Image Classification through Text Mining techniques: a Proposal

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

[1] Instituto Nacional de Astrofísica, Óptica y Eletrónica (INAOE)
LabTL, Computer Science Department
Luis Enrique Erro No. 1, C.P. 72840, Tonantzintla, Puebla, México
{pastor,mmontesg,hugojair}@inaoep.mx
[2] National University of Colombia
MindLab, Computing Systems and Industrial Engineering Department
Cra 30 No 45 03-Ciudad Universitaria, Bogotá DC, Colombia.
fagonzalezo@unal.edu.co

Abstract. Image classification is an important task for the organization and analysis of visual information. According to the literature one of the most important concepts is the *visual word*; a visual element that represents a set of visual-similar regions. The Bag-of-Visual Words (BoVW) is one of the most widely used approaches in High Level Computer Vision (HLCV). The BoVW is an histogram of the occurrence of visual words in each image, which is in some way inspired by the Bag-of-Words (BoW) used in Natural Language Processing (NLP). In spite of the success of BoVW, it has the same limitations of BoW (e.g., the overlook of the spatial context). In this research proposal we bear in mind the successful evidence of visual words in HLCV, and we take the analogy of visual-textual words to a new higher level. This is, by designing methods inspired in NLP, we aim to consider contextual (e.g., spatial, sequential), and high level (e.g., semantic) information among visual words. However, bringing NLP like approaches pose several nontrivial problems, for example: i) the definition of analogous attributes (visual-textual), ii) a suitable strategy to interpret images; documents can be read only in one direction, but in images we have a 2D plane without an specific way to *read* them, iii) the way to extract high level information (e.g., semantic). This paper presents the proposed research methodology and through preliminary results, we provide strong evidence of the feasibility of this research. For this, a popular NLP technique is used to improve the BoVW; the Bag-of-Visual n-grams (BoVN). The idea is evaluated in the challenging task of Histopathology image classification overcoming the BoVW and an state-of-the-art approach based in language models.

Key words: Visual words, n-grams, spatial context, image classification, histopathology.

1 Introduction

Nowadays there is a huge amount of images available through different media sources. In many situations all this information is useless without appropriate tools for analysis. In this regard, image classification is one of the most important tasks for the organization and exploitation of visual information for different areas. The representation of images is one of the key procedures for successful models in classification. Currently one of the most widely used approaches in the state-of-the-art of High Level Computer Vision (HLCV) tasks is the Bag-of-Visual Words (BoVW). The BoVW is somehow inspired by the Bag-of-Words (BoW) representation of text mining (see e.g., [16]). Under the BoW formulation, vocabulary vectors representing documents are built, and each element of the vector indicates the presence or absence of each word in the document. Similarly, in HLCV tasks a vocabulary of visual word is generated (clustering feature vectors representing image regions and taking the centroid of each cluster as a visual word) in order to represent images through vectors that accounts for the occurrence of visual words in each image (see Figure 3). The BoVW has been successfully used in several HLCV task including: medical image categorization [2, 3], texture and object classification [20], video retrieval [15], image retrieval [18], human activity recognition [19], etc.

Problem to solve: Notwithstanding the fact that visual words approaches (like BoVW) are widely used, they usually do not exploit the contextual (spatial relationships) and high level (e.g., semantic) information among visual words. Spatial context has proven to be useful to increase the performance of several HLCV tasks (see e.g., [5, 9]). In this direction, contextual and high level information among visual words could be captured taking the analogy visual-textual words into a complete new higher level using Natural Language Processing (NLP) approaches.

Main Objective: Designing and developing methods for image classification, which based on the concept of visual word and inspired by NLP approaches, can model contextual and high level information to improve the classification.

Main Contribution: the design of novel and effective HLCV methods inspired by NLP, that consider the properties of the image domain to exploit effectively contextual and high level information among visual words. For example, novel methods based on: n-grams (sequences of n elements), weighting schemes (weight functions for the visual elements), semantical distributional analysis, etc.

In this context, the interest of this research lies in a relatively young area, the intersection of the fields of NLP and HLCV, which has been the main subject of study of different forums and works [1,14,17,18]. To design a successful approach we focus on techniques that have proven to be highly useful in NLP. To figure out whether the best approaches in NLP have the opportunity to improve visual words methods, we begin exploring one basic, intuitive, yet effective idea of NLP; n-grams. n-grams are sequences of n elements which have proven to be very useful in text categorization tasks for capturing the context [16]. Through the achieved results, we show the feasibility of this proposal giving two main contributions: i) a method to extract n-grams from visual words, and ii) the way to effectively

use n-grams as attributes for a classifier. The proposal overcomes the traditional BoVW and an state-of-the-art approach based on language models. The rest of this paper is organized as follows. Section 2 describes the proposed research methodology. Section 3 describes the dataset. Section 4 introduces our approach. Section 5 shows and discusses preliminary results. Finally, Section 6 shows our conclusions and indicates the paths of future research.

