# From Bag-of-Visual-Words to Bag-of-Visual-Phrases using n-Grams

Glauco V. Pedrosa and Agma J. M. Traina
Instituto de Ciências Matemáticas e de Computação - ICMC
Universidade de São Paulo - USP
{gpedrosa, agma}@icmc.usp.br

***Abstract*—The Bag-of-Visual-Words has emerged as an effective modeling approach to represent local image features. It describes local image features by assigning them a visual word according to a visual dictionary. The image representation is given by the frequency of each visual word in the image, as a similar representation used in textual documents. In this paper, we present a novel approach building a high-level description using a group of words (phrases) for representing an image. We introduce the use of n-grams for image representation, based on the idea of "Bag-of-Visual-Phrases". In the field of computational linguistics, an n-gram is a phrase formed by a sequence of n-consecutive words. As analogy, we represent an image by a combination of n-consecutive visual words. We made representative experiments using three public benchmark databases of textures and nature scenes and two medical databases to demonstrate an area that can benefit from the proposed technique. Our proposed Bag-of-Visual-Phrases approach improved up to 44% the retrieval precision and up to 33% the classification rate compared to the traditional Bag-of-Visual-Words, being a valuable asset for content-based image retrieval and image classification.**

***Keywords*-Image description; CBIR; Bag-of-features; SIFT;**

## I. Introduction

Image representation plays an essential role in image categorization and retrieval applications. Research in this area has advanced in order to obtain methods that capture the semantics of the image, extracting features perceptually efficient and compact [1]. This fact should be taken into consideration to make a CBIR (Content-Based Image Retrieval) system closer to the users' expectation.

One of the state of art technique for image representation is the Bag-of-Visual-Words [2], [3], [4], [5], [6], [7], [8], also known as Bag-of-Features or Bag-of-Keypoints. This approach describes an image using a dictionary composed of different visual words. Visual words are local image patterns, which concentrate relevant semantic information about the image. As illustrated in Figure 1, after extracting the local image features, each feature is assigned to its nearest visual word according to a visual dictionary. The traditional image representation, employed by the Bag-of-Visual-Words approach, is the number of occurrences of each visual word contained in the image, as an analogy to the Bag-of-Words representation used for textual information retrieval.

A more powerful description can be obtained by grouping words [9], [10], [11], [12], once this approach can encode the arrangement between the visual words in the image space. In fact, aggregating spatial information in visual words is a promising approach for image description, because the appearance of the visual words can change profoundly when they participate in relations. Spatial information reaches a high-level semantic characterization and leads to a more meaningful image representation.

The goal of this work is to represent an image taking into consideration the relationship between the visual words, instead of considering the image as a set of isolated words. In this paper, we introduce the idea of using $n$-grams for generating *visual phrases*. The use of $n$-grams is an efficient model already used in natural language processing to represent a textual document [13]. The words are modeled such that each $n$-gram is composed of $n$ sequential words. For example, the 1-gram (unigram) representation is composed of isolated words, such as {*intelligence*, *artificial*, *computer*, *vision*, *medical*, *systems*}. By analogy, the 1-gram representation is the traditional Bag-of-Visual-Words approach. On the other hand, the 2-gram (bigram) is represented by two sequences of words, such as {*artificial intelligence*, *computer vision*, *medical systems*, *artificial systems*}, enriching the semantic and yielding a more complete representation. The core advantages of using $n$-gram models are the simplicity and the ability to scale up the content representation very effectively by simply increasing $n$.

We performed representative evaluations in several different databases. We evaluated our proposed method in benchmark databases of the image classification and retrieval area, and also in medical databases, which have shown to be an application area that can benefit from the proposed technique, since it adds more semantics to the image description. Comparative results, with respect to the traditional Bag-of-Visual-Words approach, show that the use of $n$-grams is a promising descriptor to achieve a more robust image characterization and make a CBIR system closer to users' expectation.

The rest of the paper is organized as follows. Section 2 gives the formal definitions and the background needed to follow the work, such as the motivation for the proposed work. Section 3 explains the method to represent an image in $n$-gram visual words. Section 4 presents the experimental analysis and Section 5 gives the conclusions of this paper.

