Toward a higher-level visual representation for content-based image retrieval

Ismail Elsayad, Jean Martinet, Thierry Urruty, Chabane Djeraba

Toward a higher-level visual representation for content-based image retrieval

Ismail El sayad · Jean Martinet · Thierry Urruty · Chabane Djeraba

Abstract Having effective methods to access the desired images is essential nowadays with the availability of a 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 on 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 of 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.

Keywords SURF · Content-based image retrieval · Visual words · Visual phrases · Gaussian mixture model · Spatial weighting · pLSA

I. El sayad (✉) · J. Martinet · T. Urruty · C. Djeraba
LIFL/CNRS-UMR 8022, University of Lille 1, Telecom Lille 1, Lille, France
e-mail: ismail.elsayad@lifl.fr

J. Martinet
e-mail: jean.martinet@lifl.fr

T. Urruty
e-mail: thierry.urruty@lifl.fr

C. Djeraba
e-mail: chabane.djeraba@lifl.fr

 Springer

1 Introduction

Due to the explosive spread of digital devices, the amount of 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) [29]. 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 [2]. Recently, many image retrieval systems have shown that the part-based representation for image retrieval [37] is much superior to 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 [17, 28, 32] 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. 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. [3], 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 [1]. 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. In Section 4, we describe the method for mining visual phrases from visual words to obtain the final image representation. In Section 5, we present an image similarity method based on visual words and visual phrases. We report the experimental results in Section 6, and we give a conclusion to this article in Section 7.

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 [2]. The text documents are first parsed into words. Second, the words are represented by their stems: for example "walk", "walking" and "walks" would be represented by the stem "walk". Third, 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. In the popular Vector Space Model [27], for instance, a vector represents each document, with components given namely by the frequency of occurrence of the words in the document. The search task is performed by comparing the query vector to the document vectors, and by returning the most similar documents, i.e. the documents with the closest vectors, as measured by the cosine distance.

Hammouda and Kamel [11] 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.

The analogy with images considers that an image is represented as a bag of visual words with a given topology. 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 [17, 28].

Another similar part-based image represensations that are proposed recentlty are visterms [15, 23, 24], SIFT-bags [39] blobs [7], and VLAD [14] vector representation of an image which aggregates descriptors based on a locality criterion in the feature space. The different approach is the one proposed by Morand et al. [21]. This approach introduced scalable indexing of video content by objects without parsing them into their constituent elements. Morand et al. built a descriptor based on multi-scale histograms of wavelet coefficients of objects. In this case, the performance of the whole system will be so related to how much the process of the extracting the objects is accurate. Eventhough most of these part-based image representations report remarkable experimental resuts, the bag of visual words has drawn much attention recently, as it tends to code the local visual characteristics toward object level and achieves good results in representing variable object appearances caused by changes in pose, scale and translations [16, 37].

At the syntactic level, there is a correspondence between a text document and an image, providing 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 Fig. 1 we propose the syntactic granularities of an image and a text document, and *analogies* between their constituent elements. In this analogy, pixels correspond to letters, patches to words, and group of patches to phrases.

Fig. 1 Analogy between image and text document in syntactic granularity
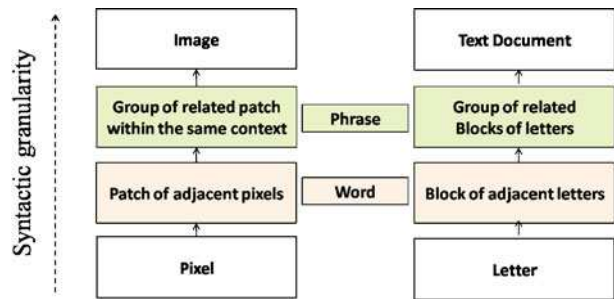
Zheng et al. [38] made an analogy between image retrieval and text retrieval, and have proposed a higher-level representation (*visual phrase*) based on the analysis of visual word occurrences to retrieve images containing desired objects. ***Visual phrases*** are defined as pairs of adjacent local image patches. The motivation of the visual phrase is to have a compact representation which has more discrimination power than the lower level (visual words). We share the same objective of designing a mid-level descriptor for representing images. However, while Zheng et al. consider adjacent pairs of patches only, the proposed approach is more general since it handles any set of items, which is more general than just pairs. In that way, we can represent more accurately the relations between objects.

Yuan et al. [36] 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 are also involved in strong association rules (to be discussed later in this article) which gives a higher representation level with more meaningful aspects.

Hoíng et al. in [13] proposed to construct another higher level representation (triplets of entities) from visual words (entities) by studying the spatial relationships between them. The proposed representation describes triangular spatial relationships with the aim of being invariant to image translation, rotation, scale, flipping, and robust to view point changes if required. Beside we share the same motivation for constructing a higher level representation, this approach lacks statistical and semantical learning for the lower level which is a pre-step to construct the higher level representation in our approach.

2.2 Weighting scheme

Inspired by the success of vector-space model in text document representation, the bag-of-visual-words approach usually converts images into vectors of visual words based on their weights. The term weighting is a key technique in information retrieval [26], and it is based on three different factors.

The first factor is the term frequency (*tf*). Terms that are frequently mentioned in individual documents, or document excerpts, appear to be useful in the recalling process. This suggests that a *tf* factor can be used as part of the term-weighting system measuring the frequency of occurrence of the terms in the document or query texts. Term-frequency weights have been used for many years in automatic indexing environments.

The second factor is the inverse document frequency (*idf*). Term frequency factors alone cannot ensure acceptable retrieval performance. Specifically, when the high frequency terms are not concentrated in a few particular documents, but instead are prevalent in the whole collection, all documents tend to be retrieved, and this affects the search precision. Hence a new collection-dependent factor must be introduced that favors terms concentrated in a few documents of a collection. The well-known inverse document frequency (*idf*) factor performs this function. The *idf* factor varies inversely with the number of documents n to which a term is assigned in a collection of N documents.

The third factor is the normalization factor. In addition to the term frequency and the inverse document frequency, the normalization factor appears useful in systems with widely varying vector lengths. In many situations, short documents tend to be represented by short term vectors, whereas much larger term sets are assigned to the longer documents. When a large number of terms are used for document representation, the chance of term matches between queries and documents is high, and hence the larger documents have a better chance of being retrieved than the short ones. Normally, all relevant documents should be treated as equally important for retrieval purposes. This suggests that a normalization factor can be incorporated into the term-weighting formula to equalize the length of the document vectors. The normalization factor converts the feature into unit length vector to eliminate the difference between short and long documents.

Yang et al. in [34] evaluated many frequency weighting schemes which are 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 a certain visual word *w* is typical among "building" images. An image containing 100 occurrences of *w* is not necessarily to be more likely a "building" image than an image containing only 25 occurrence 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.

2.3 Elimination of noisy words in bag-of-visual-words approaches

In bag-of-visual-words models for images, the vocabulary creation process, based on clustering algorithms such as k-means, is quite rude and leads to many noisy words. Such words add ambiguity in the image representation. Thus, it reduces the effectiveness of the retrieval processes. This problem has been addressed in the first video-Google paper by Sivic and Zisserman [28]: they used, as an analogy with text retrieval models, stop-lists that remove the most and least frequent words from the collection, which are supposed to be the most noisy.

