Toward A Higher-Level Visual Representation For Content-Based Image Retrieval

Ismail El sayad
LIFL UMR CNRS 8022
University of Lille1 and
Telecom-Lille1
Villeneuve d'Ascq, France
ismail.elsayad@lifl.fr

Jean Martinet
LIFL UMR CNRS 8022
University of Lille1 and
Telecom-Lille1
Villeneuve d'Ascq, France
jean.martinet@lifl.fr

Thierry Urruty
LIFL UMR CNRS 8022
University of Lille1 and
Telecom-Lille1
Villeneuve d'Ascq, France
thierry.urruty@lifl.fr

Samir Amir
LIFL UMR CNRS 8022
University of Lille1 and
Telecom-Lille1
Villeneuve d'Ascq, France
samir.amir@lifl.fr

Chabane Djeraba
LIFL UMR CNRS 8022
University of Lille1 and
Telecom-Lille1
Villeneuve d'Ascq, France
chabane.djeraba@lifl.fr

ABSTRACT

Having effective methods to access the desired images is essential nowadays with the availability of huge amount of digital images. The proposed approach is based on an analogy between content-based image retrieval and text retrieval. The aim of the approach is to build a meaningful mid-level representation of images to be used later for matching between a query image and other images in the desired database. The approach is based firstly on constructing different visual words using local patch extraction and fusion of descriptors. Secondly, we introduce a new method using *multilayer pLSA* to eliminate the noisiest words generated by the vocabulary building process. Thirdly, a new *spatial weighting* scheme is introduced that consists in weighting visual words according to the probability of each visual word to belong to each of the n Gaussian. Finally, we construct visual phrases from groups of visual words that are involved in strong *association rules*. Experimental results show that our approach outperforms the results of traditional image retrieval techniques.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]

General Terms

Design, Algorithms.

Keywords

SURF, content-based image retrieval, visual words, visual phrases, Gaussian mixture model, spatial weighting, pLSA.

MoMM2010, 8-10 November, 2010, Paris, France.

1. INTRODUCTION

Due to the explosive spread of digital devices, digital content grows rapidly. The increasing need for automatic processing, description, and structuring of large digital archives motivates to have an effective content-based image retrieval (CBIR). In typical CBIR systems, it is always important to select an appropriate representation for images. Indeed, the quality of the retrieval depends on the quality of the internal representation for the content of the visual documents [3]. Recently, many image retrieval systems have shown that the part-based representation for image retrieval is much superior over traditional global features. Indeed, one single image feature computed over the entire image is not sufficient to represent important local characteristics of different objects within the image.

Nowadays, bag-of-visual-words [15, 17, 10] has drawn much attention. Analogous to document representation in terms of words in text domain, the bag-of-visual-words approach models an image as an unordered bag of visual words, which are formed by vector quantization of local region descriptors. This approach achieves good results in representing variable object appearances caused by changes in pose, scale and translation. Despite the success of the bag-of-visual-words approach in recent studies, the precision of image retrieval is still incomparable to its analogy in text domain, i.e. the document retrieval because of many important drawbacks.

Firstly, most of the local descriptors are based on the intensity or gradient information of images, so neither shape nor color information is used. In the proposed approach, in addition to the SURF descriptor that was proposed by Bay et al. [4], we introduce a novel descriptor (Edge context) that is based on the distribution of edge points.

Secondly, since the bag-of-visual-words approach represents an image as a collection of local descriptors, ignoring their order within the image, the resulting model provides a rare amount of information about the spatial structure of the image. In this paper, we propose a new spatial weighting scheme that consists of weighting visual words according to

the probability of each visual word to belong to one of the n Gaussian in the 5-dimensional color-spatial feature space.

Thirdly, the low discrimination power of visual words leads to low correlations between the image features and their semantics. In our work, we build a higher-level representation, namely *visual phrase* from groups of adjacent words using *association rules* extracted with the *Apriori* algorithm [2]. Having a higher-level representation, from mining the occurrence of groups of low-level features (visual words), enhances the image representation with more discriminative power since structural information is added.

The remainder of the article is structured as follows: Section 2 reviews related works to the proposed approach. In Section 3, we describe the method for constructing visual words from images and mining visual phrases from visual words to obtain the final image presentation. In Section 4, we present an image similarity method based on visual words and visual phrases. We report the experimental results in Section 5, and we give a conclusion to this article in Section 6.

2. RELATED WORKS

2.1 Analogy Between Information Retrieval And CBIR

Text retrieval systems generally employ a number of standard steps in the processes of indexing and searching a text collection [3]. The text documents are firstly parsed into words. Secondly, the words are represented by their stems: for example "walk", "walking" and "walks" would be represented by the stem "walk". Thirdly, a stop list is used to filter very common words out, such as "the" and "an", which occur in most documents and are therefore not discriminating for a particular document.

