IMAGE CLASSIFICATION THROUGH TEXT MINING TECHNIQUES

By:

Adrián Pastor López Monroy

A dissertation submitted in partial fulfillment
of the requirements for the degree of:

DOCTOR OF SCIENCE IN COMPUTER SCIENCE

at

Instituto Nacional de Astrofísica, Óptica y Electrónica

February, 2017
Tonantzintla, Puebla

Supervised by:

Dr. Manuel Montes-y-Gómez, INAOE
Dr. Hugo Jair Escalante, INAOE
Dr. Fabio A. González, UNAL Colombia

ABSTRACT

Nowadays there is a huge amount of images available through different media sources. In many situations all this information is useless without the appropriate tools for analysis. In this regard, image classification is one of the most important tasks for the organization and exploitation of visual information in different areas. The representation of images is one of the key procedures for successful models in classification. According to the literature one of the most important concepts for capturing visual patterns is the *visual word*; a visual element that represents a set of visual-similar regions. In this regard, the Bag-of-Visual Words (BoVW) representation is one of the most widely used approaches in computer vision. 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).

Although the BoVW is simple and effective, facilitating its use to a wide range of problems, it inherits some well known limitations from the traditional BoW. For example, the disregarding of spatial and semantic information among visual words, which hinder the extraction of valuable visual-patterns. In this regard, the information retrieval and text mining communities have proposed several solutions for similar problems using textual features. In this thesis, we alleviate the latter limitations by taking the analogy visual-textual words into a new higher level. This is, by designing and evaluating methods inspired in NLP, we aim to capture the spatial context (e.g., spatial, sequential), and high level (e.g., semantic) information among visual words. For this purpose, we defined suitable strategies to interpret images, which allow us to obtain highly discriminative attributes and representations.

In order to capture the spatial context, we build new simple-effective visual features inspired in the popular idea of n-gram representations in NLP. For this, we propose building a codebook of multi-directional visual n-grams, and use them as attributes to represent images by means of the BoVW representation. Regarding to the semantic visual information, we propose to represent images by adapting distributional representations. These analogous *visual-textual* representations exploit statistics of visual words occurrences and co-occurrences along the dataset. Furthermore, we also proposed two novel distributional visual feature representations, which allow to capture intra-class and inter-class specific visual information. Finally, we also propose suitable strategies to jointly exploit contextual and semantic information of visual words. For this, we consider the

visual words and the new visual n-grams as different feature spaces, then we propose different fusion strategies to better integrate such visual information.

We report experimental results in several image datasets showing the effectiveness of the proposals over BoVW and other methods in the literature. We evaluate the proposed ideas in the image classification task using five different datasets. Experimental results show that the proposed strategies outperform or are competitive with; i) the traditional BoVW, ii) the BoVW using visual n-grams under traditional fusion schemes (e.g., ensemble based classifiers) and iii) other approaches in the literature for image classification that consider the spatial context and semantic information.

RESUMEN

Hoy en día existe una gran cantidad de imágenes disponibles a través de distintos medios. En muchas situaciones toda esta información es inútil sin las herramientas apropiadas para su análisis. En este sentido, la clasificación de imágenes es una de las tareas más importantes para la organización y aprovechamiento de la información visual en diferentes áreas. Con respecto a esto, la representación de imágenes es uno de los procedimientos clave para modelos exitosos en clasificación. De acuerdo a la literatura, uno de los conceptos más importantes para capturar patrones visuales es la *palabra-visual*; un elemento visual representativo de un conjunto de regiones que son visualmente similares entre sí. En este sentido, la Bolsa de Palabras Visuales (*Bag-of-Visual Words, BoVW*) es uno de los enfoques más ampliamente utilizados en visión computacional. La BoVW es un histograma de la ocurrencia de palabras visuales en cada imagen, la cual en cierto sentido está inspirada en la Bolsa de Palabras (*Bag-of-Words*) utilizada en Procesamiento de Lenguaje Natural (*Natural Language Processing, NLP*).

Aun cuando la BoVW es conceptualmente sencilla y efectiva, algo que facilita su aplicabilidad en un amplio rango de problemas, ésta hereda algunas conocidas limitaciones de la BoW. Por ejemplo, en la BoVW no se considera la información espacial y semántica entre las palabras visuales, lo cual dificulta la extracción de patrones visuales valiosos. En este contexto, las comunidades de recuperación de información y minería de texto han propuesto distintas soluciones para problemas utilizando características textuales. En esta tesis, suavizamos las ya mencionadas limitaciones llevando la analogía de palabra visual-textual a un nuevo nivel. Esto es, por medio de el diseño y evaluación de métodos inspirados en NLP, se propone capturar la información del contexto espacial (por ejemplo, espacial, secuencial), y de alto nivel (por ejemplo, la semántica) que existe entre las palabras visuales. Para este propósito, hemos definido estrategias adecuadas para interpretar imágenes, las cuales permitan obtener atributos y representaciones que sean altamente representativas de las clases en el problema.

