment [28] or vector quantisation [25] is used to encode the local features in the word space. Finally, a histogram is derived by counting the occurrence frequency of words. As known, BoW histograms are "orderless" [17] because they discard the spatial relationship between words.

2.2. Improving BoW by incorporating the spatial information

The methods used to exploit the spatial information for boosting BoW descriptors either incorporate local spatial data into words [15], [19], [24], or encode global spatial layout of words [17], [22], [31]. The local methods only capture the short-range spatial information and yield "orderless" histogram features. On the other hand, the global methods encode the long-range spatial distributions of words using the spatial partitioning of images [17] or other spatial relationship [22], [31]. In essence, the spatial partitioning based methods are designed empirically and they do not use the finer spatial layout within images. In contrast, the methods exploiting other spatial relationship between words have been studied less.

Although Dong and Chantler [12] showed that global spatial layouts can be modelled using contours, it cannot be guaranteed that contours are accurately detected. In this case, the graphs built in the 2D image plane would be a better choice. We thus model the spatial relationship between words in both the short-range and the long-range spatial extents by building a Delaunay triangulation graph [8] from the locations of each word. The spatial relationship

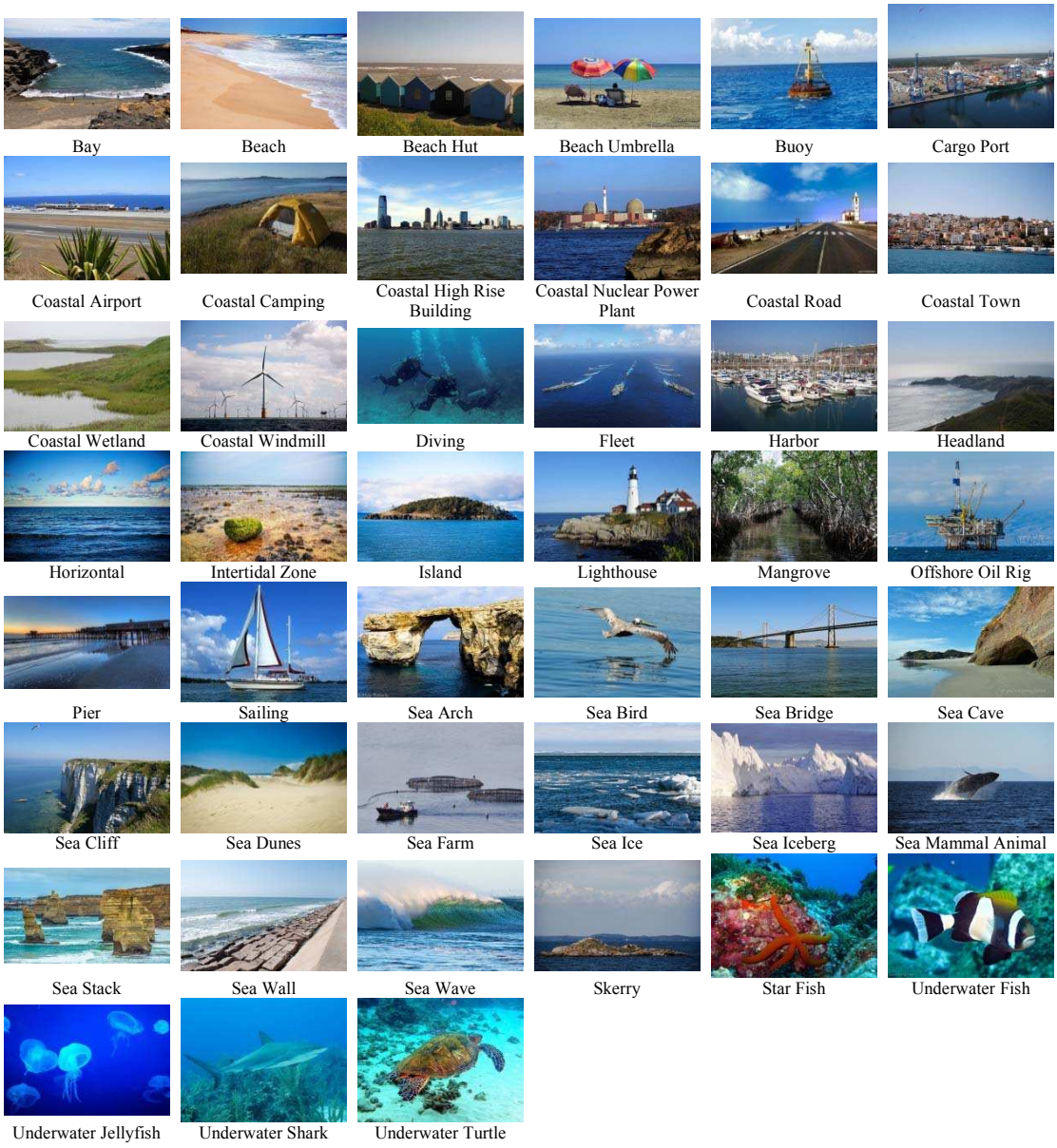


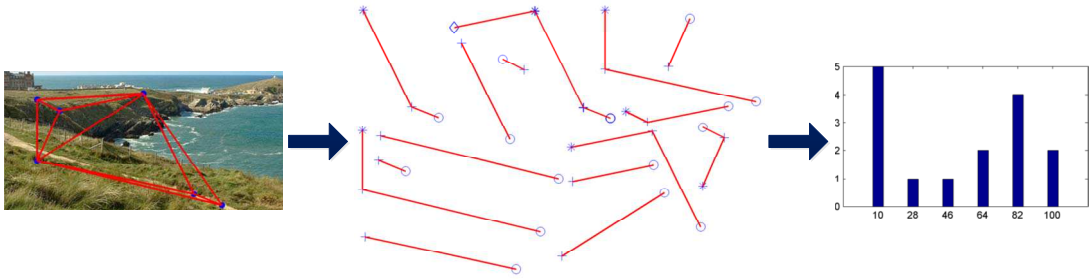
Figure 2. The example images of the 45 classes in FOSD.

between two nodes is modelled using the shortest path connecting these. Since the path often contains more than two nodes, they encode higher order spatial relationships.

3. Flickr oceanic scene dataset (FOSD)

The FOSD dataset consists of 45 classes of oceanic scene images (see Figure 2). Since some classes were difficult to source and we wanted that each class has the same size (which makes the classifier be more fairly trained), the size was fixed as 100 images. In total 4500 images are contained in this dataset. All FOSD images were sourced "in the wild", i.e. these images were derived online rather than being captured in the controlled conditions. Each image was downloaded from Flickr [2] with an "Attribution-NonCommercial-NoDerivs 2.0 Generic" (CC BY-NC-ND 2.0) license [1]. This license permits sharing and only permits sharing except that resizing is acceptable. The maximum of the height and width of images was resized to 640 pixels while the original aspect ratio was kept constant.

The FOSD dataset aims to provide various oceanic scene images for recognition. It supplies not only inter-class but also intra-class variations. The 45 classes can be divided into natural and artificial scenes. The natural scene classes include Bay, Beach, Headland, Horizontal, Intertidal Zone,