Hammouda and Kamel [9] have presented a novel phrase-based document index model, which allows an incremental construction of *a phrase-based* index of the document set with an emphasis on the efficiency of the retrieval, rather than relying only on single-term indexes. This approach has provided an efficient phrase matching that can be used to judge the similarity between documents. The combination of these two components (words and phrases) creates an underlying model for robust and accurate document similarity calculation that leads to much improved results over traditional methods.

In syntactic level, there is a correspondence between a text document and an image where an image is a particular arrangement of different pixels in a 2D space, while a text document is a particular arrangement of different letters in a 1D space. In this analogy, pixels correspond to letters, patches to words, and group of patches to phrases. Thus, a *visual word* is a local segment in an image, defined either by a region (image patch or blob) or by a reference point together with its neighborhood [15].

Zheng et al. [20] made an analogy between image retrieval and text retrieval, and have proposed a higher-level representation *visual phrase* based on the analysis of visual word occurrences to retrieve images containing desired objects. Visual phrases are defined as pairs of adjacent local image patches. We share the same objective of designing a mid-level descriptor for representing documents. However while Zheng et al. consider only adjacent pairs of patches, our proposed approach is more general since it handles any set of items. In that way, we can represent more accurately the relations between objects.

Yuan et al. [19] have proposed another higher-level lexicon, i.e. visual phrase lexicon, where a visual phrase is a spatially co-occurrent pattern of visual words. This higher-level lexicon is much less ambiguous than the lower-level one (visual words). The main contribution of this approach is to present a fast solution to the discovery of significant spatial co-occurrent patterns using frequent item set mining. On one hand, we share the same aim of designing a higher level of representation that enhances the discrimination power of the lower level. On the other hand, we went beyond mining the frequent item set by detecting the items that are not only frequent but also are involved in strong association rules (to be discussed later in this article) which gives a higher representation level with more meaningful aspects.

2.2 Weighting Scheme

Inspired by the success of the vector-space model [14], for text document representation and retrieval, the bag-of-visual-words approach usually converts images into vectors of visual words based on their frequency. Yang et al. in [1] evaluated many frequency weighting schemes that are based on different factors. Two major factors in term weighting are *tf* (term frequency) and *idf* (inverse document frequency). A third factor is the normalization factor, which converts the feature into unit length vector to eliminate the difference between short and long documents.

Many text categorization methods have used weighting schemes based on these factors, such as *tf-idf* weighting, stop word removal, and feature selection. The best weighting scheme in Information Retrieval does not guarantee good performance in *CBIR* since the count information can be noisy. Suppose that a certain visual word (w) is typical among "building" images. An image containing 100 occurrences of w is not necessarily to be more likely about building than an image containing only 25 occurrences of w, but a CBIR system trained from the first image can be mislead by the high count and will not retrieve the second image since it will be classified as a "non-building" image. For this reason, we create a weighting scheme that weights the visual words according to the spatial constitution of an image content rather than the number of occurrences.

3. VISUAL WORD AND VISUAL PHRASE CONSTRUCTION

In this section, we describe different components of the chain of processes in constructing the visual words and visual phrases. Figure 1 presents the different process starting from detecting interest and edge points till the image description of the image by visual words and phrases.

3.1 Visual Word Construction

We use the fast Hessian detector [4] to extract interest points. In addition, the canny edge detector [7] is used to detect edge points. From both sets of interest and edge points, we use a clustering algorithm to group these points into different clusters in the 5-dimensional color-spatial feature space (see the visual construction part in Figure 1). The clustering result is necessary to extract our Edge context descriptor (to be discussed later in this paper) and to estimate the spatial weighting scheme for the visual words.

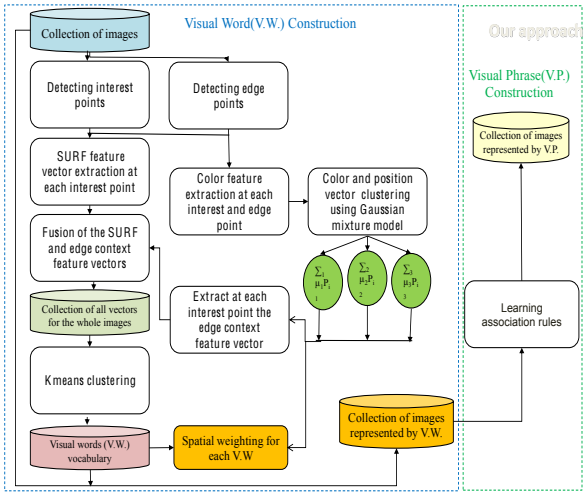

Figure 1: Flow of information in the visual document representation model.