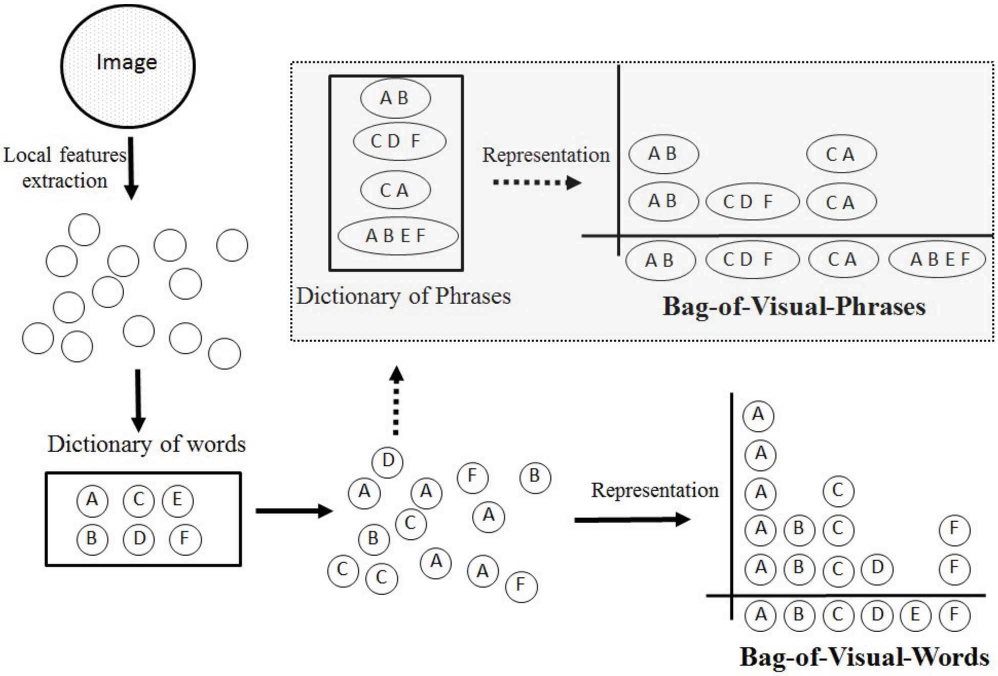


Figure 1. Scheme to represent an image in Bag-of-Visual-Word and Bag-of-Visual-Phrases.

## II. Background and motivation

The Bag-of-Visual-Words (BoVW) approach is a technique used to model the local image features. These local features are described by an *unordered* set of keypoints, where each keypoint describes representative local image features. The goal is to quantize these features using a visual dictionary.

The main idea of using visual dictionaries is to consider that the image visual patterns are similar to textual words present in textual documents. Therefore, an image is composed of visual words as a textual document is composed of textual words.

Clustering is a common method in the literature for learning a visual dictionary. A dictionary can be built by clustering the local features detected in a set of training images from the database, such as schematized in Figure 2. Formally, let $P = \{p_1, p_2, ..., p_z\}$ be the local features detected in a subset of the database image. The visual dictionary is given by a division of $P$ into $k$ distinct clusters $\pi_k = \{C_1, C_2, ..., C_k\}$, so that $C_1 \cup C_2 \cup ... \cup C_k = P$, $C_i \neq \emptyset$, and $C_i \cap C_j = \emptyset$ for $i \neq j$ and $i, j = 1...k$.

A visual word $w_i$ is the centroid of cluster $C_i$. When a new image arrives, its features are extracted and assigned to the nearest visual word $w_i$, for $i = 1...k$, where $k$ is the number of words in the visual dictionary. The image representation employing the BoVW approach is simply the normalized histogram of the quantized visual words detected in the image.

Spatial information is a very important feature for the characterization of images and objects, because the appearance of objects can deeply change when they participate in relations. One of the first work to attempt encoding geometric information with Bag-of-Visual-Words is the spatial pyramid [14], which splits the image into hierarchical cells and computes bag-of-visual-words for each cell, concatenating the results at the end. However, it is crucial that the image characterization does not depend on the placement of the image, because the features should be invariant to geometric transformations. Other works employ correlograms of visual words [15] and image splitting by linear and circular projections [16]. The method proposed in [17] encode spatial-relationship information of visual words using image space partitions to count the occurrences of the visual words in relation to the other visual words positions.

It seems plausible that grouping words might be applied successfully for enriching the BoVW representation [12], once a high semantic information level is reached. The benefits of using group of words have been proven to boost local feature matching [11]. Previous works employ phrases to model the co-occurrences of the words in local neighborhoods, using methods to encode the spatial layouts with a grid-dependent size [18]. Besides that, a big problem of these approaches is that the number of phrases can exponentially grow according to the number of words in a phrase. Thus, it is necessary to select a subset from the entire phrase set. Sophisticated mining or learning algorithms have been proposed for this selection [9], [10], but it may still be risky to discard a large portion of phrases, because some of which may be representative ones for the images.

Instead of using a dictionary formed by a huge list of phrases with different number of words, why do not use a dictionary formed by phrases with a limited but representative number of words? A representation that utilizes phrases with a limited sequence of $n$ words is denominated $n$-gram. In the

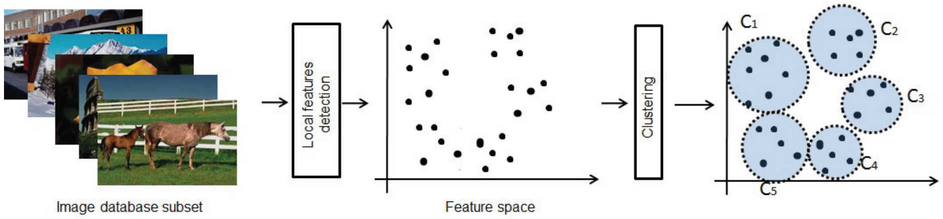


Figure 2. Process used to generate a visual dictionary. Initially, a keypoint detector is applied in a set of training images to detect representative local image points. The detected keypoints are represented by a descriptor that summarizes the information about the region around each keypoint. A dictionary can be built by clustering the keypoints detected.

fields of computational linguistics and probability, an $n$-gram is a contiguous sequence of $n$ *items* from a given sequence of text or speech [13]. An $n$-gram could be any combination of letters. However, the items in question can be phonemes, syllables, letters, words or pairs according to the application.

