- **Title:** Fusing Integrated Visual Vocabularies-Based Bag of Visual Words and Weighted Colour Moments on Spatial Pyramid Layout for Natural Scene Image Classification

- **Running head:** FIVVBOBWCMNSIC

- **Keywords:** image classification, natural scenes, bag of visual words, integrated visual vocabulary, pyramidal colour moments, feature fusion, semantic modelling.

- **Number of pages:** 24

- **Number of Tables:** 6

- **Number of Figures:** 10

- **Authors’ addresses:**

  1) Yousef Alqasrawi  
  School of Computing, Informatics and Media (SCIM)  
  University of Bradford  
  Bradford, BD7 1DP, UK  
  Phone: (00 44 7517207035)  
  Email: Y.T.N.Al-qasrawi@Bradford.ac.uk  
  Fax: (00 44 1274 235856)

  2) Daniel Neagu  
  School of Computing, Informatics and Media (SCIM)  
  University of Bradford  
  Bradford, BD7 1DP, UK  
  Phone: 5704 (00 44 1274 235704)  
  Email: D.Neagu@Bradford.ac.uk  
  Fax: (00 44 1274 235856)

  3) Peter Cowling  
  School of Computing, Informatics and Media (SCIM)  
  University of Bradford  
  Bradford, BD7 1DP, UK  
  Phone: 4005 (00 44 1274 234005)  
  Email: P.I.Cowling@Bradford.ac.uk  
  Fax: (00 44 1274 235856)
- List of figure captions:

Fig. 1 Keypoint detection and description process. The circles overlaid on the image indicate keypoints located using DoG feature detector. Each keypoint is described and stored in feature vector. Each feature vector contains 128 descriptive values, using SIFT descriptors.

Fig. 2 Visual vocabulary construction process. The left side of Features database shows universal visual vocabulary. The right side shows the proposed integrated visual vocabulary. Each class features (for class 1, 2, …, M) represents feature vectors of training images for a specific image category.

Fig. 3 (a) Sample image with circles around interest points. (b) Sky and water contain little information of interest. Red borders in (b) shows important information that helps discriminate image content.

Fig. 4 Feature fusion process on spatial pyramid layout (L=2). The left column represents histograms of BOW. The right column represents colour moments for the HSV colour space bands. The middle column represents an image at different levels overlaid with circles around interest points.

Fig. 5 Some examples of the images used for each category from the Dataset 1, Dataset 2, Dataset 3 and Dataset 4 respectively.

Fig. 6 The classification performance of IPBOW_PCM compared with different baseline methods for each scene class of Dataset 2. It is clear that in most scene classes IPBOW_PCM outperforms other methods. Gist+PCM features perform best for the open country scene class.

Fig. 7 A comparison of the average classification performance accuracy of different image representation methods on Dataset 2 (a) and Dataset 1 (b).

Fig. 8 The classification performance of our approach compared with different methods for each scene class of Dataset 1. It is clear that in most scene classes our approach outperforms other methods.

Fig. 9 Performance comparisons between our proposed approach (IPBOW_WPCM) based on SIFT features and IBOW image representation based on rgbSIFT features [47] both tested on Dataset 1 (a) and Dataset 3 (b).

Fig. 10 For each scene concept (rows), (a) shows the average of UBOW histograms (b) the average of IBOW histograms and (c) shows sample images and their corresponding IBOP histograms. We can see that most image histograms tend to belong to their average histograms. Though, some classes get confused with other classes such as river/lakes and mountains since many mountain images contain water and vice versa.
- **List of symbols used in the article**

  - $M$ number of classes
  - $C$ set of $M$ scene classes
  - $V$ set of $M$ class-specific vocabularies
  - $V_j$ set of $k$ visual words learned from training images of class $j$
  - $v_i$ $i$th visual word
  - $u_j$ $j$th visual word
  - $|V|$ size of visual vocabulary
  - $h(d)$ histogram of visual words for image $d$
  - $h_i(d)$ number of descriptors in image $d$
  - $N_d$ total number of descriptors in image $d$
  - $L$ number of levels on the spatial pyramid layout
  - $h_l^i(d_{ri})$ histogram vector of BOW for image $d$ at level $l$ and sub-region $r_i$
  - $c_l^i(d_{ri})$ colour moments vector for image $d$ at level $l$ and sub-region $r_i$
  - $m$ number of images in the training image dataset
  - $T$ real valued threshold vector
  - $T_{ri}$ average density of keypoints at level $l$ and image sub-region $r_i$ over $m$ images
  - $H(d)$ feature vector for image $d$ results from concatenation of BOW and weighted pyramidal colour moments.
  - $w$ weights vector that indicates the importance of colour information
  - $K$ kernel function

- **Mailing address for galley proofs**

  **Postal Address:**
  Yousef Alqasrawi
  School of Computing Informatics and Media
  Horton D4.24
  University of Bradford
  BD7 1DP, UK.
Fusing Integrated Visual Vocabularies-Based Bag of Visual Words and Weighted Colour Moments on Spatial Pyramid Layout for Natural Scene Image Classification

Yousef Alqasrawi, Daniel Neagu, Peter Cowling

School of Computing, Informatics and Media (SCIM), University of Bradford, Bradford, BD7 1DP, UK

Abstract

The bag of visual words (BOW) model is an efficient image representation technique for image categorisation and annotation tasks. Building good visual vocabularies, from automatically extracted image feature vectors, produces discriminative visual words which can improve the accuracy of image categorisation tasks. Most approaches that use the BOW model in categorising images ignore useful information that can be obtained from image classes to build visual vocabularies. Moreover, most BOW models use intensity features extracted from local regions and disregard colour information which is an important characteristic of any natural scene image. In this paper we show that integrating visual vocabularies generated from each image category, improves the BOW image representation and improves accuracy in natural scene image classification. We use a keypoints density-based weighting method, to combine the BOW representation with image colour information on a spatial pyramid layout. In addition, we show that visual vocabularies generated from training images of one scene image dataset, can plausibly represent another scene image dataset on the same domain. This helps in reducing time and effort needed to build new visual vocabularies. The proposed approach is evaluated over three well-known scene classification datasets with 6, 8 and 15 scene categories respectively using 10-fold cross-validation. The experimental results, using support vector machines with histogram intersection kernel, show that the proposed approach outperforms baseline methods such as Gist features, rgbSIFT features and different configurations of the BOW model.

Keywords: image classification, natural scenes, bag of visual words, integrated visual vocabulary, pyramidal colour moments, feature fusion, semantic modelling.
1. Introduction

The availability of low-cost image capturing devices and the popularity of photo-sharing websites such as Flickr and Facebook have led to a dramatic increase in the size of image collections. For efficient use of such large image collections, image categorisation, searching, browsing and retrieval techniques are required [1-3].

Scene classification has been investigated in two complementary research areas, understanding human visual perception of scenes and developing computer vision techniques for automatic scene categorisation. This entail reveals how humans are able to understand, analyse and represent scenes, which is of great research interest in many psychological studies. Knowledge from these studies can help computer vision researchers design and develop systems to close the gap between human semantic and image low-level information [4, 5]. From the computer vision viewpoint, scene classification is the task of automatically assigning an unlabelled image into one of several predefined classes (e.g., beach, coast or forest). It provides contextual information to help other processes such as object recognition, content-based image retrieval and image understanding [6, 44]. However, designing and implementing algorithms that are capable of successfully recognising image categories remains a challenging problem [7]. This is because of illumination changes, scale variations, occlusions, large variations between images belonging to the same class and small variations between images in different classes.

