Bag-of-Visual-Ngrams for Histopathology Image Classification

A. Pastor López-Monroy[1], Manuel Montes-y-Gómez[1], Hugo Jair Escalante[1],
Angel Cruz-Roa[2], Fabio A. González[2]

[1] LabTL, Computer Science Department,
Instituto Nacional de Astrofísica, Óptica y Electrónica,
Luis Enrique Erro No. 1, C.P. 72840, Pue. Puebla, México.

[2] MindLab, Computing Systems and Industrial Engineering Department,
National University of Colombia,
Cra 30 No 45 03-Ciudad Universitaria, Bogotá DC, Colombia.

ABSTRACT

This paper describes an extension of the Bag-of-Visual-Words (BoVW) representation for image classification (IC) of histophatology images. This representation is one of the most used approaches in several high-level computer vision tasks. However, the BoVW representation has an important limitation: the disregarding of spatial information among visual words. This information may be useful to capture discriminative visual-patterns in specific computer vision tasks. In order to overcome this problem we propose the use of visual n-grams. N-grams based-representations are very popular in the field of natural language processing (NLP), in particular within text mining and information retrieval. We propose building a codebook of n-grams and then representing images by histograms of visual n-grams. We evaluate our proposal in the challenging task of classifying histopathology images. The novelty of our proposal lies in the fact that we use n-grams as attributes for a classification model (together with visual-words, i.e., 1-grams). This is common practice within NLP, although, to the best of our knowledge, this idea has not been explored yet within computer vision. We report experimental results in a database of histopathology images where our proposed method outperforms the traditional BoVWs formulation.

Keywords: Visual Words, Visual N-grams, Image Classification, Histopathology

1. INTRODUCTION

Nowadays, the amount of digital images available is constantly growing. This is mainly due to the availability of cheap image-capturing devices. In order to effectively exploit this overwhelming amount of information, several methods for the automatic processing, organization, and analysis of images have been proposed. In this aspect, image classification (IC) is one of the most studied tasks regarding the organization (e.g., indexing/storing images according to predefined categories) and analysis of visual information (e.g., for automated medical diagnosis from visual imagery). The general approach consists of representing images with vectors of visual features and using standard supervised-learning methods for building a classifier.

A crucial step regarding IC is to select an appropriate image representation. Different ways of representing images have been proposed so far. One of the most used is the Bag-of-Visual-Words (BoVW) formulation [12]. The BoVW representation is inspired in the bag-of-words (BoW) representation used in text classification and information retrieval (see, e.g., [18]). The underlying idea of BoW is to represent a document by a numerical vector that indicates the presence/absence of words in a document. Likewise, in computer vision tasks, a vocabulary of visual words is first generated (usually by grouping vectors of visual features representing parts of images) and then images are represented by histograms that account for the occurrence of visual words in images. This representation has been successfully used in several high-level computer vision

Further author information: (Send correspondence to A.P.L.M.)
A.P.L.M.: E-mail: pastor@ccc.inaoep.mx
M.M.G.: E-mail: mmontesg@ccc.inaoep.mx
H.J.E.: E-mail: hugojair@ccc.inaoep.mx
A.C.R.: E-mail: aacruzr@unal.edu.co
F.A.G.: E-mail: fagonzalezo@unal.edu.co

tasks, including, medical image classification[4,17], object recognition[24], video retrieval[12], image retrieval[16] and human-activity recognition[19].

Notwithstanding the fact that BoVW is widely used, it has an important shortcoming (inherited from the traditional BoW representation): it ignores spatial relationships among visual words. Spatial context has proven to be helpful for boosting the performance in diverse computer vision tasks (see e.g., Galleguillos and Belongie[8]). Thus it is promising trying to extend the BoVWs representation to incorporate spatial information. In this direction, in this paper we propose a natural extension to the BoVW formulation: the Bag-of-Visual-$n$grams (BoVN). $n-$grams are sequences of $n-$ words, they have been widely used in NLP, in particular within text mining and information retrieval[1,14,20]. This type of representation can capture compound word-patterns, e.g., *united-states*, *very-good*, *visual-words*, etc. Similarly, we propose building codebooks of visual $n-grams$ (multidirectional sequences of visual words) and then representing images using a BoW formulation. Our hypothesis is that this representation can capture frequent spatial patterns that may help to improve the classification performance in IC.

Most of the work in BoVW has been devoted to natural images, however we chose, as a first application of our approach, to focus on automatic classification of histopathology images. Histopathological images have particularities that distinguish them from the analysis of natural images: heterogeneous rich visual content, high intra-class variability and complex mixtures of non-localized patterns. In particular, a BoVWs representation assumes that there are localized patterns (visual words) which could characterize high-level concepts in the image; this is not necessarily true for histopathology images[5]. In this paper we consider the automatic classification of histopathology images according to tissue structures (healthy or pathological) that can be recognized by visual inspection of an expert pathologist (see Figure 1). Those images are particularly challenging, mainly because their classification is related to pathological lesions, morphological and architectural features, which encompass a complex mixture of visual patterns that allow to decide about the illness presence.[5].