(a) (b) (c)

Figure 3. The pipeline of the computation of the GoW descriptor in terms of a word: (a) constructing an undirected DT graph from the locations of the word; (b) obtaining an ensemble of shortest paths connecting two locations; and (c) accumulating a histogram from the distances along the shortest paths, which encodes the spatial layout of the locations of the word.

Island, etc. In contrast, the artificial scene classes comprise Beach Hut, Buoy, Cargo Port, Coastal Road, Coastal Town, and so on. Different classes may differ in objects or even without obvious objects but with similar background, such as Beach Hut, Beach Umbrella and Beach. Besides, images contained in an individual class may be very different. These factors make recognition of this dataset challenging. Compared to existing datasets, e.g. SUN [29] and Places205 [32], the images contained in FOSD were filtered by experts and own more accurate annotations, especially for those terrain classes (e.g. Sea Stack and Sea Arch). In this context, FOSD cannot be replaced by these datasets, even if the quantity of FOSD images is relatively small. Meanwhile, we are continuously expanding FOSD.

4. The graph-of-words (GoW) image descriptor

As known, it is not practical to train a new CNN from scratch with the use of a relatively small dataset. However, encoding the words learnt from the convolutional features that are extracted using a pre-trained CNN normally outperforms the FC features extracted from the same CNN [5], [20], [30]. In this situation, deep word encoders [13], [21], [25] are preferred for small datasets. Compared to the features computed by fine-tuning a pre-trained CNN, these encoders are more generic. Yet the deep word encoders generally do not exploit the spatial layout of words whose importance has been addressed [11], [17]. Therefore, we introduce a new image descriptor exploiting the spatial layout of words. Specifically, this descriptor builds a Delaunay triangulation (DT) graph [8] from the locations of an individual word in the image plane (see Figure 3 for pipeline). The topology of the graph is represented using an ensemble of shortest paths connecting two nodes. We quantise the distances along each path into a d-bin histogram to obtain a sparse representation of the graph. The histograms obtained in terms of w words are concatenated into a $w \times d$ feature vector. This feature vector is referred to as "Graph-of-Words" or "GoW". Compared to the previous work incorporating the spatial data into BoW [15], [17], [19], [22], [24], [31], the GoW descriptor encodes not only the more complicated spatial relationship between word locations but also the longer-range spatial extent. In addition, we fuse the GoW descriptor with SPM [17], FV [21] and VLAD [13] to exploit richer image characteristics.

