Natural Scene Image Annotation Using Local Semantic Concepts and Spatial Bag of Visual Words

Yousef Alqasrawi
Faculty of Information Technology, Applied Science Private University, Amman, Jordan

Abstract: The use of bag of visual words (BOW) model for modelling images based on local invariant features computed at interest point locations has become a standard choice for many computer vision tasks. Visual vocabularies generated from image feature vectors are expected to produce visual words that are discriminative to improve the performance of image annotation systems. Most techniques that adopt the BOW model in annotating images declined favorable information that can be mined from image categories to build discriminative visual vocabularies. To this end, this paper introduces a detailed framework for automatically annotating natural scene images with local semantic labels from a predefined vocabulary. The framework is based on a hypothesis that assumes that, in natural scenes, intermediate semantic concepts are correlated with the local keypoints. Based on this hypothesis, image regions can be efficiently represented by BOW model and using a machine learning approach, such as SVM, to label image regions with semantic annotations. Another objective of this paper is to address the implications of generating visual vocabularies from image halves, instead of producing them from the whole image, on the performance of annotating image regions with semantic labels. All BOW-based approaches as well as baseline methods have been extensively evaluated on 6-categories dataset of natural scenes using the SVM and KNN classifiers. The reported results have shown the plausibility of using the BOW model to represent the semantic information of image regions and thus to automatically annotate image regions with labels.

Keywords: image annotation, natural scenes, bag of visual words, visual vocabulary, semantic modelling, semantic concepts.

1. INTRODUCTION

The availability of low-cost digital cameras and mobile devices and their attractive ability to upload and share photos to photo-sharing websites have led to a dramatic increase in the size of image collections. For efficient use of such large image collections, image annotation, searching, browsing and retrieval techniques are required [1-3]. Having images automatically annotated with labels, images can be retrieved in the same way as text documents.

Automatic image annotation is considered a promising approach to bridge the semantic gap between low-level image features and high-level semantics through extracting semantic features or concepts from images using machine learning techniques [4]. Image annotation helps to understand image visual content. In this paper, semantics are text words used in describing the visual content of an image.

In general, automatic image annotation systems learn the semantic concepts of a set of images by associating low-level features to high-level concepts. The trained system is then used to predict a set of semantic concepts to annotate previously unseen test images. Two major aspects are related with automatic image annotation: feature extraction and semantic concepts learning. Feature extraction can be performed at two levels: image level and region level. At image level, the visual content of an image is described by global features such as color histograms. However, global features do not describe different parts of an image. At region level, images are initially partitioned into rectangular blocks or regions via fixed size grid or using a segmentation algorithm. Each block or region is represented by features extracted from it, such as color and texture. Thus, an image with n blocks or regions is represented by a set of N feature vectors. Then, automatic image annotation system is trained, using a machine learning approach such as SVM, to assign each feature vector to a pre-defined category. Automatic image annotation is not an easy task as it is computationally expensive [4] and produce unrelated semantically similar regions [5]. For this reason, this work avoids image segmentation and adopt grid-based image division.

Recently, many approaches utilized local invariant features [6], local semantic concepts [7], and the bag of visual words (BOW) in many computer vision tasks and have shown improvements in the performance of scene image classification [6, 8, 9, 10-12]. There are two key phases to build an image classification system that use BOW technique. The first phase extract local features that characterize the visual content of the image. The work that is presented 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 proposed to advance the performance of the traditional BOW model. It can be classified into three main types. The first type tries to improve the visual vocabulary construction [20-24]. The second type integrates multiple features with weighting techniques [11, 13-16]. In the third type, different techniques are proposed that incorporate spatial information with the BOW and have shown improvement in the performance of scene classification tasks [17-20].
Despite that most of these approaches achieved encouraging results in many computer vision tasks, no one claimed that his approach is the best. The complexity nature of natural scenes and different arrangements of items in image content means that, images with visually similar contents from two different classes are usually mis-categorised, (e.g., ambiguity in visual appearance between coasts and river/lake categories). We suggest that many of these problems could be better addressed by constructing a unified technique that utilizes knowledge from discriminative visual vocabularies, multiple image features and their spatial arrangements [21].

This paper proposes a framework for automatically annotating image regions with semantic keywords from a predefined vocabulary of words. The framework utilizes a hypothesis presented in [22]. The hypothesis claims that local semantic concepts in natural scenes correlates with local keypoints found in image regions. The proposed hypothesis investigates the correlation between the distribution of local semantic concepts and local keypoints detected in image regions and annotated with these semantic concepts. Based on this hypothesis, concept-based bag of visual words is proposed to represent the visual content of image regions. In this paper, concept-based bag of visual words is further improved by proposing to build visual vocabularies from image halves. Extensive experiments are conducted over a natural scene dataset with six categories.

2. RELATED WORK

Due to the semantic gap between image features and human interpretation of images, automatic image annotation comes as a suggested solution to bridge this gap. It works by extracting semantic keywords from images by using machine learning approaches [4].

In general, automatic image annotation systems consider learning the semantic concepts of a set of images by associating low-level features with high-level concepts. The trained system is then used to predict a set of semantic concepts to annotate previously unseen test images. There are many different techniques to carry out image annotation. According to [23], image annotation can be classified into three types:

- Classification-based annotation: in this type, image annotation is treated as classification problem using multiple classifiers, such as SVM, to learn image semantic concepts.
- Probabilistic-based annotation: in this type, probabilistic models are developed to estimate the relation between visual image contents and semantic concepts.
- Retrieval-based annotation: in this type, semantic concepts of images that are semantically relevant to the query image are employed to annotate the query image.

For automatic image annotation, higher level semantics can be learned from image sample or image regions represented with low level features. New images are labeled with semantic labels using the trained model. Automatic image annotation can be accomplished in three ways:

- Annotate images with single label.
- Annotate images with multi labels.
- Annotate images with metadata from the web.

In the first approach, image annotation is considered as a binary classification problem. Image features are first provided to a binary classifier, such as support vector machines, artificial neural network, or decision tree. The trained classifier is then used to label a new image with a semantic label. Support vector machines are the common choice for many classification problems, such as image classification, object recognition and text classification [24-27]. It has an advantage over other classifiers that it achieves optimal class boundaries by finding the maximum distance between the hyperplane of classes. The hyperplane is trained to separate the samples of one class from other class. For multi-class classification and annotation problem, two common schemes are used: one-vs.-one and one-vs.-all. Thus, multiple SVMs are used to learn each class individually such that a test image is labeled with a decision fused from decisions of all classifiers.

Chapelle et al. [25] proposed a framework to train 14 SVM classifiers to learn 14 image semantic concepts using one-vs.-all paradigm. Only color histogram is used to represent visual image content at three color channels of HSV color space. No image segmentation is employed in their work. Each SVM is trained on a particular semantic concept where all images belong to the same semantic concept are regarded as positive samples while the others are considered as negative. A new image is classified based on a voting approach to select the classifier with maximum probability value.

To improve the classification power of SVM, Goh et al. [26] proposed a three-level classification scheme using three different sets of SVMs, one-class, two-class, and multiclass, to annotate images with semantic classes. They developed a confidence-based approach to fuse the output of classifiers propagated at different levels based on the difference between the highest two output decisions. An image is assigned to the semantic class with highest cumulative confidence. Image annotation is performed at image level without any segmentation.
Cusano et al. [24] proposed a framework to annotate image regions with seven semantic concepts: sky, skin, vegetation, snow, water, ground, and buildings. Their image annotation framework uses multi-class SVM to assign/classify image regions into one of the seven semantic concepts. All images used in their experiments are divided into overlapped tiles around each pixel and are labeled with the seven semantic concepts. Image regions are described by color histogram in the HSV color space. The classifiers are trained on random sample of tiles chosen from each semantic concept. To annotate the whole image, they proposed a threshold-based strategy to accept or reject the semantic concept based on the decision obtained from SVM classifiers.