2 Research Methodology

The research methodology is as follows (main contributions are steps 3 and 4):

1. **Identifying and obtaining copora:** To find datasets with challenging peculiarities that NLP methods could handle (e.g., contextual information).
2. **Analyzing and developing methods to extract visual words:** To identify methods to get a better analogy between visual-textual words, our initial paths considers: i) Extraction of regions through regular grids [13], and ii) Extraction of regions through key points [12].
3. **Proposing a set of new representations inspired in NLP to capture contextual information:** For this, we consider the following three approaches as the best candidates, which aims to capture contextual information at different levels:
   (a) **Sequences of visual words to capture the pure local context**: The general idea is to use sequences of elements similar to the n-grams (sequences of n words) for text mining [6]. A challenge here consists in defining a suitable way to extract such n-grams. This is because in contrast to text documents, in images we have a 2D plane, and the way to read the elements is not defined.
   (b) **Locally weighted bag of words to capture local-global context:** Using this approach a higher level of contextual information can be capture [4]. Through this representation it is possible to assign different weighs to several *parts* of a document, which in analogy could facilitates to focus in relevant image *regions*.
   (c) **n-gram graphs to capture the pure global context:** In the classic text mining graphs of words [7], nodes would represent visual words, which are connected among them by edges modeling the co-occurrence, frequency and order. Such graphs allow to capture global information about the elements in the target object (images).
4. **Proposing a set of new representations inspired in NLP to capture high level information:** For this we consider the following three approaches as the best candidates, which aims to capture high level information (e.g., semantic) when bringing to the image domain.
   (a) **Concise Semantic Analysis (CSA) of visual words:** This is an special distributional semantical representation, which through the use of low-dimensional vectors, allows to capture relationships among documents and the target classes [11]. In analogy, adapting this approach is

possible to build image vectors that highlights discriminative relationships with each target class.

(b) **Knowledge based hierarchies (*is-a* hierarchies):** These kind of hierarchies are widely used in text mining to represent semantic relationships among words of specific domains. Similarly, having a hierarchy for visual words would makes possible to capture different information. Building such hierarchies for visual words is challenging, but could be achieved in different ways (etc. using the hierarchy produced by a hierarchical clustering, using distributional semantical representations, etc.).

5. **Designing and implementing a method to integrate the information extracted by the NLP inspired approaches:** This last step involves to take advantage of different spaces of information. There are several ways in the literature to achieve this combination (e.g., classifier ensemble techniques, multiple kernel learning, etc.) [8].

In the following section, we present the work done so far. For this, we describe the initially dataset, then we explain the process to build the visual words and extract visual n-grams to improve BoVW (item 3(a) of the proposed methodology). Finally, we discuss the obtained results and future avenues of inquiry.

3 Description of the Image Collection

The proposed methods and representations in this research will be evaluated using several image collections ranging from natural to medical images (selected collections depend of the Step 1 of the methodology). For the evaluation of the sequences of visual words we initially perform experiments using an Histopathology image collection. We decide to use this kind of images, because their visual tissues structures (healthy or pathological) make them challenging. In this images, classification is related to pathological lesions and morphological-architectural features which can be captured by our proposed visual n-grams (see Figure 1).

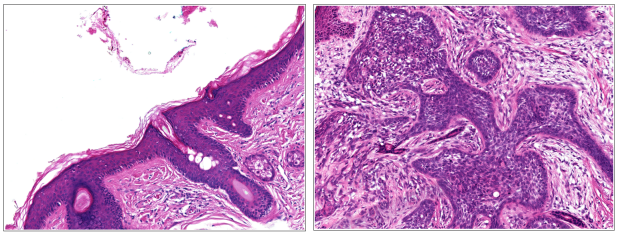

Fig. 1: Example of histopathology images from skin biopsies with healthy and pathological tissues (basal-cell carcinoma), left and right respectively.

In the evaluation we use a dataset of 1417 histopathology images, annotated by a pathologist, describing the presence of architectural features, and pathological tissues [3]. Each image might belong to one or more of 7 categories (see

distribution in Table 1). To evaluate our approach we built a binary classifier using an standard one-vs-rest approach.

Histopathology class	**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: The seven binary problems of the 1417 Histopathology image collection. The positive instances are images belonging to a target category.