In recent years, local invariant features or local semantic concepts [14] and the bag of visual words (BOW) became very popular in computer vision and have shown impressive levels of performance in scene image classification [6, 12, 15-19]. There are two main parts to build an image classification system within the BOW framework. The first relates to the extraction of features that characterise image content. The work described in this paper relates to this part. The second part is the classifier. The elements needed to build a bag of visual words are: feature detection, feature description, visual vocabulary construction and image representation, each step is performed independently of the others.

Recently, a range of new methods have been advanced to improve the performance of the conventional BOW paradigm. We can classify these methods into three main categories. The first category attempts to improve the construction of the visual vocabulary [20-24]. The second category suggests using multiple features with weighting techniques to combine features [18, 25-28]. In the third category, techniques that add spatial information over the BOW have been proven to improve the performance of scene classification tasks [29-32].

Although these approaches have achieved promising results on scene image classification tasks, there is found to be no overall best approach. The complexity of natural scenes and wide variety of arrangements of entities in images means that, images with similar visual contents from two different categories are often miscategorised, (e.g., confusion in visual appearance between coasts and river/lake classes). We believe that some of these problems could be better solved by building a unified approach that uses knowledge from discriminative visual vocabularies, multiple image features and their spatial arrangements.
In this paper, we build upon the work of Alqasrawi et al. [26] to improve categorisation performance for natural scene images. In our previous work [26], we introduced a weighting approach to fuse image colour information, with the traditional bag of visual word model for natural scene image classification. In this paper, we focus on learning discriminative visual words from image classes separately, in order to map local keypoints to most representative visual words.

In this work, we propose integrating knowledge from discriminative visual vocabularies learned from image classes, multiple image features and spatial arrangements information to improve the conventional bag of visual words, for natural scene image classification task. To improve visual vocabulary construction, visual vocabularies extracted from the training images of each scene image category are combined into a single integrated visual vocabulary. It is composed of discriminative visual words from different scene categories. The proposed approach is extensively evaluated over three well-known scene classification datasets with 6, 8 and 15 scene categories [7, 11, 29] respectively using 10-fold cross validation. We tested our work on a fourth dataset with 4 scene classes. This dataset is a subset of 8 scene classes [11] composing natural scene images with no man-made objects (refer to section 5.2). We will show in this paper that the integrated visual vocabulary is more discriminative than the universal visual vocabulary to build BOW histograms. Moreover, we show that the Keypoint Density-based Weighting (KDW) method proposed in [26] can be used effectively with the integrated visual vocabulary, to control the fusion of image colour information and BOW histograms on a spatial pyramid layout. We compare our approach to a number of baseline methods such as Gist features [11], rgbSIFT features [47] and different configurations of conventional BOW.

The rest of this paper is organised as follows. In section 2, we give an overview of related work. In section 3, we describe the three main steps needed to represent and summarise image contents. We describe our approach for merging colour information with BOW in section 4. We present our experimental setup and results in section 5 and we conclude the paper in section 6.

2. Related work

Early work in scene image classification (low-level processing) was based on low-level image features such as colour and texture, extracted automatically from the whole image or from image regions [2, 8-11]. Methods based on global image features often failed to successfully represent the high-level semantics of user perception [2, 3]. Semantic modelling (mid-level processing) uses an intermediate semantic level representation, falling between low-level image features and (high-level) image classification, in an attempt to narrow the semantic gap between low-level features and high-level semantic concepts [7, 12]. A simple way to represent a semantic concept is to partition an image into blocks and then label each block manually with one or more semantic concepts [7, 13]. However, such systems are found to be time consuming and expensive. A recent survey on state of the art techniques for scene classification that exploits features on the spatial domain, frequency domain and constrained domain is presented in [45].
Researchers in computer vision have recently started to make use of techniques based on text document retrieval, to represent image contents in image classification and retrieval systems [34]. The bag of words approach is one of these techniques and is very common in text-based information retrieval systems. The analogy between document and image is that both contain information. However, the main obstacle is how to extract semantic “visual words” from image content. In the literature, much work has been done in image/object classification and retrieval based on the bag of visual words approach. We review those approaches most strongly connected with the work of this paper below.

Zhu et al. [34] are perhaps the firsts who tried to represent images using an analogue of the bag of words approach. In their work, images are partitioned into equal size blocks which are then indexed using a visual vocabulary, whose entries are obtained from the block features. The bag of visual words approach received a substantial increase in popularity and effectiveness with the development of robust salient features detectors and descriptors such as Scale Invariant Feature Transform (SIFT) [17]. Csurka et al. [19] and Sivic et al. [35] showed how to use the bag of visual words by clustering the low-level features using the K-means algorithm, where each cluster corresponds to a visual word. To build the bag of visual words, each feature vector is assigned to its closest cluster. Perronnin et al. [23] proposed the use of adapted vocabularies by combining universal vocabularies with class-specific vocabularies. In their work, a universal visual vocabulary is learned to describe the visual features of all considered image classes. Then, class-specific vocabularies are combined with the universal vocabulary to refine accuracy. Perronnin et al’s work is an interesting contribution to the computation of distinctive visual vocabularies. However, their proposed adapted vocabulary does not show the differences between scene classes and it handles only one kind of image feature. Another contribution to build discriminative visual vocabularies has been investigated by Jurie and Triggs [36], which proposes a clustering algorithm to build a visual vocabulary. Their algorithm produces an ordered list of centers. A quantisation rule is used in such a way that patches are assigned to the first center in the list that lies within a fixed radius $r$, and left unlabelled if there is no such center. In Nilsback and Zisserman [37], three different visual vocabularies are created to represent colour, shape and texture properties of flower images. These visual vocabularies are combined to obtain a weighted joint frequency histogram. In Nister and Stewenius [22], local features extracted from images are hierarchically quantised in a vocabulary tree. It was shown that retrieval results are improved with a larger vocabulary.

Four different approaches are investigated in [38] to create compacted visual vocabulary, while retaining classification performance. The four approaches consists of 1) global visual vocabulary construction; 2) class-specific visual vocabulary construction; 3) annotating a semantic vocabulary and 4) soft-assignment of image features to visual words. These approaches were evaluated against each other on a large dataset. The experimental results suggested that best method depends upon application at hand. Also, in a different context, an approach to integrate different vocabularies has been introduced in [46] to build bag of phrases for near duplicate image detection.
A bag of visual words represents an image as an orderless collection of local features, without spatial information. Spatial pyramid matching was proposed by Lazebnik and Raginsky [14] as an extension to the orderless bag of visual words. The spatial pyramid divides the image into 1x1, 2x2, 4x4, etc. regions. Assuming a visual vocabulary is given; local features extracted from each region are quantised and then combined using a weighting scheme which depends on region level. Based on this approach, three different hierarchical subdivisions of image regions were recently proposed for recognising scene categories [32]. Most work that used BOW and spatial pyramid matching focuses mainly on intensity information analysis and discards image colour information. We believe colour information has an equally significant importance in recognising natural scene image categories. Many studies have investigated including multiple image features within the framework of BOW [18, 26, 28]. Early fusion and late fusion are the two main approaches used to combine different types of image features. In the early fusion approach, features were combined before building visual vocabularies. In late fusion, each feature has its own visual vocabulary. The fusion of colour and intensity information on the BOW paradigm is proposed by Quelhas et al. [18]. Quantised colour information and the BOW are computed over local interest regions. Although this approach has shown an improvement in classification accuracy, it has two main limitations: (1) Colour information is computed over interest regions only; and (2) No spatial information is implemented. In [28] a novel approach is proposed to recognise object categories using multiple image features. Their model, the Top-Down Colour Attention model, considers two properties for image representation: feature binding and vocabulary compactness. The first property involves combining colour and shape features at the local level, while the second property concerns separate visual vocabularies for each of the two features. Very recently, a number of visual patch weighting methods and different configurations of BOW fused with multiple image features have been investigated by Jiang et al. [39] for semantic concept detection in video images. The invariance properties and the distinctiveness of colour descriptors, such as rgbSIFT features, are studied recently by Sande at el. [47] for object and scene classification tasks. They proposed a systematic approach to provide a set of invariance properties, such as invariance to light intensity.