How useful the $n$-gram representation could be for image description based on the idea of Bag-of-Visual-Phrases? The idea of using $n$-grams in vision is quite similar to a number of previous works that combine visual words with spatial information. For example, in [12] triplets of visual words are used. The work in [19] use 2-grams ("doublets") from spatially neighboring word pairs. In [20], the authors use "higher order features" (i.e. BoVW + 2-grams, 3-grams etc), but instead of considering all possible $n$-grams, they perform feature selection to only pick relevant $n$-grams. In [21] the authors propose to do joint clustering of nearest feature pairs. In this paper, we propose a slightly different measure of similarity considering $n$-grams and BoVW, and a simple and different procedure to extract the $n$-grams from the image compared to [22] and [23]. The details of the proposed technique is presented in the next section.

## III. The proposed method

The proposed method describes an image as a Bag-of-Visual-Phrases, where the Visual Phrases are $n$-grams extracted from the image. In our proposed model, the image representation is the frequency that each $n$-gram appears in the image. In this section we explain how to extract the $n$-grams from an image and how to represent an image using Bag-of-Visual-Phrases.

### *A. Extracting the $n$-grams from an image*

In text mining, a unigram representation can be obtained by placing a small window over the text, such that we only look at one word at a time. In a similar way, a bigram can be thought of as a window that shows *two* words at a time and moving this window to the right, one word at a time, in a stepwise manner. This procedure is the same for extracting $n$-grams from a text with $n$ greater than two. To take this "window" analogy to our image representation problem, we could say that all visual $n$-grams of an image can be generated by placing a region over each visual word, such that we only look at the $n$-nearest visual words at time.

Formally, let $P = \{p_1, p_2, ..., p_m\}$ be the local features detected in an image. Each local feature $p_i$ is represented by its coordinates $(x_i, y_i)$ in the image and it is assigned to a visual word $w_j$. The $n$-Nearest Visual Word ($n$NVW) of a local feature $p_i$ is given by:

$$\begin{aligned} n\text{NVW}(p_i, n) = \{P' \subset P| \quad |P'| = n, \forall p_j \in P', \\ \forall p_r \in [P - P'] : d(p_i, p_j) < d(p_i, p_r)\} \end{aligned} \tag{1}$$

where:

$$d(p_i, p_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \tag{2}$$

The next definition is needed to give the main building block of our proposed technique.

**Definition 1**. *The proposed image representation in $n$-grams is given by the occurrences of the $n$-Nearest Visual Words considering each local feature $p_i$ detected in an image P:*

$$n\text{-gram}(P) = \{n\text{NVW}(p_i, n)\}, \text{for } i = 1...m. \tag{3}$$

*where $m$ is the number of keypoints detected in the image.*

To illustrate, Figure 3 shows an example of extracting the 2-grams from an image. For each local point, we look at its closest neighbor point. There is no need to specify a threshold distance, since we get the nearest point. We consider the pair as a 2-gram phrase. In a similar way, to extract the $n$-grams from an image, with $n > 2$, we just need to increase the number of nearest points.

### *B. Modeling the image in Bag-of-Visual-Phrases using $n$-grams*

After extracting the $n$-grams from an image, each $n$-gram is treated as a visual phrase. To model an image in Bag-of-Visual-Phrases, the next step is to count the number of visual phrases according to a dictionary of visual phrases. This dictionary of visual phrases is given by the all possible combinations of $n$-grams. However, the size of the dictionary can be exponential with respect to the number of the visual words in the vocabulary. For example, considering 100 different visual words, the number of all possible combinations of 2-grams is $100^2$.

To reduce the size of the dictionary of visual phrases we do not consider $n$-grams with repeated visual words, this means, an $n$-gram with $n$ similar consecutive visual words. Additionally, we consider inverted $n$-gram as the same visual phrase. Inverted visual phrases consists of phrases with the