4 Sequences of Visual Words: Visual *n*-grams

Our first approximation to capture the local context of visual words is through the use of one popular technique in NLP: the n-grams. For this we propouse the use n-grams of visual words to improve the BoVW. In other words, we focus in the Bag-of-Visual n-grams (BoVN). In Figure 2 we outlines each step of the process for generating the BoVN. In the first step, the training collection is used to generate the dictionary of visual words (codebook) (explained in Section 4.1). In the second step, each patch of each image is replaced by the nearest visual word in the codebook. The second step also involves the extraction of n-grams in order to build our visual n-gram codebook (explained in Section 4.2). The third step combines the visual words codebook and the visual n-gram codebook. The final codebook is used to build histograms of the visual n-grams in each image. We explain in detail the latter steps in the following subsections.

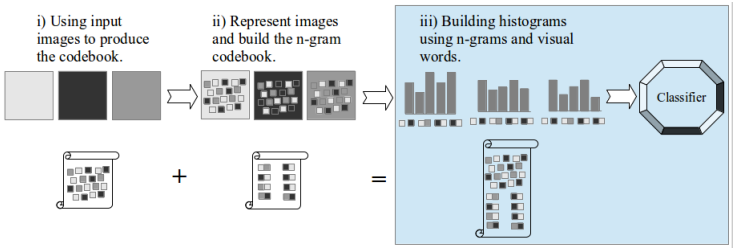

Fig. 2: Image Representation through Bag-of_Visual-Ngrams.

4.1 Construction of the Visual Words Codebook

In Figure 3 (A), we show the extraction of visual words using a standard BoVW formulation. In the step 1 and 2, we use a regular-grid-based patch extraction. In step 3, we represent each patch using the discrete cosine transform (DCT) applied to each channel of the RGB color space. We merge the 64 coefficients from each of the three channels to get the final descriptor. In the last step, the codebook is built using cluster centroids (using a 400-Means algorithm) of the training patch descriptors. Those settings are supported by previous closely-related studies using Histopathology images, showing better performance than other configurations (including SIFT and raw-patches, and several k values) [2].

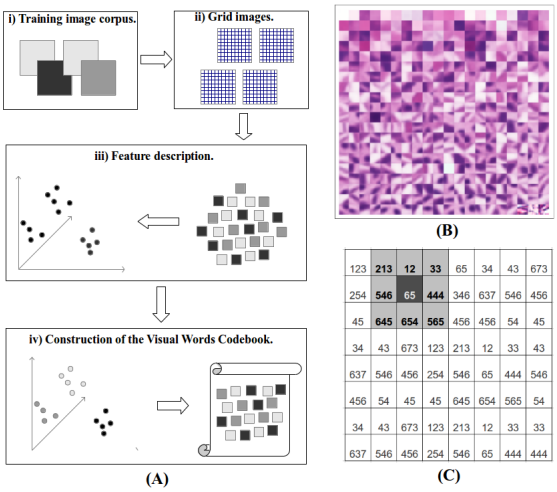

Fig. 3: (A) The process to build a visual word codebook. (B) Example of a generated visual word codebook. (C) The process to build visual n-grams using an sliding window. For the dark path (65) the extracted n-grams are: 65-12, 65-213, 65-546, 65-645, 65-654, 65-565, 65-444, 65-33.

4.2 Extraction of visual n-grams

To capture spatial relationships among visual words, we inspired our idea in the use of word n-grams for text classification tasks. They are sequences of n consecutive words that helps to maintain semantic relationships between words, which allows to represents as one attribute concepts like "cold war". Nonetheless in image domain, the extraction of visual n-grams face some additional issues. For example, a document can be read only in one direction, but sequences of image descriptors can be extracted horizontally, vertical, at an angle of θ degrees, etc.). Another problem is to determine the right direction to interpret each visual

n-gram. For example, 3-grams in text normally can be interpreted correctly only in one direction (say, "the human being", but not "being human the"). On the other hand, visual 3-grams can have the same order but different orientation if the image is rotated. Therefore, the two descriptor sequences d_a-d_b-d_c and d_c-d_b-d_a might be the same pattern. In this work, we consider such patterns the same, making them rotation invariant.

To construct visual n-grams we apply the following effective approach. Consider a document containing the codeword matrix for each image (see (C) in Figure 3), the main idea is to produce n-grams ignoring the orientation in which they appear. For this, we iterate over each item $a_{i,j}$ of the matrix A and we extract neighbors in a straight fashion. Thus, we build horizontal,vertical and diagonal sequences using items between the current item $a_{i,j}$ 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 approximation to text n-grams produces eight possible n-grams for each position in the matrix. Finally, each n-gram is normalized to be "read" only in one direction. For example, in the visual n-gram codebook, a trigram 21-61-73 is indexed as the same item than 73-61-21. For image classification we build feature vectors using a BoVN. This is, using the learned codebook, each image is represented by an histogram of the occurrence of found visual n-grams. We use a Support Vector Machine (SVM) using the default settings of Weka framework. We used a SVM because it has shown to be effective in similar histology image problems finding visual patterns [2, 3].