3. Location-Aware Image Semantic Representation

In this section we introduce the main steps needed to construct four forms of bag of visual words which will be used in this work to represent visual image content:

- Universal BOW (UBOW) based on universal visual vocabulary.
- Integrated BOW (IBOW) based on class-specific visual vocabularies.
- Universal Pyramid BOW (UPBOW) similar to UBOW but on spatial pyramid layout.
- Integrated Pyramid BOW (IPBOW) similar to IBOW but on spatial pyramid layout.

The following subsections details all steps required to build these four BOW image representations. In each case we will consider how we extract and describe local features
from images, build universal and integrated visual vocabularies and map local features to the closest visual words on spatial pyramid layout.

3.1. Local invariant points detection and description

In this work we use the Difference of Gaussian (DoG) point detectors and SIFT descriptors [17] to detect and describe local interest points or patches from images. Generally, these methods show good performance compared to other methods in the literature [40]. The DoG detector has properties of invariance to translation, rotation, scale and constant illumination changes. Once local invariant points are defined, SIFT descriptors are used to capture the structure of the local image patches and are defined as local histograms of edge directions computed over different parts of the patch. Each patch is partitioned into 4x4 parts and each part is represented by a histogram of 8 orientations (bins) that gives a feature vector of size 128. In this paper we use the binaries provided at [41] to detect DoG local points and to compute the 128-D real valued SIFT descriptors from them. This process is described in Fig 1. Features extracted from all images are stored in Features Database. In section 3.2 we describe how this is used to build visual vocabularies.

![Fig. 1 Keypoint detection and description process. The circles overlaid on the image indicate keypoints located using DoG feature detector. Each keypoint is described and stored in feature vector. Each feature vector contains 128 descriptive values, using SIFT descriptors.](image)

3.2. Visual Vocabulary Construction

In this section, we describe how to learn both the universal visual vocabulary and the proposed integrated visual vocabulary that will be used in the rest of this paper. To obtain the visual vocabulary, we use feature vectors (SIFT features) stored in image Features Database as described in section 3.1. All feature vectors from all training images on the dataset are
quantized, using the \( k \)-means algorithm, to obtain \( k \) centroids or clusters. These centroids represent visual words. The \( k \) visual words constitute the universal visual vocabulary. This vocabulary is used to build the UBOW and the UPBOW. For the integrated visual vocabulary, SIFT features from all training images of each scene class are clustered into \( k \) visual words. More formally: Let \( C = \{ C_1, C_2, \ldots, C_M \} \) be the set of \( M \) scene classes considered. Let \( V = \{ V_1, V_2, \ldots, V_M \} \) be the set of \( M \) class-specific vocabularies. Each \( V_j = \{ V_{j1}, V_{j2}, \ldots, V_{jk} \} \) is a set of \( k \) visual words learned from all training images of class \( j \). We call \( V \) the integrated visual vocabulary. This vocabulary is used later to build IBOW and IPBOW.

The rationale behind building an integrated visual vocabulary is to try to find more specific discriminative visual words from each image class in order to avoid interference with other classes. In the universal visual vocabulary, visual words that belong to a specific concept (e.g., foliage) may be assigned to a cluster or visual word of a different concept (e.g., rock). We believe that our integrated visual vocabulary may be robust enough to incorporate naturally existing intra-class variations to discriminate between different image classes. For example, building visual vocabulary for coasts scene images would contain informative information about water, sand and sky, in contrast to other scene classes such as mountains. Fig. 2 shows details of how to construct both kinds of visual vocabularies. We will show later in the experimental results section how the distribution of the mean of all IBOW of training images are different and more informative and discriminative than the UBOW generated from universal visual vocabulary (see Fig. 10 for differences between universal and integrated visual vocabularies).

**Fig. 2** Visual vocabulary construction process. The left side of Features database shows universal visual vocabulary. The right side shows the proposed integrated visual vocabulary. Each class features (for class 1, 2, ..., \( M \)) represents feature vectors of training images for a specific image category.

### 3.3. Summarizing image content using the BOW

The Bag of Visual Words provides a summary of image contents. In section 3.1, we discussed feature detection and description of image content. Section 3.2 showed how to build universal and integrated visual vocabularies from image local features. For simplicity, we refer to both IBOW and UBOW more generally as BOW (which differ in which kind of
vocabulary we use to build them). To build the BOW histogram, each image SIFT descriptor is assigned to the index of the nearest cluster in the visual vocabulary. The visual words in the context of this paper refer to the cluster centres (centroids) produced by the $k$-means clustering algorithm. Let $V$ denote the set of all visual words produced from the clustering step over a set of local point descriptors $V = \{v_i| i = 1, \ldots |V|\}$, where $v_i$ is the $i$-th visual word (or cluster) and $|V|$ is the size of the visual vocabulary. We use a vocabulary of size 200 for both the universal visual vocabulary and for class specific visual vocabulary. In the case of the integrated visual vocabulary, $|V|$ is $200M$ (where $M$ is the number of classes). Experiments showed no improvements in performance beyond 200 [29]. The set of all SIFT descriptors for each image $d$ is mapped into a histogram of visual words $h(d)$ at image-level, such that:

$$h_i(d) = \sum_{j=1}^{N_d} f_d^{(i)} , i = 1, \ldots |V|$$

\[
f_d^{(i)} = \begin{cases} 
1, & \|u_j - v_i\| \leq \|u_j - v_l\|, \quad l = 1, \ldots |V| \text{ and } i \neq l \\
0, & \text{otherwise}
\end{cases}
\]

where:

$h_i(d)$ is the number of descriptors in image $d$ having the closest distance to the $i$-th visual word $v_i$ and $N_d$ is the total number of descriptors in image $d$. $f_d^{(i)}$ is equal to one if the $j$-th descriptor $u_j$ in image $d$ is closest to visual word $v_i$ among other visual words in the vocabulary $V$.

### 3.4 Spatial pyramid Layout

Although the orderless bag of visual words approach is widely used and has made a noticeable increase of performance in object/scene image modelling, it seems likely that we can enhance it for scene recognition tasks by incorporating spatial information. Spatial pyramid matching [29] works by repeatedly subdividing an image into increasingly finer sub-regions and then computing histograms of local patches found inside each image sub-region. An image is represented as a concatenation of weighted histograms at all levels of divisions. In this paper, spatial pyramid layout refers to representing images by placing a sequence of increasingly coarser grids over an image. Here we did not penalise local histograms of BOW as described in [29, 32], since the experiments showed that to do so decreases the classification accuracy of our system.