4.1. The GoW descriptor

4.1.1 Learning deep convolutional words

We use the convolutional features extracted using Places-CNN [32] that was trained using Places205 [32] and the Alex-Net [16] architecture to learn words. The advantage of this model [32] over that [16] trained using ImageNet [9] has been shown for scene recognition [32]. We used the MatConvNet library [27] and imported the Places-CNN model [32] into the compatible format. Given an image, it is first resized into six scales: 2^s ($s \in$ {-2, -1.5, -1, -0.5, 0, 0.5}). Then, each resized image is subtracted by the mean colour and is fed to CNN to obtain 256-D feature vectors at the last convolutional layer (Conv5). Meanwhile, the location where each feature vector is extracted is mapped from different scales into the original image. Finally, 64-D feature vectors are obtained by applying L_2 normalisation, PCA, whitening and L_2 normalisation to those feature vectors in turn. A subset ($\approx$2 million) is randomly selected from the features extracted from the dataset. The Gaussian Mixture Model is used to compute the FV [21] descriptor while k-means is used to learn words for other word encoders. We use vector quantisation to map the features into the word space.

4.1.2 Constructing DT graphs from a word map

Once the convolutional features extracted from an image are quantised into words, we obtain a word map in which each location is assigned a word label. In terms of each word, we build a DT graph [8] from the labelled locations. In total, w (the number of words) DT graphs are built. The reasons for constructing a graph for each word rather than building it for all the words include that: (1) it is not

possible to build a DT graph from the locations sampled in the form of a grid where local features are extracted as DT requires that three points are not parallel; and (2) the computational cost for representing shortest paths is reduced from $(N_{W_1}+\cdots+N_{W_w})(N_{W_1}+\cdots+N_{W_w}-1)/2$ to $N_{W_1}(N_{W_1}-1)/2+N_{W_2}(N_{W_2}-1)/2+\cdots+N_{W_w}(N_{W_w}-1)/2$ (N_{W_i} is the number of the locations of the i-th word) as only the intra-word spatial relationship is considered. Figure 3 (a) shows an example of the DT graph built from the locations of a word. As can be seen, the spatial layout of the locations is retained.

4.1.3 Disassembling a DT graph into shortest paths

In essence, the DT graph [8] is a series of edges connecting two adjacent nodes. Given a DT graph, we model the spatial relationship between both two adjacent and two non-adjacent nodes using the shortest path between these nodes. As a result, the DT graph is disassembled into an ensemble of shortest paths (see Figure 3 (b)). Given a DT graph of the word W_i (i=1, 2, …, w) containing N_{W_i} nodes, $N_{W_i}(N_{W_i}-1)/2$ shortest paths are derived using Dijkstra's algorithm [10] in total. Each shortest path connects two terminal nodes and each pair of nodes is connected by a unique shortest path. Hence, the spatial layout of the locations of a word is retained in an ensemble of shortest paths. For w words, we obtain w ensembles of shortest paths respectively (as a DT graph is built for each word).

4.1.4 Encoding the shortest paths

Each shortest path is encoded as the total distance traversing along the path, i.e. the sum of the distances between two adjacent nodes on the path. Given a DT graph [8], the distances computed for each shortest path are quantised into a d-bin histogram (see Figure 3 (c)). This histogram is used to represent the graph. Especially, the bins are normalised by the length of the image's diagonal to diminish the influence of varying image resolutions.

4.1.5 Generating GoW feature vectors

The w bins of the BoW histogram is used to weight the w shortest path distance histograms computed from w DT graphs [8], respectively. This processing preserves the occurrence frequency of the words. When the w distance histograms are symbolized as DH_i ($i \in \{1,2,\dots,w\}$) and the bins of the BoW histogram are denoted as $BOWH_i$, each weight β_i is computed as:

$$\beta_i = BOWH_i / \sum_d DH_i,\ i \in \{1,2,\dots w\}. \qquad (1)$$

The weighting operation is implemented as the production of β_i and DH_i. The w weighted distance histograms are concatenated into a single feature vector, namely, "GoW".