Multilevel SVM sets are also explored in [27] to annotate images with semantic concepts using both global features and local features extracted from the whole image and image regions, respectively. For region-based features, images are divided into 5 non-overlapping blocks. Both features are used in two different sets of SVMs. This approach also differs from the previous one in the way how predictions are fused to get the semantic label for the new image.

Single label image annotation approaches improve image retrieval by just typing the keywords related to the semantic concepts. So, there is no need to do image matching.

Multiple image annotation refers to approaches that assign more than one label to the whole image. To make images understandable by humans it is important to represent the semantic structure of images for semantic image annotation. The choice of the semantic keywords or concepts is dependent on the domain, user needs and the application. For multiple image annotation, this work focuses on using constrained vocabulary of a small size to locally annotate natural scene images with semantic concepts.

As has been mentioned earlier, single label image annotation is a classification problem. In contrast, multiple labels image annotation seeks to annotate image with multiple keywords, either to image regions or part of them. Barnard and Forsyth [28] proposed a generative hierarchical clustering model to assign multiple labels to an image through mapping labels to regions. Their model learns the joint statistics of words and regions and build word-region co-occurrence matrix to attach words to regions. Instead of using hierarchical clustering, Duygulu et al. [29] proposed a translation model to map regions to words. Both models used normalized cuts algorithm [30] to extract regions from images.

Aksoy et al. [31] addressed the relations between labeled regions. They proposed a visual grammar to represent the relation between image regions to reduce the semantic gap. Bayesian classifiers are used to label image regions then the grammar model is used to classify images into scene categories.

Mojsilovic et al. [32] addressed the problem of the semantic gap by introducing semantic indicators (sky, water, skin,...etc) based on experiments with human subjects. Global and local visual features are extracted from images and quantized into regions. These regions are names by human subjects. These semantic indicators are used then for scene categorization.

Recently, Makadia et al. [33] argue that most complicated state-of-the-art automatic image annotations techniques lack of comparisons with simple baseline measures to justify the need for such complicated models. They proposed a set of baseline methods, using color and texture, to automatically annotate images with keywords based on nearest neighbors. For color features, color histograms are generated from the three components of images represented in RGB, HSV and LAB color spaces, respectively. Gabor filters and Haar wavelets are used to represent image textures. To calculate nearest neighbors, three distance measure are jointly combined and evaluated using KL-divergence, L1-distance, and L2-distance. For test images, predicted labels are assigned to whole image without specifying to which region the labels refer to. Many surveys on automatic image annotation are available in literature and most recently are carried out by [4, 34].

3. NATURAL SCENE DATASET

This paper studies the task of automatically annotating image regions with local semantic keywords or concepts. To accomplish this task, we use a natural scene dataset of 700 images categorized into six scene categories [5]. The six categories are river/lakes, coasts, plains, forests, mountains, and sky/clouds. Authors of this dataset manually annotated image regions with one of the nine semantic concepts: sky, water, grass, trunks, foliage, field, rocks, flowers and sand. For validation and experimental work, images were partitioned into a matrix of (10x10) regions which produce 70000 regions (see Figure 1). In the original work of Vogel and Schiele [5], image regions with multiple semantic concepts were excluded in training and testing local semantic classifiers. Images in the dataset are organized into two folders. The first folder contains natural scene images. The second folder contains text files, each file contains semantic concepts of the corresponding natural scene images in the first folder. Assume that I=\{i_1,i_2,...,i_N\} be the set of images from the first folder and T=\{t_1,t_2,...,t_N\} is the set of text files, where N=700. A text file t_j contains a list of annotations for an image i_j, where j=1,2,...,N. The distribution of local semantic concepts in each scene category is depicted in Table 1.
Table 1: Sizes of the nine local concept classes located in each scene category.

|          | Sky   | Water | Grass | Trunks | Foliage | Field | Rocks | Flowers | Sand |
|----------|-------|-------|-------|--------|---------|-------|-------|---------|-------|
| Coasts   | 2960  | 4526  | 430   | 32     | 1284    | 194   | 1922  | 46      | 825   |
| Rivers/lakes | 1728  | 2826  | 273   | 82     | 2629    | 204   | 1553  | 12      | 0     |
| Forests  | 335   | 39    | 465   | 1419   | 6154    | 309   | 47    | 31      | 0     |
| Plains   | 2879  | 14    | 1008  | 36     | 904     | 2649  | 330   | 1894    | 897   |
| Mountains| 4416  | 54    | 649   | 56     | 2335    | 735   | 7401  | 62      | 23    |
| Sky/Clouds| 2978  | 34    | 78    | 0      | 33      | 97    | 57    | 0       | 0     |
| # of image regions | 15290 | 7293 | 3503 | 1025 | 13709 | 4188 | 11310 | 2049  | 1745 |
| OVERALL  |       |       |       |       |         |       |       |         | 60718 |

4. LOCAL SEMANTIC CONCEPTS VS. LOCAL KEYPOINTS

This section reviews our hypothesis proposed in [22] which investigates the correlation between local semantic concepts and local keypoints, based on their distributions over all image regions. Local semantic concepts are labels assigned to image regions. Using (10x10) fixed grid layout, an image is divided into 100 image regions. Low-level features, such as color and texture, are normally used to represent the visual content of image regions. These features are not invariant to different changes of the visual contents. Thus, to provide invariance to changes in illumination, rotation, etc., interest points can be used [35]. Interest points or local keypoints correspond to image structures that are considered important.

Using sparse representation approach, a keypoint detector, such as DoG detector [6], locates local keypoints that contain distinctive information in their surrounding area and should be invariant to geometric transformations. These keypoints are then described using robust and informative features such as SIFT.

Having the ground truth semantic labels provided with the natural scene dataset, and the local keypoints provided by an interest point detector, two facts can be drawn:

1. Each region in an image can be described by two corner locations; the top left corner \((x_1, y_1)\) and the bottom right corner \((x_2, y_2)\). The \((x_2-x_1) \times (y_2-y_1)\) is the size of the given region.

2. Applying Keypoint detector to an image would generates a list of coordinates of all the keypoints detected in the image. The detector also generates more features that describes the area around each detected keypoint, such as orientation and scale. We are interested on the coordinates of keypoints.

Now, we have data about the coordinates of all image regions and coordinates of all detected keypoints. The assumption is; the distribution of semantic keywords, used to annotate regions, may correlates with the distribution of detected keypoints in a particular scene category. This hypothesis can be justified as: for each natural scene category and semantic concept, we count the number of keypoints with coordinates located in the areas of all image regions labelled with the same semantic concept. If the distribution of semantic concepts is similar or close to the distribution of local keypoints then there is a possible correlation between them.
Figure 2 shows the distribution of local keypoints located in image regions of each semantic concept and over all scene categories.

The correlation between semantic concepts and local keypoints is shown in Figure 3. Also, the distributions of local semantic concepts and local keypoints over each of the six natural scenes are depicted in Figure 4 and Figure 5. The distribution of local semantic concepts at the upper and lower halves of all images is shown in Figure 6 whereas the distribution of local keypoints in image regions labelled with the semantic concepts is shown in Figure 7.
Figure 5: Distribution of keypoints located in regions of each semantic concept and over each scene category.

Figure 6: Distribution of image regions located in the upper and lower halves of images.

Figure 7: Distribution of local keypoints found in image regions in the upper and lower halves of images.
5. IMAGE ANNOTATION FRAMEWORK

In this section, a framework for annotating image regions with local semantic concepts based on bag of visual words is presented. The framework consists of two parts as depicted in Figure 8.