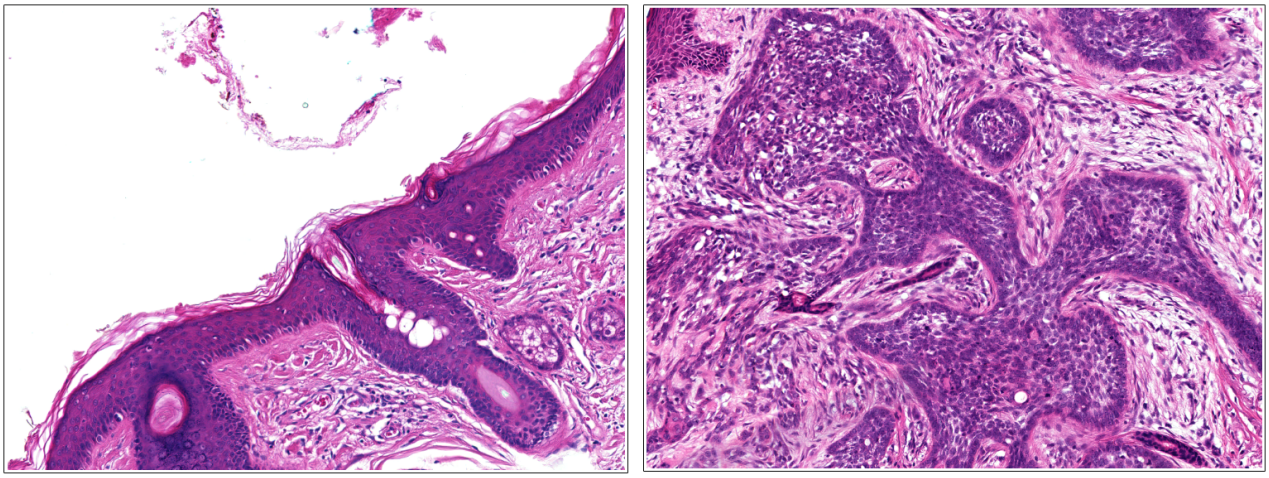


Figure 1. Example of histopathology images from skin biopsies with healthy (ephitelium) and pathological tissues (morpheaform basal-cell carcinoma), left and right respectively, used for basal-cell carcinoma diagnosis.

The main contributions of this work are twofold. On the one hand, we introduce the use of n-grams under the BoVW formulation for IC; where n-grams are used as attributes for a classification model. On the other hand, we show that the BoVN can outperform the performance of the BoVW approach for the classification of histopathology images.

The rest of this paper is organized as follows. The next section reviews related work on visual words. Section 3 introduces the use of visual n-grams. Section 4 explains the experimental settings we considered. Section 5 reports the experimental results we obtained, and Section 6 presents conclusions derived from our study and outlines future work directions.

2. RELATED WORK

This section reviews relevant work for this paper. We begin discussing the BoVW approach and its use in the computer vision literature; then we analyze works exploiting spatial information under the BoVW formulation. Finally, we outline

the basic idea of the n-gram model in NLP and how it could be applied in IC.

The BoVW approach was first introduced by Sivic and Zisserman tackling the problem of video retrieval [12]. Because of the good performance in that task, the BoVW approach quickly became popular and began to expand into other fields of computer vision such as IC [4,5,7], image retrieval [6,16], object recognition [24] and human-activity recognition [19]. The success of this approach could be easily explained through an analogy with the BoW used in text classification tasks. In this context, the image regions play the role of words which could be highly discriminative to identify a particular class/topic [24]. For example, in a object recognition, a wheel (or part of it) could be highly predictive for a car category.

The usual way to obtain visual words is as follows: i) a set of regions/parts are extracted from images, these regions are represented by feature vectors; ii) feature vectors are clustered, the centroids of the clustering process are considered as visual words to build a vocabulary/codebook; iii) visual words are used as attributes and are used to represent images (e.g. an histogram of the visual words that each image contains).

A limitation of the BoVW approach is that it disregards spatial relationships among words. Several works have tackled the problem of incorporating information from the relationships between visual words. For example, Jamieson et al. [10] represented relations using a graph of visual words in order to describe logos in sports photos. Other efforts have brought ideas from other areas such as NLP. For example in image retrieval, Zheng et al. [25] proposed the idea of visual phrases by pairwise grouping close or overlapping (according to a threshold) keypoint regions. Since the latter implies to test all keypoints in a one-vs-rest fashion, they test only on those frequent keypoints in the image dataset. In other works, Yuan et al. [22],[23] took advantage of the use of k-nearest neighbors algorithm to group visual words and building visual phrases of different lengths in order to get relevant information. In video data mining, visual phrases have also been used for obtaining the principal objects and characters in a video by clustering on viewpoint invariant configurations [13]. In other work, Quack et al. [11] have explored local sets of visual words to detect frequent and disctinctive features for object classes, this provides the option to use the method for object recognition or as a feature selector. Other approaches have used Language Models (LM) in order to capture spatial information. A language model is a popular technique used in NLP to model sequences of words. Previous works using LMs for computer vision tasks, perform several steps before training a LM [21], for example; the use of co-currence and proximity information of neighbor visual words. The latter is because a LM needs to "read" the visual words in some direction. For example, Tirilly et al. [15], used principal component analysis to project visual descriptors in a particular direction-axes, then inducing a sequence of visual words. Word sequences are classified using a Language Model Classifier (LMC). The LMC builds a LM for each class using the training documents. For testing, they measure the probability of belonging to each LM, and predict the class of the most probable. In Section 5, we compare the performance of our proposal to that approach.