### 4. Pyramidal fusing of BOW and image colour information

In this section, we show how to model image semantic information based on merging BOW and colour moments using a spatial pyramid layout. The motivation of our approach is that most techniques that use BOW rely only on intensity information extracted from local
invariant points and neglect colour information which seems likely to help in recognition performance for scene image categories. We can see in Fig.3 an image with circles around a rather dense set of interest points produced by DoG detectors. We see that interest points are not uniformly distributed across the image, but rather are clustered on salient regions in the scene. In natural scene images, colour information has a significant effect in discriminating image areas such as sky, water and sand. Hence, we believe that merging colour information and the BOW will be significant in modelling image visual semantics.

Fig. 3 (a) Sample image with circles around interest points. (b) Sky and water contain little information of interest. Red borders in (b) shows important information that helps discriminate image content.

Fig. 4 Feature fusion process on spatial pyramid layout (L=2). The left column represents histograms of BOW. The right column represents colour moments for the HSV colour space bands. The middle column represents an image at different levels overlaid with circles around interest points.
In our previous work [26], we proposed the Keypoint Density-based Weighting method (KDW) for merging colour information and BOW over image sub-regions at all granularities on the spatial pyramid layout. The KDW method aims to regulate how important colour information is in each image sub-region before fusing it with BOW. The spatial pyramid layout (refer to Fig. 4) works by splitting an image into increasingly coarser grids to encode spatial locations of image local points. Hence an image with \( L = 2 \) levels, will have three different representations with \( \sum_{l=0}^{L_l}(2^l)^2 = 21 \) image sub-regions overall. Each image sub-region is represented by a combination of the BOW and a weighted colour moments vector of size 6 on the HSV colour space (2 for Hue, 2 for Saturation and 2 for Value). Both colour moments and the BOW histogram are normalised to be unit vectors before the merging process. An image with number of levels \( L = 2 \) and a visual vocabulary of size 200 will produce a vector of dimension 4326.

We formulate our proposed approach below:

Let \( L \) denotes the number of levels, \( l = 0,1,...,L \), needed to represent an image \( d \) on the spatial pyramid layout, i.e., an image \( d \) will have a sequence of \( L \) grids of increasingly finer granularity. Let \( h^l(d_r) \) and \( c^l(d_r) \) denote a histogram vector of BOW computed using equation (1) and the colour moments vector respectively. Both are computed from an image \( d \) at level \( l \) and sub-region \( r, i = 1,..,(2^l)^2 \).

The concept of Keypoint Density-based Weight (KDW): Colour moment vector \( c^l(d_r) \) is assigned a high weight on image sub-regions that have a keypoint density below threshold \( T^l_r \). Colour information will be less important in image sub-regions with high number of local interest points. The threshold \( T \) is a real valued vector. Each component represents the average density of keypoints (number of keypoints) at specific image sub-region over all training images. We proposed the keypoint density-based weight as:

\[
T^l_r = \frac{1}{m} \sum_{j=1}^{m} h^l(d_r^j)
\]

(3)

where \( m \) is the number of images in the training image dataset. The components of the threshold vector, which is the average keypoint density of all images at specific sub-regions and granularity, help in making a decision about the importance of colour information at specific image sub-region. The unified feature vector \( H(d) \) for image \( d \) is a concatenation of weighted colour moments and BOW at all levels and over all granularities:

\[
H(d) = \left\{ (h^0(d_r), w^0_d c^0(d_r)), (h^1(d_r), w^1_d c^1(d_r)),..., h^l(d_r), w^l_d c^l(d_r)) \right\}
\]

(4)
\[ w_i^j = \begin{cases} 
1, & \sum_{j=1}^{k_1} h_j^i(d_e) < T_e^i \\
0.5, & \text{otherwise}
\end{cases} \quad (5) \]

We should notice that the values of weights \( w \) are non-negative numbers to indicate the importance of colour information. We aim to cause images from the same category to be close, and images from different categories to be far away in the new image representation. Values for the weights have been obtained empirically during learning the support vector machine SVM classifiers [43]. We should notice that weight values are highly dependent on the threshold vector obtained from equation (3). Here, we use the proposed integrated visual vocabulary described in section 3.2 and the spatial pyramid layout to generate \( \text{IBOW} \) and \( \text{IPBOW} \) histograms. Fusing weighted pyramidal colour moments (WPCM) with the \( \text{IBOW} \) and \( \text{IPBOW} \) histograms using equation (4) we obtain improved image representation (\( \text{IBOW} + \text{W_PCM} \) and \( \text{IPBOW} + \text{W_PCM} \)). We assume that building visual vocabularies from individual scene classes could produce more discriminative visual words than using universal vocabulary. To justify this assumption, all UBOW and IBOW histograms generated from training images of Vogel’s dataset [7], described in section 5.2, have been averaged to see the distribution of both BOWs in each scene class. Fig. 10 shows the difference between both averages. Some sample BOW histograms are shown and they tend to be similar or close to their average vector.

5. Experimental setup

The first part of this section presents the support vector machine (SVM) classifier and the protocol we follow in all our experiments. Next, we describe the origin and composition of datasets we use in our experiments. Experimental results are then reported with discussion. We use the confusion matrix to assess the performance of all considered experiments.

5.1. Scene classifier

Multi-class classification is done using a support vector machine (SVM) with a histogram intersection kernel. We use SVM in our study as they have been empirically shown to yield higher classification accuracy in scene and text classification tasks [18, 26, 28]. Variations in the classification accuracy are possible due to our choice of SVM kernel function. In this work, we use the histogram intersection kernel. Many studies in image classification observe that the histogram intersection SVM kernel is very useful. Moreover, histogram intersection has been shown more effective than the Euclidean distance in supervised learning tasks [24, 29]. Odone et al. [42] proved that histogram intersection is a Mercer kernel and thus can be used as a similarity measure in kernel based methods. Given two BOW histograms \( h(d_1) \) and \( h(d_2) \), the histogram intersection kernel is:

\[ K(h(d_1), h(d_2)) = \sum_i \min(h(d_1)_i, h(d_2)_i) \quad (6) \]

The protocol we follow for each of the classification experiments was as follows: All experiments have been validated, using 10-fold cross validation where 90% of all images are
selected randomly for learning the SVM and the remaining 10% are used for testing. The procedure is repeated 10 times such that all images are actually tested by the SVM classifier. The average of the results over the 10 splits yields the overall classification accuracy. To implement the SVM method we used the publicly available LIBSVM software [43], dedicated to Matlab, where all parameters are selected based on 10-fold cross validation on each training fold. We use one-against-one multi-classification approach that results in \[
\frac{M(M - 1)}{2}
\] two-class SVMs for \(M\) scene classes.

5.2. Image datasets

There are many image datasets available in the computer vision literature, but most of them are dedicated to object detection and categorisation tasks. Performance of the proposed scene classification approach is tested on two types of image datasets: a dataset with natural scene images only, which is our main concern, and datasets with heterogeneous images including different kind of images. The reason for choosing natural scene images is that they generally are difficult to categorise in contrast to object-level classification because natural scenes constitute a very heterogeneous and complex stimulus class [4]. Also, we considered scene images that constitute artificial objects to allow fair and straightforward comparison with state-of-the-art scene classification methods. Four datasets were used in our experiments:

Dataset 1: This dataset, kindly provided by Vogel et al. [7], contains natural scene images only with no man-made objects. It contains a total of 700 colour images of resolution 720 × 280 and distributed over 6 categories. The categories and number of images used are: coasts, 142; rivers/lakes, 111; forests, 103; plains, 131; mountains, 179; sky/clouds, 34. One challenge in this image dataset is the ambiguity and diversity of inter-class similarities and intra-class differences which makes the classification task more challenging.