Yang et al. have pointed at the inefficiency of stop-lists method and proposed several measures usually used in feature selection for text retrieval: document frequency (DF), x2 statistics (Chi^2), Mutual information (MI), and Point wise Mutual information (PMI). These selection measures remove the most uninformative words determined by each criterion. The vocabulary is reduced by 70% and the mean average precision has dropped merely by 5%, but after it drops at a much faster rate. This shows that feature selection is an effective technique in image retrieval but with some limitation. In comparison, a vocabulary in text categorization can be reduced by up to 98% without loss of classification precision [35].

Tirilly et al. in [30] have introduced another method to eliminate presumed useless visual words. This method aims at eliminating the noisiest words generated by the vocabulary building process, using the standard probabilistic latent semantic analysis (*pLSA*). The Standard *pLSA* was originally devised by Hofmann [12] in the context of text document retrieval, where words constitute the elementary parts of documents. The key concept of the *pLSA* model is to map the high dimensional word distribution vector of a document to a lower dimensional topic vector (also called aspect vector).

Therefore *pLSA* introduces a latent, i.e. unobservable, topic layer between the documents (i.e. images here) and the words. It is assumed that each document consists of a mixture of multiple topics and that the occurrences of words (i.e. visual words in images) is a result of the topic mixture. This generative model is expressed by the following probabilistic model:

$$P\left(d_i/w_j\right) = P\left(d_i\right) \sum_k P\left(z_k/d_i\right) P\left(w_j/z_k\right) \tag{1}$$

where $P(d_i)$ denotes the probability of a document d_i of the database to be picked, $P(z_k/d_i)$ the probability of a topic z_k given the current document, and $P(w_j/z_k)$ the probability of a word w_j given a topic. The model is graphically depicted in Fig. 2. Ni denotes the number of words which each of the M documents consists of. The eliminating process proposed by Tirilly et al. shows that this technique improves the performance of classifiers by eliminating only one third of the words. We share the same methodology but we used a *multilayer pLSA* rather than the standard *pLSA*. Eliminating the ambiguous visual word is a pre-step before learning the association rules to construct the visual phrases.

2.4 Association rules

Association rules learning is a popular and well researched analogy for discovering interesting relations between variables in large databases. They are popular in sale transaction analysis, especially for market basket analysis.

Fig. 2 Standard pLSA-model
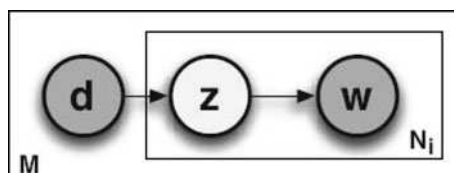
Haddad et al. [10] discussed how to use **association rules**, to discover knowledge about **relations between terms** without any pre-established thesaurus, hierarchical classification or background knowledge. They used these relations between terms to expand queries and they showed how it could be advantageous for information retrieval.

Given a set of items and a set of transactions, the ***confidence*** between two item sets (X and Y) can be defined as the chance these two items occur within the same transaction. The ***support*** can be defined as the percentage of transactions containing both item sets. A rule is evaluated as strong if its confidence exceeds a ***confidence threshold*** and its support exceeds a ***support threshold***.

Given a set of documents D, the problem of mining association rules is to discover all the rules whose support and confidence are greater than some pre-defined minimum support and minimum confidence. Although a number of algorithms are proposed improving various aspects of association rule mining, Apriori by Agrawal et al. [1] remains the most commonly used algorithm. Haddad et al. [10] have applied association rules to text analysis. Their work aims at extracting the terminology from a text corpus by using patterns applied after a morphological analysis. The terminology is structured with automatically extracted dependencies relations. This extracted terminology enables a more precise description of the documents.

Association rules have been used subsequently for discovering relevant patterns in several types of data, namely to extract phrases from text. An approach called Phrase Finder is proposed to construct a collection-dependent association thesauri automatically using large full text document collections. The association thesaurus can be accessed through natural language queries in INQUERY,, an information retrieval system based on the probabilistic inference network. The main idea is to build this description with phrases after finding some statistical associations between individual words in the collection.

Martinet and Satoh [20] adapted the definition of association rules to the context of perceptual objects, for merging strongly associated features, in order to get a more compact representation of the data. The building of the mid-level representation is done by iteratively merging features corresponding to frequently occurring patterns (likely to correspond to physical objects), which are involved in strong association rules.

Our approach is inspired by this approach since we construct the representation space of visual phrases from visual words that are involved in strong association within the same local context. The new representation space of visual phrases enables a better separation of images, that is to say that index terms in the new space have a higher discriminative power, and consequently are likely to yield a more precise search.

3 Visual word construction

In this section, we describe different components of the chain of processes in constructing the visual words. Figure 3 presents the different process starting from detecting interest and edge points till the image description of the image by visual words before introducing the higher level of representation. We use the fast Hessian detector [3] to extract interest points. In addition, the canny edge detector [6] is
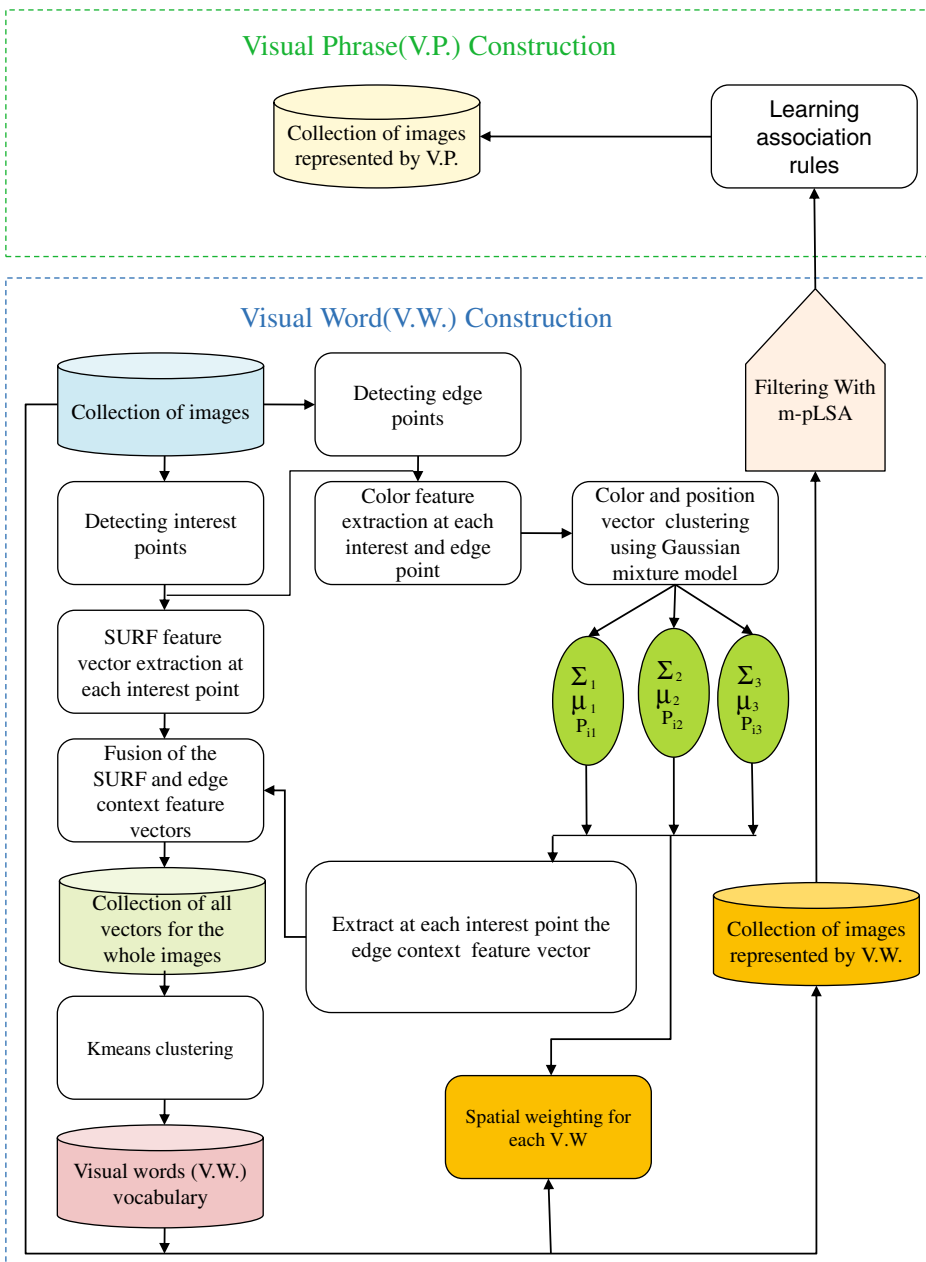
Visual Phrase(V.P.) Construction