Para capturar el contexto espacial, se han construido nuevas características visuales que son simples-efectivas, y que están inspiradas en la idea popular de las representaciones de n-gramas de NLP. Para esto, se proponen construir un vocabulario de n-gramas visuales multi-direccionales, y usarlos como atributos para representar a las imágenes por medio de la representación de BoVW. Con respecto a la información semántica, se propone representar

a las imágenes por medio de la adaptación de representaciones distribucionales. Estas representaciones toman ventaja de las ocurrencias y co-ocurrencias de palabras visuales a lo largo de todo el conjunto de datos. Además, también se proponen estrategias adecuadas para tomar ventaja conjunta de la información contextual y semántica entre las palabras visuales. Para ello, se consideran a las palabras visuales y los n-gramas visuales como espacios de características distintas, luego entonces se proponen diferentes estrategias de fusión que permitan obtener una mejor integración de la información visual.

Se reportan resultados experimentales en distintas colecciones de imágenes que muestran la efectividad de las propuestas sobre la BoVW y otros métodos en la literatura. Las ideas propuestas se han evaluado en la tarea de clasificación de imágenes utilizando cinco diferentes conjuntos de imágenes. Los resultados experimentales muestran que las estrategias propuestas superan o son competitivas con; i) la BoVW tradicional, ii) la BoVW utilizando n-gramas visuales bajo algún esquema tradicional de fusión de información (por ejemplo, clasificadores basados en ensambles) y iii) otros enfoques en la literatura para clasificación de imágenes que consideran el contexto espacial y la información semántica.

AGRADECIMIENTOS

Al Consejo Nacional de Ciencia y Tecnología (CONACYT), por el apoyo otorgado a través de la beca no. 243957. Así como al INAOE por todas las facilidades prestadas durante mi estancia académica.

A mis asesores, Dr. Manuel Montes y Gómez, Dr. Hugo Jair Escalante y Dr. Fabio A. González quienes con su conocimiento, experiencia y buen carácter me acompañaron a lo largo de mis estudios.

A mis sinodales, Dr. José Francisco Martínez Trinidad, Dr. Miguel Arias Estrada, Dr. Jesús A. González Bernal, Dr. Leopoldo Altamirano Robles y Dr. Roberto Paredes Palacios por sus observaciones y comentarios.

A mi familia, amigos y personas cercanas a mí por su apoyo constante e incondicional. Finalmente, pero no menos importante, a mis compañeros del INAOE y todos aquellos que creyeron en mí apoyándome, animándome, escuchándome.

DEDICATORIA

Para Dios,
porque siempre está conmigo
guiándome en el camino.

Para todos los que creyeron en mí,
apoyándome, animándome, escuchándome,
siempre para alcanzar mis sueños.

CONTENTS

ABSTRACT i

RESUMEN iii

AGRADECIMIENTOS v

I Introduction 1

1 INTRODUCTION 3
- 1.1 Working Hypothesis . . . 5
- 1.2 Main Objective . . . 6
- 1.3 Contributions . . . 6
- 1.4 Document Outline . . . 7

II Theoretical Background Review 9

2 TEXT AND IMAGE CLASSIFICATION BACKGROUND 11
- 2.1 Basic Concepts on Text Classification . . . 11
  - 2.1.1 The Bag of Words Representation . . . 12
  - 2.1.2 Relevant Textual Features . . . 13
- 2.2 The Bag of Visual Words Representation . . . 14
- 2.3 Ensemble Based Classifiers . . . 15
  - 2.3.1 Diversity in the Search Space . . . 15
  - 2.3.2 Decision Making in Ensembles . . . 17

3 RELATED WORK ON CONTEXTUAL AND SEMANTIC VISUAL INFORMATION 19
- 3.1 Contextual information under the analogy visual-textual words . . . 19
- 3.2 Semantic information under the analogy visual-textual words . . . 21
- 3.3 Relevant Fusion Information Strategies . . . 22