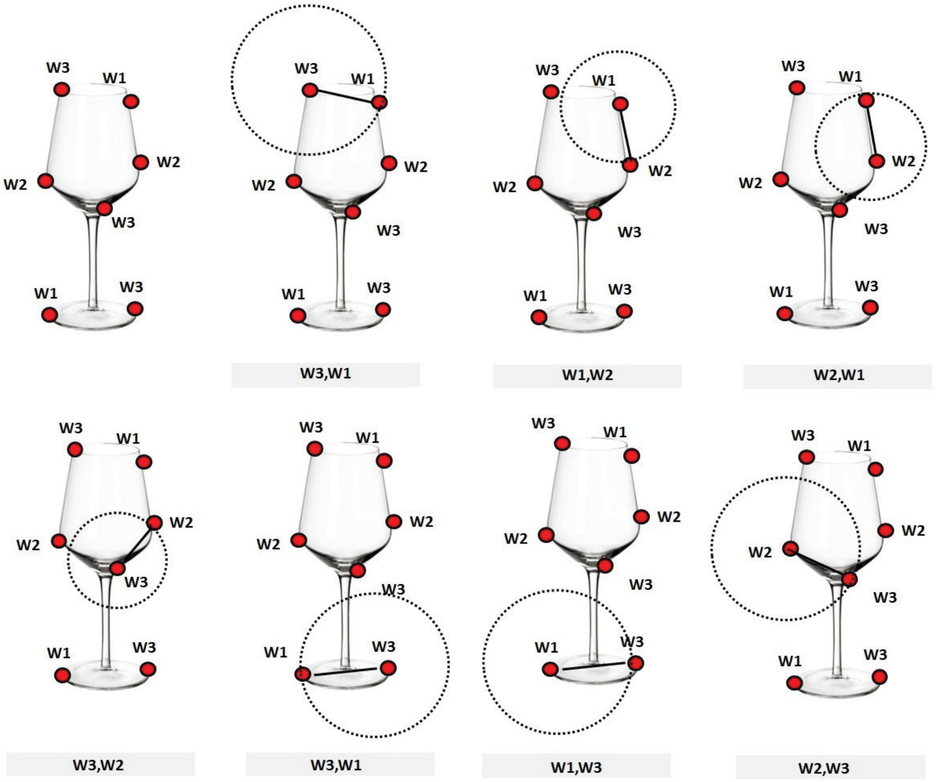
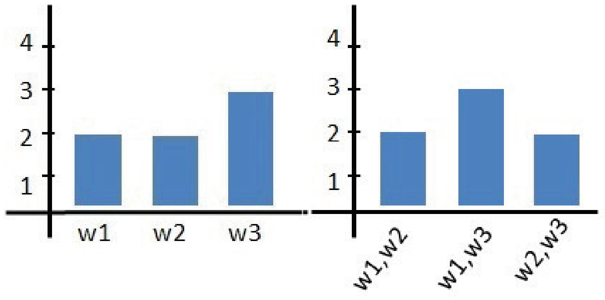


Figure 3. Example of the process used to extract 2-grams from an image. For each keypoint we look at its closest neighbor point. To extract $n$-grams from an image, with $n > 2$, we just need to increase the number of nearest points.



(a) (b)

Figure 4. Representation of the image in fig. 3 by: (a) Bag-of-Visual-Words, (b) Bag-of-Visual-Phrase considering 2-gram phrases and the proposed dictionary.

same visual words, where the visual words appear in different orders. With these two restrictions the dictionary of visual phrases can be reduced to 50% or more.

To illustrate, the 2-gram phrases generated by a dictionary formed by the words $\{w_1, w_2, w_3\}$ is $\{\{w_1, w_2\}, \{w_1, w_3\}, \{w_2, w_3\}\}$. Figure 4 shows the representation of the image from Figure 3 considering the traditional Bag-of-Visual-Words and the proposed method of Bag-of-Visual-Phrases using 2-gram phrases.

### *C. Calculating the images' distances*

To measure the dissimilarity (distance) between two images $A$ and $B$, we take advantage of the distance between the histograms of Bag-of-Visual-Words and Bag-of-Visual-Phrases of both images.

Let $h_A$ and $H_A$ be the normalized histograms of Bag-of-Visual-Words and Bag-of-Visual-Phrases of image $A$, respectively, and $h_B$ and $H_B$ of image $B$. The distance used to measure the dissimilarity between the images $A$ and $B$ is given by:

$$\text{Distance}(A, B) = ||h_A - h_B||_1 + ||H_A - H_B||_1 \qquad (4)$$

where $||.||_1$ is the $L_1$ distance.

Two images A and B are considered similar when $\text{Distance}(A, B) \rightarrow 0$.

## IV. Experimental results

In this section, we report representative experimental results performed to evaluate the effectiveness of the image representation technique proposed in this work. We compared our method with the traditional Bag-of-Visual-Words representation, which has only the frequency of occurrence of the words in the image. The Bag-of-Visual-Words is by analogy the 1-gram representation.

We seek a vocabulary of visual words which will be invariant to changes in viewpoint and illumination. For this, we used the SIFT descriptor [24], which is one of the most widely used descriptor to extract local image features. SIFT descriptors, based on histograms of local orientation, has some tolerance to illumination change.

The experiments was conducted to evaluate the accuracy of 2-gram and 3-gram representations compared to 1-gram representation. The tests was performed in two different tasks: image retrieval and image classification.