In all of the previous works the authors have proposed interesting extensions to the BoVW representation, besides in all of these works are reported improvements over the standard BoVW representation. However, these proposals do not correspond necessarily to the way in which sequences of words are processed in NLP for boosting the performance. For instance, LMs are rarely used within NLP for text categorization. In this paper we adopt a representation that has proven to be very helpful for text categorization and apply it to IC. We focus in the idea of n-grams, which are merely sequences of n words [14]. Specifically, we propose using the Bag-of-Visual-$n$grams (BoVN) to represent images with the goal of improving the classification performance. A distinctive feature of our proposal is that we use n-grams as features for a classification model (in combination with visual words, i.e., 1-grams). This particular setting has proved to be very effective in text classification, outperforming LMs, straight BoW and other novel representations [1,14,20]. One should note, however, that the extension of BoVW to BoVN is not straightforward, in text mining, documents have a single spatial direction, whereas images lie in a 2D plane. Thus, we need a suitable way to extract the aforementioned spatial information.

3. REPRESENTING IMAGES THROUGH VISUAL N-GRAMS

In this section we explain in detail the Bag-of-Visual-$n$grams (BoVN) approach. In Figure 2 we show the general process for generating the BoVN. In the first step we take the whole (training) images and extract the visual words using the standard procedure outlined in Section 2. In this work we divide images into a grid and extract features from each patch [4]. In a second step, each patch of each image is replaced by the nearest visual word from the codebook generated in step 1 (Section 3.1). The third step involves the extraction of n-grams in order to build our visual n-gram codebook (explained in Section 3.2). In the fourth and final step we use the visual words codebook plus the visual n-gram codebook in order to get a final codebook. We use our final codebook to build histograms of the visual n-grams for each image. Each of these steps are described in the rest of this section.

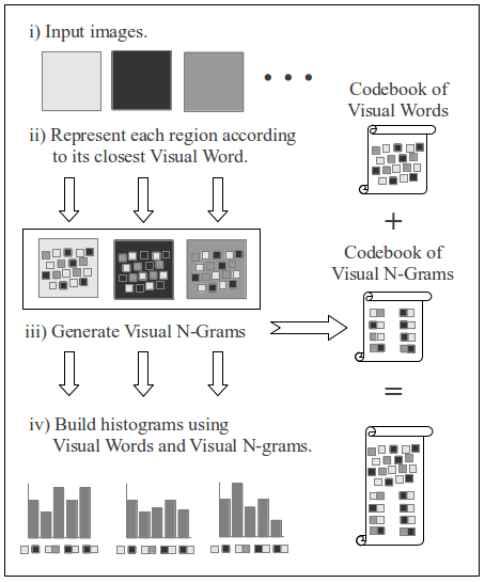


Figure 2. Image Representation through Bag-of_Visual-Ngrams.

3.1 Construction of the Visual Words Codebook

In this section we explain the first stage through the Bag-of-Visual-Ngrams (BoVN) approach. In Figure 3, we show the process to extract the visual words for an image collection using the standard BoVW formulation. We start extracting small patches from the images. For this, we use a regular-grid-based extraction. This is done by partitioning images using a regular grid, and taking each grid item as a patch of fixed size, see step ii) in Figure 3.