In the first part, DoG detector and SIFT descriptor are employed to find and represent local keypoints in all images of the dataset. Each SIFT descriptor is a vector of size 128-D. For each natural scene category, k-means algorithm, based on the Euclidean distance metric, is applied to all descriptors to build visual vocabularies. These vocabularies are then aggregated to form an integrated visual vocabulary of size (K×M) where K is the vocabulary size and M is the number of scene categories. We followed the same approach mentioned in [21] in building visual vocabularies for the six scene categories dataset mentioned in Section 3.

In the second part, image regions are represented using BOW, i.e., a histogram is generated for each region where each bin corresponds to the frequency of occurrence of each visual word in that image region. In this paper, bag of visual words (BOW) generated from an image region is called Concept-based Bag of Visual Words (CBOW). After representing all image regions using concept-based bag of visual words, and noting that image regions are labeled with semantic concepts, the CBOW histograms are used to learn classifiers to label image regions located in the test dataset with local semantic concepts.

Figure 8: Flow diagram of the proposed image annotation framework.

5.1. Image visual vocabulary construction

The visual vocabulary on which CBOW rely on is of great importance. Building visual vocabularies from each scene category have shown improvements in the performance of natural scene classification task. To apply the same approach in building visual vocabularies to local semantic concepts, the number of visual vocabularies becomes larger. This is because of the fact that the number of local semantic concepts are usually larger than the number of scene categories these semantic concepts
belong to. Therefore, it is worth to explore the influence of applying visual vocabularies obtained from natural scene categories (globally obtained from natural categories) to map SIFT features located in image regions (locally used to represent regions) to the indices of visual words and thus build CBOWs histograms.

Another choice of improving the power of visual vocabulary is to incorporate spatial information about local keypoints when clustering their features. Natural scene images, such as coast, contain semantic concepts that usually appear in common places. For example, the sand concept can be found at the bottom of an image whereas the sky and water concepts appear at the top. Also, building visual vocabularies using k-means algorithm may lead to group dissimilar local keypoints in the same cluster. To avoid this problem, images are partitioned into two halves of equal size: upper and lower. An image of size W×H is divided into two halves, each has a dimension of (W/2)×(H/2). In this case, two integrated visual vocabularies are generated: Upper integrated visual vocabulary and Lower integrated visual vocabulary.

Upper integrated visual vocabulary is generated by clustering SIFT features located at the upper half of all training images. Lower integrated visual vocabulary is generated by clustering SIFT features located at the lower half of all training images. This approach has two main advantages:

1. Create more representative visual vocabularies about natural scenes that benefits from the spatial information of the keypoints and at the same time reduce ambiguities between clusters resulted from the clustering step.
2. Reduces the clustering time while maintaining or improving the discriminative power of the clusters. Preliminary experimental work has shown that quantizing local features at different spatial levels reduce the clustering time as well as improving the discriminative power of visual words or clusters. But, it increases the dimensionality of the BOW.

Comparing the annotation performance of the proposed framework using universal visual vocabulary, integrated visual vocabulary and upper and lower visual vocabularies allows us to analyse indirectly the dependence of bag of visual words representation on spatial locations of local keypoints. Also, to study the influence of using global visual vocabularies to represent image regions.

5.2. Image region representation

Two types of features are used to represent the visual content of image regions. In the first type, image regions are represented using concept-based bag of visual words (CBOW). These CBOW are generated using different visual vocabularies, as described in the previous section. In the second type, multiple low-level features are used as baseline methods and also to expand the discriminative power of CBOW.

5.3. Concept-based bag of visual words (CBOW)

In this paper, bag of visual words BOW generated at region level are called concept-based bag of visual words (CBOW). Each image in the dataset is spliced into 10x10 sub-regions. Having images portioned into 10x10 regions, the next step is to build CBOW histogram for each image region. To do that, first, local keypoints are automatically detected in the whole image. Second, local keypoint features are extracted from each region which is defined around those keypoints. Third, construct visual vocabularies from detected descriptors. Fourth, map local keypoint to the index of the most similar visual word and their frequencies are counted to construct the CBOW for each image region r.

There are four kinds of visual vocabularies constructed and used in this paper to build the CBOW from image regions. We aim to compare and analyze their performances to annotate image regions. This is mentioned in the experimental work section. These visual vocabularies are:

- Universal visual vocabulary: To obtain the visual vocabulary, we use feature vectors (SIFT features) stored in image features database. All feature vectors from all training images on the dataset are quantized, using the k-means algorithm, to obtain centroids or clusters. These centroids represent visual words. Theses visual words constitute the universal visual vocabulary. This visual vocabulary contains 200 visual words.

- Integrated visual vocabulary: For the integrated visual vocabulary, SIFT features from all training images of each scene class are clustered into visual words. This visual vocabulary contains 1200 visual words (number of scene classes × 200 clusters).

- Universal Upper and Lower visual vocabulary: the upper visual vocabulary contains 200 visual words and the lower visual vocabulary contains 200 visual words.

- Integrated Upper and Lower visual vocabulary: the upper visual vocabulary contains 1200 visual words and the lower visual vocabulary contains 1200 visual words.
Let \( I = \{i_1, i_2, ..., i_N\} \) be the set of all images in the dataset and that each image \( i_k \) is divided into 10x10 sub-images such that \( u = [u_1, u_2, ..., u_{100}] \), where \( u_r \) is the \( r \)-th region in image \( i_k \). In Algorithm 1, universal visual vocabulary generated for the natural scene dataset with 6 scene categories is used to build CBOWs from image regions. This vocabulary has been used in [21] for image classification. This vocabulary contains 200 visual words generated by clustering all SIFT features extracted from training images.

Algorithm 1: An algorithm to build CBOW from images regions using universal visual vocabulary.

**Input:** Use the universal visual vocabulary \( V \) to construct CBOW from local image regions as follows:

**For all images** \( I = \{i_1, i_2, ..., i_N\} \) in the dataset **Do**:

a. Apply DoG feature detection technique to locate interest points in image \( i_k \).

b. Extract SIFT features (128-D) from each located keypoint.

c. For each image region \( \hat{i}_k \) in the image \( i_k \) **Do**:

   Quantize all SIFT descriptors located in region \( \hat{i}_k \) into one of the \( M \) visual words, where \( M \) is the size of the visual vocabulary \( V \). An image region \( \hat{i}_k \) is then represented as a histogram \( \hat{h}_k \) of the frequencies of visual words.

end

**Output:** a collection of \( \hat{h}_k \) histograms, where each \( \hat{h}_k \) is the CBOW for image \( k \) at region \( r \).

In Algorithm 2, the integrated visual vocabulary generated by clustering all SIFT features of training images over each scene category is used to build CBOW. This vocabulary has been used in [21] for the same dataset. It is important to study the influence of building visual vocabularies from local features located in parts of an image rather the whole image and compare their performances with the traditional visual vocabularies.

Algorithm 2: An algorithm to build CBOWs from images regions using integrated visual vocabulary.

**Input:** Use the integrated visual vocabulary \( V \) to construct CBOWs from local image regions as follows:

**For all images** \( I = \{i_1, i_2, ..., i_N\} \) in the dataset **Do**:

a. Apply DoG feature detection technique to locate interest points in image \( i_k \).

b. Extract SIFT features (128-D) from each located keypoint.

c. For each image region \( \hat{i}_k \) in the image \( i_k \) **Do**:

   Quantize all SIFT descriptors located in region \( \hat{i}_k \) into one of the \( M \) visual words, where \( M \) is the size of the visual vocabulary \( V \). An image region \( \hat{i}_k \) is then represented as a histogram \( \hat{h}_k \) of the frequencies of visual words.

end

**Output:** a collection of \( \hat{h}_k \) histograms, where each \( \hat{h}_k \) is the CBOW for image \( k \) at region \( r \).
As mentioned earlier, clustering local descriptors located at the upper halves of images may generate a better quality clusters, i.e., their members are more semantically similar. It aims to reduce inter-class similarity and increase intra-class similarity throughout clustering features at image parts.