4.2. The fused GoW descriptors

The FV [21] and VLAD [13] descriptors normally perform better than BoW [25] as they are computed from the difference between words and local features. However, the GoW descriptor is calculated by counting the occurrence frequency of the shortest paths which comprises words. Therefore, it is not fair to directly compare it with FV and VLAD. In addition, GoW does not utilise spatial partitioning that the SPM [17] method uses. We thus fuse the GoW descriptor with the FV, VLAD or SPM descriptors, to exploit richer characteristics. Specifically, we fuse the probability outputs of the classifiers trained using GoW and FV, VLAD or SPM. In contrast, the fusion of features is more computationally expensive as it requires training a new classifier. Given an image, if the probabilities that two classifiers produce are $prob_i^1$ and $prob_i^2$ ($i \in \{1,2,\dots,c\}$, c is the number of classes) respectively, the fused probability is computed as $prob_i = \sqrt{prob_i^1 \cdot prob_i^2}$. The class label of the image is decided based on $prob_i$. When FV, VLAD and SPM are applied, the fused descriptors are named "GoW&F", "GoW&V" and "GoW&S" respectively.

5. Oceanic scene recognition

The GoW descriptor and its fused versions were compared with five baseline word encoders. Two sets of words learnt from the HoG [7] and SIFT [18] features, which were computed using the VLFeat library [4], respectively, were also tested. In this section, we first briefly introduce the implementation details of the five word encoders. Then, we describe the experimental setup. Finally, we report our experimental results in detail.

5.1. Implementation notes of five baseline word encoders

Bag-of-Words (BoW) The BoW [25] feature vectors were L_1 normalised because the size of images varies.

Spatial Pyramid Matching (SPM) Three levels (Levels 0, 1 and 2) of spatial pyramids were used for the SPM [17] method. The BoW features extracted at different pyramid levels were weighted in the same way as that Lazebnik et al. [17] proposed.

Fisher Vectors (FV) The signed square-rooting and L_2 normalisation were applied to the FV [21] feature vectors.

Vector of Locally Aggregated Descriptors (VLAD) The signed square-rooting and global L_2 normalisation were used for VALD [13].

Pairs of Identical Words Angle Histogram (PIWAH) The pairwise angles were computed using 60% randomly sampled word locations in each image in order to boost the computational speed. As described in [14], angle histograms were quantised into nine bins.

We learnt 50, 100 and 200 words from deep convolutional, HoG [7] and SIFT [18] features. We did not learn more words as the dimensionality of the SPM [17], FV [21] and VLAD [13] features could become high which restricts their applicability.

Local Features	HoG			SIFT			CNN		
w	50	100	200	50	100	200	50	100	200
Recog. Rate (%)	42.55 ±1.26	44.52 ±0.79	**46.26 ±0.95**	36.53 ±1.14	40.25 ±0.74	**43.47 ±0.99**	65.35 ±0.76	71.23 ±0.56	**75.02 ±0.70**

Table 1. The mean and standard deviation of the oceanic scene recognition rates obtained using the HoG [7], SIFT [18] and CNN [32] based GoW descriptors ($d = 9$) when 50, 100 and 200 words are used.

5.2. Experimental setup

We randomly divided the 100 images of each FOSD class into two equal-sized subsets. One subset was used for training while the other one was used for testing. Ten different splits of training and test images were randomly generated. Support Vector Machines (SVM) [3], [6] was used as classifier. The linear kernel [3] ($C = 10$) was used for FV [21] and VLAD [13] while the histogram intersection kernel [17] ($C = 128$ and $r = 0$) was utilised for the histogram based descriptors. The recognition rate (%) was used as performance metric which measures the percentage of the number of correctly recognised test images over the number of all test images.

5.3. Experimental results

The experiment was conducted in four schemes. First, we tested the GoW descriptor with varying parameters. Second, we compared it with two histogram methods: BoW [25] and PIWAH [14]. Third, we examined the fused GoW descriptors. Finally, we investigated the performance of GoW in terms of each oceanic class.

5.3.1 Effects of different parameters on GoW

When the length of shortest paths was quantised into d = 9 bins, we tested GoW using 50, 100 and 200 words learnt from three types of local features. Table 1 reports the recognition rates obtained. It can be seen that: (1) the CNN-based GoW descriptor performed much better than its HoG [7] and SIFT [18] based counterparts. This is independent on the number of words used; and (2) the GoW descriptor performed better when 200 words were used than that it did when 50 or 100 words were used. Although the better performance may be obtained using more words, the computational cost will be more expensive.