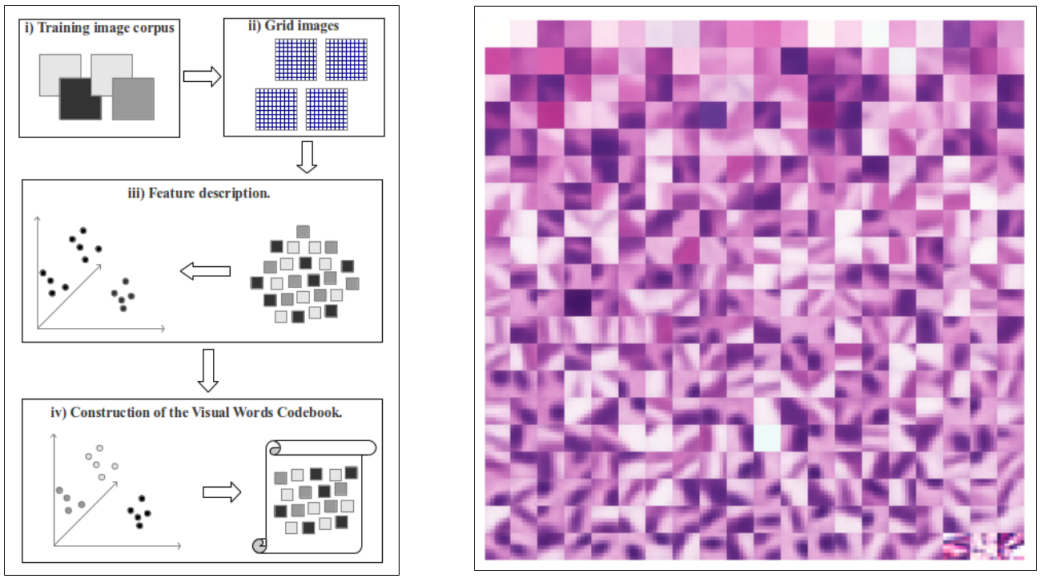


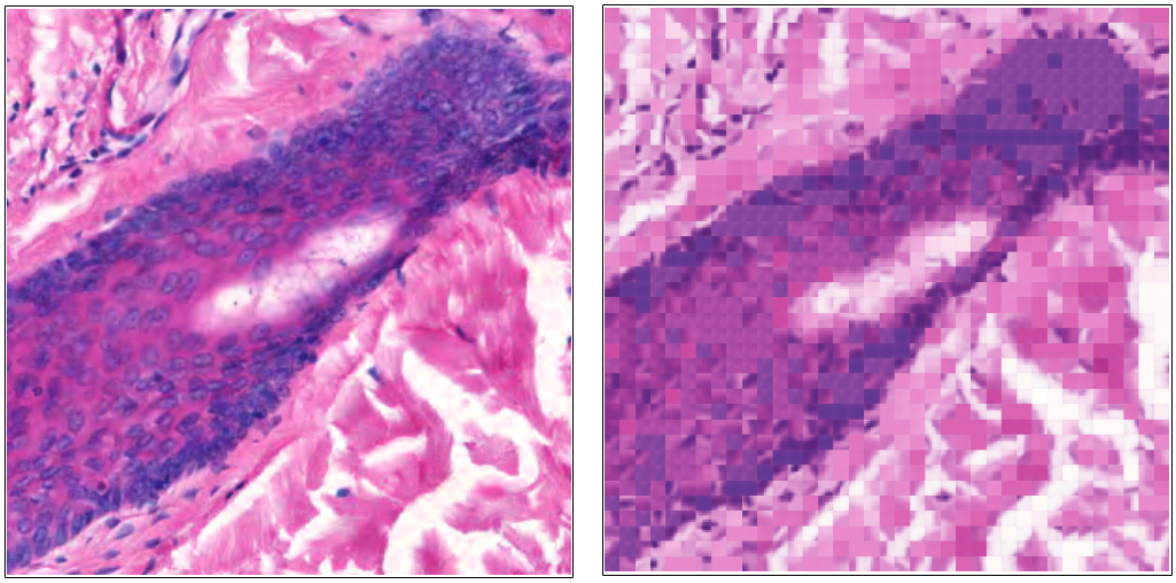
Figure 3. (a) The process to build a Visual Word codebook. (b) Example of a generated Visual Word codebook.

The next step consists in representing each extracted patch by a set of features. Among the wide variety of image descriptors in the literature, we use the discrete cosine transform (DCT) applied to each channel of the RGB color space by patch. The descriptor is built merging the 64 coefficients from each one of the three channels. We considered these features because in previous studies they have outperformed alternative representations (including SIFT features and raw-

patches)[4,5,7]. However, other types of feature-descriptors could be considered as well. This process is the third step in Figure 3.

The fourth and last step in the process is the construction of the visual dictionary or visual word codebook. The codebook is built by clustering all patch descriptors extracted from the image collection. This is done using a simple k-Means algorithm with $k = 400$, our choice was supported by a preliminary study[5]. In this process, all similar patch descriptors in the training set are grouped together independently of the source image. In this way, the k-means algorithm is used in this work to find a set of centroids which represent our visual words, which are labeled by an id and placed in the codebook. This last step is illustrated in the fourth step of Figure 3.

To represent images using the latter codebook, each image is gridded and each image patch is replaced by its closest visual word in the codebook (see Figure 4). In this way, each image is represented by a histogram that accounts for the occurrence of visual words (from the learned codebook) in the image. In the next section, we show how to use the aforementioned codebook in order to construct visual n-grams.



(a) (b)

Figure 4. Example of a represented image using the Visual Word codebook. (a) Original image. (b) Visual Words representation.

3.2 Extraction of visual n-grams

In this section we present the second stage to build our visual n-grams. As already mentioned, we assume that there is a visual codebook which we will use to represent images.

The idea for capturing spatial relationships is inspired by the use of n-grams in the area of text classification. In that area, word n-grams are sequences of n-consecutive words. That kind of simple sequences helps to maintain the semantic relationships between words. Because of that, the bag of n-grams representation can take advantage of concepts like "White House" which would be represented as one attribute in the bag of words. However, the extraction of visual $n-$grams from images is not as simple as in the case of text. While a text document can be read only in one direction, sequences of image descriptors can be obtained in many different ways (e.g., looking for sequences horizontally, vertical, at an angle of θ degrees, etc.), which pose a problem in the extraction of n-grams for images, see Figure 5. Another problem is to determine which is the right direction to interpret a given n-gram. A simple example using text are 3-grams composed by the same words but different order, those 3-grams normally have different meanings. For example, the 3-gram "united_states_of" is high probable that it refers to the country, while the 3-gram "of_states_united" surely refers to other things. On the other hand, visual n-grams that have the same order but different orientation (e.g., if an image is rotated), like 12-65-654 and 654-65-12 in Figure 5, may be related to the same pattern. In this work, we consider both patterns, 12-65-654 and 654-65-12, the same $n - gram$. In this way the generated visual $n-$grams are rotation invariant.