5 Preliminary Results on Image Clasification using Sequences of Visual Words

In the evaluation we use an stratified 10 fold cross validation (10FCV) and report the averaged F-Measure and Area Under roc Curve (AUC) of the seven binary problems. We have used several scenarios, for term weighting; binary (BIN), and term frequency (TF). The former focuses only in the presence/absence of the element, the latter in the weighted presence/absence. For size patches we have: 8x8, and 16x16. Finally, as text mining evidence suggest [16], k n-grams includes k $(n-1)$-$grams$, k $(n-2)$-$grams$, $\ldots$, k (1)-$grams$. Finally, it is worth knowing that we perform a set of specific experiments that highlights the general usefulness of visual n-grams, but a detailed study of the parameterization and other applications can be found in [10].

First Experiment: Table 2, thought the first row, shows the results of evaluating the traditional BoVW (Unigramas). There the 8 size patch using TF weighting obtains the best results. The 8x8 size patch seems to be a good size of resolution to cover the biological structure of cells, which confirms results reported of other works in this dataset [2]. On the other hand, the TF weighting best results suggest that, in general is a good choice the accounting of visual patterns rather than focusing only in their presence/absence.

In Table 2 we also present results of BoVN and an state-of-the-art approach under the same conditions. Since the number of possible n-grams are of hun-

dreds of thousands, we analyze the performance influence in the performance of BoVN (testing with values from one thousand to ten thousand of features), getting that 2500 bigrams are a good balance between the dimensionality and the performance of our approach. The results show evidence of the usefulness of n-grams, showing that, every experiment using visual bigrams outperforms unigrams. Also in Table 2 we show results of another classical approach for visual words in the literature; language models. In this paper we have implemented a Language Model Classifier (LMC) as the one used in [17]. The goal is to compare the BoVN with other approaches in HLCV that also take advantage of the contextual information. To train language models, we have used exactly the same software and parameters (*Carnegie Mellon Statistical Language Modeling Toolkit*). As can be seen from Table 2, at least for this problem and under the same conditions, LMC does not get better performance than BoVN. In part, this might be because language models rely in probabilistic, where the unbalanced data represents a common problem to build accurate models for positive classes.

Performance of BoVN					
		8x8		*16x16*	
Approach	*Criteria*	*Bin*	*TF*	*Bin*	*TF*
BoVW (1grams)	FM	48.27	58.90	47.63	52.33
	AUC	67.74	72.27	67.56	68.89
BoVN (1+2grams)	FM	59.50	**64.31**	56.67	56.09
	AUC	72.54	**76.03**	70.46	71.17
LMC (1+2+3grams)	FM	53.0		48.31	
	AUC	69.89		72.21	

Table 2: Experiments using Uni-Bi-grams (sequences of visual words), two kinds of term weighting (TF and BIN) and two different size patches (8 and 16).

In other set of experiments we also analyze the performance considering higher order n-grams (e.g., 3grams, 4grams, etc.). Results show that the best setting is $1+2$grams, which is somewhat expected because it is well known that for a higher n-grams more instances are required to find those sequences [16]

Second Experiment: In Table 3 using an 10FCV, we analyze the detailed performance for each class. Thus, for BoVW and BoVN, we use the best settings for each method in Table 2. Results in Tables 3 shows that 1+2grams overcomes 1grams in classes 1, 3, 4, and 5. The class 1 is the most important, because it is the only one related with cancer diagnosis. Images in class 1 present structural tumor cells having large and darker nuclei, which are accurately characterized by visual bigrams. Visual words (1-grams) are competitive or better in classes 2, 6 and 7 (none of them related with cancer diagnosis). Such classes are in opposite ends, either by the lack of structured spatial visual elements (classes 2 and 6) that make bigrams to lose their advantage, or because the contextual information of visual words are much more global rather than local (class 6). We

think those problems need more instances and explore other parameters (e.g. patch sizes, size of sequences, or alternative descriptors).

Detailed F-Measure by class			
Class	*(a) 1grams*	*(b) 1+2grams*	*(b-a) 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

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 3: Detailed experiments per class using Unigrams versus Uni-Bi-grams. In column"(b-a) gain/loss" we show the gain or loss caused by the use of bigrams.