3.3.1 Early Fusion and Late Fusion Strategies . . . 22
3.3.2 Intermediate Fusion Strategies . . . 23

III Contributions **25**

4 EXPLOITING THE CONTEXTUAL VISUAL INFORMATION **27**
4.1 Image Classification through Visual n-grams and MKL . . . 28
4.1.1 Construction of the Visual Words Codebook . . . 29
4.1.2 Extracting visual n-grams . . . 30
4.1.3 Exploiting the jointly use of Visual words and Visual n-grams . . . 31
4.2 Image Collections . . . 34
4.3 Experimental settings . . . 36
4.3.1 Statistical significance of results . . . 38
4.4 Experiments and Results . . . 38
4.4.1 Bag-of-Visual-Words versus Bag-of-Visual-Ngrams . . . 38
4.4.2 Strategies to Exploit Visual n-grams . . . 41
4.4.3 Bag-of-Visual n-grams and the Spatial Pyramid Representation (SPR) . . . 48
4.5 Final Remarks . . . 51

5 EXPLOITING THE SEMANTIC VISUAL INFORMATION **53**
5.1 Distributional Term Representations, from text to images . . . 53
5.1.1 Image Occurence Representation (IOR) . . . 55
5.1.2 Visual-Feature Co-occurence Representation (VCOR) . . . 55
5.1.3 Class Occurrence Representation (COR) . . . 56
5.2 New Distributional Visual-Feature Representations . . . 57
5.2.1 Subclass Occurrence Representation (SOR) . . . 57
5.2.2 Group Occurrence Representation (GOR) . . . 59
5.3 Visual features used as terms . . . 60
5.3.1 Visual words . . . 60
5.3.2 Visual bigrams . . . 61
5.4 Experimental Settings . . . 62
5.5 Experiments and Results . . . 63
5.5.1 Distributional representations on visual words . . . 63
5.5.2 Distributional representations on Visual Words and Visual n-grams . . . 65
5.6 Final Remarks . . . 67

IV General Conclusion **69**

6 GENERAL CONCLUSIONS **71**
6.1 Scientific Publications 73

A SUPPLEMENTARY RESULTS **3**

B FINE GRAIN CLASSIFICATION RESULTS **7**

List of Figures

2.1 The Bag-of-Visual-Words framework in computer vision. 16

4.1 Image Representation through Bag-of-Visual-Ngrams. 28
4.2 The process to build a visual word codebook. 29
4.3 Example of a represented image using the Visual Word codebook. 30
4.4 The process to build a Visual n-grams using a sliding window. 31
4.5 Image samples of the image collections. 37
4.6 Example of an image related with cancer diagnosis and its new relevant visual features. 45
4.7 Sample instances to expose the image characteristics of the Birds dataset. . . . 47
4.8 Example of a Mandarin duck image and its new relevant visual bigrams. 47
4.9 Sample instances to expose the image characteristics of the Butterflies dataset. . 48

5.1 The typical two stages of the Distributional Term Representations framework. . 54
5.2 Subclass Occurrence Representation (SOR). 58
5.3 Group Occurrence Representation (GOR). 60
5.4 Remider of the n-grams of visual words. 61

B.1 Three images of different aircraft variants in the FGVC-Aircraft dataset. 8
B.2 Three images of different bmw cars in the BMW-10 dataset. 8
B.3 Three images of different dog breeds in the Stanford Dogs dataset. 8

List of Tables

4.1 Representative MKL algorithms in the literature. 35
4.2 Image collections used for the evaluation of visual n-grams. 36
4.3 F-Measure results for visual words vs visual n-grams. 39
4.4 F-Measure results for visual words vs visual n-grams. 40
4.5 Strategy 1: Single Kernel (Linear | Intersection) - Several Spaces. 42
4.6 Strategy 2: Several Kernels (Linear + Intersection) - Single Space in all datasets. 43
4.7 Strategy 2: Several (Linear + Intersection) Kernels - Single Space in the Histopathology dataset. 45
4.8 Strategy 2: Several (Linear + Intersection) Kernels - Single Space in the Birds dataset. 46
4.9 Strategy 2: Several (Linear + Intersection) Kernels - Single Space in the Butterflies dataset. 48
4.10 F-Measure results for RBMKL vs SPR. 50
4.11 F-Measure results for SPR extended with visual n-grams and MKL. 50

5.1 Distributional Terms Representations using SIFT-based and DCT-based visual words. 64
5.2 Distributional Terms Representations using SIFT-based and DCT-based visual n-grams. 66

A.1 F-Measure: Distributional Terms Representations using SIFT-based and DCT-based visual words. 4
A.2 F-Measure: Distributional Terms Representations using SIFT-based and DCT-based visual n-grams. 5

B.1 Preliminary evaluation for Fine Grained Classification 10

Part I

Introduction

CHAPTER 1

INTRODUCTION

Nowadays the huge amount of digital information available is constantly growing. Much of this information are images generated by image-capturing devices in a wide variety of different domains. All this vast amount of images could be exploited for the benefit of several practical applications, which makes important to have automated tools to assist their analysis. In general, Image Classification (IC) aims to organize images according to predefined categories. IC is one of the most important tasks regarding the organization and analysis of visual information. There are several methods for IC, but the traditional approach consists in representing images with vectors of visual features, and then building classification models using supervised learning algorithms (Csurka et al., 2004).