123	**213**	**12**	**33**	65	34	43	673
254	**546**	**65**	**444**	346	637	546	456
45	**645**	**654**	**565**	456	456	54	45
34	43	673	123	213	12	33	43
637	546	456	254	546	65	444	546
456	54	45	45	645	654	565	54
34	43	673	123	213	12	33	33
637	546	456	254	546	65	444	444

Figure 5. The process to build a Visual n-grams using a sliding window. For the dark path (65) the extracted n-grams are: 65-12, 65-213, 65-546, 65-645, 65-654, 65-565, 65-444, 65-33.

In order to construct visual n-grams we apply the following simple but effective approach. First of all remember that, we have a document containing the codeword matrix for each image (see Figure 5). Algorithm 1 shows our approach to build Visual N-Grams. The main idea is to produce n-grams ignoring the orientation in which they appear. To construct n-grams we iterate over each element $a_{i,j}$ of the matrix A (lines 2 and 3) and we extract neighbors in a straight fashion (lines 4 to 12). That is, we extract sequences using items between the items $a_{i,j}$ and $a_{i+k,j+h}$, if and only if they are part of the straight line joining $a_{i,j}$ and $a_{i+k,j+h}$. This leads to obtain n-grams in horizontal, vertical and diagonal directions. The "if ... exists" (lines 5 to 12) lines in Algorithm 1 are needed because even for elements in corners and edges we are attempting to extract all possible neighbors (this help us to keep things simpler and easier to explain). The latter condition leaves us with eight possible n-grams for each position in the matrix. Finally, each n is normalized to be read just in one way and consequently indexed as the same item in our new visual n-gram codebook.

Algorithm 1 get *no-orientation* n-grams

Require: **A** (a matrix of $x \times y$ containing the codewords), **n** (the length of the required visual sequences)
Ensure: $L = (S_1, \dots S_k); k = 1...l$ (a list of found sequences)

1: $n = n - 1$
2: **for** $i = 0$ **until** x **do**
3: **for** $j = 0$ **until** y **do**
4: Create an empty temporal sequence list TSL
5: if $(a_{i,j}, \dots, a_{i-n,j})$ exists, append to TSL
6: if $(a_{i,j}, \dots, a_{i-n,j-n})$ exists, append to TSL
7: if $(a_{i,j}, \dots, a_{i,j-n})$ exists, append to TSL
8: if $(a_{i,j}, \dots, a_{i+n,j-n})$ exists, append to TSL
9: if $(a_{i,j}, \dots, a_{i+n,j})$ exists, append to TSL
10: if $(a_{i,j}, \dots, a_{i+n,j+n})$ exists, append to TSL
11: if $(a_{i,j}, \dots, a_{i,j+n})$ exists, append to TSL
12: if $(a_{i,j}, \dots, a_{i+n,j+n})$ exists, append to TSL
13: **for** each sequence item E **in** TSL **do**
14: **if** $e_0 > e_n$ **then**
15: reverse(E)
16: **end if**
17: **end for**
18: Append elements in TSL to L
19: $j++$
20: **end for**
21: $i++$
22: **end for**

Once we have the visual n-gram codebook we proceed with the image representation. For this, each image is represented by a histogram of the occurrence of visual n-grams found in the image.

3.3 Image Classification

To perform image classification, we represent an image by its BoVN and use such histograms as feature vectors for training a classifier. We use a Support Vector Machine (SVM) using the default settings for the Sequential Minimal Optimization algorithm of Weka[9]. We used a SVM because among other approaches it has shown to be effective using the BoVW representation[2]. Furthermore, SVM has been used in similar histology image problems in order to find visual patterns[4,7].

4. EXPERIMENTAL SETTINGS

For evaluating the use of BoVN approach we consider a dataset of histopathology images, annotated by a pathologist, describing the presence of one architectural or morphological features, and pathological tissues[5,7]. Images correspond to RGB color images at 10X magnification stained with Hematoxilyn-eosin (H&E) from skin tissues diagnosed as healthy (collagen, epidermis, hair follicle, eccrine glands, sebaceous glands and inflammatory infiltrate), or by presence of basal-cell carcinoma (BCC) (which is the only one related with cancer diagnosis).

We take a subset from original histopathology dataset[7], which is composed of 1,417 histopathology images of 300x300 pixels, where each one might belong to one or more of 7 different categories related with morphological and architectural structures, from healthy or pathological tissues, for BCC diagnosis. In Table 1 we show the histopathology image distribution for each category.

Histopathology image	**positives**	**negatives**
1. basal-cell carcinoma	518	899
2. collagen	1238	179
3. epidermis	147	1270
4. hair follicle	118	1299
5. eccrine glands	126	1291
6. sebaceous glands	136	1281
7. inflammatory infiltrate	99	1318

Table 1. Histopathology image distribution for each category.

For evaluating the proposed approach, we built a binary classifier for each category. To accomplish this, we take as positive instances images belonging to the target category, and the rest as negative (i.e., a standard one-vs.-rest approach). As can be seen, each binary problem is imbalanced, in particular for classes 3-7, which makes the problem even more challenging.