6 Conclusions

The interest of this research lies in the fields of HLCV and NLP. The underlying motivation is to improve state-of-the-art visual words approaches (such as the BoVW) through methods that takes the analogy visual-textual words into a new higher level. For this, we consider contextual (spatial) and high level (semantic) information, which is overlooked by several approaches like BoVW. Since the use of the contextual information is a common factor in NLP tasks, as an initial approach, we propose the natural extension to BoVW; the use on visual n-grams as attributes (the BoVN). Our results suggest strong evidence of the usefulness of BoVN in Histophalogy images. To the best of our knowledge, n-grams have never been extracted as we propose; in a similar way they boost NLP tasks, and subsequently use them as feature vectors for a classifier. Future research paths include bringing ideas to capture contextual information in a more global way, and extracting high level information among visual words.

References

1. Bruni, E., Tran, K.N., Baroni, M.: Multimodal distributional semantics. Journal of Artificial Intelligence Research (49), 1–47 (2014)
2. Cruz-Roa, A., Caicedo, J.C., González, F.A.: Visual pattern mining in histology image collections using bag of features. Artificial intelligence in medicine 52, 91–106 (2011)
3. Díaz, G., Romero, E.: Micro-structural tissue analysis for automatic histopathological image annotation. Microscopy Research and Technique 75, 343–358 (2012)

4. Escalante, H.J., Solorio, T., Montes-y Gómez, M.: Local histograms of character n-grams for authorship attribution. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. pp. 288–298 (2011)
5. Galleguillos, C., Belongie, S.: Context based object categorization: A critical survey. Computer Vision and Image Understanding 114, 712–722 (2010)
6. García-Hernández, R.A., Martínez-Trinidad, J.F., Carrasco-Ochoa, J.A.: A new algorithm for fast discovery of maximal sequential patterns in a document collection. In: Computational Linguistics and Intelligent Text Processing, pp. 514–523. Springer (2006)
7. Giannakopoulos, G., Mavridi, P., Paliouras, G., Papadakis, G., Tserpes, K.: Representation models for text classification: a comparative analysis over three web document types. In: Proceedings of the 2nd International Conference on Web Intelligence, Mining and Semantics. p. 13. ACM (2012)
8. Kuncheva, L.: Combining pattern classifiers. Wiley Press, New York pp. 241–259 (2005)
9. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2006. vol. 2, pp. 2169–2178. IEEE (2006)
10. López-Monroy, A.P., Montes-y Gómez, M., Escalante, H.J., Cruz-Roa, A., González, F.A.: Bag-of-visual-ngrams for histopathology image classification. In: IX International Seminar on Medical Information Processing and Analysis, vol. 8922, p. 89220P. SPIE (2013)
11. López-Monroy, A.P., Montes-y-Gómez, M., Villaseñor-Pineda, L., Carrasco-Ochoa, J.A., Martínez-Trinidad, J.F.: A new document author representation for authorship attribution. In: Mexican Conference in Pattern Recognition. pp. 283–292. Springer (2012)
12. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision 60(2), 91–110 (2004)
13. Nowak, E., Jurie, F., Triggs, B.: Sampling strategies for bag-of-features image classification. In: Computer Vision–ECCV 2006, pp. 490–503. Springer (2006)
14. Quack, T., Ferrari, V., Leibe, B., Van Gool, L.: Efficient mining of frequent and distinctive feature configurations. In: IEEE 11th International Conference on Computer Vision. pp. 1–8. IEEE (2007)
15. Sivic, J., Zisserman., A.: Video google: A text retrieval approach to object matching in videos. In: Proceedings of the International Conference on Computer Vision (2003)
16. Tan, C.M., Wang, Y.F., Lee, C.D.: The use of bigrams to enhance text categorization. Information processing and management 38, 529–546 (2002)
17. Tirilly, P., Claveau, V., Gros, P.: Language modeling for bag-of-visual words image categorization. In: ACM Proceedings of the 2008 international conference on Content-based image and video retrieval. pp. 249–258 (2008)
18. Tirilly, P., Claveau, V., Gros, P.: A review of weighting schemes for bag of visual words image retrieval. Tech. rep., Technical report, TEXMEX - INRIA - IRISA (2009)
19. Wang, H., Ullah, M.M., Klaser, A., Laptev, I.: Evaluation of local spatio-temporal features for action recognition. In: Proceedings of the British Machine Vision Conference. pp. 1–11 (2009)
20. Zhang, J., Marszalek, M., Lazebnik, S., Schmid, C.: Local features and kernels for classification of texture and object categories: a comprehensive study. International Journal of Computer Vision 73, 213–238 (2007)