Although images can be divided to any number of tiles, only two parts from images are considered in this work; the upper half and the lower half. In other words, 50% of the image content, the upper part, is used to build the upper visual vocabulary whereas the other 50%, the lower part, is used to build the lower visual vocabulary. Upper and lower visual vocabularies are generated at two levels. The first level considers building upper and lower visual vocabularies from SIFT descriptors of all scene categories whereas in the second level visual vocabularies are generated from SIFT descriptors of each scene category, as shown in Figure 9.

Figure 9: Upper and lower visual vocabularies construction. Images in the middle are samples from all training images used in the construction process.
Algorithm 3: An algorithm to build CBOWs from images regions, at upper halve, using universal visual vocabulary.

**Input:** Use the *upper universal visual vocabulary* $V$, to construct CBOWs from local image regions as follows:

For all images $I = \{i_1, i_2, ..., i_N\}$ in the dataset Do:

- a. Apply DoG feature detection technique to locate interest points in image $i_k$.
- b. Extract SIFT features (128-D) from each located keypoint.
- c. For each image region $\hat{t}_k$ in the upper halve of image $i_k$, where $r=1,2, ..., 50$ Do:
  
  Quantize all SIFT descriptors located in region $\hat{t}_k$ into one of the $M$ visual words, where $M$ is the size of the visual vocabulary $V$. An image region $\hat{t}_k$ is then represented as a histogram $\hat{h}_k$ of the frequencies of visual words.

end

**Output:** a collection of $\hat{h}_k$ histograms, where each $\hat{h}_k$ is the CBOW for image $k$ at region $r$.

Having all visual vocabularies generated from upper and lower halves of images, the next step is to use them to build CBOW from image regions. Algorithms 3 and 4 illustrate the required steps to build CBOW from images regions at upper halves of images.

Algorithm 4: An algorithm to build CBOWs from images regions, at upper halve, using integrated visual vocabulary.

**Input:** Use the *upper integrated visual vocabulary* $V$, to construct CBOWs from local image regions as follows:

For all images $I = \{i_1, i_2, ..., i_N\}$ in the dataset Do:

- a. Apply DoG feature detection technique to locate interest points in image $i_k$.
- b. Extract SIFT features (128-D) from each located keypoint.
- c. For each image region $\hat{t}_k$ in the upper halve of image $i_k$, where $r=1,2, ..., 50$ Do:
  
  Quantize all SIFT descriptors located in region $\hat{t}_k$ into one of the $M$ visual words, where $M$ is the size of the visual vocabulary $V$. An image region $\hat{t}_k$ is then represented as a histogram $\hat{h}_k$ of the frequencies of visual words.

end

**Output:** a collection of $\hat{h}_k$ histograms, where each $\hat{h}_k$ is the CBOW for image $k$ at region $r$. 

11
To avoid repeating the same algorithms for the lower halves, lower visual vocabularies and image regions at the lower halves (r=51, 51,…, 100) are replaced with the corresponding ones used in Algorithms 3 and 4.

5.4. Local from Global CBOW

Constructing CBOW from visual vocabularies generated from each scene category allows us to study the relationship between natural scene categories and local semantic concepts. The principle of building integrated visual vocabulary is to quantize local features from each scene category. Thus, in the case of local semantic concepts, integrated visual vocabularies should be generated from local features located in image regions labeled with these semantic concepts.

The main aim of this section is to analyze the relationship between visual vocabularies, generated from scene categories, and CBOWs histograms generated from local image regions. The analysis is based on the distribution of all CBOWs histograms generated from image regions labeled with each of the semantic concepts. Also, we show that integrated visual vocabularies are more suitable to represent local features in image regions rather than the universal visual vocabulary. Furthermore, the distributions of CBOWs generated from upper and lower integrated visual vocabularies are analyzed.

To analyze the relationship between a visual vocabulary and the semantic concepts, all CBOW histograms generated from a visual vocabulary are summed up. Given that all image regions are annotated with semantic concepts and that each image region is represented by a CBOW then it is possible to sum up all CBOWs histograms for each semantic concept. For example, the distribution of all CBOW generated from image regions labeled with the semantic concept sky, using universal visual vocabulary, is shown in Figure 10. The same figure shows the distributions of all CBOW for water, grass, trunks, foliage, field, rocks, flowers, and sand. These distributions do not show any relation of the natural scene categories with the local semantic concepts.

In contrast, Figure 11 shows CBOW distributions of image regions using integrated visual vocabulary. The integrated visual vocabulary contains 1200 visual words in which each 200 visual words represent 200 clusters generated from clustering all local descriptors located in images of specific scene category.

Figure 10: Sum of the CBOW histograms obtained using universal visual vocabulary at image level.

Figure 11: Sum of the CBOW histograms obtained using integrated visual vocabulary at image level.
The first 200 visual words represent clusters for the natural scene coasts, and the other 1000 visual words are clusters for the natural scene categories river/lakes, forests, plains, mountains and sky/clouds, respectively. This figure depicts interesting relationship between local semantic concepts and the natural scene categories. For example, large numbers of the local keypoints, found in image regions labeled with the semantic concept sky, are assigned to the last 200 visual words of the CBOW histograms. The last 200 visual words are clusters belonging to the scene category sky/clouds. And it is usual to see sky areas in sky/clouds natural scenes. Another interesting indication in the figure is the plot for the semantic concept water. For this concept, large numbers of the keypoints are assigned to the first 200 visual words which actually belong to the natural scene coasts, and this is natural to have water in natural scene coasts. The concept water is also available in the river/lakes scene category depicted in the same plot. The same figure contains the distributions for the semantic concept sand. Many of the keypoints are located in the first 200 visual words and visual words (600-799).

To this end, there is a relationship between natural scene categories and the semantic concepts they contain which are analyzed using the distributions of CBOW obtained using the integrated visual vocabulary and as has been discussed above. Moreover, these relationships confirm the hypothesis, presented earlier, in a way that image regions labeled with a semantic concept contain keypoints that are more likely to appear in the distributions of the CBOWs histograms for that semantic concept. Furthermore, Figure 11 shows the plausibility of using visual vocabularies, generated from local descriptors at image level, to build CBOWs. This will be confirmed in the experimental results section.

In the case of building visual vocabularies from local keypoints descriptors located at the upper and lower halves of images, the same approach is followed to analyze their relationship with the local semantic concepts. For the upper halves of images, two visual vocabularies are generated: universal and integrated visual vocabularies from which CBOW histograms are generated. For the lower halves of images, two visual vocabularies are also constructed: universal and integrated visual vocabularies.

![Figure 12: Sum of the CBOW histograms obtained using universal visual vocabulary at the upper half of the images.](image)

![Figure 13: Sum of the CBOW histograms obtained using universal visual vocabulary at the Lower half of the images.](image)

Similar to using universal visual vocabulary in Figure 10, Figure 12 and Figure 13 show that the distributions of the CBOW over the nine semantic concepts do not show any indication from the distributions of the local semantic concepts on the upper and lower halves of images. Therefore no relationship can be inferred from both halves, though for image annotation, the performance of upper and lower CBOWs may still outperform CBOW at whole image level. This will be discussed in the experimental work section.
To reduce the ambiguity between visual words and improve their discriminative power, integrated visual vocabularies are generated from image halves. This is confirmed by analyzing the distributions of local keypoints in the CBOW over the nine semantic concepts and at both image halves. In Figure 14, the semantic concept grass is usually appear in the lower halves of images but could also appear at the upper halve of the image, such as images of plains and mountains scenes. This is shown in the grass plot, where many of the local keypoints are indexed to the visual words (600-799) and (800-1000). These visual words represent clusters of plains and mountains scene categories.

For the water semantic concept, many of the keypoints are assigned to the visual words of the natural scene coasts. But, for the same semantic concept, water appears more in the lower halves of coasts and river/lakes scenes categories, as shown in Figure 15.