Collection of images represented by V.P. ← Learning association rules

Visual Word(V.W.) Construction

Collection of images → Detecting edge points

Detecting interest points

Color feature extraction at each interest and edge point → Color and position vector clustering using Gaussian mixture model

SURF feature vector extraction at each interest point

Σ_1 μ_1 P_{i1} Σ_2 μ_2 P_{i2} Σ_3 μ_3 P_{i3}

Fusion of the SURF and edge context feature vectors

Collection of all vectors for the whole images

Extract at each interest point the edge context feature vector

Collection of images represented by V.W.

Kmeans clustering

Visual words (V.W.) vocabulary

Spatial weighting for each V.W

Filtering With m-pLSA

Fig. 3 Flow of information in the visual document representation model

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 Fig. 3). 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.

3.1 Gaussian mixture model

In this approach, based on the Gaussian Mixture Model (GMM) [5], 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: $f_1, ..., f_m$ can be extracted.

The set of points is assumed to be a mixture of n Gaussian in the 5-dimensional color-spatial feature space and the Expectation-Maximization (EM) [8] algorithm is used to iteratively estimate the parameter set of the Gaussians. The parameter set of the Gaussian mixture is: $\theta = \{\mu_i, \Sigma_i, P_i\}, i = 1, ..., n$ where μ_i, Σ_i, and P_i are the mean, the covariance, and the prior probability of the ith Gaussian cluster respectively.

By applying Bayes theorem at each E step, we estimate the probability of a particular feature vector f_j belonging to the i th Gaussian according to the outcomes from the last maximization step as the following:

$$P\left(g_j = i \mid f_j, \theta_t\right) = \frac{P\left(f_j \mid g_j = i, \theta_t\right) P\left(g_j = i \mid \theta_t\right)}{P\left(f_j\right)} \tag{2}$$

$$P\left(f_j\right) = \sum_{k=1}^{n} P\left(f_j \mid g_j = k, \theta_t\right) P\left(g_j = i \mid \theta_t\right) \tag{3}$$

In which g_j denotes the Gaussian which f_j comes from and θ_t is the parameter set at the t th iteration.

At each M-step, the parameter set of the n Gaussians is updated toward maximizing the log -likelihood, which is:

$$Q(\theta) = \sum_{j=1}^{m} \sum_{i=1}^{n} P\left(g_j = i \mid f_j, \theta_t\right) ln\left(P\left(f_j \mid g_i = i, \theta_t\right) P\left(g_j \mid \theta_t\right)\right) \tag{4}$$

At the final step of the EM algorithm, we obtain all the parameters needed to construct our set of Gaussians and each point is assigned to one of the Gaussians.

3.2 Extracting and describing local features

In our approach, we use the SURF low-level feature descriptor that describes how the pixel intensities are distributed within a scale-dependent neighborhood of each interest point detected by the Fast-Hessian. This descriptor is similar to the SIFT one [19], but Bay et al. have used integral images [31] in conjunction with filters known as Haar wavelets in order to increase the robustness and decrease the computation time. Haar wavelets are simple filters which can be used to find gradients in the x and y directions. The extraction of the descriptor can be divided into two distinct tasks.

Firstly, the square regions centered around each interest point are constructed, and oriented along the orientation at the interest point. The size of this window is 20 times as big as the scale of the detected interest point and the region is split up regularly into smaller 4 × 4 square sub-regions (see Fig. 4).

Secondly, Haar wavelet response in horizontal direction (*dx*) and in vertical direction (*dy*) are summed up over each sub-region and form a first set of entries to the feature vector. In order to bring in information about the polarity of the intensity changes, the sum of the absolute values of the *dx* and *dy* responses are also extracted. Hence, each sub-region has a four-dimensional descriptor vector *v* for its underlying intensity structure as the following:

$$v = \left(\sum dx, \sum dy, \sum |dx|, \sum |dy| \right) \tag{5}$$

This results in a descriptor vector for all 4 × 4 sub-regions with 64 dimensions.

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. [4] 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
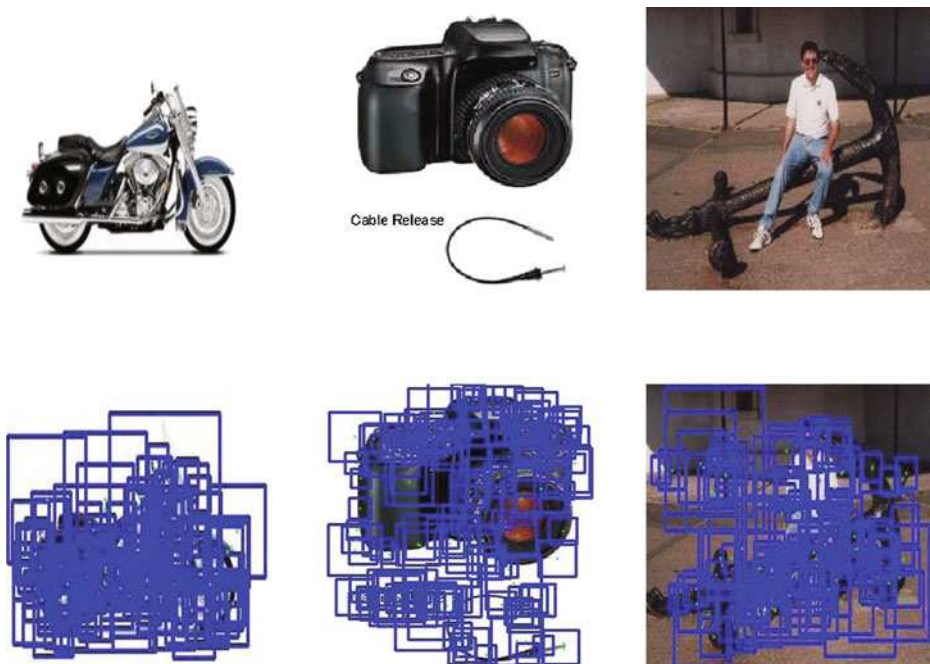

Fig. 4 Examples of images after Surf features extraction

distribution over relative positions is a robust, compact, and highly discriminative descriptor.