3.1.1 *Extracting And Describing Local Features*

In this approach, based on the Gaussian Mixture Model (GMM) [6], we model the color and position feature space for set of interest and edge points. The Gaussian mixture model used to extract the Edge context descriptor and to construct our novel spatial weighting scheme.

Firstly, a 5-dimensional color-spatial feature vector, built from the 3 dimensions for RGB color plus 2 dimensions (x, y) for the position, is created to represent each interest and edge point. In an image with m interest/edge points, a total of m 5-dimensional color-spatial feature vectors: $Z_1, ..., Z_m$ can be extracted.

In our approach, we use the SURF low-level feature descriptor which is 64 dimensional vector that describes the distribution of pixel's intensities within a scale-dependent neighborhood of each interest point detected by the Fast-Hessian. This descriptor is similar to the SIFT one [12], but Bay et al. have used integral images [16] in conjunction with filters known as Haar wavelets in order to increase robustness and decrease the computation time.

In addition to the *SURF* descriptor, we introduce a novel *Edge context descriptor* at each interest point detected by the Fast-Hessian, based on the distribution of the edge points in the same Gaussian (by returning to the 5-dimensional color-spatial feature space). Our descriptor is inspired by the shape context descriptor proposed by Belongie et al. [5] with respect to the extracted information from edge point distribution. Describing the distribution of these points enriches our descriptor with more information, rather than the intensity described by SURF. Moreover, the distribution over relative positions is a robust, compact, and highly discriminative descriptor.

As shown in Figure 2, vectors from each interest point in the 2D spatial image space are drawn point to all other edge points (that are within the same cluster in 5-dimensional color-spatial feature space). Then the Edge context descriptor for each interest point is represented as a histogram of 6 bins for R (*magnitude* of the drawn vector from the interest point to the edge points) and 4 bins for θ (*orientation angle*).

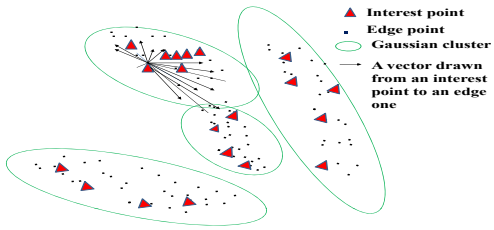

Figure 2: Extraction of the Edge context descriptor in the 2D spatial space where the points are already clustered before in the 5-dimensional color-spatial Gaussian space.

For this novel descriptor many invariances are applied.

Firstly, invariance to translation is intrinsic to the Edge context definition since the distribution of the edge points is measured with respect to fixed interest point.

Secondly, invariance for scale is achieved by normalizing the radial distance by a mean distance between the whole set of points within a single Gaussian in the 5-dimensional color-spatial feature space.

Thirdly, invariance for rotation is achieved by measuring all angles relative to the tangent angle of each interest point.

Following the visual construction part in Figure 1, after extracting the Edge context feature, fusion between this descriptor and the SURF descriptor is performed. This fused feature vector is composed of 88 dimensions (64 from SURF + 24 from the Edge context descriptor). Hence, the new feature vector describes information on the distribution of the intensity and the edge points of the image. It enriches our image representation with more local information.

3.1.2 *Quantizing the Local Features*

Visual words are created by clustering the fused feature vectors (SURF + Edge context feature vector) in order to form a visual vocabulary. Quantization of the features into visual words is performed by using a vocabulary tree [13] in order to support large vocabulary size. The vocabulary tree is computed by repeated k-means clusterings that hierarchically partition the feature space.

This hierarchical approach overcomes two major problems of the traditional direct k-means clustering in cases where k is large. Firstly clustering is more efficient during visual word learning and secondly the mapping of visual features to discrete words is way faster than using a plain list of visual words. Finally,we map each feature vector of an image to its closest visual word. Therefore we query the vocabulary tree for each extracted feature, and the best matching visual word index is returned. Figure 4 shows an example of a fused feature vector assigned int a discrete visual word 6.