Dataset 2: This dataset is a subset of the Oliva and Torralba [11] dataset. It constitutes images of natural scene categories with no artificial objects, which are semantically similar to images in Dataset 1 and is distributed as follows: coasts, 360; forest, 328; mountain, 374; open country, 410. The total number of images in this dataset is 1472.

Dataset 3: This dataset contains heterogeneous image categories. It consists of a total of 2688 colour images, 256x256 pixels, and distributed over 8 outdoor categories. The categories and number of images used are: coast, 360; forest, 328; mountain, 374; open country, 410; highway, 260; inside city, 308; tall building, 356; street, 292. This dataset is created by Oliva and Torralba [11] and is available online at http://cvcl.mit.edu/database.htm.

Dataset 4: This dataset, provided by Lazebnik et al. [29], contains 15 heterogeneous natural scene image categories. All images in this dataset are grayscale images (i.e., no colour images). It contains different kind of images and the average image size is 300x250 pixels. Images are distributed over categories as follow: highway, 260; inside city, 308; tall building, 356; street, 292; suburb, 241; forest, 328; coast, 360; mountain, 374; open country, 410; bedroom, 216; kitchen, 210; living room, 289; office, 215; industrial, 311; store, 315. The
first 8 categories were from Oliva and Torralba [11] and the first 13 were from Li and Perona [12]. Fig. 5 depicts sample images from the four datasets aforementioned with different image classes.

![Dataset Images](image-url)

**Fig. 5** Some examples of the images used for each category from the Dataset 1, Dataset 2, Dataset3 and Dataset 4 respectively.

5.3. **Feature extraction**

In this work, we used Matlab to conduct all experiments. As we mentioned earlier, in the experiments we perform 10-fold cross validation in order to achieve more accurate performance estimation. The binaries provided by [40] are used to detect and describe local
keypoints using DoG and SIFT as parameters. Further we extracted basic rgbSIFT features with local keypoint detection and standard parameters using the Color Descriptor software provided by [47]. To build different visual BOW histograms, SIFT and rgbSIFT features extracted from training images are used to build 10 visual vocabularies, one for each fold. Gist features are extracted from images using the implementation provided by Oliva and Torralba at (http://people.csail.mit.edu/torralba/code/spatialenvelope/).

5.4. Experimental results

In this section, we conduct extensive experiments to empirically evaluate the performance of our proposed approach and compare it to the existing baseline and BOW models for natural scene image categorisation tasks. We present four sets of experiments each corresponds to one of the datasets mentioned in section 5.2. In the first experiments, we tested the performance of our proposed approach on colour natural scene images with no artificial objects, i.e., Dataset 1 and Dataset 2 respectively. The performance of our proposed approach is compared with Gist features, improved Gist features and different configurations of BOW models generated from SIFT features and rgbSIFT features. All visual vocabularies employed in our experiments are generated separately for SIFT and rgbSIFT features. In the second experiments, we test our proposed approach on Dataset 3 which contains different kind of images and scene categories. We intend to find out how our approach performs on heterogeneous set of scene classes. In the third experiment, we use grayscale images, Dataset 4, to test the performance of our proposed approach on large number of images and scene categories. In the last experiment, we investigate the possibility of using visual vocabularies generated from Dataset 1 to produce IBOW for Dataset 2.

Firstly, we present the classification performance of Gist features and improved Gist features (i.e., Gist with pyramidal colour moments) tested on Dataset 1 and 2. The Gist descriptor [11] uses a low dimensional representation of the scene which does not require any segmentation process. A bank of Gabor filters are employed in the frequency domain and tuned to different orientations and scales. The image is divided into a 4x4 grid for which orientation histograms are computed. The Gist features produce a vector of dimension 512. Further details can be found in [11]. The results published by Oliva and Torralba [11] are based on eight scene classes, so to compare their approach to ours we use their code to repeat their experiments for classifying the chosen four scene classes, i.e., Dataset 2. Tables 1 and 2 depict the classification results of using the Gist features on Dataset 1 and 2. What is interesting is that although scene classes in Dataset 2 are similar in their visual semantics to the corresponding scene classes in Dataset 1, the results for the former dataset are significantly better than for the latter. It seems that the classes in Dataset 1 do not exhibit consistent properties as detected by Gist. To improve the Gist features, we propose to integrate image colour information to Gist features by fusing pyramidal colour moments (PCM, L=2) with Gist features. This combination of image features has resulted in an improvement in the classification performance and thus supporting the significance of pyramidal colour moments approach. Detailed results are depicted in Fig. 6 and 8 for class specific classification performance using Gist features compared to other approaches. Fig. 7(a) and 7(b), report the average classification results of (Gist) and (Gist+PCM)
representations on both datasets. It is clear that adding pyramidal color moments to the Gist features outperformed the classification performance of using Gist features alone.

Table 1
The first part of this table shows the confusion matrix of our proposed approach (IPBOW_PCM) with no weighting tested on Dataset 2. The diagonal bold values are the average classification rate of each image category. The overall classification accuracy is 88.7% and is clearly outperforms the Gist features shown in the second part.

<table>
<thead>
<tr>
<th></th>
<th>Coast</th>
<th>Forest</th>
<th>Mountain</th>
<th>Open country</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coast</td>
<td>0.90</td>
<td>0.00</td>
<td>0.03</td>
<td>0.08</td>
</tr>
<tr>
<td>Forest</td>
<td>0.00</td>
<td>0.96</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>Mountain</td>
<td>0.02</td>
<td>0.02</td>
<td>0.89</td>
<td>0.07</td>
</tr>
<tr>
<td>Open country</td>
<td>0.08</td>
<td>0.04</td>
<td>0.06</td>
<td>0.82</td>
</tr>
<tr>
<td><strong>Overall accuracy rate</strong></td>
<td></td>
<td></td>
<td></td>
<td><strong>88.7</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>IPBOW_PCM</th>
<th>Gist [11]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coast</td>
<td>0.90</td>
<td>0.88</td>
</tr>
<tr>
<td>Forest</td>
<td>0.96</td>
<td>0.93</td>
</tr>
<tr>
<td>Mountain</td>
<td>0.89</td>
<td>0.84</td>
</tr>
<tr>
<td>Open country</td>
<td>0.82</td>
<td>0.75</td>
</tr>
<tr>
<td><strong>Overall accuracy rate</strong></td>
<td><strong>84.2</strong></td>
<td></td>
</tr>
</tbody>
</table>

Secondly, we present the classification performance of using pyramidal colour moments fused with the proposed IBOW and IPBOW, using the KDW weighting method, to represent image contents. We used integrated visual vocabularies to build IBOW from the whole image and IPBOW from image sub-regions as discussed in section 3.3. The pyramidal colour moments were fused with IBOW and IPBOW using our weighting method, to obtain two new image representations; IBOW_WPCM and IPBOW_WPCM. We can observe from table 1 that adding spatial information and colour moments to the IPBOW improves the classification performance. Table 1 indicates clearly that our approach, excluding the weighting technique, outperform Gist features by +4.5%. This is mainly because Gist features do not contribute colour information and spatial layout which provides informative features for scene classification task. In Table 2, our approach to represent image content (IPBOW+WPCM) outperforms others’ work [7, 11, 18] and improves upon our earlier work [26] by (+4.4%). This provides empirical evidence that the integrated visual vocabulary provides more informative visual words than the universal visual vocabulary. Also, the results show how the weighting influences the performance of the IBOBow and thus improves the classification results. Despite this, Gist features still performs very well in some classes such as ‘river/lakes’ and ‘sky/clouds’ classes which are most difficult for our approach to recognise.