As shown in Fig. 5, 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*). For this novel descriptor many invariances is 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.
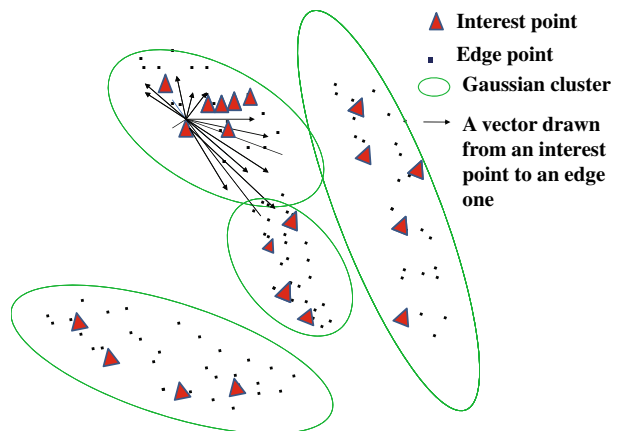
Following the visual construction part in Fig. 3, after extracting the Edge context feature, the fusion between this descriptor and the SURF descriptor is performed. This fused feature vector is composed of 88 dimensions (64 from SURF + 24 from the Edge context descriptor). Hence, the new feature vector describes the information on the distribution of the intensity and the edge points of the image. It enriches our image representation with more local information.

3.2.1 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 [22] 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 related to traditional direct k-means clustering in cases where k is large. Firstly, clustering is more efficient

Fig. 5 Extraction of the Edge context descriptor in the 2D spatial space where the points have already been clustered before in the 5-dimensional color-spatial Gaussian space

▲ Interest point
▪ Edge point
◯ Gaussian cluster
⟶ A vector drawn from an interest point to an edge one

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 6 shows an example of a fused feature vector assigned into a discrete visual word 6.

3.3 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. [18] proposed a multilayer multi-modal probabilistic Latent Semantic Analysis (*mm-pLSA*). The proposed approach (*mm-pLSA*) has two modes: one mode for visual words and the 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:

$$P\left(I/w_l\right) = \sum_{i=1}^{P} \sum_{j=1}^{V} P(I) P\left(z_i^t/I\right) P\left(z_j^v/z_i^t\right) P\left(w_l/z_j^v\right) \tag{6}$$

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) being the same as the total number of categories of
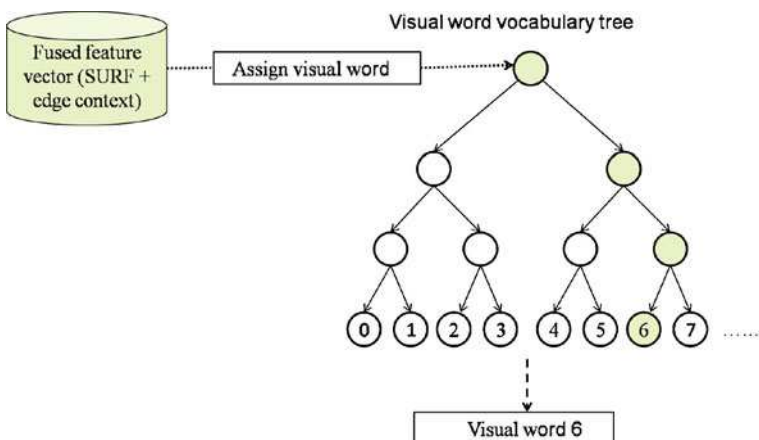

Fig. 6 Example of assigning a fused feature vector into a discrete visual word

the image dataset. The total number of the 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 are categorized as irrelevant. After that, we eliminated all visual words whose probability $P(w_l/z_j^v)$ is low with respect 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 7 shows examples of images after eliminating ambiguous visual words. The experiments reported in Section 5 show that this technique improves the performance of 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

Fig. 7 Examples of images after filtering the noisy visual words using *m-pLSA*

Before filtering   After filtering
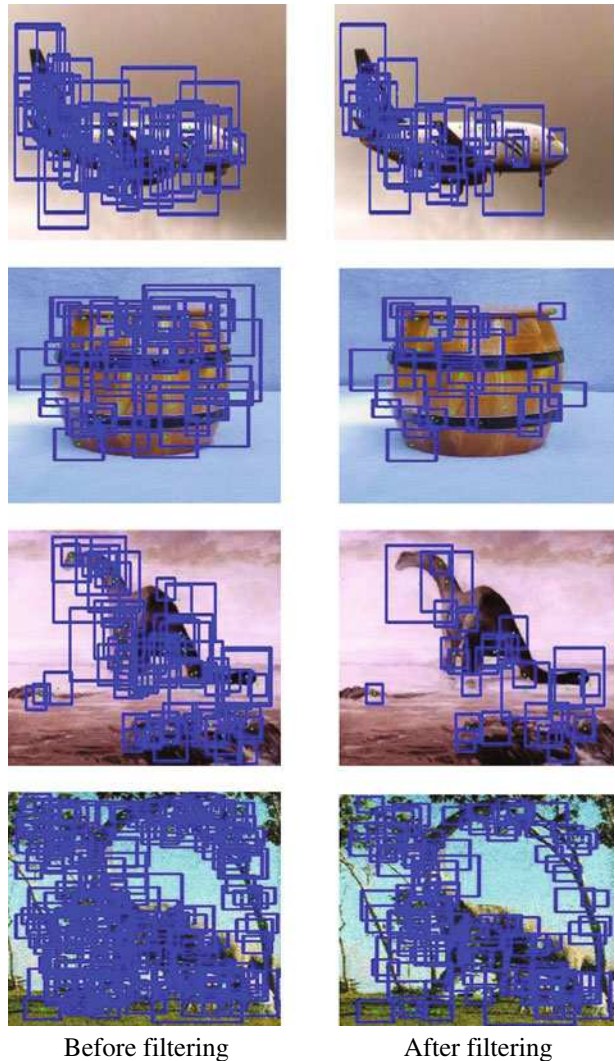
higher-level aspects. This is very natural since images consist of multiple objects and belong to different categories.

3.4 Spatial weighting for the visual words

To perform the **spatial weighting**, we innovate a new scheme which differs from the **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 g_i. Therefore, the weighted term frequency $(Tf_{w_l g_i})$ of a visual word w_l with respect to a Gaussian g_i is defined as follows:

$$Tf_{w_l \beta_i} = \sum_{m=1}^{n_i} P(g_i / f_m) \tag{7}$$

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:

$$Tf_{w_l} = \sum_{i=1}^{n_{w_l}} \left(Tf_{w_l g_i} \right) / n_{w_l} \tag{8}$$

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

$$If_{w_l} = ln \frac{n}{n_{w_l}} \tag{9}$$

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

$$Sw_{w_l} = Tf_{w_l} \times If_{w_l} \tag{10}$$

4 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
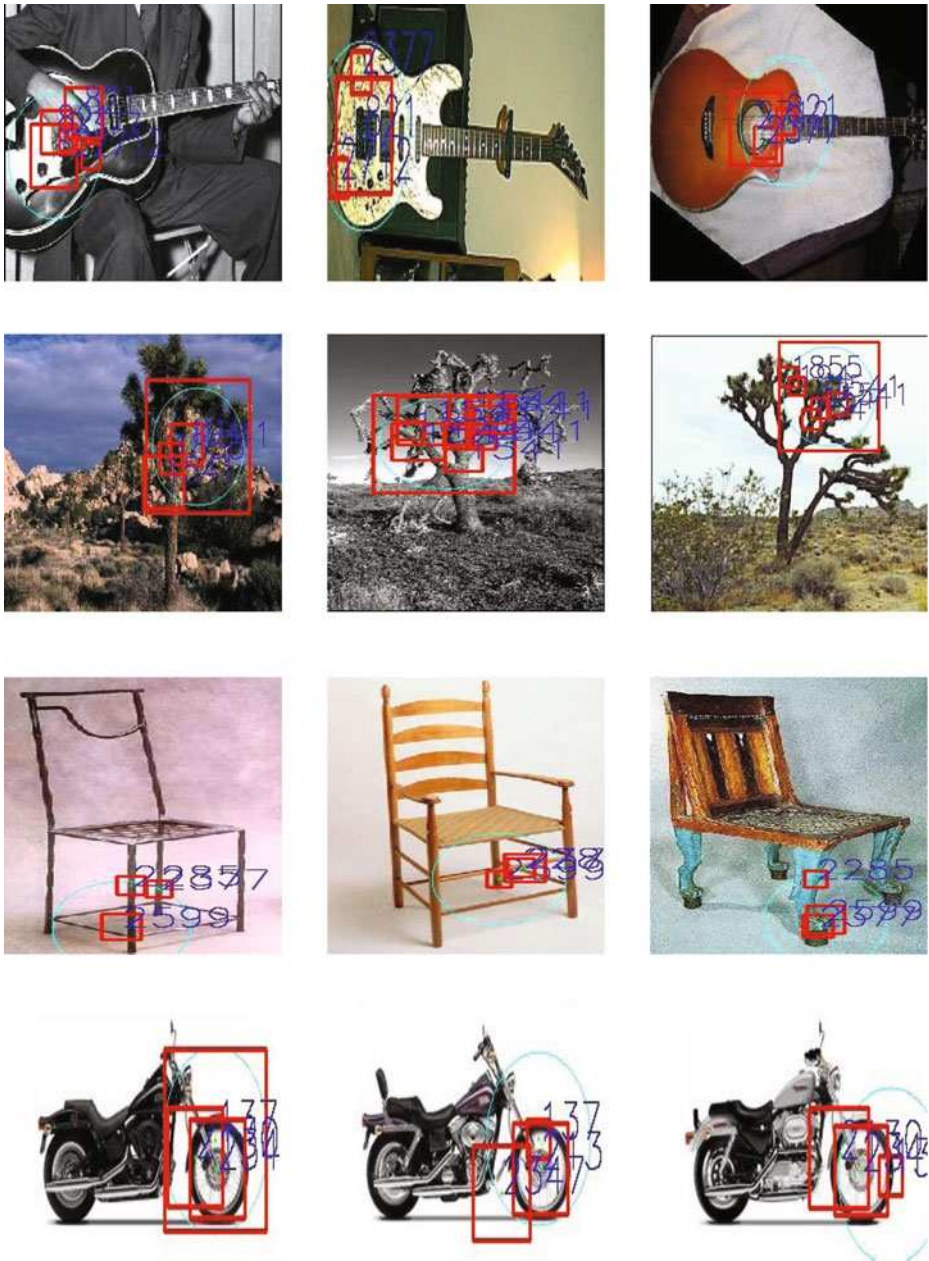
Fig. 8 Examples of visual phrases corresponding to four different categories (electric guitar, brain, joshua tree, and rhino). The square resembles a local patch, which denotes one of the visual words, and the circle around the center of the patch denotes the local context

as visual phrase.This higher level of representation is will increase the low discrimination power of the visual words and enhance the the retrieval performance of the system.

4.1 Mining association rules

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 Fig. 8). 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 [1] 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 the 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 Fig. 8 shows an example of a visual phrase that contains three 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.

4.2 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 phrases. The key is the *base 36* of c where c is the concatenating of the constituent visual words indexes after sorting. Figure 9 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* (base 36 of 621775119120652130). This internal representation scheme offers us several important benefits.
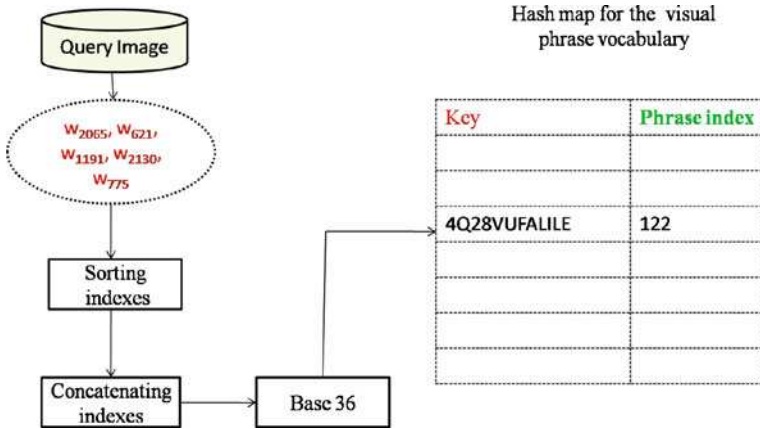
Fig. 9 An example of five visual words ($w_{2065}, w_{621}, w_{1191}, w_{2130}, w_{775}$) mapped to a visual phrase p_{122}

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.

5 Image representation, indexing and retrieval

Given the proposed image representation discussed in Section 4, we describe here how images are represented, indexed and retrieved .

5.1 Image representation

The traditional Vector Space Model [27] of Information Retrieval [25] is adapted to our representation, and used for similarity matching and retrieval of images. The following doublet represents each image in the model:

$$I = \begin{cases} W_i \\ P_i \end{cases} \tag{11}$$

where W_i and P_i are the vectors for the word and phrase representations of a document respectively:

$$W_i = \left(w_{1,i}, \ ..., \ w_{n_w,i} \right) \ , \ P_i = \left(p_{1,i}, \ ..., \ p_{n_p,i} \right) \tag{12}$$

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, 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.

5.2 Image indexing

In our approach, we use an inverted file [33] 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.

5.3 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 to evaluate the contribution of words and phrases. The similarity measure between a query I_q and a candidate image I_c is estimated with:

$$sim\left(I_q, I_c\right) = (1 - \alpha)\,RSV\left(W_c, W_q\right) + (\alpha)\,RSV\left(P_c, P_q\right) \tag{13}$$

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.

6 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 [38].

6.1 Dataset and experimental setup

The image dataset used for these experiments is the Caltech101 Dataset1 [9]. It contains 8,707 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 randomly selecting 10 images from each class (1,010 images). The rest of the data set was used as a training dataset. The query images are picked from the test data set during the experiment. The visual word vocabulary size (K) = 3,000 and the visual phrase vocabulary size is 960.

Firstly, we run experiments with a similarity matching parameter $\alpha = 0$ 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.

6.2 Evaluation criteria: MAP, MAPC and AP

The evaluation criteria here is the mean average precision (MAP), which is the mean of average precision (AP) of each query. The AP is the sum of the precisions at each relevant hit in the retrieval list, divided by the total number of relevant images in the collection. AP is defined as:

$$AP = \frac{\sum_{r=1}^{R} \times rel(r)}{T} \tag{14}$$

where r is image rank, R is the total number of images retrieved, $Prec(r)$ is the precision of retrieval list cut-off at rank r, $rel(r)$ is an indicator (0 or 1) of the relevance of rank r, and T is the total number of relevant images in the corpus. The average precision is an ideal measure of retrieval quality, which is determined by the overall ranking of relevant images. Intuitively, MAP gives higher penalties to fault retrievals if they have higher position in the ranking list. This is rational, as in practice, searchers are more concerned with the retrieved results in the top.