3.1.3 *Filtering The Noisy Visual Words*

In this section, we introduce another method to eliminate presumed useless visual words. This method aims at eliminating the most noisy words generated by the vocabulary building process, using *multilayer pLSA*. Lienhart et al.[11] proposed a multilayer multimodal probabilistic Latent Semantic Analysis (*mm-pLSA*). The proposed approach (*mm-pLSA*) has two modes: one mode for visual words and the

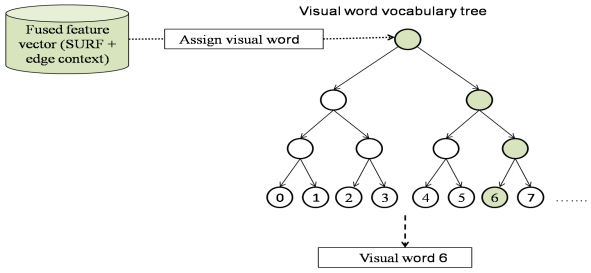

Figure 3: Example of assigning a fused feature vector into a discrete visual word.

other one for image tags. We used only the visual word mode. In the *multilayer pLSA* (*m-pLSA*), we have two different hidden topics.

- Top-level latent topics z_i^t.
- Visual latent topics z_j^v.

This generative model is expressed by the following probabilistic model:

$$P(I/w_l) = \sum_{i=1}^{P}\sum_{j=1}^{V} P(I)P(z_i^t/I)P(z_j^v/z_i^t)P(w_l/z_j^v) \quad (1)$$

where $P(I)$ denotes the probability of a an image I of the database to be picked, $P(z_i^t/I)$ the probability of a top-level topic z_i^t given the current image, $P(z_j^v/z_i^t)$ the probability of a visual latent topic z_j^v given a z_i^t and $P(w_l/z_j^v)$ the probability of a visual word w_l given a z_j^v.

We assigned one top-level latent topic per category of images, the total number of top-level latent topics (P) is the same as the total number of categories of the image dataset. The total number of visual concepts is V where $V < P$. We categorized visual concepts according to their joint probabilities with all top-level latent topics $P(z_j^v/z_i^t)$. All visual concepts whose joint probability to all top-level latent concepts are lower than a given threshold is categorized as irrelevant. After that, we eliminated all visual words whose probability $P(w_l/z_j^v)$ is low to a given threshold for every relevant visual concept, since they are not informative for any relevant visual concept. Therefore, we propose to keep only the most significant words for each relevant visual concept. Figure 4 shows examples of images after eliminating ambiguous visual words. Experiments reported in Section 5 show that this technique improves the performance of the image retrieval. An important aspect of this model is that every image consists of one or more visual aspects, which in turn are combined to one or more higher-level aspects. This is very natural since images consist of multiple objects and belong to different categories.

3.1.4 Spatial Weighting For The Visual Words

To perform the ***spatial weighting***, we innovate a new scheme which differs from ***tf-idf*** weighting scheme. Suppose that in an image, there are local descriptors obtained from the interest point set belonging to the same Gaussian and assigned to a visual word w_l, where $1<l<K$ and K is the number of visual words in the visual vocabulary. The sum of the probabilities of salient point occurrences will indicate the contribution of visual word w_l to a Gaussian β_i. Therefore,

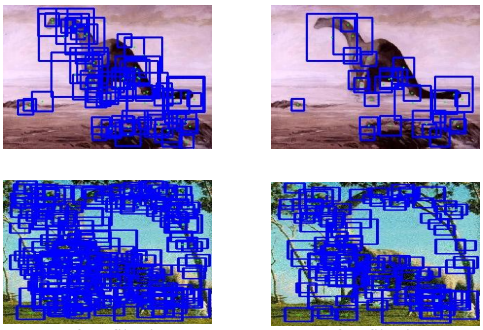

Figure 4: Examples of images after filtering the noisy visual words using *m-pLSA*.

the weighted term frequency ($Tf_{w_l\beta_i}$) of a visual word w_l with respect to a Gaussian β_i is defined as follows:

$$Tf_{w_l\beta_i} = \sum_{m=1}^{n_i} P(\beta_i/Z_m) \quad (2)$$

The average weighted term frequency (Tf_{w_l}) of w_l with respect to an image I where w_l occurs in n_{w_l} Gaussian is defined as follows:

$$Tf_{w_l} = \sum_{i=1}^{n_{w_l}} (Tf_{w_l\beta_i})/n_{w_l} \quad (3)$$

The weighted inverse Gaussian frequency of w_l with respect to an image I with n Gaussian is defined as follows:

$$If_{w_l} = ln\frac{n}{n_{w_l}} \quad (4)$$

The final spatial weight of the visual word w_l is defined by the following formula:

$$Sw_{w_l} = Tf_{w_l} \times If_{w_l} \quad (5)$$

3.2 Visual Phrase Construction

Before proceeding to the construction phase of visual phrases for the set of images, let us examine phrases in text. A phrase can be defined as a group of words functioning as a single unit in the syntax of a sentence and sharing a common meaning. For example, from the sentence "*James Gordon Brown is the Prime Minister of the United Kingdom and leader of the Labor Party*", we can extract a shorter phrase "*Prime Minister*". The meaning shared by these two words is the governmental career of James Gordon Brown.