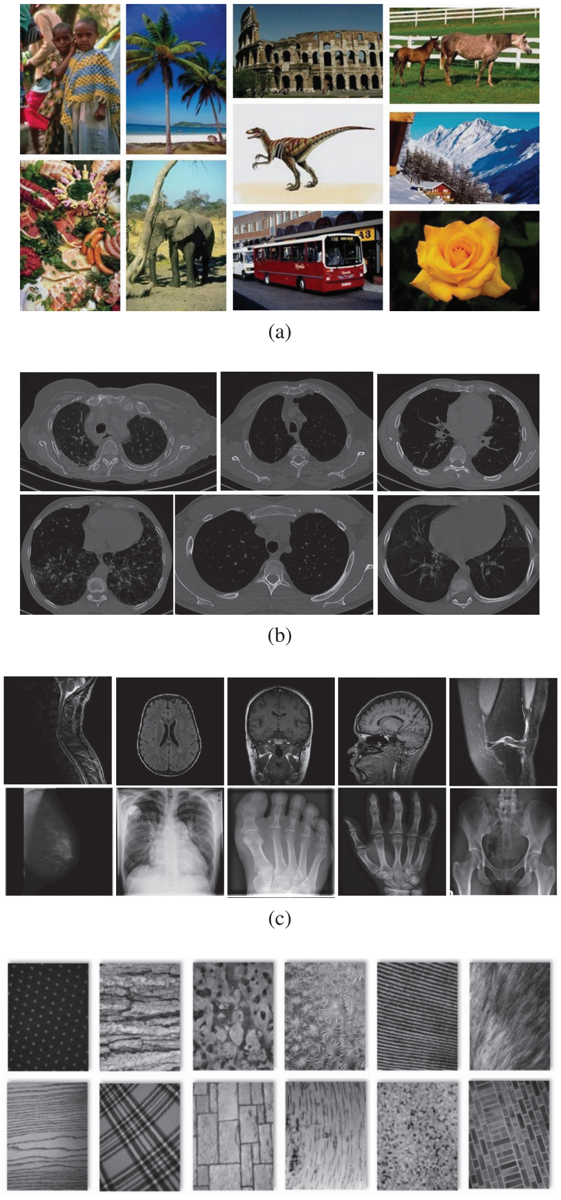
### *A. Image retrieval evaluation*

We performed an image retrieval evaluation on four different image databases, two public databases and two medical ones:

- Corel1000 database [1]: consists of images of natural scenes. It is composed of 1000 images divided into 10 classes, 100 images in each class. Figure 5a shows an image for each class of this database;
- Lung database [2]: consists of CT images of lung, composed of 234 images divided into 6 classes (39 images in each class), classified according to the Lung findings: Emphysema, Honeycombing, Interlobular Septal, Healthy, Consolidation and Ground-glass. Figure 5b shows an image for each class of this database;
- Medical Image Exams database [3]: it contains 2,200 medical images of X-ray and MRI, classified according to body part and type of cut, being composed of 11 distinct classes with 200 images in each class. Figure 5c shows images samples from this database.
- Texture database [25]: includes surfaces composed of materials such as wood, marble and fur under varying viewpoints, scales and illumination conditions. This database consists of 1,000 images comprising 40 samples of 25 different textures. Figure 5d shows images samples from this database.

The comparative retrieval performance was evaluated using the mean of Average Precision (mAP). The Precision is the ratio of the number of relevant retrieved images to the total number of retrieved images. The Average Precision computes the average value of Precision at each ranking position where a relevant image is retrieved. The top value is 1, which means that all relevant images were retrieved in the first positions of the ranking. The mAP determine the mean of Average Precision considering each image as query. The best value of mAP is 1.

Table I presents the average mAP values for the evaluated databases. The 2-gram representation presented the best results in three databases. In general, the 2-gram boosted the precision in 5% in these databases. Except for the Corel1000 database, the 3-gram presented the best result. In this database, the 3-gram representation increased 13% the precision compared with the traditional 1-gram representation.



(a)

(b)

(c)

(d)

Figure 5. Sample images from the: (a) Corel1000 database, (b) Lung database, (c) Medical Image Exams database , (d) Texture database.

Considering each class individually, we can see where the results were more significant. Table III presents the mAP results for each class of the Lung database and Tables II, IV and V summarize the results from representative classes of the other three databases, where the difference between the methods were considered significant. Figures 6a, 6b, 6c and 6d present the Precision values for each class of the evaluated databases.

For the class Emphysema in the Lung database and for the class Knee in the Medical Image Exams database, the 2-gram representation had a gain up to 17% in Precision. For the class Africa in the Corel1000 database, the 2-gram representation improve the precision in 26%. Considering the Class 1 of the Texture database the 3-gram representation achieved a gain of 44% in precision.

[1] available at: http://wang.ist.psu.edu/docs/related/

[2] provide by the Clinical Hospital with our university.

[3] provide by the Clinical Hospital with our university.