In addition to the MAP, we estimate the mean average precision for a given class ($MAPC$). It is defined as the mean of the AP of each query within a given class. This evaluation criteria helps to compare the performance of the different classes within the given data set.

6.3 Assessment of the spatial weighting performance

6.3.1 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 to the relation between the average number of visual words in each class and the corresponding retrieval performance. Figure 10 plots ($MAPC$) 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 MAPC after filtering.

On the 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 visual words. Tables 1 and 2 show a clear difference in the average number of visual words between the classes that highly perform and the classes that have poor performance.
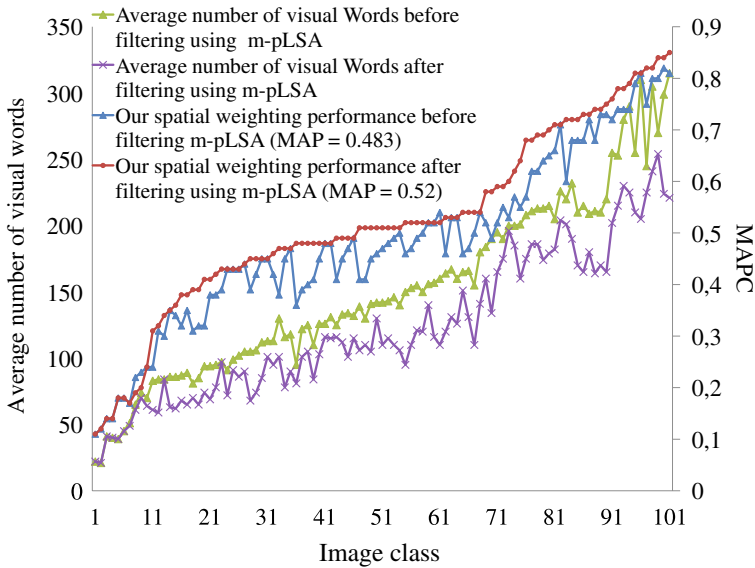
Fig. 10 Evaluation of the performance of the proposed approach and the average number of visual words on each class of images after filtering the noisy visual words

Where AVG denotes the average number of visual words in each class before filtering, the average number of visual words in each class.

This huge difference in the average number of visual words due to the interest point detector, as we mentioned before that we use the Fast Hessian detector which is faster comparing to the 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.

6.3.2 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 11 shows the

Table 1 The 10 classes with the highest MAPC

Class	AVG before filtering	AVG after filtering	MAPC before filtering	MAPC after filtering
Wild-cat	253	215	0.74	0.78
Brain	280	230	0.74	0.78
Accordion	290	225	0.74	0.79
Euphonium	255	210	0.79	0.81
Dragonfly	310	205	0.81	0.81
Crocodile	245	225	0.75	0.82
Gerenuk	305	241	0.8	0.82
Crocodile-head	270	254	0.8	0.84
Tick	299	224	0.82	0.84
Trilobite	315	221	0.81	0.85

| Table 2 The 10 classes with the lowest MAPC	Class	AVG before filtering	AVG after filtering	MAPC before filtering	MAPC after filtering
	Mayfly	22	22	0.11	0.11
	Leopards	21	21	0.12	0.12
	Sea-horse	41	41	0.14	0.14
	Pyramid	40	40	0.14	0.14
	Anchor	39	39	0.18	0.18
	Bass	45	45	0.18	0.18
	Wrench	51	49	0.17	0.17
	Revolver	65	61	0.22	0.19
	Flamingo-head	74	70	0.23	0.2
	Mandolin	70	64	0.24	0.24

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 MAPC after 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, anchor, bass. Introducing the Edge context for retrieving images from such classes is significant.

6.3.3 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 [28]. The experiment on the traditional Bag-of-Visual-Words is done after the choice of the optimal number of clusters $K2 = 3,500$ in order to be
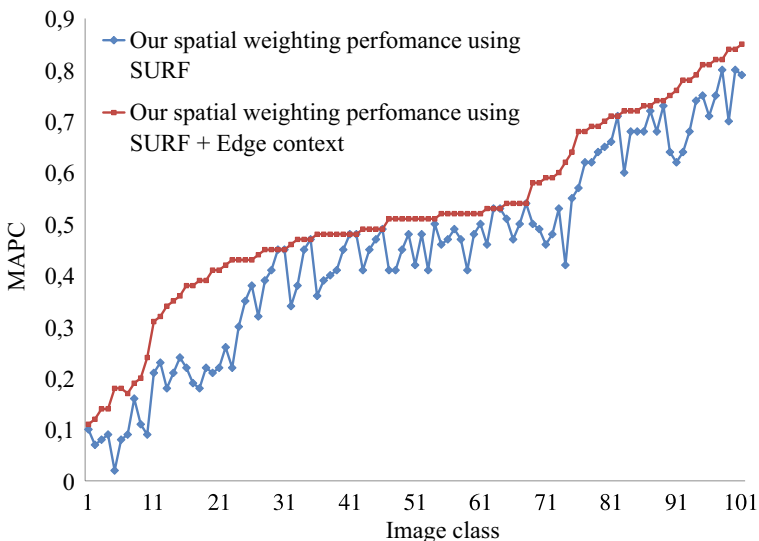
Fig. 11 Effect of fusing our Edge context descriptor with SURF

fair. Figure 12 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 MAPC 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 six image classes (pyramid, revolver, dolphin, metronome, lotus, sunflower and stegosaurus) out of 101 in the used data set. It noticed that there is remarkable change in illumination within the images of the six categories especially the Dolphin category. This can be noticed as SURF limitation in handling illumination changes.

The *MAP* of the traditional approach is 0.39 which is less than the proposed approach (0.520). Having this difference over a data set containing 101 classes emphasizes the good performance of the proposed approach.

6.4 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 its influence on the overall performance of the system. In addition, we compare our approach to Zheng et al.'s approach.

6.4.1 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 13 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
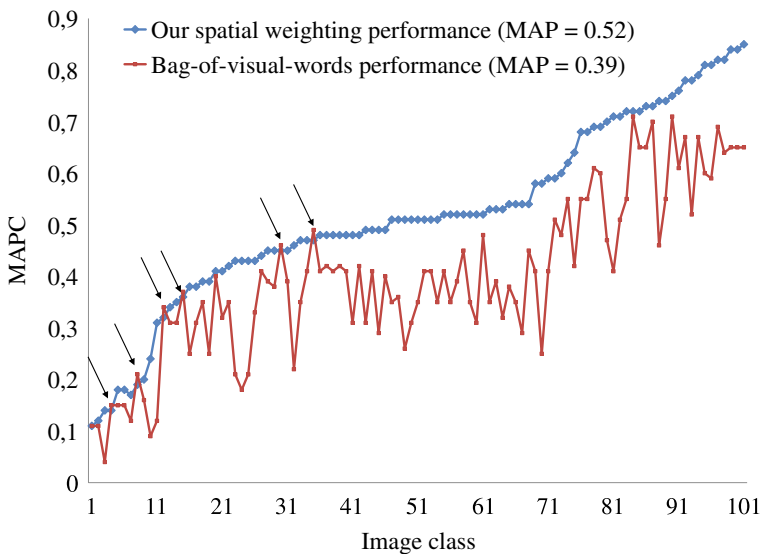

Fig. 12 Comparison between the spatial weighting approach performance and the traditional bag-of-visual-words

scenario in which only visual words are used (($\alpha = 0$)). However, the combination of both yields better results than using words or phrases separately. Figure 14 shows some examples of the retrieved images for different query images. The left part of the figure shows the query images, and the right one shows the retrieved images. For each query image, 10 retrieved images are represented where the first five are the top five retrieved images using ($\alpha = 0.5$), and the others are the 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.