Analogous to documents, which are particular arrangements of words in 1D space, images are particular arrangements of patches in 2D space. Such patches standing alone have low discrimination power. They are not independent but are likely to belong to the same physical object with each other and consequently, they are likely to have the same conceptual interpretation.

The inter-relationships among patches encode important information for our perception. Applying association rules, we used both the patches themselves and their inter-relationships

to obtain a higher-level representation of the data known as visual phrase.

In the proposed approach, the visual phrase is constructed from group of non-noisy visual words that share strong association rules and are located within the same local context (see the green circles in Figure 5). All local patches are within the same context whenever the distances between their centers are less or equal to a given threshold. Considering the set of all visual words (visual vocabulary) $W = \{w_1, w_2, ..., w_k\}$, D is a database (set of images I), $T = \{t_1, t_2, ..., t_n\}$ is the set of all different sets of visual words located in the same context.

An association rule is a relation of an expression $X \Rightarrow Y$, where X and Y are sets of items. The properties that characterize association rules are:

- The rule $X \Rightarrow Y$ holds in the transaction set T with support s if s % of transaction in T contain X and Y.
- The rule $X \Rightarrow Y$ holds in the transaction set T with confidence c if c % of transactions in T that contain X also contain Y.

Given a set of documents D, the problem of mining association rules is to discover all strong rules, which have a support and confidence greater than the pre-defined minimum support (*minsupport*) and minimum confidence (*minconfidence*). Although a number of algorithms have been proposed to improve various aspects of association rule mining, Apriori [2] remains the most commonly used because of its efficiency comparing to others. Since the aim is to discover the inter-relationships between different visual words, we consider the following:

- W denotes the set of items.
- T denotes the set of transactions.
- X and Y can be sets of one or more of frequent visual words that are within the same context.

After mining the whole transactions and finding the association rules, all visual words located in the same context and involved in at least one strong association rule will form the visual phrase. Figure 8 shows examples of visual phrases corresponding to four different categories (*electric guitar, brain, joshua tree, and rhino*). The upper part of the Figure 5 shows an example of a visual phrase that contains 3 different visual words whose indexes are 821, 2377, and 2712. These visual words are strongly correlated with each other since they are involved in strong association rules.

3.2.1 *Representation Scheme For Visual Phrase vocabulary*

For the purpose of indexing and retrieval, we need an efficient representation scheme to describe and store the visual phrase vocabulary. We devise a simple but efficient method based on *hashing*. A hash map that contains the indexes for all visual phrases is constructed to map groups of visual words (that are involved in at least one strong association rule and are within the same local context in a given query image) to visual phrase. The key is the *base 36* of c where c is the concatenating of the constituent visual words indexes after sorting. Figure 6 represents an example of five visual words(w_{2065},w_{621},w_{1191},w_{2130},w_{775}) mapped to a visual phrase p_{122} that has a hash key = *4Q28VUFALILE*

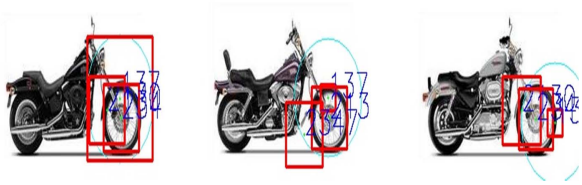

Figure 5: Examples of visual phrases corresponding to four different categories (electric guitar, brain, joshua tree, and rhino). The square resembles a local patch, which denotes one of the visual words, and the circle around the center of the patch denotes the local context .

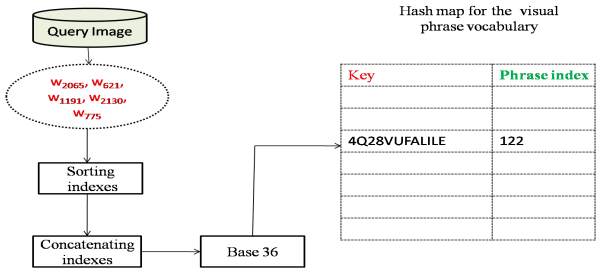

Figure 6: An example of five visual words (w_{2065},w_{621},w_{1191},w_{2130},w_{775}) mapped to a visual phrase p_{122}.

(base 36 of 621775119120652130). This internal representation scheme offers us several important benefits.

Firstly, the hash mapping of the visual words to visual phrases is way faster than using a plain list of visual phrases and it is also better from the binary search. For instance, binary search can locate an item in a sorted table of n items with $log_2 n$ key comparisons. Therefore, this hash map will be more efficient than binary search since no comparison with other items is needed.