We have performed several experiments for each classification problem. In those experiments, to extract image patches from BCC dataset we have gridded images using two settings: with patches of 8x8 pixels (which we denoted as 8) and with patches of 16x16 pixels (which we denoted as 16). For each experiment we have used a 10-fold cross validation. It is worth noting that, in our n-gram experiments a setting of order n includes all n-grams of lower or equal order than n. The feature combination was done in that way because that is the way that n-grams have shown to improve text classification[1,14,20] (we also performed experiments with separated representations but obtained worse results, confirming the results reported in text classification tasks). Furthermore, we have 400 unigrams and different number of n-grams for each different value of n (from 1 to 4). The latter means that, in an experiment of 3-grams ($1 + 2 + 3grams$) we have combined 400 unigrams plus x bigrams plus x 3-grams features for our BoVN. Moreover, it is worth mentioning that we have normalized each set of attributes in an individual way. Finally, for the experiments we have tested two of the most used term weighting schemes in Text Classification. The first one is term frequency which we denote as TF. TF weighting consist in using the histogram values in a feature vector, but normalized by the total number of items in the instance. On the other hand, boolean weighting scheme build feature vectors replacing each value v in the histogram by 1 if $v > 0$ and by 0 if not.

Along Section 5 we will explain more details about each experiment such as: the way we measured the performance and other specific conditions for each experiment.

5. EXPERIMENTS AND RESULTS

In this section we explain the purpose and details for each experiment. We have chosen the most relevant experiments for analizing different properties of the use of visual-n-grams for IC. The best result of each set of experiments is put in bold. In the following tables we report the average (over the classes) of F_1-Measure (FM) and the average area under ROC curve (AUC) of the seven binary problems in BCC histopathology image collection.

5.1 Bag-of-Visual-Words versus Bag-of-Visual-Ngrams

In these initial experiments, we have used four different scenarios with different settings. For term weighting we have: i) binary (BIN), and ii) term frequency (TF). For size patches we have: i) 8x8 (8), and ii) 16x16 (16). We show the averaged F-Measure and AUC of the seven binary problems in our histology image collection.

The first experiment considers all visual words contained in images. The aim of this experiment is to determine the classification effectiveness using the traditional BoVW under different conditions. Experiments in Table 2 show that the 8 size patch using TF weighting obtains the best results. Which is somewhat expected since it is related with a good size of resolution to cover the biological structure of cells[4]. On the other hand, we think TF weighting in general is good choice due to the better accounting of visual patterns (remember that binary weighting only see for the presence).

Visual Words		
Config	*FM*	*AUC*
Bin-8	48.27	67.74
Bin-16	47.63	67.56
TF-8	**58.59**	**72.27**
TF-16	52.33	68.89

Table 2. Experiments using Visual Words (Unigrams) through two kinds of term weighting (TF and BIN) and two different size patches (8 and 16).

Once we know the maximum performance that can be obtained under the visual words formulation we analyze the performance of BoVN. First we study the performance of BoVN using different numbers $n-$grams, that is, we use the top-x $n-$grams for building our $n-gram$ codebook. This reduction is necessary because the number of initial n-grams is of the order hundreds of thousands. The following experiment presents a study of how much the dimensionality influences the classification performance. Table 3 shows the results obtained with the best settings varying the number of the most frequent bigrams. From this experiment we can see that using 2500 bigrams results in the highest performance, just sightly better than the experiment using 5000. Furthermore, it can be seen that there is a difference of almost 6% in the 64.31% averaged F-Measure obtained by the unigrams-bigrams, against the 58.59% averaged F-Measure in experiments using unigrams (which is the traditional BoVW). We think this is because of the pairs of visual words are finding good visual patterns, which in some way reinforce some evidence in text categorization. Knowing this evidence, we will take advantage of it in the following experiments.

Frequency threshold					
Config	*1.5K*	*2.5K*	*5K*	*7.5K*	*10K*
1+2grams	63.73	**64.31**	64.03	62.24	61.63

Table 3. Experiments using Bigrams to analyze the impact of dimensionality.

Experiments in Table 4 use visual bigrams in order to examine its behavior under the same conditions that visual words. From Table 4 we can see that the weighting TF using patches of size 8, again outperforms the other settings. Also note that every configuration of visual bigrams get better results than the traditional visual words. This clearly shows that, in general way, and under same conditions combining as attributes n-grams of visual words are better than use only the visual words.

The last experiment of this section is shown in Table 5. This table presents the results of the experiments in a BoVN approach for visual n-grams using different values of n. We perform experiments using unigrams (which are the traditional visual words and our baseline) to Tetragrams. The purpose of these experiments is to expose whether considering n-grams of higher order than 1, could improve the performance of the classifier. From results in Table 5 we can figure out that the

Unigrams vs Uni+Bigrams				
Config	*F-Measure*		*AUC*	
	1grams	*1+2grams*	*1grams*	*1+2grams*
Bin-8	48.27	59.50	67.74	72.54
Bin-16	47.63	56.67	67.56	70.46
TF-8	58.9	**64.31**	72.27	**76.03**
TF-16	52.33	56.09	68.89	71.17

Table 4. Experiments using sequences of Visual Words (Uni-Bi-grams) through two kinds of term weighting (TF and BIN) and two different size patches (8 and 16).

best setting is $1 + 2$grams. This can be due to the following reasons. The first one is related with the size of the sequences: it is well known that the higher n for n-grams, the higher number of instances are required to find that sequences of length n[14]. The second one is related with the high dimensionality: using longer sequences produces large vocabularies, which also produce sparse feature vectors since long sequences are more difficult to find.