6.4.2 Comparing the proposed approach performance with Zheng et al.'s 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 one since we share the same objective by introducing a higher-level visual representation, and the data set used by this approach is the same as ours. It is obvious from the results displayed in Fig. 13 that our approach globally outperforms this approach. Visual phrases, in Zheng et al. approach, are defined as pairs of adjacent local image patches and are constructed based on the analysis of visual words occurrences to retrieve images containing desired objects. However, we define the visual phrases as a set of visual words that are strongly correlated.

These results can be a good contribution for the effectiveness of the association rules as a frequent mining technique in image retrieval. Moreover, these results
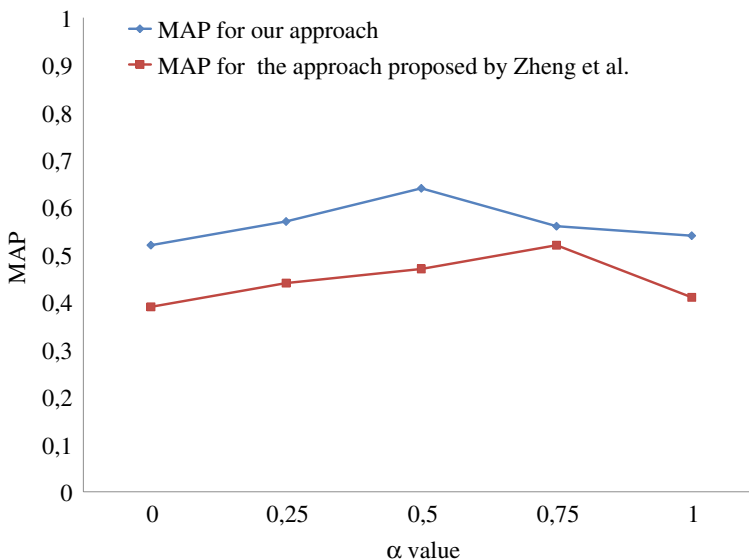

Fig. 13 Contribution of visual words and visual phrases in our approach and in Zheng et al. approach
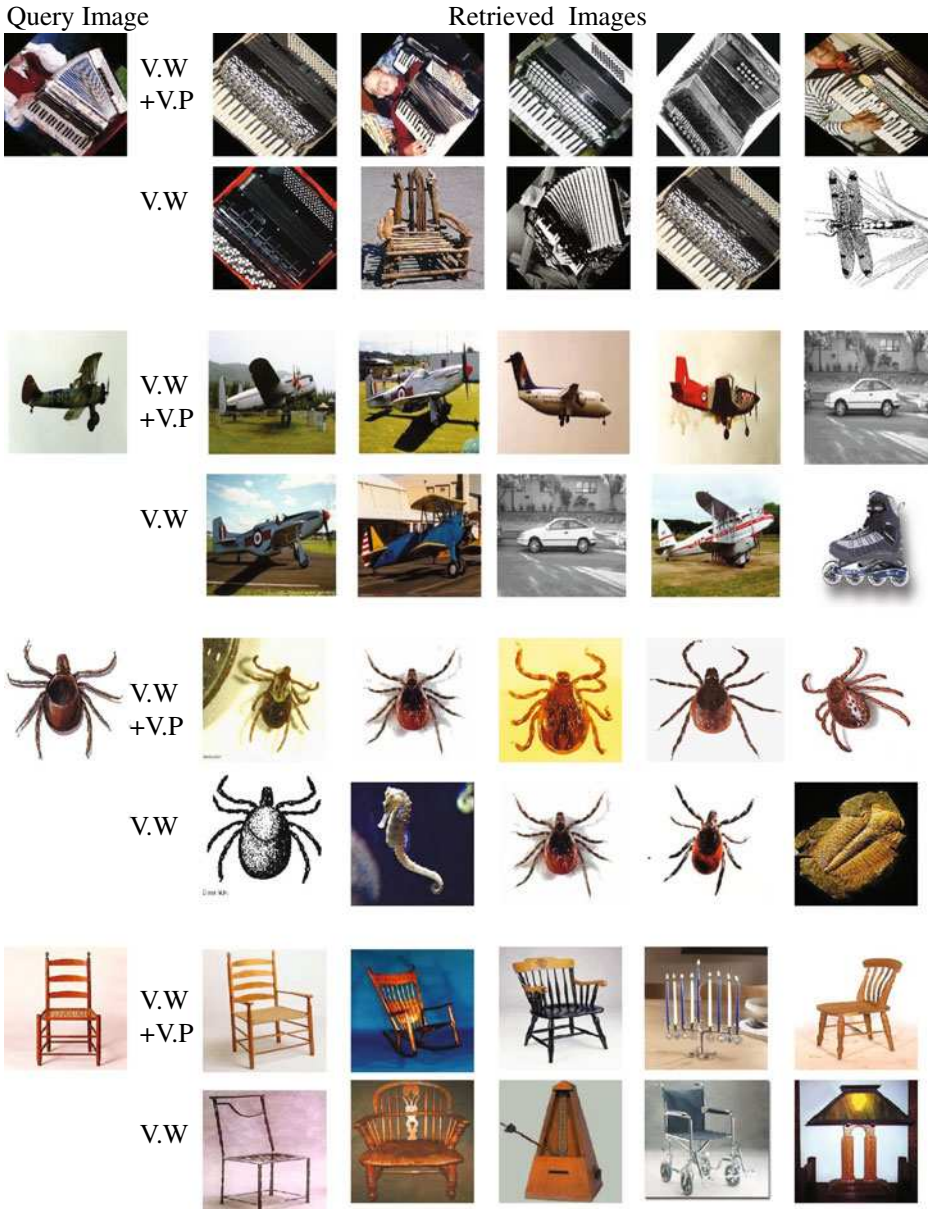
Fig. 14 Examples of retrieved images based on (visual words + visual phrases) and visual words alone

contribute also to 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 results. However, the

best text techniques cannot guarantee the same result in images since an image is a particular arrangement of different pixels in a 2D space and its spatial arrangement should be taken in consideration.

7 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 the 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 the 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.