Secondly, the choice of 36 is convenient and compact in that the digits can be represented using the Arabic numerals 0-9 and the Latin letters A-Z. Thus, we allocate less memory.

4. IMAGE REPRESENTATION, INDEXING AND RETRIEVAL

Given the proposed image representation discussed in Section 3, we describe here how images are represented, indexed and retrieved .

4.1 Image Representation

The traditional Vector Space Model [14] of Information Retrieval [14] is adapted to our representation, and used for similarity matching and retrieval of images. The following doublet represents each image in the model:

$$I = \begin{cases} \vec{W}_i \\ \vec{P}_i \end{cases} \quad (6)$$

where $\vec{W}_i$ and $\vec{P}_i$ are the vectors for the word and phrase representations of a document respectively:

$$\vec{W}_i = (w_{1,i}, ..., w_{n_w,i}) \ , \ \vec{P}_i = (p_{1,i}, ..., p_{n_p,i}) \quad (7)$$

Note that the vectors for each level of representation lie in a separate space. In the above vectors, each component represents the weight of the corresponding dimension. We used the ***spatial weight scheme*** defined in Section 3.1, for the words and the standard ***td.idf-weighting scheme*** for the phrases. Thus, we map images into documents and we apply document retrieval techniques to image retrieval.

4.2 Image Indexing

In our approach, we use an inverted file [18] to index images. The inverted index consists of two components: one includes indexed visual words and visual phrases, and the other includes vectors containing the information about the spatial weighting of the visual words and the occurrence of the visual phrases.

4.3 Similarity Measure And Retrieval

After representing the query image as a doublet of visual words and phrases, we consult the inverted index to find candidate images. All candidate images are ranked according to their similarities to the query image. We have designed a simple measure that allows evaluating the contribution of words and phrases. The similarity measure between a query I_q and a candidate Image I_c is estimated with:

$$sim(I_q, I_c) = (1-\alpha)RSV(\vec{W}_c, \vec{W}_q) + (\alpha)RSV(\vec{P}_c, \vec{P}_q) \quad (8)$$

The Retrieval Status Value (RSV) of 2 vectors is estimated with the cosine distance. The non-negative parameter α is to be set according the experiment runs in order to evaluate the contribution between visual words and visual phrases.

5. EXPERIMENTS

This section describes the set of experiments we have performed to explore the performance of the proposed methodology. Firstly, we investigate the performance of the proposed approach and the average number of visual words on each class of images after filtering the noisy visual words. Secondly, we evaluate the effect of fusion Edge context descriptor with SURF. Thirdly, we compare the performance of the spatial weighting scheme with the traditional bag-of-visual-words. Fourthly, we evaluate the contribution between visual words and visual phrases. Finally, we compare our approach to Zheng et al. approach [20]. To evaluate the performance, we compute the average precision (AP) for the top 20 retrieved images for each class of images and the mean average precision (MAP).

5.1 Dataset And Experimental Setup

The image dataset used for these experiments is the Caltech101 Dataset1 [8]. It contains 8707 images, which includes objects belonging to 101 classes. The number of images in each class varies from about 40 to about 800 with an average of 50 images. For the various experiments, we construct the test data set by selecting randomly 10 images from each class (1010 images). The query images are picked from this test data set during the experiment. The visual word vocabulary size (K)=3000 and the visual phrase vocabulary size is 960.

Firstly, we run experiments with a similarity matching parameter α=0 in order to compare our spatial weighting scheme with other approaches. Then, we evaluate the contribution between words and phrases by running the experiments several times with different values of α.

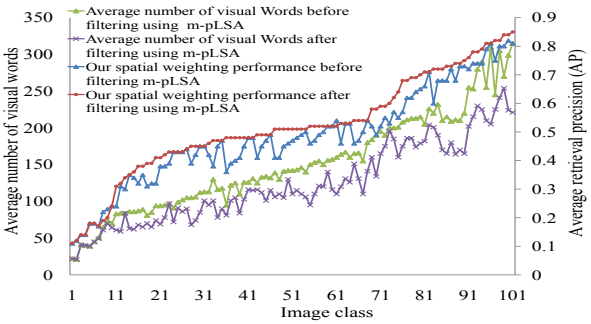

Figure 7: Evaluation of the performance of the proposed approach and the average number of visual words on each class of images after filtering the noisy visual words.