The GoW descriptor was further quantised into d = 3, 6, 9, 12, 15 and 18 distance bins when 200 words were used. To reduce the computational cost, DT graphs [8] were built using 60% randomly sampled locations of each word. Figure 4 shows the results obtained using the GoW descriptors computed from HoG [7], SIFT [18] and CNN features [32]. It can be observed that: (1) the CNN-based GoW descriptor greatly outperformed the HoG and SIFT based GoW descriptors; (2) the HoG-based GoW performed better than the SIFT-based GoW; and (3) the GoW descriptor performed the best using nine bins.

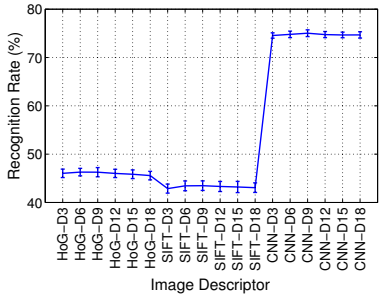


Figure 4. The mean and standard deviation of the oceanic scene recognition rates obtained using the GoW descriptor when 200 words and different bins of shortest path distances are used. Here, three types of local features are examined.

Proportion	10%	30%	60%	100%
Recog. Rate (%)	71.73±0.81	74.14±0.67	**75.02±0.70**	74.92±0.76

Table 2. The mean and standard deviation of the oceanic scene recognition rates obtained using GoW ($d = 9$ and $w = 200$) when different proportions of the locations of each word are used.

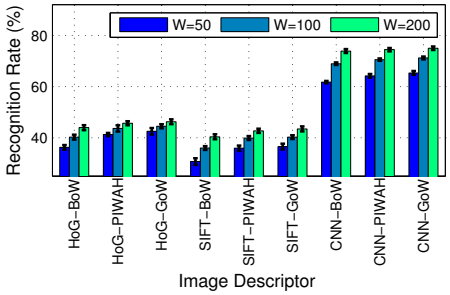


Figure 5. The mean and standard deviation of the oceanic scene recognition rates obtained using BoW [25], PIWAH [14] and GoW (d = 9) when different numbers of words are used. Here, three different types of local features are tested.

We also tested the CNN-based GoW descriptor (d = 9 and w = 200) using different proportions of word locations. Table 2 lists the recognition rates obtained when p = 10%, 30%, 60% and 100% locations of each word were used. It can be seen that the GoW descriptor performed the best when 60% locations of each word were used.

5.3.2 Comparison with BoW and PIWAH

Since GoW was computed based on histograms, we compared it (d = 9 and p = 60%) with two histogram descriptors: BoW [25] and PIWAH [14]. Figure 5 shows the average recognition rates derived using these descriptors. As can be seen, the GoW descriptor produced the better performance than BoW and PIWAH when different numbers of words were used. Again, the CNN [32] based descriptors outperformed their HoG [7] and SIFT [18] based counterparts with a large margin. Table 3 further reports the best average recognition rates obtained using the

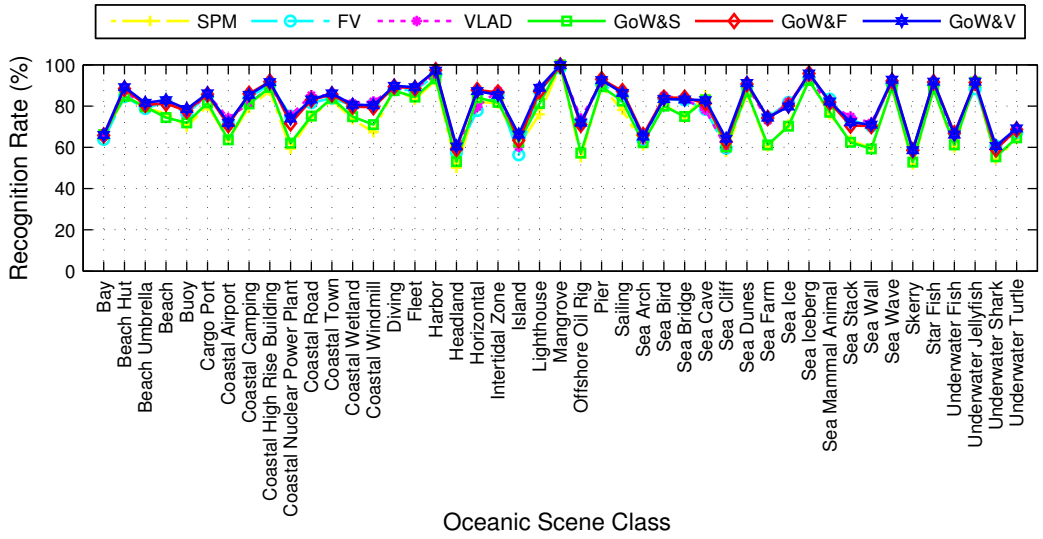


Figure 7.The average recognition rates (%) derived using six different descriptors that are computed using deep convolutional features on 45 FOSD scene classes. The d value is set as 9 for the three fused GoW descriptors (whose performances are plotted in solid lines).