![Figure 14: Sum of the CBOB histograms obtained using integrated visual vocabulary at the Upper half of the images.](image1)

![Figure 15: Sum of the CBOB histograms obtained using integrated visual vocabulary at the Lower half of the images.](image2)

It is interesting to use Figure 14 and Figure 15 to analyze the differences between the distributions of the CBOW at upper and lower halves. For example, local keypoints located in image regions of the sand semantic concept appear in the lower halve more than the upper halve which gives an indication of the importance of building multi-level visual vocabularies from image halves.

5.5. Multiple features

Motivated by the importance of using color information and textural features to describe the visual contents of natural scene images, this section presents a number of features that are used to improve the performance of CBOW image region representation and therefore will improve the performance of natural scene annotation task. Beside the CBOW, three types of features are chosen in this work. The first two features are devoted to represent color whereas the third feature is devoted to represent texture. Color histogram and color moments are used to extract color information from image regions. Discrete Wavelet Transform (DWT) is used to extract the textural features from image regions. To represent the visual contents of image regions, color moments, color histogram and DWT are used. Different combinations of these features are used without any weighting approach. Features are directly propagated to form a single feature vector.
In this paper, the wavelet decomposition is performed at image region level using 2 levels of decomposition using two-dimensional DWT functions available in MATLAB. Since images are converted into HSV color space, each image thus has three components: H, S and V. So, DWT is applied to each component of an image region separately. A feature vector of length 18 (3 components × 6 energy measure) is therefore constructed from the three sub-bands (LH, HL, HH) at all resolutions and for each image region.

5.6. Prototypical local semantic concept representation

The idea of prototypical semantic concept representation is to associate local semantic concepts with instances that are considered to be prototypical with regards to their visual information [36]. Based on this idea, prototypical local semantic concepts are learned from visual information extracted from image regions. For the natural scene dataset used in this paper, there are nine semantic concepts. For each semantic concept, visual features extracted from image regions labelled with this semantic concept are averaged to obtain a local semantic prototype that represents instances of this semantic concept, hence this prototype does not necessarily to be member of the corresponding semantic concept. It is only a summary representation of each semantic concept. This result in generating nine local semantic prototypes learned from instances (visual features) of the nine semantic concepts. These prototypes will be used in the next section to annotate image regions using KNN approach.

6. EXPERIMENTAL SETUP

The first part of this section presents local semantic concept annotators, to annotate image regions, using KNN and SVM classifiers. It also presents the protocol that is followed to conduct the experiments. Experimental results are then reported with some discussion. The performance of local image annotators is assessed using the average precision produced from the confusion matrix.

6.1. Local Semantic Annotators

Image annotation at region level can be considered as a supervised classification problem. Each image region needs to be classified into one of predefined classes. In this paper, two classifiers are employed for the annotation task: the nearest neighbour approach (KNN) and support vector machines (SVM). The protocol used for all experimental work is as follows:

Using the KNN classifier, all the experiments have been validated using 10-folds cross validation where 90% of all image regions are selected randomly for generating the local semantic prototypes and the remaining 10% are used for testing. Using SVM classifier, the same approach has been used where the 90% of all image regions used to generate the semantic concept prototypes are also used to train the SVM classifier whereas the 10% of image regions used in KNN are used for testing. The procedure is repeated 10 times for both classifiers such that an image region appears once in the testing part over the ten folds. This paper use the well-known LIBSVM [37] tool to implement the SVM classifier. The 10-fold cross validation technique is performed in each training set to choose SVM parameters. The Histogram intersection kernel is used to learn the SVM.

For the KNN approach, it is a pre-requisite to decide the similarity metric and the number of neighbors (K) that need to be visited to decide to which class an input instance belongs to, using a voting technique. Using the nine local semantic prototypes presented in Section 5.6, the KNN classifier needs to assign an input feature vector, which represents the visual content of input image region, to one of the local semantic concepts, thus K=1.

Algorithm 5: A step-by-step algorithm for generating semantic concept prototypes and how they are used to annotate image regions with local semantic concepts using KNN approach.

**Input:** (1) A collection of BOW histograms representing all image regions in the image dataset.

(2) A confusion matrix mat of size \( N \times N \), where \( N \) is the number of semantic concepts. Initialize mat to zeros.

**For** \( i = 1 \) to 10 **Do**:

\[
\begin{align*}
\text{a.} & \quad \text{Randomly select 90\% of all BOWs histograms for training and 10\% for testing.} \\
\text{b.} & \quad \text{For each semantic concept } c_j, j=1,2,..., N, \text{ generate local semantic concept prototype } P_j \text{ by averaging all BOWs from the training set.} \\
\text{c.} & \quad \text{Use KNN classifier, using the Euclidean distance and } K=1, \text{ to find similarities between the BOWs in the test set and the local semantic prototypes } P. \text{ An image region, represented by a BOW histogram, is assigned to the semantic concept } j \text{ such that the semantic prototype } P_j \text{ is the most similar prototype to the BOW of this image region.} \\
\text{d.} & \quad \text{Compare the labels assigned to image regions from the test set with the ground truth labels. Report the results in the confusion matrix mat.}
\end{align*}
\]

**Output:** Confusion matrix mat.
For simplicity, in this section the bag of visual words BOW and IBOW histograms are used to refer to the concept-based bag of visual words CBOW generated using the universal visual vocabulary and the integrated visual vocabulary, respectively. In other words, instead of using CBOW and CIBOW, the authors simply refer to them by BOW and IBOW. This section explicitly refers to the BOWs and IBOWs generated from image regions at the upper and lower halves in the experimental results. Algorithm 5 presents the work flow of annotating image regions with local semantic concepts using KNN and local semantic prototypes. The same algorithm is used to conduct different experiments using the local semantic prototypes approach.

Algorithm 6 presents the work flow of annotating image regions with local semantic concepts using KNN and local semantic prototypes generated from the upper and lower halves of images.

6.2. Experimental results

In this paper, four sets of experiments are conducted. The first two sets consider using universal and integrated visual vocabularies, obtained in [21], to build BOW and IBOW histograms from image regions. As mentioned at the beginning of the previous section, BOW and IBOW refer to concept-based BOW generated using both visual vocabularies. The last two sets of experiments investigate generating visual vocabularies from image halves to build BOW and IBOW for image regions at each half separately. For fair comparisons, different types of features are included in the experiments and their annotation performances are reported.

Algorithm 6: A step-by-step algorithm for generating semantic concept prototypes at the upper and lower halves of images and how they are used to annotate image regions with local semantic concepts using KNN approach.

**Input:** (1) A collection of BOW histograms representing all image regions at the upper half of images in the dataset.
(2) A collection of BOW histograms representing all image regions at the lower half of images in the dataset.
(3) Two confusion matrices mat_upper and mat_lower, both of size $N \times N$, where $N$ is the number of semantic concepts. Initialize mat_upper and mat_lower to zeros.

For $i = 1$ to 10 Do:

a. Randomly select 90% of all BOWs histograms, generated from the upper halves, for training and 10% for testing.

b. For each semantic concept $c_j$, $j = 1, 2, \ldots, N$, generate local semantic concept prototype $P_{j, \text{upper}}$ by averaging all BOWs from the training set.

c. Use KNN classifier, using the Euclidean distance and $K=1$, to find similarities between the BOWs in the test set and the local semantic prototypes $P_{j, \text{upper}}$. An image region at the upper half, represented by a BOW histogram, is assigned to the semantic concept $j$ such that the semantic prototype $P_{j, \text{upper}}$ is the most similar prototype to the BOW of this image region.

d. Compare the labels assigned to image regions from the test set with the ground truth labels. Report the results in the confusion matrix mat_upper.

end

For $i = 1$ to 10 Do:

a. Randomly select 90% of all BOWs histograms, generated from the lower halves, for training and 10% for testing.

b. For each semantic concept $c_j$, $j = 1, 2, \ldots, N$, generate local semantic concept prototype $P_{j, \text{lower}}$ by averaging all BOWs from the training set.

c. Use KNN classifier, using the Euclidean distance and $K=1$, to find similarities between the BOWs in the test set and the local semantic prototypes $P_{j, \text{lower}}$. An image region at the lower half, represented by a BOW histogram, is assigned to the semantic concept $j$ such that the semantic prototype $P_{j, \text{lower}}$ is the most similar prototype to the BOW of this image region.

d. Compare the labels assigned to image regions from the test set with the ground truth labels. Report the results in the confusion matrix mat_lower.

end

**Output:** Confusion matrix mat, where $mat = (mat_{\text{upper}} + mat_{\text{lower}})$
Three types of features are included: color histogram, color moments and DWT. These features are used separately to represent the visual content of image regions. Also, these features are linearly integrated with the BOW and IBOW histograms to study their influence on the discriminative power of BOW by including color and textural features.