Experiments with n-grams	
Config	*FM*
1grams	58.59
1+2grams	**64.31**
1+2+3grams	62.69
1+2+3+4grams	61.34

Table 5. Experiments using sequences of Visual Words (from Unigrams to Tetragrams) to analyze the impact of sequence length.

5.2 Detailed analysis per class for unigrams and bigrams for visual words

Below we present the detailed results per class. Those experiments consider the best setting in Table 2 using visual unigrams against the best setting in Table 4 using visual bigrams.

In Table 6 and 7 we show respectively the obtained F-Measure and AUC for the positive class in each category. For these experiments, we performed a 10-fold cross validation using unigrams and bigrams in each of the seven binary problems. Also in the "(b-a) gain/loss" column we expose the gain or loss (in F-Measure or AUC) caused by the use of bigrams.

Detailed F-Measure by class			
Class	*(a)*	*(b)*	*(b-a)*
	1grams	*1+2grams*	*gain/loss*
1	86.10	90.70	4.6
2	94.80	95.50	0.7
3	74.40	83.40	9.0
4	36.80	50.80	14.0
5	35.80	52.50	16.7
6	48.00	43.60	-4.4
7	34.20	33.70	-0.5

Table 6. Detailed experiments per class using Visual Words (Unigrams) versus sequences of Visual Words (Uni-Bi-grams).

Experiments in Tables 6 and 7 shows that by using visual bigrams higher improvements (fourth column) are obtained, by comparing to unigrams, for classes 1, 3, 4, 5, which are respectively basal-cell carcinoma, epidermis, hair follicle and eccrine glands. From these classes the most important is the first one, this is because it is the only one related with cancer diagnosis. We think that good results in class 1 were obtained because much part of these images are characterized by tumor cells having large and darker nuclei, which are successfully retained in visual bigrams. On the other hand, visual words seem to be very competitive or better in classes 2, 6 and 7, which are collagen, hair follicle and inflammatory infiltrate (none of those related with cancer diagnosis) are images which structured spatial relations of its components are

Detailed AUC by class			
Class	*(a) 1grams*	*(b) 1+2grams*	*(b-a) gain/loss*
1	89.00	92.50	3.5
2	76.40	79.20	2.8
3	84.00	90.90	6.9
4	62.60	68.80	6.2
5	62.80	71.60	8.8
6	68.70	66.90	-1.8
7	62.40	62.30	-0.1

Table 7. Detailed experiments per class using Visual Words (Unigrams) versus sequences of Visual Words (Uni-Bi-grams).

not enough to be capture by using bigrams due its simplicity (collagen or inflammatory infiltrate) or by high complexity and greater visual variability (hair follicle). We think that these problems would need more instances and explore other parameters (e.g. patch sizes, large sequences of visual words, or alternative descriptors) in order to get appropriated visual n-grams that could improve the visual words.

5.3 Comparison against other typical approaches

In addition to the SVM using the visual words and visual n-grams as attributes, we also present experiments using another classical approach for visual words; language models. As explained in Section 2, language models have been used in previous works [15,21] for building classifiers. In this paper, we have implemented a language model classifier as the one used in [15], which is based on the *CMU-Cambridge Statistical Language Modeling Toolkit v2* [3]. The goal is to compare our proposal to alternative methodologies that also use sequences of visual words. The language model classifier uses $1 + 2 + 3grams$ (configurations up to $10 - grams$ were tested) remaining parameters of the software were left by default (e.g., smoothing good turing discount and backoff). The language model classifier works as follows:

- For each binary problem, it takes the training documents and builds two model languages (one for positive class and one for the negative).
- For each test document, it measures the distance (using the probability chain rule) against the positive and negative model and it assigns the closest category.

Table 8 shows the results of the experiments comparing the language model classifier (LMC) and visual n-grams classifier. From this experiment, it can be seen that at least for this problem and under the same conditions the LMC does not provide better performance than visual n-grams. We think this is due to the high unbalanced data, which provides very few documents to build accurate language models for some positive classes. Moreover, since language models rely in probabilistic bases, the unbalanced data represents a common problem.

LMC vs Uni+Bigrams				
Config	F-Measure		AUC	
	LMC	1+2grams	LMC	1+2grams
TF-8	53.0	**64.31**	69.89	**76.03**
TF-16	48.31	56.09	72.21	71.17

Table 8. Experiments using sequences of Visual Words (Uni-Bi-grams) compared with a LMC.

6. CONCLUSIONS

In this paper we have proposed an extension to the standard BoVW formulation for image representation. Our extension focused on extracting sequential visual patterns and use them as attributes for a classification model. It is worth noting that, although the general idea of n-grams in visual words has been explored for works related to information retrieval,

language models, and feature selection, to the best of our knowledge it has never been used as an attributes for machine learning algorithms in image classification tasks. In this way, as a first application we have used a histopathology image collection. In order to get our sequential visual patterns we extract n-grams using ideas from NLP, but taking into account the particularities of the image domain. Experimental results under different conditions and settings have shown that the use of visual n-grams as attributes has proven to be useful since it allows to improve the effectiveness of the classification versus the traditional BoVW approach and LMC. We believe this is because the method is finding visual patterns, which could be difficult to obtain from individual visual words. Future work includes applying the visual n-grams based-representation to other computer vision tasks that use BoVW as representation. Besides, we would like to study the impact that the order in which $n-$grams are built have into the classification performance.