Local Features	HoG			SIFT			CNN		
Descriptor	BoW	PIWAH	GoW	BoW	PIWAH	GoW	BoW	PIWAH	GoW
Recog. Rate (%)	43.97 ±0.98	45.67 ±0.79	**46.26 ±0.95**	40.38 ±0.98	42.72 ±0.83	**43.47 ±0.99**	73.92 ±0.77	74.46 ±0.68	**75.02 ±0.70**

Table 3. The mean and standard deviation of the oceanic scene recognition rates derived using BoW [25], PIWAH [14] and GoW ($d = 9$) when 200 words are used.

three descriptors with 200 words.

5.3.3 Performance of the fused GoW descriptors

Given that d = 9 bins are quantised for GoW when 60% locations are used for each word, we compare the GoW&S, GoW&F and GoW&V descriptors calculated using different numbers of words that are learnt from three types of local features against the corresponding SPM [17], FV [21] and VLAD [13] descriptors respectively. In Figure 6, the average recognition rates obtained using these descriptors are shown. It can be seen that: (1) each of the enhanced GoW descriptors normally outperformed the original descriptor. In other words, the GoW descriptor is complementary to the SPM [17], FV [21] and VLAD [13] descriptors; (2) each of the CNN [32] features based descriptors significantly outperformed its HoG [7] and SIFT [18] based counterparts; and (3) the highest average recognition rate 80.64% was derived using the GoW&V descriptor when 200 words were used. Besides, the results derived using FV, VLAD, GoW&F and GoW&V were better than the average recognition rate: 78.45%±0.68 that was produced by the features extracted at the penultimate fully-connected layer of Places-CNN [32]. Table 4 reports the best average recognition rates derived using the six descriptors that were computed from three types of local features.

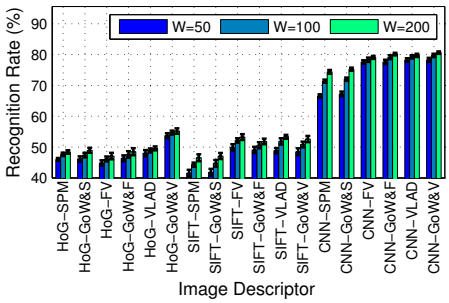


Figure 6. The mean and standard deviation of the oceanic scene recognition rates obtained using the enhanced GoW descriptors: GoW&S, GoW&F and GoW&V and the baselines: SPM [17], FV [21] and VLAD [13] when different numbers of words learnt from three types of local features are used.

Local Features	HoG		SIFT		CNN	
Descriptor	SPM	GoW&S	SPM	GoW&S	SPM	GoW&S
Recog. Rate (%)	48.43±0.57	**48.96±0.80**	46.54±1.12	**47.09±0.99**	74.41±0.61	**75.28±0.50**
Descriptor	FV	GoW&F	FV	GoW&F	FV	GoW&F
Recog. Rate (%)	47.08±1.05	**48.51±1.16**	**53.25±0.94**	51.82±0.90	79.10±0.49	**80.17±0.44**
Descriptor	VLAD	GoW&V	VLAD	GoW&V	VLAD	GoW&V
Recog. Rate (%)	49.58±0.63	**55.20±0.88**	**53.33±0.65**	52.60±1.03	79.74±0.53	**80.64±0.39**

Table 4. The mean and standard deviation of the oceanic scene recognition rates obtained using the enhanced GoW descriptors: GoW&S, GoW&F and GoW&V and three baselines: SPM [17], FV [21] and VLAD [13] when 200 words learnt from different local features are used.

features.