The representation of images is a key procedure for IC, thus a number of different approaches have been proposed so far. The Bag-of-Visual Words (BoVW) (Sivic and Zisserman., 2003; Csurka et al., 2004) representation is one of the most used approaches because of its simplicity and effectiveness, achieving outstanding performance in several computer vision tasks, for example: medical image classification (Tommasi et al., 2007; Cruz-Roa et al., 2011a), category level scene classification (Fei-Fei and Perona, 2005), object recognition (Zhang et al., 2007), video retrieval (Sivic and Zisserman., 2003), image retrieval (Tirilly et al., 2009), human-activity recognition (Wang et al., 2009), etc. The core idea behind the BoVW is very similar to the Bag-of-Words (BoW) representation used in text mining tasks (see, e.g., (Turney and P., 2010)). On the one hand, under BoW each document is represented with a vector, taking each word in the vocabulary as an attribute. On the other hand, the BoVW precomputes a vocabulary of visual words from the training dataset (e.g., clustering vectors of relevant visual features representing parts of images), then the BoVW represents images with vectors that account for the presence/absence of visual words in images (e.g., histograms of visual words). [1]

Although the BoVW representation has been used by many researchers, most of the work

[1] In this analogy, visual words play the role of words to identify a particular class/topic (Zhang et al., 2007). Thus, the pure presence/absence of specific visual features can provide valuable information for discriminating between target classes. For example, in face recognition, an eye (or part of it) could be highly informative to recognize a face.

involving BoVW has been devoted to the analysis of new visual descriptors[2]. In fact, there is a few amount of research approaching the well known deficiencies of BoW/BoVW representations. In this aspect, we can differentiate the following relevant issues that could compromise the BoW/BoVW applicability in several domains:

- *The disregarding of spatial context among visual words:* The BoVW representation is an histogram of occurrences of visual words, then no spatial information among extracted features is considered. In specific computer vision tasks, spatial context properly exploited, has been useful to improve the performance of several approaches (Galleguillos and Belongie, 2010; Krapac et al., 2011)[3].

- *The assumption of independence among visual words:* No relationship, other than mere occurrence, among visual words is captured by the representation. This is, the representation does not capture associations between visual words and other elements in the problem. For example the distributional semantic information, which could be very useful to capture discriminative visual-patterns. For example, the distribution of visual word occurrences and co-occurrences along the dataset (Lavelli et al., 2004).

- *The high dimensionality and sparseness:* images usually are represented in a large vector space of length equal to learned-vocabulary size, which can hinder the use of some learning methods (Joachims, 1998), to increase the runtime performance (Joachims, 1999), in some cases obfuscating the interpretation of the representation (Phan et al., 2008; Sriram et al., 2010). Also, images usually contain a small subset of features (visual words), resulting in sparse representations that make difficult to interpret and build accurate models for some image classification problems.

In this thesis we introduce novel approaches to represent and classify images, which to some extent, overcome the discussed drawbacks for BoVW in image classification. For example, regarding to the first one (*the disregarding of spatial context among visual words*) it is promising to enrich the BoVW by using spatial information in order to improve the representativeness of the visual elements. Thus, we begin adapting one intuitive and 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 (Tan et al., 2002). For facing the second problem (*the assumption*

[2]Novel visual descriptors have enhanced the classification performance using approaches to improve the representation of visual information, then building more representative visual words (Tirilly et al., 2009).

[3]Most of the time at the cost of requiring higher computational resources, the spatial context have been captured in several ways; for example, computing relative (Tirilly et al., 2008) or absolute (Lazebnik et al., 2006) spatial configurations of visual words, or integrating the distance and angle information among specific visual words (Krapac et al., 2011; Cao et al., 2010)

of independence among visual words), we are interested in semantic relationships between visual features, in order to represent images using the distribution of visual words along the instances in the training set or target classes in the problem. These semantic attributes usually are few, but they are rich in representativeness, which also faces the third issue (*the high dimensionality and sparseness*). In order to extract such attributes, we exploit the distributional hypothesis, which states that words (visual words in our case) with similar meanings tend to occur in the same contexts (Sahlgren, 2008). Among the most relevant strategies exploiting the distributional hypothesis are the Distributional Term Representations (DTRs) (Lavelli et al., 2004). Thus, through DTRs we build instance representations that consider contextual information by means of term occurrences and co-occurrences. By computing such statistics, DTRs can produce enriched representations that help to overcome, the BoVW shortcomings. Furthermore, we devise two novel distributional strategies that learn appropriated groups of images to compute better suited distributional features. Finally, in this thesis we also propose the jointly use of visual words and its contextual and semantic visual information to represent images. First of all, we propose Multiple Kernel Learning (MKL)[4] intermediate fusion strategies to jointly exploit the use of visual words and our version of visual n-grams, which can be seen as representing images under different feature spaces (e.g., visual words and visual n-grams). After that, in Chapter 6 we describe practical strategies to better exploit contextual and semantic information by integrating visual words and visual n-grams into the DTRs.