The first set of experiments investigates image region annotation using local semantic prototypes and the KNN classifier. Algorithm 5 is used to generate local semantic prototypes from BOW histograms and to annotate new image regions with semantic concepts using the KNN classifier. The same algorithm can be applied to generate local prototypes for different features. The only thing that needs to be changed in the algorithm is to replace BOW with the features extracted from image regions. For example, replacing BOW with IBOW histograms will generate local semantic prototypes from IBOW and then the KNN classifier is used to annotate new image region, represented by an IBOW histogram, with one of the predefined semantic labels. Another example is to use color histogram to represent the visual content of image regions. In this work similar to [5], three histograms are obtained from an image region represented in the HSV color space. The first histogram (36-bins) is obtained from the Hue component of the image region, the second histogram (32-bins) is obtained from the S component while the third histogram (16-bins) is obtained from the V component. These three histograms are concatenated to form a single HSV color histogram of 84-bins. Algorithm 5 is used again but BOW histograms are now replaced with the HSV color histograms from which local semantic prototypes are generated and used by the KNN classifier to annotate new image regions.

For multiple features, local semantic prototypes are first generated for each feature separately and then for each semantic concept the prototypes of the different feature types are aggregated to form a single local semantic prototype for the corresponding local semantic concepts. For example, the natural scene dataset used in this paper is composed of nine semantic concepts. Suppose that color histogram (84-D feature vector) and BOW (200-D feature vector) are the features to be concatenated. Firstly, Algorithm 5 is used to generate nine local semantic prototypes for the color histogram and the same algorithm is used to generate another nine local semantic prototypes for the BOW histograms. For each semantic concept, the local semantic prototype produced from the color histograms is concatenated with the local semantic prototype produced from the BOW histograms. This will generate nine local semantic prototypes each of length (84+200 = 284-D). The same approach can be applied for more than two feature types.

To this end, in the first experimental set, 14 experiments are conducted using the following features and their combinations for image region annotation: IBOW, UBOW, Color histogram (ColHist), color moments (Mom), DWT (Wav), IBOW+Mom, UBOW+Mom, IBOC+ColHist, UBOW+ColHist, IBOW+Wav, UBOW+Wav, IBOW+ColHist+Wav, UBOW+ColHist+Wav, and ColHist+Wav.

In the second set of experiments, SVM is used to annotate image regions. Here, local semantic prototypes are not used. For IBOW histograms, SVMs are trained on 90% of all image regions in the dataset, represented by their IBOW histograms. The trained SVMs are then used to assign labels to the remaining image regions in the dataset. For multiple features, each feature vector is first normalized to a unit length vector and then they are aggregated into a single feature vector. The SVMs are then used for training and testing. Similar to the first set of experiments, 14 experiments are conducted using the same features and their combination as aforementioned.

Figure 16 presents the performance of the first two sets of experiments. Bars in blue show the performance of image region annotation using local semantic prototypes with the KNN classifier. Bars in red show the performance of image region annotations using SVMs. It is worth to remind that IBOW and UBOW histograms are generated from visual vocabularies constructed at scene level and not concept level, the local from global approach. Thus, it is interesting to see their performances for image region annotation. It is obvious that SVMs outperform the local semantic prototypes in all types of features and their combinations. Also, IBOW histograms outperform UBOW in both sets of experiments. Adding color and textural features to the UBOW and IBOW histograms improved the accuracy of image region annotation.

Figure 16: Accuracies of annotating images with the nine semantic concepts using KNN and SVM classifiers.
The performance of IBOW histograms has gained the best annotation accuracy when color histograms and DWT are added to them. Nevertheless, integrating low-level features such as color histogram with the DWT has also gained a good annotation performance compared with more complicated features such as IBOW and UBOV histograms. However, they have not got the best annotation results but they can be used to improve the performance of the IBOW and UBOV histograms. The annotation performance for each semantic concept resulted from the 14 experiments using both SVM and local semantic prototypes are shown in Table 2 and Table 3. Each row contains the accuracies of each semantic concept. These values are the diagonal elements of the confusion matrix generated from each experiment. The last column from both tables is used to generate the results shown in Figure 16.

Table 2: (KNN and semantic prototypes) Accuracies of each experiment (row), where elements in each row are the diagonal elements of the confusion matrix resulted from each experiment. Accuracy is generated based on 10-folds CV.

| Sky   | Water | Grass | Trunks | Foliage | Field | Rocks | Flowers | Sand | Acc.  |
|-------|-------|-------|--------|---------|-------|-------|---------|------|-------|
| IBOW  |       |       |        |         |       |       |         |      | 88.54 |
| UBOV  |       |       |        |         |       |       |         |      | 86.61 |
| CollHist |     |       |        |         |       |       |         |      | 85.89 |
| Mon   |       |       |        |         |       |       |         |      | 86.24 |
| Wav   |       |       |        |         |       |       |         |      | 84.68 |
| IBOW+Mon |   |       |        |         |       |       |         |      | 44.61 |
| UBOV+Mon |  |       |        |         |       |       |         |      | 47.70 |
| CollHist+Mon | |       |        |         |       |       |         |      | 47.33 |
| IBOW+Wav |   |       |        |         |       |       |         |      | 84.68 |
| UBOV+Wav |  |       |        |         |       |       |         |      | 86.47 |
| CollHist+Wav | |       |        |         |       |       |         |      | 86.32 |
| IBOW+CollHist+Wav | |       |        |         |       |       |         |      | 85.67 |
| UBOV+CollHist+Wav | |       |        |         |       |       |         |      | 86.72 |
| CollHist+Wav | |       |        |         |       |       |         |      | 84.69 |

The last two sets of experiments are dedicated to study the influence of building visual vocabularies from local keypoints located at the upper and lower halves of images. It aims to improve the quality of visual words generated from clustering features of local keypoints at upper and lower halves of images.

Two visual vocabularies are generated from the upper halves of images; universal visual vocabulary and integrated visual vocabulary. Image regions located at the upper halves of images are represented by UBOV and IBOW histograms generated using the upper universal and integrated visual vocabularies, respectively. Another two visual vocabularies are generated from the lower halves of images. Image regions located at the lower halves of images are represented by UBOV and IBOW
histograms generated using the lower universal and integrated visual vocabularies. To use multiple features with the UBOW and IBOW, the same low-level features used in the first two sets are employed here.

In the third set of experiments Algorithm 6 is used to generate local semantic prototypes for the upper half and local semantic prototypes for the lower half. Also, 14 experiments are conducted using this algorithm. Local semantic prototypes at the upper half are used by KNN to annotate test image regions located at the upper half of images. Results of annotations are compared with the ground truth and then reported in mat_upper confusion matrix. The same procedure is applied to the lower half with results reported in another confusion matrix, mat_lower. Both matrices are added together to get a single confusion matrix. Results of annotating images at the upper and lower halves, using upper and lower local semantic prototypes and the KNN classifier, is shown in blue bars in Figure 17.