5.3.4 Performance on each oceanic scene class

Figure 7 displays the average recognition rates for 45 FOSD scene classes obtained using SPM [17], FV [21], VLAD [13], GoW&S, GoW&F and GoW&V that were

w	50			100			200		
Descriptor	BoW	PIWAH	GoW	BoW	PIWAH	GoW	BoW	PIWAH	GoW
CR (%)	84.58 ±0.54	85.54 ±0.31	**86.93 ±0.33**	88.38 ±0.22	89.06 ±0.29	**89.30 ±0.38**	90.93 ±0.46	91.16 ±0.35	**91.46 ±0.41**

Table 5. The mean and standard deviation of the aerial scene recognition rates derived using BoW [25], PIWAH [14] and GoW ($d = 9$) with different numbers of words.

w	50		100		200	
Descriptor	SPM	GoW&S	SPM	GoW&S	SPM	GoW&S
CR (%)	86.27±0.46	**87.31±0.37**	89.20±0.37	**89.89±0.40**	90.61±0.36	**91.42±0.48**
Descriptor	FV	GoW&F	FV	GoW&F	FV	GoW&F
CR (%)	94.37±0.17	**94.41±0.13**	94.26±0.28	**94.31±0.19**	94.44±0.25	**94.87±0.20**
Descriptor	VLAD	GoW&V	VLAD	GoW&V	VLAD	GoW&V
CR (%)	93.17±0.28	**93.79±0.38**	92.75±0.32	**93.79±0.27**	93.46±0.33	**94.33±0.37**

Table 6. The mean and standard deviation of the aerial scene recognition rates derived using the fused GoW descriptors: GoW&S, GoW&F and GoW&V ($d = 9$) and three baselines: SPM [17], FV [21] and VLAD [13] when different numbers of words learnt from deep convolutional features are used.

computed based on deep convolutional features [32]. It can be seen that the GoW&V descriptor performs better than, or at least comparably to, the other descriptors on all the 45 classes. Generally speaking, the "Mangrove" class is the easiest one for all the six descriptors while "Headland" and "Skerry" are the most difficult classes for these descriptors.

6. Aerial scene recognition

We further generalised the GoW descriptor and its fused versions to aerial scene recognition. The *UC Merced Land Use* dataset [31] was used. In total 21 classes are included in this dataset and each class contains 100 images. Different objects and spatial patterns can be found in these images. Following the original experimental setup [31], we performed a five-fold cross-validation scheme using SVM [6]. However, we repeated this scheme for ten times rather than only once as that Yang and Newsam [31] did. The mean and standard deviation of the classification rates (CR, %) were used as performance metric. Since the resolution of the images is 256×256 pixels, which is smaller than that of FOSD images, we used all the locations of each word to compute the GoW descriptor. The other setup was kept constant as described in Section 5.

Table 5 reports the results obtained using the BoW [25], PIWAH [14] and GoW descriptors. It can be observed that GoW always outperformed its counterparts. However, the gap between the performances obtained using GoW or PIWAH and that derived using BoW became smaller when more words were used. This finding is consistent with the observation obtained by Khan et al. [14]. Compared to the results obtained using the SIFT-based BoW method [31], the CNN-based BoW performed better (88.38% vs. 71.86%, $w = 100$).

In Table 6 we further compare the GoW&S, GoW&F and GoW&V descriptors computed using different numbers of deep words against the corresponding SPM [17], FV [21] and VLAD [13] descriptors respectively. As can be seen, the fused GoW descriptor always produced the better performance than the corresponding original descriptor. In addition, our best performance 94.87±0.20 is higher than that obtained using the features extracted at the penultimate fully-connected layer: 91.35±0.38.

7. Conclusions and future work

In this paper, we first introduced a novel task: Oceanic Scene Recognition. This task can be used to monitor the oceanic environment by analysing the images acquired using a series of cameras that are placed at different locations. To this end, we collected a new image dataset, i.e. *Flickr Oceanic Scene Dataset* (FOSD), which can provide the labelled scene data that are usually not covered by existing data sets.

Restricted by the relatively small size of this dataset, it is not feasible to train a CNN from scratch. Since encoding visual words learnt from the convolutional features in a pre-trained CNN normally yields better results than the FC features calculated from the same CNN [5], [20], [30], and considering the existing encoders rarely use the spatial layout of words, we proposed a new image descriptor encoding the higher order spatial relationship between the locations of the same word. The spatial relationship is encoded in the form of the shortest paths extracted from the Delaunay triangulation (DT) graph [8] that are built from those locations. The proposed descriptor is referred to as "Graph-of-Words" or "GoW" for short. This descriptor outperformed two histogram descriptors: BoW [25] and PIWAH [14] in the oceanic and aerial scene recognition tasks. Also, three enhanced GoW descriptors were obtained by fusing the probability outputs of the classifiers trained using the GoW and the SPM [17], FV [21] or VLAD [13] features respectively. The three descriptors outperformed their individual counterparts in different conditions. These promising results should be attributed to the fact that the GoW descriptor exploits both the short-range and long-range higher-order spatial layouts of words.