For evaluating this research work we focus on automatic classification using five different image collections: Histopathology, Birds, Butterflies, Scenes and a subset of CalTech-101. These image collections have special particularities like: heterogeneous rich visual content, high intra-class variability and complex mixtures of non-localized structural patterns. In particular, a BoVWs representation assumes that there are localized patterns (visual words) which could characterize high-level concepts in the image. Nevertheless, in this thesis we go beyond exploiting the usefulness of spatial context and semantic information captured by text mining techniques, in order to encompass a complex mixture of visual patterns that allow to decide about the class. The experimental evaluation reported in this thesis suggests positive evidence that the proposed approaches are good alternatives to other approaches reported in the literature.

1.1 Working Hypothesis

The main hypothesis for this research is stated as follows:

The analogy visual-textual words in computer vision can be better exploited by adapting

[4]MKL exploits similarity kernel functions to delegate the construction of a new combined kernel function to an algorithm (Gönen and Alpaydın, 2011; Alioscha-Pérez et al., 2012).

NLP approaches, which consider contextual (e.g., spatial relationships) and high level (e.g., semantic) information among visual words, in order to propose novel and effective methods for the image classification task.

1.2 Main Objective

To design and develop methods for image classification, which based on the concept of visual word and inspired by NLP approaches, allow to model contextual and semantic information to improve the classification. In order to achieve this, we define the following specific objectives:

1. To analyze and develop the appropriated methods to extract visual words that better exploit the analogy between visual-textual words.
2. To propose a new representation inspired in visual n-grams in order to take advantage of the contextual information among visual words.
3. To propose a set of new representations inspired in distributional representations in order to model the semantic visual information in images.
4. To design and implementing strategies to integrate the information extracted by the NLP inspired approaches by taking advantage of spatial contextual and semantic visual information.

1.3 Contributions

The main contributions of this thesis are [5]:

1. A new method to capture the spatial context based on the extraction and exploitation of n-grams of visual words.
2. Novel methods to capture semantic information inspired in the distributional hypothesis, which adapt Distributional Term Representations (DTRs) for image classification.
3. Two novel distributional visual feature representations, that automatically obtain better suited representations according to each image domain.
4. Fusion strategies to handle the combined use of contextual and semantic visual information, which allow the construction of classification models.

[5] In Section 6.1 we list the scientific publications derived from this thesis.

1.4 Document Outline

In this thesis we consider the successful evidence of visual words in the BoVW, but we intend to take the analogy of visual-textual words into a new higher level. Thus, in this research work we are interested in exploiting this visual-textual feature analogy into other successfully representations from the text mining community. The remaining of the thesis document is organized as follows:

- In Part II, there are two chapters presenting the relevant background elements, with the aim of making this thesis as self-contained as possible. This part contains the following chapters:
  - In chapter 2, we describe some of the most relevant concepts in text and image classification for this research.
  - In chapter 3, we present some of the most relevant works in the literature exploiting NLP approaches to model visual information.
- Part III organizes the main contributions of this thesis, which are related to contextual and semantic visual information.
  - In chapter 4 we present the proposed approaches to capture the contextual visual information. For this, we describe the formulation of visual ɳ-grams and intermediate fusion strategies to improve the image classification.
  - In chapter 5 We describe the proposals to capture the semantic visual information. For this, we present several strategies for exploiting the distributional semantic information.
- Part IV has the chapter 6, which outlines the main conclusions and future work of this research.

Part II

Theoretical Background Review

CHAPTER 2

TEXT AND IMAGE CLASSIFICATION BACKGROUND

The interest of this research lies in a relatively young area, the intersection of the fields of NLP and computer vision (called Vision and Language), which has been the main subject of study of different forums and works (Ferraro et al., 2015). In this regard, to design successful approaches based on visual words, we focus on techniques that have proven to be highly useful in NLP. Thus, to figure out whether the best approaches in NLP have the opportunity to improve visual words methods, we begin by exploring the most relevant NLP techniques for capturing spatial context and semantic information. In this regard, it is also necessary to outline the most relevant concepts on computer vision, which benefit this bridge between textual and visual modeling. In this way, we introduce the two broad topics of interest: i) text classification and ii) image classification. Section 2.1 presents the main framework in text classification, which encompasses the main textual features and conventional classification methods. In a similar way, Section 2.2 introduces relevant concepts related to visual words construction, and the aforementioned BoVW framework in image classification.