7. ACKNOWLEDGEMENTS

This work was partially funded by Colciencias under the project "Automatic Annotation and Retrieval of Radiology Images Using Latent Semantics" and by LACCIR R1212LAC006 under the project "Multimodal image retrieval to support medical case-based scientific literature search". López-Monroy thanks for doctoral scholarship CONACyT-Mexico 243957. Finally, Cruz-Roa also thanks for doctoral grant supports Colciencias 528/2011 and "An Automatic Knowledge Discovery Strategy in Biomedical Images" DIB-UNAL/2012.

References

[1] Bekkerman, R. and Allan, J. (2004). Using bigrams in text categorization. Technical report, Department of Computer Science, University of Massachusetts, Amherst.

[2] Boiman, O., Shechtman, E., and Irani, M. (2008). In defense of nearest-neighbor based image classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition 2008. CVPR 2008*, pages 1–8.

[3] Clarkson, P. and Rosenfeld, R. (1997). Statistical language modeling using the cmu-cambridge toolkit. In *Proceedings of EUROSPEECH*, volume 97, pages 2707–2710. International Speech Communication Association Rhodes,, Greece.

[4] Cruz-Roa, A., Caicedo, J. C., and González, F. A. (2011a). Visual pattern mining in histology image collections using bag of features. *Artificial intelligence in medicine*, 52:91–106.

[5] Cruz-Roa, A., Díaz, G., Romero, E., and González, F. A. (2011b). Automatic annotation of histopathological images using a latent topic model based on non-negative matrix factorization. *Journal of Pathology Informatics*, 4.

[6] Csurka, G., Dance, C. R., Fan, L., Willamowski, J., and Bray, C. (2004). Visual categorization with bags of keypoints. In *International Workshop on Statistical Learning in Computer Vision, ECCV*, volume 1, page 22.

[7] Díaz, G. and Romero, E. (2012). Micro-structural tissue analysis for automatic histopathological image annotation. *Microscopy Research and Technique*, 75:343–358.

[8] Galleguillos, C. and Belongie, S. (2010). Context based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114:712–722.

[9] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. (2009). The weka data mining software: An update. *SIGKDD Explorations*, 11.

[10] Jamieson, M., Fazly, A., Dickinson, S., Stevenson, S., and Wachsmuth, S. (2007). Learning structured appearance models from captioned images of cluttered scenes. In *IEEE 11th International Conference In Computer Vision, 2007. ICCV 2007.*, pages 1–8.

[11] Quack, T., Ferrari, V., Leibe, B., and Van Gool, L. (2007). Efficient mining of frequent and distinctive feature configurations. In *IEEE 11th International Conference on Computer Vision, 2007. ICCV 2007.*, pages 1–8. IEEE.

[12] Sivic, J. and Zisserman., A. (2003). Video google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision, ICCV*.

[13] Sivic, J. and Zisserman, A. (2004). Video data mining using configurations of viewpoint invariant regions. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, volume 1, pages I–488. IEEE.

[14] Tan, C. M., Wang, Y. F., and Lee, C. D. (2002). The use of bigrams to enhance text categorization. *Information processing and management*, 38:529–546.

[15] Tirilly, P., Claveau, V., and Gros, P. (2008). Language modeling for bag-of-visual words image categorization. In *ACM Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 249–258.

[16] Tirilly, P., Claveau, V., and Gros, P. (2009). A review of weighting schemes for bag of visual words image retrieval. Technical report, Technical report, TEXMEX - INRIA - IRISA.

[17] Tommasi, T., Orabona, F., and Caputo, B. (2007). Image annotation task: an svm-based cue integration approach. In *Working notes of the 2007 CLEF Workshop*.

[18] Turney, P. and P., P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

[19] Wang, H., Ullah, M. M., Klaser, A., and Laptev, I. (2009). Evaluation of local spatio-temporal features for action recognition. In *Proceedings of the British Machine Vision Conference*, pages 1–11.

[20] Wang, S. and Manning, C. D. (2009). Baselines and bigrams: Simple, good sentiment and topic classification. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*, pages 90–94.

[21] Wu, L., Li, M., Li, Z., Ma, W. Y., and Yu, N. (2007). Visual language modeling for image classification. In *ACM Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 115–124.

[22] Yuan, J., Wu, Y., and Yang, M. (2007). Discovery of collocation patterns: from visual words to visual phrases. In *IEEE In Computer Vision and Pattern Recognition, 2007. CVPR 2007*, pages 1–8.

[23] Yuan, J., Yang, M., and Wu, Y. (2011). Mining discriminative co-occurrence patterns for visual recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011*, pages 2777–2784. IEEE.

[24] Zhang, J., Marszalek, M., Lazebnik, S., and Schmid, C. (2007). Local features and kernels for classification of texture and object categories: a comprehensive study. *International Journal of Computer Vision*, 73:213–238.

[25] Zheng, Q. F., Wang, W., and Gao, W. (2006). Effective and efficient object-based image retrieval using visual phrases. In *ACM Proceedings of the 14th annual ACM international conference on Multimedia*, pages 77–80.