![Figure 17: Accuracies of annotating images with the nine semantic concepts using KNN and SVM classifiers. BOWs and IBOWs are generated from image regions using visual vocabularies generated from the upper and lower halves of images.](image)

It is obvious from the figure that generating local semantic prototypes from image halves improved the annotation accuracies in all experiments and for most features. The IBOW histograms are still the best in representing image regions. Adding textural features to IBOW has slightly improved the annotation and works better than adding color information. More details about the annotation accuracies of each of the experiments are shown in Table 4.

Table 4: (KNN and semantic prototypes at upper and lower halves) Accuracies of each experiment (row), where elements in each row are the diagonal elements of the confusion matrix (upper+lower) resulted from each experiment. Accuracy is generated based on 10-folds CV at each halves.

|                | Sky  | Water | Grass | Trunks | Foliage | Field | Rocks | Flowers | Sand | Acc  |
|----------------|------|-------|-------|--------|---------|-------|-------|---------|------|------|
| IBOW           | 99.92| 91.80 | 84.04 | 77.72  | 69.98   | 56.97 | 89.06 | 93.22   | 67.74 | 79.73|
| UBOW           | 94.09| 29.77 | 68.23 | 55.38  | 55.39   | 36.73 | 91.41 | 67.40   | 47.45 | 67.56|
| CoHist         | 57.37| 53.37 | 60.29 | 29.41  | 29.11   | 43.61 | 22.74 | 48.71   | 43.78 | 41.85|
| Mum            | 84.20| 26.22 | 5.48  | 38.38  | 15.54   | 31.06 | 13.60 | 46.17   | 44.47 | 24.13|
| Wav            | 82.25| 23.34 | 22.24 | 41.32  | 48.34   | 16.79 | 16.98 | 50.17   | 28.14 | 44.50|
| IBOW+Mom       | 60.13| 60.77 | 87.70 | 80.12  | 74.63   | 61.77 | 91.60 | 93.51   | 75.99 | 73.16|
| UBOW+Mom       | 44.23| 40.61 | 48.07 | 64.55  | 48.21   | 36.24 | 61.85 | 66.18   | 61.26 | 50.39|
| IBOW+CoHist    | 64.91| 68.83 | 85.95 | 68.55  | 61.65   | 67.88 | 72.07 | 76.87   | 72.44 | 68.11|
| UBOW+ColHist   | 64.08| 60.40 | 80.53 | 62.58  | 59.49   | 61.91 | 70.21 | 71.40   | 64.01 | 64.75|
| IBOW+Wav       | 86.71| 69.37 | 85.00 | 83.73  | 82.23   | 62.63 | 90.19 | 98.85   | 75.42 | 81.70|
| UBOW+Wav       | 85.00| 40.44 | 69.57 | 62.65  | 69.39   | 42.74 | 81.96 | 69.89   | 56.91 | 69.84|
| IBOW+CoHist+Wav| 62.87| 70.22 | 85.94 | 72.94  | 63.29   | 71.06 | 68.69 | 76.77   | 79.83 | 67.80|
| UBOW+ColHist+Wav| 62.07| 63.83 | 81.67 | 66.15  | 60.94   | 85.47 | 68.09 | 72.85   | 64.14 | 65.03|
| CoHist+Wav     | 53.08| 56.56 | 61.55 | 37.66  | 34.13   | 51.03 | 22.88 | 46.80   | 59.72 | 43.75|

The last set of experiments (14 experiments) is conducted on different features using SVM classifier. SVMs are trained and tested on image regions at the upper and lower halves of images. At the upper half, image regions represented by the IBOW histograms or other features are used to train SVM which in turn used later to annotate test image regions at the upper half of images. Results are reported in a confusion matrix mat_upper. The same approach is applied for image region at the lower half of images. Results are reported in another confusion matrix mat_lower. Both matrices are add to each other to get final confusion matrix. The red bars in Figure 17 show annotations accuracies using different features. The best result is achieved using IBOW combined with ColHist and Wav features. It reports 94.02% annotation accuracy. More details are shown in Table 5.
Table 5: (SVM at upper and lower halves) Accuracies of each experiment (row), where elements in each row are the diagonal elements of the confusion matrix (upper+lower) resulted from each experiment. Accuracy is generated based on 10-folds CV at each halves.

|       | Sky  | Water | Grass | Trees | Foliage | Field | Rocks | Flowers | Sand | Acc. |
|-------|------|-------|-------|-------|---------|-------|-------|---------|------|------|
| IBOw  | 90.44| 82.96 | 99.46 | 99.42 | 73.53   | 68.55 | 99.55 | 99.90   | 62.35| 87.17|
| UBOW  | 80.50| 23.46 | 44.48 | 80.12 | 73.07   | 67.26 | 92.49 | 82.49   | 74.15| 74.94|
| CalibE| 94.36| 65.94 | 58.15 | 10.77 | 34.85   | 90.45 | 75.65 | 71.45   | 66.85| 76.08|
| Mem   | 85.23| 94.00 | 1.68  | 11.68 | 47.31   | 4.92  | 77.30 | 20.35   | 24.30| 48.45|
| Way   | 94.72| 96.02 | 9.31  | 24.33 | 81.88   | 36.95 | 32.25 | 18.01   | 12.84| 62.61|
| IBOw+Mem| 85.39| 82.83 | 88.07 | 75.45 | 92.57   | 90.76 | 93.36 | 81.16   | 80.26| 88.44|
| UBOW+Mem| 86.17| 41.82 | 50.76 | 57.59 | 82.23   | 67.86 | 99.52 | 99.65   | 80.80| 78.65|
| IBOw+CalibE| 95.86| 65.58 | 90.49 | 87.75 | 96.21   | 89.57 | 94.24 | 85.33   | 95.90| 93.61|
| UBOW+CalibE| 88.80| 66.55 | 93.01 | 73.97 | 76.53   | 72.04 | 94.61 | 83.94   | 79.89| 84.57|
| IBOw+Way | 94.99| 84.33 | 90.66 | 89.42 | 78.53   | 68.55 | 99.53 | 90.90   | 88.08| 87.86|
| UBOW+Way | 86.15| 85.36 | 90.15 | 73.97 | 75.07   | 67.26 | 94.18 | 79.06   | 74.15| 83.99|
| IBOw+CalibE+Way| 96.51| 65.58 | 99.49 | 87.75 | 96.21   | 99.57 | 94.24 | 92.63   | 95.90| 94.62|
| UBOW+CalibE+Way| 88.01| 41.23 | 59.32 | 58.09 | 73.84   | 62.49 | 99.72 | 99.65   | 86.82| 78.04|
| CalibE+Way| 98.91| 84.82 | 67.88 | 59.46 | 51.72   | 72.35 | 81.97 | 77.89   | 74.50| 85.97|

7. CONCLUSION

This paper has presented the problem of image annotation at image region level. The task was to assign labels to image regions generated from a regular grid. The paper has addressed using BOW model to represent image regions. A framework for image region annotation has been proposed. The relationship between the distributions of local semantic concepts and local keypoints located in image regions labelled with these semantic concepts are studied in detail. Also, this paper has investigated using visual vocabularies generated from natural scene classes to represent local semantic concepts, the local from global approach. Generating visual vocabularies from image halves were also investigated to generate BOW histograms. An extensive experimental work has been conducted using different features and classifiers to annotate image regions with semantic concepts. Our experimental results shows the plausibility of local from global approach for image region annotation as well as the discriminative power of using visual vocabularies from image halves. It showed an improved annotation results using IBOW combined with low-level features.

ACKNOWLEDGEMENTS

The author is grateful to the Applied Science Private University, Amman, Jordan, for the full financial support granted to this research. The author would like to thank Dr. Julia Vogel for providing us access to the natural scene image dataset and for valuable discussion.