2.1 Basic Concepts on Text Classification

The amount of information available on Internet is overwhelming, and much of it is plain text (e.g., books, journals, e-mails, blogs, source code, etc.). In this context, several issues and applications related to text classification have emerged, for example: authorship attribution (Stamatatos, 2009), author profiling (Schler et al., 2006), opinion mining (Pang and Lee, 2008), etc. From a computational perspective, the text classification task consists in learning (based on specific textual features) about one or more document classes, in order to automatically identify them in future texts. In this context, the text classification can be stated as an standard single-labeled multiclass classification problem (Sebastiani, 2008).

2.1.1 The Bag of Words Representation

Most of the text classification tasks can be approached as standard classification problems. This means that they can be stated as single-labeled multiclass classification problem, where the target groups of documents represent the classes to discriminate. Several standard machine learning methods have been used to face the identification of each target class. In this regard, the framework of most relevant strategies are divided in the following three key procedures: i) feature extraction, ii) document representation and iii) classification. In this section we review the second one: the representation of documents. Currently, one of the most effective and simple approaches is the Bag of Words (BoW) (Sebastiani, 2008; Stamatatos, 2009; Nguyen et al., 2013) representation. The BoW representation builds document vectors using textual features; for example, taking each word in the corpus vocabulary as an attribute. In this way, BoW represents documents with feature vectors, and assigns a value to each feature (Pavelec et al., 2008). This value could be Boolean (1 or 0), frequencies computed from the analysis of the corpus, or a specific weighting scheme. BoW representations have been used for thematic classification, authorship attribution, spam filtering, plagiarism detection, etc.

In spite that BoW has been widely used, there are some important drawbacks that could compromise its applicability. For example, BoW representations do not maintain any order or relation among the textual features, which could give valuable information and improve the representativeness. Another problem with BoW occurs in realistic scenarios where there are large vocabularies, but few training data and imbalanced classes. The latter causes that BoW representations tend to favor majority classes, when in fact each document can actually belong to any class (Stamatatos, 2008). Moreover, BoW representations require huge computational resources to classify large sets of documents with huge vocabularies, which could be impractical in some situations (e.g., author profiling for socialmedia, where there are hundreds of texts belonging to specific author profiles) (Solorio et al., 2011). In order to address the main drawbacks of BoW, other kinds of representations have been used according to each classification problem. For instance, the use of tensors (Plakias and Stamatatos, 2008) for representing stylistic properties of texts, graphs based representations for placing relevant features in the same neighbourhood, or even rule based representations in order to define feature logic relevance. Most of the latter alternative representations consider relationships between terms, however the classification performance usually is sightly better than BoW at the cost of requiring excessive computational resources. Furthermore, most of the alternative representations do not provide a solution for the problem of information dispersion and high dimensionality.

2.1.2 Relevant Textual Features

In text classification tasks each specific problem has its own particularities and should be approached accordingly (Abbasi and Chen, 2008). In this regard, one of the most important factors to achieve effective solutions relies in textual features. For example, while in thematic classification the most valuable features in BoW are the content words, in authorship attribution the most important textual features are non-thematic; since the main goal is to model the writing style of each author (Stamatatos, 2009). In the following subsections, we briefly present the most relevant textual features in the literature.

Lexical Features

Lexical features generally consider the text as a sequence of tokens. In general, tokens could be words, numbers, punctuation marks, acronyms, etc. In this context, it is possible to define several lexical features based in such sequences of tokens. For example, the frequency of tokens, frequency of sequences of n tokens (called n-grams), measuring the length of words, sentences or even paragraphs. Moreover, there also exists strategies to measure rates of spelling errors, vocabulary richness or lexical repetitiveness (Miranda-García and Calle-Martín, 2005). Lexical features have one important advantage, most of them could be extracted by using simple existing tools such as tokenizers.[1]

Character Based Features

From a general perspective, this character based features considers the text as a sequence of characters. In this regard, it is possible to define textual features based on a number of statistics, ranging from specific rates of individual elements to interesting sequences of characters. For example, one of the most popular character feature are n-grams, which are mere sequences of n characters. Throughout character n-grams, the approaches are able to capture content, stylistic and contextual information (Stamatatos, 2009). Furthermore, n-grams usually are more robust to spelling errors than traditional approaches based on word tokenizers. For example, consider a socialmedia document containing the words *Brasil* and *Brazil*; conventional lexical word approaches would consider them as two different features, when in fact both represent exactly the same concept besides an spelling error. On the other hand, using an approach based on character 3-grams we can extract the following attributes; *Bra*, *ras*, *raz*, *asi*, *azi*, *sil* and *zil*. The latter means that we hold the thematic information in *Bra*, but also the spelling errors in *sil* and *zil*. In spite of the advantages for some character base features, there also exists some

[1] Note that some languages, such as Chinese, requires more sophisticated and specific analysis tools, such as: lemmatizers, sentences splitters, thesaurus, spelling correctors, etc.

disadvantages regarding to lexical features. For instance, the high dimensionality and sparseness in the final representation.