Table 2
The first part of this table shows the confusion matrix of our proposed approach (IPBOW_WPCM) tested on dataset 1. The diagonal bold values are the average classification rate of each image category. The overall classification accuracy is 73.7%. The second part of this table reports results of other approaches on the same dataset. It is obvious that our approach outperforms other approaches reported in the literature.

<table>
<thead>
<tr>
<th></th>
<th>Coast</th>
<th>River/lake</th>
<th>Forest</th>
<th>Plain</th>
<th>Mountain</th>
<th>Sky/cloud</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coast</td>
<td>72.54</td>
<td>8.45</td>
<td>2.11</td>
<td>5.63</td>
<td>11.27</td>
<td>0.00</td>
</tr>
<tr>
<td>River/lake</td>
<td>18.02</td>
<td>49.55</td>
<td>10.81</td>
<td>5.41</td>
<td>15.32</td>
<td>0.90</td>
</tr>
<tr>
<td>Forest</td>
<td>1.94</td>
<td>3.88</td>
<td><strong>90.29</strong></td>
<td>1.94</td>
<td>1.94</td>
<td>0.00</td>
</tr>
<tr>
<td>Plain</td>
<td>9.16</td>
<td>4.58</td>
<td>6.11</td>
<td><strong>64.89</strong></td>
<td>14.50</td>
<td>0.76</td>
</tr>
<tr>
<td>Mountain</td>
<td>6.15</td>
<td>2.79</td>
<td>1.68</td>
<td>3.91</td>
<td><strong>84.36</strong></td>
<td>1.12</td>
</tr>
<tr>
<td>Sky/cloud</td>
<td>5.88</td>
<td>2.94</td>
<td>0.00</td>
<td>5.88</td>
<td>0.00</td>
<td><strong>85.29</strong></td>
</tr>
<tr>
<td><strong>Overall accuracy rate</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>IPBOW_WPCM</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coast</td>
<td>72.54</td>
<td>54.93</td>
<td>59.9</td>
<td>69.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>River/lake</td>
<td>49.55</td>
<td>49.55</td>
<td>41.6</td>
<td>28.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Forest</td>
<td>90.29</td>
<td>83.50</td>
<td>94.1</td>
<td>85.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Plain</td>
<td>64.89</td>
<td>58.02</td>
<td>43.8</td>
<td>62.6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mountain</td>
<td>84.36</td>
<td>74.30</td>
<td>84.3</td>
<td>77.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sky/cloud</td>
<td>85.29</td>
<td>85.29</td>
<td>100</td>
<td>76.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Overall accuracy rate</strong></td>
<td><strong>73.7%</strong></td>
<td><strong>65.3%</strong></td>
<td><strong>67.2%</strong></td>
<td><strong>66.7%</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Refer to Fig. 6, it can be seen that our approach works very well in the first three classes, but the performance degrades for the ‘open country’ scene class against (Gist+PCM) image representation. Furthermore, in Fig.8, the performance of our approach on ‘river/lakes’ and
scene classes have gained comparable results against other approaches and outperformed them on the other four classes. The overall performance results of our approach against other methods are shown in Fig. 7(a) and 7(b).

Fig. 6 The classification performance of IPBOW_PCM compared with different baseline methods for each scene class of Dataset 2. It is clear that in most scene classes IPBOW_PCM outperforms other methods. Gist+PCM features perform best for the open country scene class.

Our proposed approach is also compared with colour by design methods. We used rgbSIFT [47] features, extracted from training images, to generate integrated visual vocabularies. Each rgbSIFT feature is a vector of 384-D (SIFT features of 128-D are extracted from RGB image bands respectively). An image is then represented as a histogram counting the number of keypoints characterised by rgbSIFT that belongs to a specific vocabulary index. The average of the 10 accuracy rates using 10-fold cross validation is used to measure the performance of all experiments as mentioned in section 5.4. Fig. 9(a) compares the average classification rate of our proposed approach and rgbSIFT features tested on Dataset 1. The results confirm the effectiveness of our approach compared with rgbSIFT in image classification task. Our proposed model achieved better results on five image categories out of 6 while rgbSIFT features performed better in recognising Plains category.

Moreover, this paper investigates the influence of applying visual vocabularies generated from one image dataset to generate BOW from another image dataset. We hypothesise that visual words that exist in a specific image class are similar to those in another class with same visual semantic features. For example, ‘coasts’ class in Dataset 1 contains visual words
that are similar in semantic to visual words in 'coasts' class of Dataset 2. We used integrated visual vocabularies generated from Dataset 1 to index visual patches of images in Dataset 2. Fig. 6 and 7(a) show our findings, where VIBOW stands for BOW produced by applying integrated visual vocabularies generated from Dataset 1. Although classification results are lower than our approach it is found to be better than Gist features. The results show the plausibility of using visual vocabularies of one dataset to generate BOW for another dataset within the same domain. On the other hand, this is not applicable if images in datasets are different in their visual appearance and semantic content. In order to test the performance of our proposed approaches on more heterogeneous image categories we conducted extensive experiments on Datasets 3 and 4. Table 3 depicts the confusion matrix of our proposed model tested on 8-scene categories, i.e., Dataset 3. In terms of average classification accuracy rate, we achieved 88.28% which is comparable to other baseline approaches as depicted in table 4. The results show an improvement on the classification performance of IBOW over UBOW and how our weighting method influences the overall classification rate. Fig. 9(b) compares the average classification rate of our approach with rgbSIFT features.

**Table 3:**
Confusion matrix of eight class dataset, Dataset 3, based on our proposed approach. Rows and columns corresponds to correct and predicted classes respectively. The diagonal bold values are the average classification rate of each image category. The overall system accuracy is 88.28% and is comparable to other state-of-the-art image classification approaches.

<table>
<thead>
<tr>
<th></th>
<th>Coast</th>
<th>Forest</th>
<th>Highway</th>
<th>Inside city</th>
<th>Mountain</th>
<th>Open country</th>
<th>Street</th>
<th>Tall building</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coast</td>
<td>90.83</td>
<td>0.28</td>
<td>1.11</td>
<td>0.00</td>
<td>2.22</td>
<td>5.00</td>
<td>0.00</td>
<td>0.56</td>
</tr>
<tr>
<td>Forest</td>
<td>0.00</td>
<td>95.73</td>
<td>0.00</td>
<td>0.00</td>
<td>2.13</td>
<td>2.13</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Highway</td>
<td>6.15</td>
<td>0.00</td>
<td>81.54</td>
<td>3.46</td>
<td>1.15</td>
<td>2.31</td>
<td>2.31</td>
<td>3.08</td>
</tr>
<tr>
<td>Inside city</td>
<td>0.00</td>
<td>0.32</td>
<td>0.00</td>
<td>88.96</td>
<td>0.00</td>
<td>0.32</td>
<td>3.25</td>
<td>7.14</td>
</tr>
<tr>
<td>Mountain</td>
<td>2.14</td>
<td>0.80</td>
<td>0.00</td>
<td>0.00</td>
<td>90.91</td>
<td>5.61</td>
<td>0.53</td>
<td>0.00</td>
</tr>
<tr>
<td>Open country</td>
<td>7.07</td>
<td>5.61</td>
<td>0.73</td>
<td>0.00</td>
<td>6.10</td>
<td>80.49</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>Street</td>
<td>0.00</td>
<td>0.00</td>
<td>1.03</td>
<td>6.16</td>
<td>0.68</td>
<td>0.00</td>
<td>86.64</td>
<td>5.48</td>
</tr>
<tr>
<td>Tall building</td>
<td>2.53</td>
<td>0.28</td>
<td>0.00</td>
<td>3.93</td>
<td>1.40</td>
<td>0.56</td>
<td>0.56</td>
<td>90.73</td>
</tr>
</tbody>
</table>

**Table 4:**
Average classification accuracy rate (%) on Dataset 3 using universal and integrated visual vocabularies with different BOW configurations with/out pyramid colour moments.