5.2 Assessment Of The Spatial Weighting Performance

5.2.1 *Evaluation Of The Performance Of The Proposed Approach And The Average Number Of Visual Words After Filtering The Noisy Visual Words*

In this section we show the influence of filtering noisy visual words based on the m-pLSA and we contribute on the relation between the average number of visual words in each class and the corresponding retrieval performance. Figure 7 plots the average retrieval precision (AP) for our spatial weighing approach before and after filtering. In addition, it plots the corresponding average number of visual words for each class. For a clearer presentation, we arrange the 101 classes from left to right in the figure with respect to the ascending order of their AP after filtering.

On one hand, it is obvious from the results displayed that the performance has slightly improved after filtering especially in the classes that have huge amounts of words comparing to others that have small amounts. On the other hand, there is a variation of retrieval performance among all 101 classes, and this variation is related to the average number of the visual words. Figure 7 shows a clear difference in the average number of visual words between the classes that highly perform and the classes that have poor performance.

The number of visual words on an image depends on the interest point detector, as we mentioned before that we use the Fast Hessian detector which is faster comparing to others. The computational time for detecting the interest points is reduced by using image convolutions based on integral images. Having these convolutions decreases the number of detected interest points and this contributes as a limitation for Fast Hessian in images with rare texture.

5.2.2 *Effect Of Fusing Our Edge Context Descriptor With SURF*

As we mentioned in Section 2, we employ the SURF descriptor and introduce a novel Edge context descriptor in order to extract local features. Figure 8 shows the empirical investigation on the influence of fusing local descriptors on the dataset. As we have done in the previous section, we arrange the 101 classes from left to right in the figure with respect to the ascending order of their average retrieval precision af-

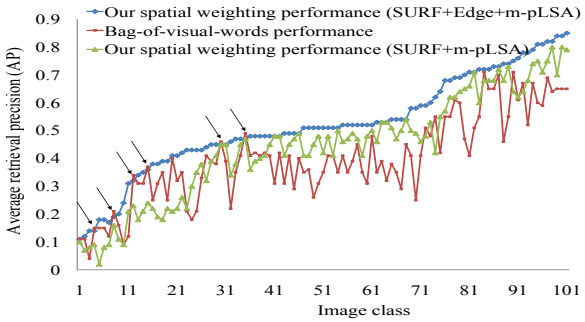

Figure 8: Effect of fusing our Edge context descriptor with SURF comparing to the traditional bag-of-visual-words

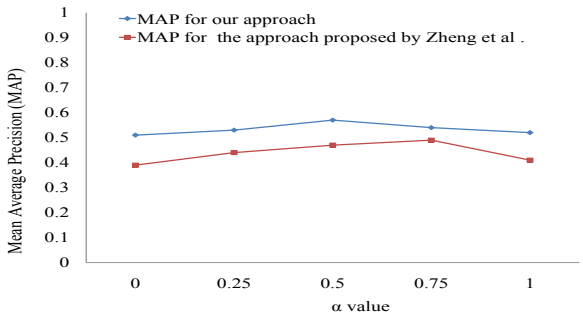

Figure 9: Contribution of viusal words and viusal phrases in our approach and in Zheng et al. approach.

ter filtering and fusing local descriptors. It is obvious from the results displayed, that the spatial weighting approach performs better with the two local descriptors merged especially for the classes that have few number of visual words like leopards sea-horse, human face, achor, bass. Introducing the Edge context for retrieving images from such classes is significant.

5.2.3 *Comparing The Spatial Weighting Performance With The Traditional Bag-Of-Visual-Words*

We compare the proposed spatial weighting scheme to the traditional **bag-of-visual-words** approach [15]. The experiment on the traditional Bag-of-Visual-Words is done after the choice of the optimal number of clusters $K2$=3500 in order to be fair. Figure 8 shows the average retrieval performance for the two approaches. Similarly as before, we arrange 101 classes from left to right with respect to the ascending order of their average retrieval precision in order to get a clearer representation.

It is obvious from the results displayed, that the spatial weighting approach globally outperforms the traditional approach except for 6 image classes out of 101 in the used data set. The 6 classes are **dolphin, revolver, metronome, lotus, pyramid, sunflower and stegosaurus**. Having this difference over a data set containing 101 classes emphasizes the good performance of the proposed approach.

5.3 Evaluation Of Constructing The Higher-Level Visual Representation (Visual Phrase)

In the previous section, we demonstrated the good performance of the spatial weighting approach. In this section we will explain the importance of introducing the visual phrase and it is influence on the over all performance of the system. In addition, we compare our approach to Zheng et al. approach.