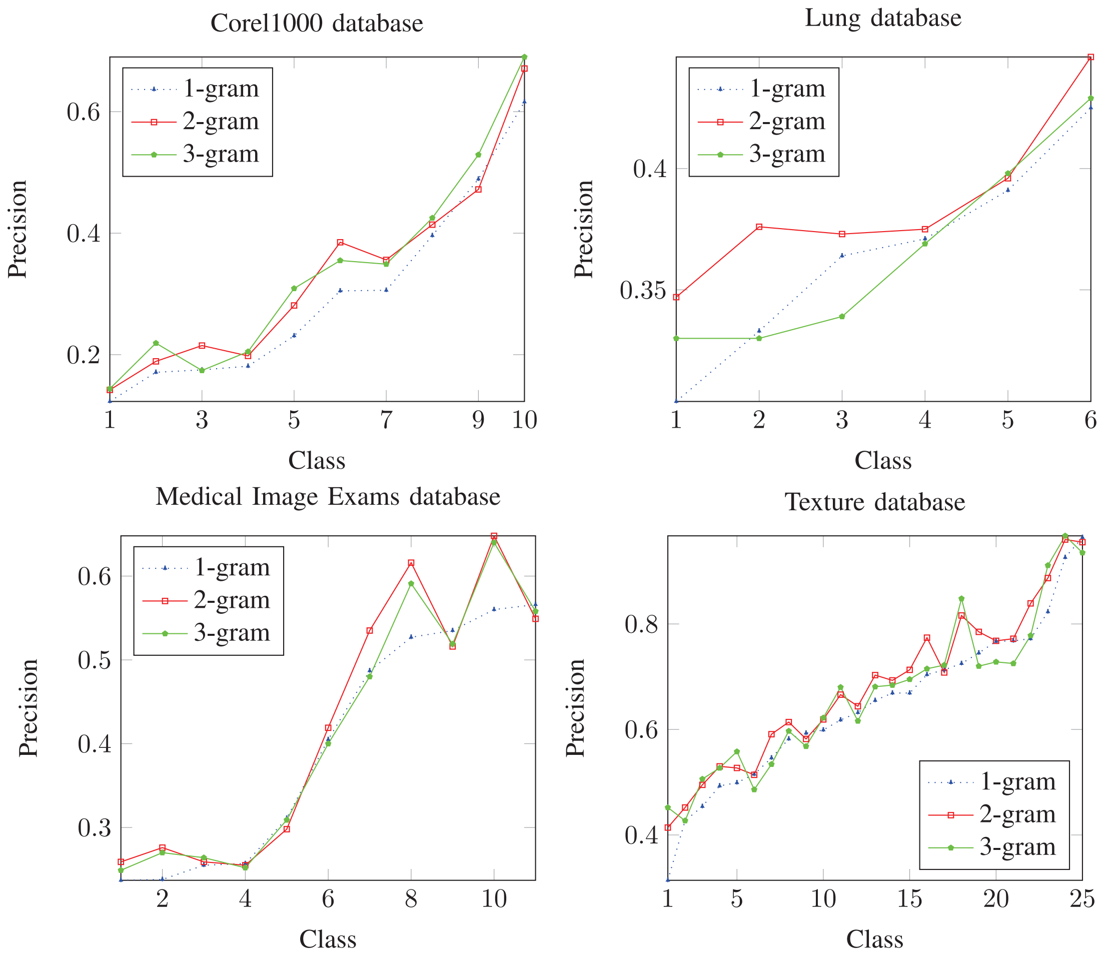


Figure 6. Precision values for each class of the evaluated databases.

Table I
AVERAGE RESULTS FOR THE EVALUATED DATABASES.

| *Database* | *Image representation* | | |
|---|---|---|---|
| | *1-gram* | *2-gram* | *3-gram* |
| Corel1000 | 0.299 | 0.332 | **0.340** |
| Lung | 0.365 | **0.385** | 0.366 |
| Medical Exams | 0.398 | **0.421** | 0.412 |
| Texture | 0.648 | **0.681** | 0.667 |

In general, when a class is complex, this means, when it has more visual details, the $n$-gram representation with a large value of $n$ tends to have better results. However, the value of $n$ affects the dictionary size, such that more visual phrases are being analyzed and this can affect the retrieval performance.

Considering the overall performance for all the evaluated databases, the 2-gram and 3-gram achieved a gain in Precision compared to the traditional 1-gram representation. These results demonstrated that the high-level image feature proposed in this work was able to improve the retrieval system and making a CBIR closer to the users' expectation.

Table II
PRECISION RESULTS FOR SOME CLASSES OF THE COREL1000 DATABASE

| *Class* | | *Image Representation* | | |
|---|---|---|---|---|
| | | *1-gram* | *2-gram* | *3-gram* |
| 2 | Food | 0.171 | 0.189 | **0.219** |
| 3 | Horses | 0.175 | **0.215** | 0.174 |
| 5 | Flower | 0.231 | 0.281 | **0.309** |
| 6 | Africa | 0.305 | **0.385** | 0.355 |
| 7 | Elephant | 0.306 | **0.356** | 0.349 |
| 9 | Mountain | 0.489 | 0.472 | **0.529** |
| 10 | Dinosaur | 0.616 | 0.671 | **0.690** |

### *B. Image classification evaluation*

We performed an image classification evaluation using the 15-scenes database [14]. This database (Fig. 7) is composed of 4485 images categorized in fifteen different scenes. Each category has 200 to 400 images, and average image size is 300x250 pixels. The major sources of the pictures in the database include the COREL collection, personal photographs, and Google image search. This is one of the most complete

Table III
PRECISION RESULTS FOR THE LUNG DATABASE

| *Class* | | *Image representation* | | |
|---|---|---|---|---|
| | | *1-gram* | *2-gram* | *3-gram* |
| 1 | Emphysema | 0.304 | **0.347** | 0.330 |
| 2 | Honeycombing | 0.333 | **0.376** | 0.330 |
| 3 | Interlobular Septal | 0.364 | **0.373** | 0.339 |
| 4 | Healthy | 0.371 | **0.375** | 0.369 |
| 5 | Consolidation | 0.391 | 0.396 | **0.398** |
| 6 | Ground-glass | 0.425 | **0.446** | 0.429 |

Table IV
PRECISION RESULTS FOR SOME CLASSES OF THE MEDICAL IMAGE EXAMS DATABASE

| *Class* | | *Image representation* | | |
|---|---|---|---|---|
| | | *1-gram* | *2-gram* | *3-gram* |
| 1 | Chest | 0.237 | **0.259** | 0.249 |
| 2 | Brain Axial | 0.238 | **0.276** | 0.270 |
| 4 | Hand | 0.257 | **0.258** | 0.252 |
| 5 | Foot | 0.311 | 0.298 | **0.319** |
| 6 | Brain Coronal | 0.405 | **0.419** | 0.400 |
| 7 | Breast | 0.487 | **0.535** | 0.480 |
| 8 | Knee | 0.527 | **0.616** | 0.591 |
| 10 | Abdomen | 0.560 | **0.648** | 0.640 |
| 11 | Brain Sagittal | **0.566** | 0.549 | 0.558 |

scene category database used in the literature thus far.