<table>
<thead>
<tr>
<th></th>
<th>UBOW</th>
<th>UBOW_PCM</th>
<th>IBOW</th>
<th>IBOW_PCM</th>
<th>IPBOW</th>
<th>IPBOW_PCM</th>
<th>IPBOW_WPCM</th>
<th>Gist</th>
<th>rgbSIFT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average</td>
<td>(74.74)</td>
<td>(81.10)</td>
<td>(79.50)</td>
<td>(84.90)</td>
<td>(82.25)</td>
<td>(85.23)</td>
<td>(88.28)</td>
<td>(87.31)</td>
<td>(87.54)</td>
</tr>
</tbody>
</table>
For Dataset 4, we tested our proposed approach on grayscale images for both training and testing. In this case, our pyramidal colour moments represents only the information that are available in single image band i.e., no colour information. First and second moments are computed from all image sub-regions at pyramidal layout with $L=2$ and resulted in a vector of size 42-D. The confusion matrix, depicted in Table 5, illustrate the performance of our proposed approach. We achieved 81.03% overall classification rate which is higher than traditional BOW with universal vocabularies. We compared the performance of universal BOW and integrated BOW with different configurations. Results are reported in Table 6. Also, our approach is comparable to the results obtained by Battiato et al. [32] where they achieved 79.43% classification rate on the same dataset.
Fig. 9 Performance comparisons between our proposed approach (IPBOW\_WPCM) based on SIFT features and IBOW image representation based on rgbSIFT features [47] both tested on Dataset 1 (a) and Dataset 3 (b).

Fig. 10 For each scene concept (rows), (a) shows the average of UBO\_W histograms (b) the average of IBOW histograms and (c) shows sample images and their corresponding IBOP histograms. We can see that most image histograms tend to belong to their average histograms. Though, some classes get confused with other classes such as river/lakes and mountains since many mountain images contain water and vice versa.
Table 5: Confusion matrix of Dataset 4 based on our proposed approach. Rows and columns correspond to correct and predicted classes respectively. The diagonal bold values are the average classification rate of each image category. The overall system accuracy is 81.03% and is comparable to other state-of-the-art image classification approaches.

<table>
<thead>
<tr>
<th></th>
<th>Suburb</th>
<th>Coast</th>
<th>Forest</th>
<th>Highway</th>
<th>Inside city</th>
<th>Mountain</th>
<th>Open country</th>
<th>Street</th>
<th>Tall building</th>
<th>Office</th>
<th>Bedroom</th>
<th>Industrial</th>
<th>Kitchen</th>
<th>Living room</th>
<th>Store</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suburb</td>
<td>92.95</td>
<td>0</td>
<td>0.41</td>
<td>0</td>
<td>0.41</td>
<td>0.41</td>
<td>0.41</td>
<td>1.66</td>
<td>0</td>
<td>1.24</td>
<td>0</td>
<td>2.9</td>
<td>0</td>
<td>0</td>
<td>Store</td>
</tr>
<tr>
<td>Coast</td>
<td>0</td>
<td>91.39</td>
<td>0.28</td>
<td>1.94</td>
<td>0</td>
<td>1.39</td>
<td>4.44</td>
<td>0.28</td>
<td>0</td>
<td>0.28</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Store</td>
</tr>
<tr>
<td>Forest</td>
<td>0</td>
<td>0</td>
<td>96.95</td>
<td>0</td>
<td>0</td>
<td>0.91</td>
<td>2.13</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>Store</td>
</tr>
<tr>
<td>Highway</td>
<td>0</td>
<td>5.77</td>
<td>0</td>
<td>82.69</td>
<td>1.54</td>
<td>0.77</td>
<td>2.69</td>
<td>1.54</td>
<td>1.92</td>
<td>0</td>
<td>0</td>
<td>1.92</td>
<td>0</td>
<td>0</td>
<td>1.15</td>
</tr>
<tr>
<td>Inside city</td>
<td>0</td>
<td>0.32</td>
<td>0.32</td>
<td>0</td>
<td>73.7</td>
<td>0</td>
<td>2.92</td>
<td>5.84</td>
<td>0.32</td>
<td>0.32</td>
<td>4.87</td>
<td>1.95</td>
<td>1.3</td>
<td>8.12</td>
<td>Store</td>
</tr>
<tr>
<td>Mountain</td>
<td>0</td>
<td>3.21</td>
<td>2.14</td>
<td>0</td>
<td>0</td>
<td>90.11</td>
<td>4.01</td>
<td>0</td>
<td>0.27</td>
<td>0</td>
<td>0.27</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Open country</td>
<td>0.24</td>
<td>6.1</td>
<td>2.93</td>
<td>1.22</td>
<td>0</td>
<td>4.15</td>
<td>84.88</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.49</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Street</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1.71</td>
<td>6.16</td>
<td>0.34</td>
<td>0</td>
<td>81.16</td>
<td>4.79</td>
<td>0</td>
<td>0</td>
<td>3.77</td>
<td>0.34</td>
<td>1.71</td>
<td>Store</td>
</tr>
<tr>
<td>Tall building</td>
<td>0</td>
<td>1.69</td>
<td>0.56</td>
<td>1.12</td>
<td>2.25</td>
<td>1.4</td>
<td>0.56</td>
<td>1.12</td>
<td>84.83</td>
<td>0.28</td>
<td>0.84</td>
<td>3.37</td>
<td>0.56</td>
<td>0.28</td>
<td>1.12</td>
</tr>
<tr>
<td>Office</td>
<td>0</td>
<td>0.47</td>
<td>0</td>
<td>0.93</td>
<td>0</td>
<td>0</td>
<td>0.47</td>
<td>78.14</td>
<td>6.05</td>
<td>0.93</td>
<td>4.19</td>
<td>7.44</td>
<td>1.4</td>
<td>7</td>
<td>Store</td>
</tr>
<tr>
<td>Bedroom</td>
<td>0</td>
<td>1.85</td>
<td>0</td>
<td>0.93</td>
<td>1.39</td>
<td>0.46</td>
<td>0</td>
<td>3.24</td>
<td>4.63</td>
<td>57.87</td>
<td>5.56</td>
<td>3.24</td>
<td>18.06</td>
<td>2.78</td>
<td>Store</td>
</tr>
<tr>
<td>Industrial</td>
<td>0.64</td>
<td>2.57</td>
<td>0</td>
<td>1.93</td>
<td>5.14</td>
<td>3.86</td>
<td>0.64</td>
<td>1.29</td>
<td>6.43</td>
<td>0.64</td>
<td>2.25</td>
<td>64.63</td>
<td>0.64</td>
<td>2.89</td>
<td>6.43</td>
</tr>
<tr>
<td>Kitchen</td>
<td>0</td>
<td>0.48</td>
<td>0.48</td>
<td>6.19</td>
<td>0</td>
<td>0</td>
<td>2.38</td>
<td>5.24</td>
<td>4.29</td>
<td>2.86</td>
<td>65.71</td>
<td>9.52</td>
<td>2.86</td>
<td>5.19</td>
<td>Store</td>
</tr>
<tr>
<td>Living room</td>
<td>0</td>
<td>0.69</td>
<td>0</td>
<td>0</td>
<td>1.04</td>
<td>0</td>
<td>0.35</td>
<td>0.35</td>
<td>1.38</td>
<td>2.77</td>
<td>6.23</td>
<td>3.46</td>
<td>6.23</td>
<td>72.32</td>
<td>5.19</td>
</tr>
<tr>
<td>Store</td>
<td>0</td>
<td>0</td>
<td>0.95</td>
<td>0</td>
<td>5.4</td>
<td>0.95</td>
<td>0.95</td>
<td>1.27</td>
<td>0.95</td>
<td>0.95</td>
<td>1.9</td>
<td>1.59</td>
<td>3.81</td>
<td>81.27</td>
<td>Store</td>
</tr>
</tbody>
</table>
Table 6:
Classification results on Dataset 4 using universal and integrated visual vocabularies with different configurations of BOW to represent visual content.