5.3.1 *Evaluation of the Contribution of Visual Words and Phrases*

We combine visual phrase and visual word representations by varying the parameter α used in the similarity matching approach. Figure 9 plots the MAP for different values of α over all 101 classes. When considering only visual phrases in the similarity matching ($\alpha = 1$), the MAP is slightly better than the scenario in which only visual words are used (($\alpha = 0$)). However, the combination of both yields better

Figure 10: Examples of retrieved images based on (visual words + visual phrases) and visual words alone.

results than using words or phrases separately. Figure 10 shows some examples of the retrieved images for different query images. The left part of the figure shows the query images and the right shows the retrieved images. For each query image, 10 retrieved images are represented where the first 5 is top five retrieved images using ($\alpha = 0.5$) and the others are top five retrieved images using ($\alpha = 0$).

The explanation for the need of both (word and phrase) in the representation is that there are some images, which are not texture-rich like human face, stop sign or umbrella pictures. In these images, the number of detected interest points is small which leads to a very small number of phrases. From this study, we conclude that visual phrase alone can not capture all the similarity information between images

and the visual word similarity is still required.

5.3.2 Comparing The proposed Approach Performance With Zheng et al. Approach

We compare the proposed approach to another approach that has been introduced by Zheng et al. as an effective and efficient object-based image retrieval system. We compare our approach to this since we share the same objective by introducing a higher-level visual presentation and the data set used by this approach is same like ours. It is obvious from the results displayed in Figure 9, that our approach globally outperforms this approach. Visual phrases, in Zheng et al. approach, are defined as pairs of adjacent local image patches and constructed based on the analysis of visual words occurrences to retrieve images containing desired objects. However,we define the visual phrases as set of visual words that are strong correlated.

This results can be a good contribution for the effectiveness of the association rules as a frequent mining technique in image retrieval. Moreover, these results contribute also the effectiveness of other techniques that are mentioned in the proposed methodology like spatial weighting, Edge context descriptor and elimination of noisy visual words using m-pLSA. Zheng et al. made an analogy between information retrieval and image retrieval which shows good result. However, the best text techniques cannot guarantee the same result in images since image is a particular arrangement of different pixels in a 2D space and its spatial arrangement should be taken in consideration.

6. CONCLUSION

We successfully drew out an analogy between the techniques in text retrieval and image retrieval by building visual words and phrases. Visual words were constructed by extracting features based on fusion of descriptors. A new "spatial weighting" technique has been introduced, which enhances the basic 'bag of-visual-words' approach by using spatial relations. Finally, we devised methods to construct visual phrases based on association rule technique. Our experimental studies showed that a combined use of words and phrases could perform better than using them separately. It also showed good performance compared to similar recent approaches.

In our future work, we will investigate the usage of such model on proposing computer vision solutions like human behavior analysis from video. We will work on further justification based on other datasets such as TRECVID and NUS-WIDE.

7. REFERENCES

[1] J. Y. 0003, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo. Evaluating bag-of-visual-words representations in scene classification. In *Multimedia Information Retrieval*, pages 197–206, 2007.

[2] R. Agrawal, T. Imielinski, and A. N. Swami. Mining association rules between sets of items in large databases. In P. Buneman and S. Jajodia, editors, *SIGMOD Conference*, pages 207–216. ACM Press, 1993.

[3] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[4] H. Bay, A. Ess, T. Tuytelaars, and L. J. V. Gool. Speeded-up robust features (surf). *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[5] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(4):509–522, 2002.

[6] J. A. Bilmes. A gentle tutorial on the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *U.C. Berkely*, 1997.

[7] J. Canny. A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 8(6):679–698, November 1986.

[8] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.*, 106(1):59–70, 2007.

[9] K. M. Hammouda and M. S. Kamel. Efficient phrase-based document indexing for web document clustering. *IEEE Trans. Knowl. Data Eng.*, 16(10):1279–1296, 2004.

[10] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *ICCV*, pages 604–610, 2005.

[11] R. Lienhart, S. Romberg, and E. Hörster. Multilayer plsa for multimodal image retrieval. In *CIVR*, 2009.

[12] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[13] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *CVPR (2)*, pages 2161–2168, 2006.

[14] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.

[15] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477. IEEE Computer Society, 2003.

[16] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, volume 1, pages I–511–I–518 vol.1, 2001.

[17] J. Willamowski, D. Arregui, G. Csurka, C. R. Dance, and L. Fan. Categorizing nine visual classes using local appearance descriptors. In *In ICPR Workshop on Learning for Adaptable Visual Systems*, 2004.

[18] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigabytes: Compressing and Indexing Documents and Images, Second Edition*. Morgan Kaufmann, 1999.

[19] J. Yuan, Y. Wu, and M. Yang. Discovery of collocation patterns: from visual words to visual phrases. In *CVPR*, 2007.

[20] Q.-F. Zheng and W. Gao. Constructing visual phrases for effective and efficient object-based image retrieval. *TOMCCAP*, 5(1), 2008.