Semantic Features

The extraction of semantic features from unstructured text are probably one the most challenging problems in NLP. Semantic features refers to the meaning, sense, interpretation or coherence of the target textual elements. In order to achieve this, it is necessary to perform a deep analysis of the textual elements. Even though it is an open problem in NLP, there exist some interesting approaches for modelling the semantic information. Most of the approaches in the literature exploit principles of the distributional hypothesis. The distributional hypothesis, states that words with similar meanings occur in the same contexts (Sahlgren, 2008). In this regard, the extraction of semantic features is restricted to strategies for representing the contextual words that models the meaning of words. Thus, semantic modelling assumes that words close in meaning tend to occur in similar contexts, and therefore, uses occurrence and co-occurrence information to associate words and measure their contribution to the automatically generated concepts (topics) (Lavelli et al., 2004).

The Latent Semantic Analysis (LSA) is another well known technique to model the semantic information. In LSA, terms and documents are represented into the same feature space. This is usually named the latent space, where documents and terms are projected to produce a reduced topic-based representation (Deerwester et al., 1990; Dumais, 2004). For this, LSA is built from a term-document matrix $\mathbf{M}$ where m_{ij} represents the term frequency of the word i in document j. Thus, LSA uses the Singular Value Decomposition (SVD) to decompose $\mathbf{M}$ as follows:

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\mathsf{T}} \tag{2.1}$$

Where the $\mathbf{\Sigma}$ values are called the singular values and $\mathbf{U}$ and $\mathbf{V}$ are the left and right singular vectors respectively. $\mathbf{U}$ and $\mathbf{V}$ contain a reduced dimensional representation of words and documents respectively. $\mathbf{U}$ and $\mathbf{V}$ emphasize the strongest relationships and remove the noise (Landauer et al., 1998). In other words, it makes the best possible reconstruction of the $\mathbf{M}$ matrix with the less possible information (Landauer et al., 2013).

2.2 The Bag of Visual Words Representation

The BoVW is a well established technique that discretizes the image content producing visual units that can be considered words in text. The BoVW instance-vector exposes the association of such visual units (visual words) in each image, usually through a normalized histogram indicating the presence/absence of each visual unit. For example, in object recognition the presence of a

wheel (or part of it) can provide valuable information for recognizing a car. The BoVW strategy was proposed for the video retrieval task (Sivic and Zisserman., 2003), and due to its outstanding performance and simplicity, it quickly became popular and began to expand into other fields of computer vision (Cruz-Roa et al., 2011a,b; Díaz and Romero, 2012; Csurka et al., 2004; Tirilly et al., 2009; Zhang et al., 2007; Wang et al., 2009).

There are several ways to implement the BoVW, but the general framework is shown in Figure 2.1: i) a set of selected training visual regions are represented under some visual descriptor; ii) patch descriptors are clustered, and each cluster centroid is considered a visual word in a codebook; iii) using the precomputed codebook, the visual regions of each image are replaced by the *id* of the closest visual word in the codebook. The BoVW representation is then obtained by building histograms of the visual words occurrences in the images.

2.3 Ensemble Based Classifiers

In this part we explain one of the reference techniques we used to combine different kind of features in Chapters 4 and 5. The main idea behind ensemble based classifiers consists in building a collective prediction scheme based on multiple classifiers. The aim of the ensemble model is to achieve better performance than each of its individual classifiers (Rokach, 2009). For this reason one of the most important questions in ensemble learning is: *Do several classifiers can be integrated to build a better one?*. In this regard, the answer depends of the conditions of the problem and the strategies to model the feature space into the ensemble classifier.

2.3.1 Diversity in the Search Space

According to the literature in ensemble models, the idea of combining several classifiers is a key process that should consider several elements (Rokach, 2009), but one of the most important is the following:

- *Diversity among predictions:* This consist in having diversity predictions among the members of the ensemble system. Typically this is included as a component which is responsible for the generation of diversity among member classifiers.