<table>
<thead>
<tr>
<th>Scene</th>
<th>U-BOW</th>
<th>U-BOW_PCM</th>
<th>I-BOW</th>
<th>I-BOW_PCM</th>
<th>I-BOW_PCM</th>
<th>I-BOW_PCM</th>
<th>I-BOW_PCM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Suburban</td>
<td>87.97</td>
<td>88.38</td>
<td>92.95</td>
<td>95.02</td>
<td>91.70</td>
<td>92.95</td>
<td>92.95</td>
</tr>
<tr>
<td>Coast</td>
<td>76.11</td>
<td>78.33</td>
<td>84.72</td>
<td>86.39</td>
<td>86.39</td>
<td>88.61</td>
<td>91.39</td>
</tr>
<tr>
<td>Forest</td>
<td>91.46</td>
<td>94.21</td>
<td>91.16</td>
<td>93.29</td>
<td>92.68</td>
<td>95.43</td>
<td>96.95</td>
</tr>
<tr>
<td>Highway</td>
<td>64.23</td>
<td>70.77</td>
<td>79.62</td>
<td>81.54</td>
<td>78.85</td>
<td>80.77</td>
<td>82.69</td>
</tr>
<tr>
<td>Inside city</td>
<td>57.47</td>
<td>62.34</td>
<td>70.45</td>
<td>72.73</td>
<td>70.45</td>
<td>72.08</td>
<td>73.70</td>
</tr>
<tr>
<td>Mountain</td>
<td>75.13</td>
<td>78.88</td>
<td>87.17</td>
<td>88.77</td>
<td>87.43</td>
<td>87.17</td>
<td>90.11</td>
</tr>
<tr>
<td>Open country</td>
<td>60.00</td>
<td>64.88</td>
<td>76.83</td>
<td>79.76</td>
<td>78.05</td>
<td>79.27</td>
<td>84.88</td>
</tr>
<tr>
<td>Street</td>
<td>63.36</td>
<td>71.58</td>
<td>75.34</td>
<td>80.14</td>
<td>76.37</td>
<td>78.08</td>
<td>81.16</td>
</tr>
<tr>
<td>Tall building</td>
<td>60.11</td>
<td>65.17</td>
<td>77.25</td>
<td>80.06</td>
<td>80.62</td>
<td>83.43</td>
<td>84.83</td>
</tr>
<tr>
<td>Office</td>
<td>68.84</td>
<td>73.95</td>
<td>77.67</td>
<td>80.47</td>
<td>79.07</td>
<td>78.14</td>
<td>78.14</td>
</tr>
<tr>
<td>Bedroom</td>
<td>31.02</td>
<td>35.19</td>
<td>55.36</td>
<td>56.48</td>
<td>55.56</td>
<td>57.87</td>
<td>57.87</td>
</tr>
<tr>
<td>Industrial</td>
<td>39.87</td>
<td>44.05</td>
<td>52.41</td>
<td>56.59</td>
<td>55.63</td>
<td>57.23</td>
<td>64.63</td>
</tr>
<tr>
<td>Kitchen</td>
<td>46.19</td>
<td>50.00</td>
<td>55.71</td>
<td>61.43</td>
<td>62.38</td>
<td>64.29</td>
<td>65.71</td>
</tr>
<tr>
<td>Living room</td>
<td>50.52</td>
<td>54.33</td>
<td>57.09</td>
<td>61.25</td>
<td>62.98</td>
<td>67.82</td>
<td>72.32</td>
</tr>
<tr>
<td>Store</td>
<td>60.63</td>
<td>67.94</td>
<td>67.94</td>
<td>74.92</td>
<td>69.84</td>
<td>78.10</td>
<td>81.27</td>
</tr>
</tbody>
</table>

| Accuracy (%) | 63.08 | 67.56 | 74.34 | 77.44 | 76.05 | 78.31 | 81.03 |

6. Conclusion

In this paper, we have presented a unified framework to classify natural scene images into one of a number of predefined scene classes. Our work is based on the bag of visual words (BOW) image representation scheme. The proposed framework improved BOW image representation model in two ways: (1) It generates discriminative visual vocabularies by integrating visual vocabularies learned from class-specific data; (2) It fuses image colour information with intensity-based BOW using a spatial pyramid layout. The fusion has been done using the keypoints density-based weighting (KDW) method. One of the drawbacks of using a universal visual vocabulary is that similar visual patches may be clustered into different clusters and thus lose their information. We investigated different configurations of BOW and compared their performance on three natural scene datasets. Also, we made an improvement to the well-known intensity-based Gist features by adding pyramidal colour moments in an early fusion approach. We have shown that integrated BOW (I-BOW) and pyramidal colour moments (PCM) weighted on spatial pyramid layout (IPBOW+WPCM) outperformed other baseline approaches. Experimental results showed that building integrated visual vocabulary provides better performance than the conventional universal visual vocabulary. Moreover, it is obvious that building integrated visual vocabulary is faster than universal visual vocabulary, since the clustering algorithm will deal with less feature vectors and it will probably converge faster. We have also shown that visual vocabularies of one dataset could be used to generate BOW for another dataset with acceptable classification performance. We did not focus on weighting BOW visual words since our initial experiments has shown that BOW weighting techniques such as term-frequency inverse document frequency (TF-IDF) has gained lower classification performance in our datasets [26]. Further investigation needs to be done to include feature selection algorithms [48] to reduce the influence of noisy data. We believe that building visual vocabularies from pyramid regions could generate better visual vocabularies, leading to more accurate classification and reduces clustering time. Investigating the effect of colour casting, due to different acquisition conditions, on image classification task is of possible future work.
Acknowledgements

The first author acknowledges the financial support received from the Applied Science University in Jordan. The authors would like to thank Dr. Julia Vogel for providing us access to the natural scene image dataset and for valuable discussion. Also, we would like to thank Dr. Paul Trundle for reviewing our paper and for his valuable comments, and the reviewers for their constructive feedback on our work.

References