A multi-class classification was done with a support vector machine (SVM) trained using the one-versus-all rule: a classifier is learned to separate each class from the rest, and a test image is assigned to the label of the classifier with the highest response.

Table VI shows the classification rate for the experiments using 100 images per class for training and the rest for testing (the same setup as [14]). The 2-gram representation presented the best classification rate with 48%, followed by the 3-gram representation (43%). Compared with the 1-gram, the 2-gram representation had a gain of 11% in accuracy. This result demonstrate that the 2-gram representation encode more discriminative features than the 1-gram representation.

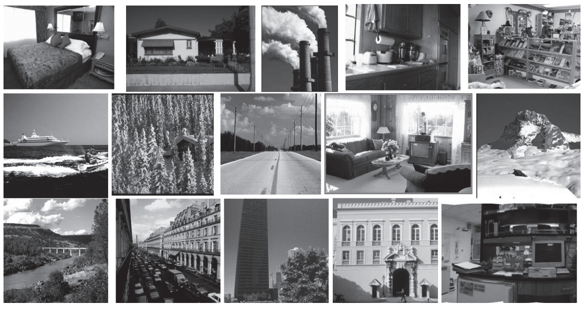


Figure 7. Example images from the 15-scenes category database.

Table V
PRECISION RESULTS FOR SOME CLASSES OF THE TEXTURE DATABASE.

| *Class* | | *Image representation* | | |
|---|---|---|---|---|
| | | *1-gram* | *2-gram* | *3-gram* |
| 1 | Class 1 | 0.314 | 0.414 | **0.452** |
| 2 | Class 2 | 0.427 | **0.452** | 0.427 |
| 4 | Class 4 | 0.493 | 0.530 | **0.527** |
| 5 | Class 5 | 0.499 | 0.527 | **0.558** |
| 6 | Class 6 | **0.515** | 0.514 | 0.486 |
| 7 | Class 7 | 0.546 | **0.591** | 0.534 |
| 9 | Class 9 | 0.593 | **0.582** | 0.568 |
| 10 | Class 10 | 0.599 | 0.619 | **0.622** |
| 11 | Class 11 | 0.618 | 0.666 | **0.680** |
| 13 | Class 13 | 0.655 | **0.703** | 0.681 |
| 15 | Class 15 | 0.669 | **0.713** | 0.695 |
| 16 | Class 16 | 0.704 | **0.774** | 0.715 |
| 18 | Class 18 | 0.725 | 0.816 | **0.848** |
| 19 | Class 19 | 0.745 | **0.785** | 0.720 |
| 22 | Class 22 | 0.772 | **0.839** | 0.778 |
| 23 | Class 23 | 0.823 | 0.887 | **0.911** |
| 24 | Class 24 | 0.926 | 0.960 | **0.967** |
| 25 | Class 25 | **0.964** | 0.955 | 0.935 |

Table VI
AVERAGE CLASSIFICATION RATE FOR THE 15-SCENES DATABASE.

| *Representation* | *Average Classification* |
|---|---|
| 1-gram | 0.428 |
| 2-gram | **0.476** |
| 3-gram | 0.434 |

Analyzing each class individually, Table VII presents the classification rates for each class of this database. For three classes, the 3-gram presented the best results. The 2-gram had a gain of 33% in accuracy for the class Kitchen and the 3-gram had a gain of 20% for the class MITforest compared with the traditional 1-gram approach.

In this evaluation task (image classification), we can see the same behavior that the previous task (image retrieval). The 2-gram and 3-gram representations presented an overall performance better than 1-gram representation. These results indicate that the 2-gram and 3-gram add more information in the image feature description, being a valuable asset to improve the image analysis.

## V. CONCLUSION

In this paper, we have introduced a novel modeling approach for representing images. The new approach represents an image by taking into consideration the relationship between its visual words. The proposed method is based on the idea of Bag-of-Visual-Phrases, which has a higher level of semantic characterization compared to the traditional Bag-of-Visual-Words. The image is represented as a collection of *visual* *phrases*, instead of considering the image as a set of isolated visual words.

Our proposed method uses a dictionary composed of visual phrases with a fixed number of words. Such representation is an analogy to the popular $n$-gram representation used for textual representation.

We have conducted experiments in five different databases. Three of the databases are public and employed as benchmarks in the image retrieval and classification community. The others two are composed of medical images to demonstrate an area that can benefit from the proposed technique. The results have shown that bigram and trigram dictionaries are sufficient to boost the retrieval and classification accuracy.

Our proposed novel modeling approach enriches the Bag-of-Visual-Words representation and the obtained results indicate that it can become a powerful and promising descriptor for image representation, and can contribute to the content-based image retrieval and image classification field.