In future work, the GoW descriptor may be further exploited based on the deep convolutional features extracted from a fine-tuned CNN. We believe that the proposed method will also boost the performance of these fine-tuned features.

Acknowledgement

Junyu Dong was supported by the National Natural Science Foundation of China (NSFC) (No. 61271405, 41576011) and the Ph.D. Program Foundation of Ministry of Education of China (No. 20120132110018).

References

[1] Creative Commons, https://creativecommons.org/

[2] Flickr, https://www.flickr.com/

[3] LIBSVM, http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[4] VLFeat, http://www.vlfeat.org/

[5] M. Cimpoi, S. Maji, I. Kokkinos, and A. Vedaldi. Deep Filter Banks for Texture Recognition, Description, and Segmentation. Int'l J. Computer Vision, 118(1):65-94, 2016.

[6] C. Cortes, and V. Vapnik. Support-vector networks. Machine Learning, 20(3):273-297, 1995.

[7] N. Dalal and B. Triggs. Histograms of Oriented Gradients for Human Detection. In Proc. CVPR, 2005.

[8] B. Delaunay. Sur la sphère vide," Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques et naturelles, 6:793-800, 1934.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In Proc. CVPR, 2009.

[10] E.W. Dijkstra. A note on two problems in connexion with graphs. Numerische Mathematik, 1:269-271, 1959.

[11] X. Dong and M. J. Chantler. The Importance of Long-Range Interactions to Texture Similarity. In Proc. CAIP, 2013.

[12] X. Dong and M. J. Chantler. Perceptually Motivated Image Features Using Contours. IEEE Transactions on Image Processing, 25(11):5050-5062, 2016.

[13] H. Jegou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid. Aggregating Local Image Descriptors into Compact Codes. IEEE Trans. Pattern Anal. Mach. Intell., 34(9):1704-1716, 2012.

[14] R. Khan, C. Barat, D. Muselet, and C. Ducottet. Spatial orientations of visual word pairs to improve Bag-of-Visual-Words model. In Proc. BMVC, 2012.

[15] J. Krapac, J. Verbeek, and F. Jurie. Modeling spatial layout with fisher vectors for image categorization. In Proc. ICCV, 2011.

[16] A. Krizhevsky, I. Sutskever, and G. E. Hinton. ImageNet Classification with Deep Convolutional Neural Networks. In Proc. NIPS, 2012.

[17] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In Proc. CVPR, 2006.

[18] D. G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. Int. J. Comput. Vision, 60(2):91-110, 2004.

[19] N. Morioka, and S. Satoh. Building Compact Local Pairwise Codebook with Joint Feature Space Clustering. In Proc. ECCV, 2010.

[20] J. Y. Ng, F. Yang, and L. S. Davis. Exploiting local features from deep networks for image retrieval. In Proc. CVPRW, 2015.

[21] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In Proc. ECCV, 2010.

[22] T. Quack, V. Ferrari, B. Leibe, and L. Van Gool. Efficient Mining of Frequent and Distinctive Feature Configurations. In Proc. ICCV, 2007.

[23] A. Quattoni and A. Torralba. Recognizing Indoor Scenes. In Proc. CVPR, 2009.

[24] J. Sánchez, F. Perronnin, and T. de Campos. Modeling the spatial layout of images beyond spatial pyramids. Pattern Recognition Letters, 33(16):2216-2223, 2012.

[25] J. Sivic, and A. Zisserman. Video Google: a text retrieval approach to object matching in videos. In Proc. ICCV, 2003.

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In Proc. CVPR, 2015.

[27] A. Vedaldi and K. Lenc. MatConvNet - Convolutional Neural Networks for MATLAB. In Proc. ICMM, 2015.

[28] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained Linear Coding for Image Classification. In Proc. CVPR, 2010.

[29] J. Xiao, K. A. Ehinger, J. Hays, A. Torralba, and A. Oliva. SUN Database: Exploring a Large Collection of Scene Categories. Int'l J. Computer Vision, 8(4):1635-1650, 2015.

[30] A. B. Yandex and V. Lempitsky. Aggregating Local Deep Features for Image Retrieval. In Proc. ICCV, 2015.

[31] Y. Yang, and S. Newsam. Spatial pyramid co-occurrence for image classification. In Proc. ICCV, 2011.

[32] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning Deep Features for Scene Recognition using Places Database. In Proc. NIPS, 2014.