References

1. Agrawal R, Imielinski T, Swami AN (1993) Mining association rules between sets of items in large databases. In: Buneman P, Jajodia S (eds) SIGMOD conference. ACM Press, pp 207–216
2. Baeza-Yates RA, Ribeiro-Neto BA (1999) Modern information retrieval. ACM Press / Addison-Wesley
3. Bay H, Ess A, Tuytelaars T, Gool LJV (2008) Speeded-up robust features (surf). Comput Vis image Underst 110(3):346–359
4. Belongie S, Malik J, Puzicha J (2002) Shape matching and object recognition using shape contexts. IEEE Trans Pattern Anal Mach Intell 24(4):509–522
5. Bilmes J (1997) A gentle tutorial of the EM algorithm and its application to parameter estimation for Gaussian mixture and hidden Markov models. Tech Rep TR-97-021, ICSI
6. Canny J (1986) A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell 8(6):679–698
7. Carson C, Thomas M, Belongie S, Hellerstein JM, Malik J (1999) Blobworld: a system for region-based image indexing and retrieval. In: Huijsmans DP, Smeulders AWM (eds) VISUAL. Lecture notes in computer science, vol 1614. Springer, pp 509–516
8. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the em algorithm. J R Stat Soc Ser B 39(1):1–38
9. Fei-Fei L, Fergus R, Perona P (2007) 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
10. Haddad M, Chevallet JP, Bruandet MF (2000) Relations between terms discovered by association rules. In: 4th European conference on principles and practices of knowledge discovery in databases PKDD'2000, workshop on machine learning and textual information access, Lyon France
11. Hammouda KM, Kamel MS (2004) Efficient phrase-based document indexing for web document clustering. IEEE Trans Knowl Data Eng 16(10):1279–1296
12. Hofmann T (2001) Unsupervised learning by probabilistic latent semantic analysis. Mach Learn 42(1–2):177–196
13. Hoíng NV, Gouet-Brunet V, Rukoz M, Manouvrier M (2010) Embedding spatial information into image content description for scene retrieval. Pattern Recognit 43(9):3013–3024

14. Jégou H, Douze M, Schmid C, Pérez P (2010) Aggregating local descriptors into a compact image representation. In: IEEE conference on computer vision & pattern recognition. http://lear.inrialpes.fr/pubs/2010/JDSP10
15. Jeon J, Manmatha R (2004) Using maximum entropy for automatic image annotation. In: In Proc. CIVR, pp 24–32
16. Jing F, Li M, Zhang L, Jiang Zhang H, Zhang B (2003) Learning in region-based image retrieval. In: In Proceedings of the IEEE international symposium on circuits and systems. Springer, pp 206–215
17. Jurie F, Triggs B (2005) Creating efficient codebooks for visual recognition. In: ICCV, pp 604–610
18. Lienhart R, Romberg S, Hörster E (2009) Multilayer plsa for multimodal image retrieval. In: CIVR '09: Proceeding of the ACM international conference on image and video retrieval. ACM, New York, NY, USA, pp 1–8
19. Lowe DG (2004) Distinctive image features from scale-invariant keypoints. Int J Comput Vis 60(2):91–110
20. Martinet J, Satoh S (2007) A study of intra-modal association rules for visual modality representation. In: Content-based multimedia indexing, 2007. CBMI '07. International workshop on, pp 344–350. doi:10.1109/CBMI.2007.385432
21. Morand C, Benois-Pineau J, Domenger JP, Zepeda J, Kijak E, Guillemot C (2010) Scalable object-based video retrieval in hd video databases. Image Commun 25(6):450–465
22. Nistér D, Stewénius H (2006) Scalable recognition with a vocabulary tree. In: CVPR (2), pp 2161–2168
23. Quelhas P, Monay F, Odobez JM, Gatica-Perez D, Tuytelaars T, Gool LJV (2005) Modeling scenes with local descriptors and latent aspects. In: ICCV, pp 883–890
24. Quelhas P, Monay F, Odobez JM, Gatica-Perez D, Tuytelaars T (2007) A thousand words in a scene. IEEE Trans Pattern Anal Mach Intell 29(9):1575–1589
25. van Rijsbergen CJ (1979) Information retrieval, 2nd edn. Butterworths, London
26. Salton G, Buckley C (1988) Term-weighting approaches in automatic text retrieval. Inf Process Manag 24(5):513–523
27. Salton G, Wong A, Yang CS (1975) A vector space model for automatic indexing. Commun ACM 18(11):613–620
28. Sivic J, Zisserman A (2003) Video google: a text retrieval approach to object matching in videos. In: ICCV. IEEE Computer Society, pp 1470–1477
29. Smeulders AWM, Member S, Worring M, Santini S, Gupta A, Jain R (2000) Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22:1349–1380
30. Tirilly P, Claveau V, Gros P (2008) Language modeling for bag-of-visual words image categorization. In: CIVR, pp 249–258
31. Viola PA, Jones MJ (2001) Rapid object detection using a boosted cascade of simple features. In: CVPR (1), pp 511–518
32. Willamowski J, Arregui D, Csurka G, Dance CR, Fan L (2004) Categorizing nine visual classes using local appearance descriptors. In: In ICPR workshop on learning for adaptable visual systems
33. Witten IH, Moffat A, Bell TC (1999) Managing gigabytes: compressing and indexing documents and images, 2nd edn. Morgan Kaufmann
34. Yang J, Jiang YG, Hauptmann AG, Ngo CW (2007) Evaluating bag-of-visual-words representations in scene classification. In: Multimedia information retrieval, pp 197–206
35. Yang Y, Pedersen JO (1997) A comparative study on feature selection in text categorization. In: ICML, pp 412–420
36. Yuan J, Wu Y, Yang M (2007) Discovery of collocation patterns: from visual words to visual phrases. In: CVPR
37. Zhang J, Marszalek M, Lazebnik S, Schmid C (2007) Local features and kernels for classification of texture and object categories: a comprehensive study. Int J Comput Vis 73(2):213–238
38. Zheng QF, Gao W (2008) Constructing visual phrases for effective and efficient object-based image retrieval. TOMCCAP 5(1)
39. Zhou X, Zhuang X, Yan S, Chang SF, Hasegawa-Johnson M, Huang TS (2008) Sift-bag kernel for video event analysis. In: ACM Multimedia, pp 229–238

Ismail El sayad is pursuing his doctoral studies in the University of Lille 1, France, that he joined in 2008. He holds a Master of Science in Computer Engineering from Duisburg-Essen University, Germany, and a Bachelor of Science in Computer Engineering from the Lebanese International University in Beirut, Lebanon. His research interests focus on multimedia mining, indexing and image semantic understanding.

Jean Martinet has joined the University of Lille 1, France, in 2008, where he is Assistant Professor in Computer Sciences. He has visited the National Institute of Informatics (NII), Japan, for a two-year postdoctoral research fellowship. He joined the video processing project at NII in 2005 after being awarded a fellowship from the Japan Society for the Promotion of Science (JSPS). He was formerly a Ph.D. student during three years at Université Joseph Fourier, Grenoble France. He received his B.Sc. and M.Sc. degrees in 1999 and 2001 from Université Joseph Fourier. His research interests include image and video analysis, indexing and retrieval.

Thierry Urruty is a post doctoral researcher at the University of Lille 1, France. He finished his PhD end of 2007 in multimedia indexing. In 2008, he has been a research assistant in the Information Retrieval Group at the University of Glasgow. His research interests are data mining, video indexing and retrieval and more recently computer vision.

Chabane Djeraba is professor of Computer Science at University Lille1, France, since 2003. He lead a research group, in the Computer Science Laboratory of Lille (LIFL UMR Lille1-CNRS n° 8022) and Telecom Lille1. His current domain of research focuses on extraction of information from videos, application to human behaviors. He has been assistant and associate professor from 1994 to 2003 in Polytechnic School of Nantes University, France. He obtained a PhD of Computer Science in 1993, Claude Bernard University, France. He obtained a master degree of Computer Science Applied to Social Sciences, in 1990, Pierre-Mendes France University of Grenoble, France. And finally, in 1989, Engineer degree in computer science, Ecole nationale Superieure d'Informatique (ESI), previously called INI, Algiers, Algeria. He published several papers, including two books, in 2010 "Multi-Modal User Interactions in Controlled Environments", Springer, and in 2008, "Mathematical Tools for Data Mining", Springer.