Table VII
CLASSIFICATION RESULTS FOR EACH CLASS OF THE 15-SCENE DATABASE. THE HIGHEST RESULTS FOR EACH CLASS ARE SHOWN IN BOLD.

| *Class* | *Representation* | | |
|---|---|---|---|
| | *1-gram* | *2-gram* | *3-gram* |
| bedroom | 0.250 | **0.319** | 0.216 |
| CALsuburb | 0.787 | **0.901** | 0.872 |
| industrial | 0.199 | **0.246** | 0.235 |
| kitchen | 0.245 | **0.327** | 0.227 |
| livingroom | 0.217 | **0.233** | 0.211 |
| MITcoast | **0.473** | 0.400 | 0.408 |
| MITforest | 0.781 | 0.860 | **0.943** |
| MIThighway | **0.419** | 0.394 | 0.263 |
| MITinsidecity | 0.452 | **0.486** | 0.453 |
| MITmountain | 0.438 | **0.529** | 0.453 |
| MITopencountry | 0.365 | **0.384** | 0.297 |
| MITstreet | 0.328 | **0.432** | 0.313 |
| MITtallbuilding | 0.430 | 0.480 | **0.484** |
| PARoffice | 0.522 | **0.583** | 0.552 |
| store | 0.521 | 0.567 | **0.591** |

## ACKNOWLEDGEMENTS

The authors acknowledge the financial support of FAPESP, CNPq, CAPES and INCT INCod.

## REFERENCES

[1] M. S. Lew, N. Sebe, and J. P. Eakins, "Challenges of image and video retrieval," in *International Conference on Image and Video Retrieval (CIVR)*, 2002, pp. 1–6.

[2] J. C. Caicedo, A. Cruz-Roa, and F. A. Gonzalez, "Histopathology image classification using bag of features and kernel functions," in *Conference on Artificial Intelligence in Medicine*, ser. Lecture Notes in Computer Science, vol. 5651, 2009, pp. 126–135.

[3] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *International Journal of Computer Vision*, vol. 87, no. 3, pp. 316–336, 2010.

[4] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 2161–2168.

[5] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[6] M. M. Rahman, S. K. Antani, and G. R. Thoma, "Biomedical cbir using "bag of keypoints" in a modified inverted index," in *International Symposium on Computer-Based Medical Systems*, ser. CBMS '11, 2011, pp. 1–6.

[7] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *International Conference on Computer Vision*, vol. 2, Oct. 2003, pp. 1470–1477.

[8] J. Wang, Y. Li, Y. Zhang, H. Xie, and C. Wang, "Boosted learning of visual word weighting factors for bag-of-features based medical image retrieval," in *Image and Graphics (ICIG), 2011 Sixth International Conference on*, 2011, pp. 1035–1040.

[9] L. Torresani, M. Szummer, and A. W. Fitzgibbon, "Learning query-dependent prefilters for scalable image retrieval," in *IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '09, 2009, pp. 2615–2622.

[10] J. Yuan, Y. Wu, and M. Yang, "Discovery of collocation patterns: from visual words to visual phrases," in *IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '07. IEEE Computer Society, 2007, pp. 18–23.

[11] Y. Zhang, Z. Jia, and T. Chen, "Image retrieval with geometry-preserving visual phrases," in *IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '11. IEEE, 2011, pp. 809–816.

[12] L. C. Zitnick, J. Sun, R. Szeliski, and S. Winder, "Object instance recognition using triplets of feature symbols," in *Tech. Report, Microsoft Research*, 2007.

[13] C. Y. Suen, "n-gram statistics for natural language understanding and text processing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 2, pp. 164–172, 1979.

[14] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, ser. CVPR '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 2169–2178.

[15] S. Savarese, J. Winn, and A. Criminisi, "Discriminative object class models of appearance and shape by correlatons," in *In IEEE Computer Vision and Pattern Recognition*, 2006, pp. 2033–2040.

[16] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang, "Spatial-bag-of-features," in *CVPR'10*, 2010, pp. 3352–3359.

[17] O. A. B. Penatti, E. Valle, and R. da Silva Torres, "Encoding spatial arrangement of visual words," in *CIARP*, 2011, pp. 240–247.

[18] Y. Jiang, J. Meng, and J. Yuan, "Grid-based local feature bundling for efficient object search and localization," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, 2011, pp. 113–116.

[19] J. Sivic, B. Russell, A. Efros, A. Zisserman, and W. Freeman, "Discovering objects and their location in images," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 1, 2005, pp. 370–377 Vol. 1.

[20] D. Liu, G. Hua, P. Viola, and T. Chen, "Integrated feature selection and higher-order spatial feature extraction for object categorization," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1–8.

[21] N. Morioka and S. Satoh, "Building compact local pairwise codebook with joint feature space clustering," in *ECCV 2010*, ser. Lecture Notes in Computer Science, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Springer Berlin Heidelberg, 2010, vol. 6311, pp. 692–705.

[22] T. Li, T. Mei, I.-S. Kweon, and X.-S. Hua, "Contextual bag-of-words for visual categorization," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 21, no. 4, pp. 381–392, 2011.

[23] P. Tirilly, V. Claveau, and P. Gros, "Language modeling for bag-of-visual words image categorization," in *Proceedings of the 2008 international conference on Content-based image and video retrieval*, ser. CIVR '08. New York, NY, USA: ACM, 2008, pp. 249–258.

[24] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[25] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine regions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 1265–1278, 2005.