REFERENCES

[1] Y. Rui, T.S. Huang and S.F. Chang, Image retrieval: Current techniques, promising directions, and open issues. Journal of visual communication and image representation, vol. 10, No. 1, March 1999, pp. 39-62.
[2] Y. Liu, D. Zhang, G. Lu and W.Y. Ma, A survey of content-based image retrieval with high-level semantics. Pattern Recognition, Vol. 40, No. 1, January 2007, pp. 262-282.
[3] R. Datta, D. Joshi, J. Li, and J.Z.Wang, Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys, Vol. 40, No. 2, April 2008, pp. 1-60.
[4] ZHANG, D., ISLAM, M. M. & LU, G., A review on automatic image annotation techniques. Pattern Recognition, 45, 2012, 346-362.
[5] VOGEL, J. & SCHIELE, B., A semantic typicality measure for natural scene categorization. Lecture notes in computer science. Springer. 2004, 195-203.
[6] LOWE, D., Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 60, 2004, 91-110.
[7] A. Bosch, X. Munoz and R. Marti, Which is the best way to organize/classify images by content?. Image and vision computing, Vol. 25, No. 6, 2007, pp. 778-791.
[8] P. Quelhas, F. Monay, J.M. Odobez, D. Gatica-Perez, T. Tuytelaars and L. Van Gool, Modeling scenes with local descriptors and latent aspects. Proc. of IEEE International Conference on Computer Vision ICCV, Beijing, China, October 17-21, 2005, pp. 883–890.

[9] L. Fei-Fei and P. Perona, A bayesian hierarchical model for learning natural scene categories. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, San Diego, CA, USA, June 20-26, 2005, pp. 524-531.

[10] D. Gokalp and S. Aksoy, Scene classification using bag-of-regions representations. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Minneapolis, Minnesota, USA, June 18-23, 2007, pp.1-8.

[11] P. Quelhas and J. Odobez, Natural scene image modeling using color and texture visterns. Proc. of International Conference on Image and Video Retrieval, CIVR, Lecture Notes in Computer Science, Tempe, AZ, USA, July 13-15, 2006, pp. 411-421.

[12] G. Csurka, C. Dance, L. Fan, j. Willamowski and Bray C., Visual categorization with bags of keypoints. Proc. of ECCV workshop on Statistical Learning in Computer Vision, Czech Republic, May 11-14, 2004, pp. 59–74.

[13] Y.G. Jiang, C.W. Ngo and J. Yang, Towards optimal bag-of-features for object categorization and semantic video retrieval. Proc. of the 6th ACM international Conference on Image and Video Retrieval, CIVR, Amsterdam, The Netherlands, July 9-11, 2007, pp. 494-501.

[14] Y. Alqasrawi, D. Neagu and P. Cowling, Natural Scene Image Recognition by Fusing Weighted Colour Moments with Bag of Visual Patches on Spatial Pyramid Layout. Proc. of the 9th international Conference on intelligent Systems Design and Applications, ISDA, IEEE Computer Society, Pisa, Italy, Nov. 30 -Dec. 2 2009, pp. 140-145.

[15] J. Yang, Y.G. Jiang, A.G. Hauptmann and C.W. Ngo, Evaluating bag-of-visual-words representations in scene classification. Proc. of the 9th ACM international Conference on Multimedia information Retrieval, ACM MIR, University of Augsburg, Germany, September 28-29, 2007, pp. 197-206.

[16] F. Khan, J. van de Weijer and M. Vanrell, Top-Down Color Attention for Object Recognition. Proc. of 12th IEEE International Conference on Computer vision, ICCV, Kyoto, Japan, September 27 - October 4, 2009, pp. 979-986.

[17] S. Lazebnik, C. Schmid and J. Ponce, Beyond bags of features: Spatial pyramid matching for recognising natural scene categories. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, New York, USA, June 17-22, 2006, pp. 2169-2178.

[18] A. Bosch, A. Zisserman and X. Munoz, Representing shape with a spatial pyramid kernel. Proc. of the 6th ACM international Conference on Image and Video Retrieval, CIVR, Amsterdam, The Netherlands, July 9-11, 2007, pp. 401-408.

[19] C.H. Lampert, M.B. Blaschko and T. Hofmann, Beyond sliding windows: Object localization by efficient subwindow search. Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Anchorage, Alaska, USA, June 24-26, 2008, pp. 1-8.

[20] S. Battiato, G. Farinella, G. Gallo and D. Ravi, Exploiting Textons distributions on spatial hierarchy for scene classification. EURASIP Journal on Image and Video Processing, special issue on Multimedia Modeling, January 2010, pp.1-13.

[21] Yousef Alqasrawi, Daniel Neagu, Peter I. Cowling: Fusing integrated visual vocabularies-based bag of visual words and weighted colour moments on spatial pyramid layout for natural scene image classification. Signal, Image and Video Processing, Springer 7(4): 2013, 759-775.

[22] Yousef Alqasrawi, Neagu, Daniel, Investigating the relationship between the distribution of local semantic concepts and local keypoints for image annotation, 14th UK Workshop on Computational Intelligence (UKCI), IEEE computer Society, 2014, pp.1-7.

[23] JA-HWUNG, S., CHIEN-LI, C., CHING-YUNG, L. & TSENG, V. S., Effective Semantic Annotation by Image-to-Concept Distribution Model. IEEE Transactions on Multimedia, 13, 2011, 530-538.

[24] CUSANO, C., CIOCCA, G. & SCHETTINI, R., Image annotation using SVM. Proceedings of Internet imaging IV, SPIE, 2004, 330-338.

[25] CHAPELLE, O., HAFFNER, P. & VAPNIK, V. N., Support vector machines for histogram-based image classification. IEEE Transactions on Neural Networks, 1999, 1055-1064.

[26] GOH, K. S., CHANG, E. Y. & LI, B., Using one-class and two-class SVMs for multiclass image annotation. IEEE Transactions on Knowledge and Data Engineering, 17, 2005, 1333-1346.

[27] QI, X. & HAN, Y., Incorporating multiple SVMs for automatic image annotation. Pattern Recognition, 40, 2007, 728-741.
[28] BARNARD, K. & FORSYTH, D., Learning the semantics of words and pictures. Eighth IEEE International Conference on Computer Vision, ICCV, 2001, 408-415.

[29] DUYGULU, P., BARNARD, K., FREITAS, J. F. G. D. & FORSYTH, D. A., Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. Proceedings of the 7th European Conference on Computer Vision-Part IV, ECCV, 2002, 97-112.

[30] SHI, J. & MALIK, J., Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI, 22, 2000, 888-905.

[31] AKSOY, S., KOPERSKI, K., TUSK, C., MARCHISIO, G. & TILTON, J. C., Learning Bayesian classifiers for scene classification with a visual grammar. IEEE Transactions on Geoscience and Remote Sensing, 43, 2005, 581-589.

[32] MOJSILOVIC, A., GOMES, J. & ROGOWITZ, B., Semantic-friendly indexing and querying of images based on the extraction of the objective semantic cues. International Journal of Computer Vision, 56, 2004, 79-107.

[33] MAKADIA, A., PAVLOVIC, V. & KUMAR, S., Baselines for image annotation. International Journal of Computer Vision, 90, 2010, 88-105.

[34] TOUSCH, A.-M., HERBIN, S. & AUDIBERT, J.-Y., Semantic hierarchies for image annotation: A survey. Pattern Recognition, 45, 2012, 333-345.

[35] MIKOLAJCYK, K. & SCHMID, C., A performance evaluation of local descriptors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 27, 2005, 1615-1630.

[36] VOGEL, J. & SCHIELE, B., Semantic modeling of natural scenes for content-based image retrieval. International Journal of Computer Vision, 72, 2007, 133-157.

[37] CHANG, C.-C. & LIN, C.-J. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST), 2, 2011, 1-27. Software available at: http://www.csie.ntu.edu.tw/~cjlin/libsvm.