The generation of diversity is one of the key elements for designing ensemble methods (Tumer and Ghosh, 1996; Krogh and Vedelsby, 1995; Kuncheva, 2004; Maimon and Rokach, 2002). It worth noting that, although diversity generation is a key component in most of the state-of-the-art methods, in pattern classification field there is no a definitive and accepted theory explaining *how* and *why* the diversity positively contributes to the performance of the final

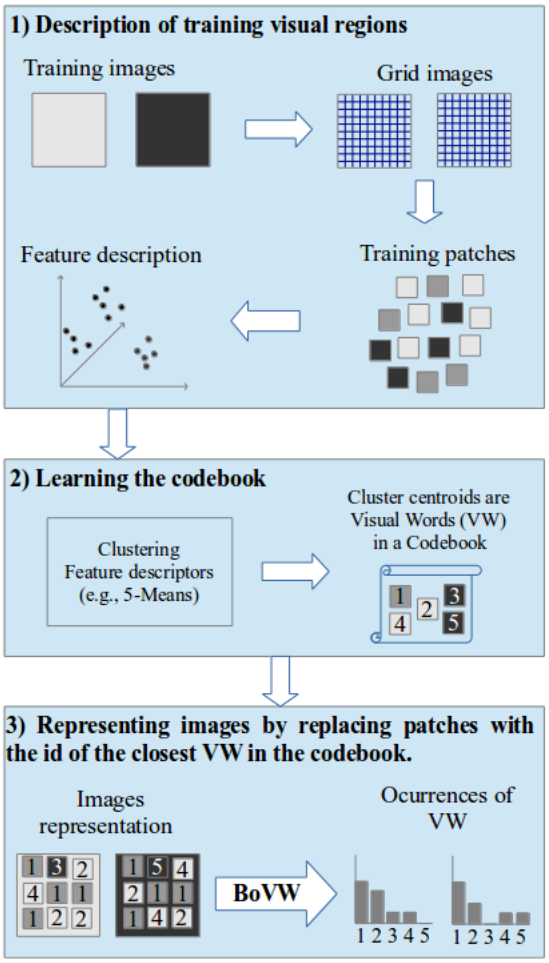

Figure 2.1: In Step 1 a set of image patches are represented under some visual descriptor. In Step 2 patch descriptors are clustered considering cluster centroids as visual words in a codebook. In Step 3 the visual regions of target images are replaced by the id of the closest visual word in the codebook. The BoVW of each image corresponds to its histogram of visual words.

ensemble(Brown et al., 2005). In this regard, (Rokach, 2009) summaries five of the most used strategies to generate the aforementioned diversity into ensemble systems:

- *Manipulating the training samples:* The idea is that each member of the ensemble is trained using different samples or projections of the dataset.
- *Manipulating the inducer:* This consist in manipulating the way the members are build. For example, different configurations of the hyper-parameters are used to build the classifier. Another alternative is to use different inducers to build the member classifiers, for example the use of different kind of classifiers (e.g., neural networks and decision trees) into the ensemble system.

- *Manipulating the representation of the target attribute (the class):* The idea is that each member classifier should focus in a different concept. Typically, the class attribute is replaced by a function such that the new target domain of the class attribute is smaller than the original.

- *Partitioning the search space:* The main idea is that each member classifier explores a different subset of the whole search space. For instance, the training of several member classifiers using all instances, but represented under different subsets of the feature space. In this way, each classifier have its own view of the dataset.

- *Hybridization:* The idea is to obtain diversity by combining any of the latter strategies. For example, using several kinds of classifiers and manipulating the search space.

2.3.2 Decision Making in Ensembles

There are also some other relevant elements for ensemble design. For example, the independence or dependence among members, which means that each member can set its own opinion based on different criteria but always using its private knowledge. On the one hand, dependent methods builds new members of the ensemble by using the performance information of the members in previous iterations (Provost and Kolluri, 1999; Freund and Schapire, 1996). The key idea is that new classifiers should focus in the misclassified instances (classification errors). Typically, this methods are guided by using an instance selection, also known as *Boosting*. On the other hand, in independent ensemble methods the training data is transformed into several subsets to train each member (Rokach, 2009). One of the advantages of this methods is that they can be easily parallelized using several base classifiers. This ensemble method is also known as bootstrap aggregating or simply *Bagging*. One important characteristic of *Bagging* ensembles is that usually it is possible to build a better classification system than each of its individual members. This is especially true for the algorithms that produce classifiers significantly different, if the dataset or its representation has been altered (Breiman, 1996).

In this thesis we are interested in generating the diversity through the manipulation of the feature space. This is, the instances are represented under different feature spaces (e.g., the contextual and semantic features). Thus we are interested in *Bagging approaches* where the subsets of features can be disjoint or not, but most of the time all individual predictions are fed into a collective decision method (e.g., a weighting vote). The combination of individual predictions is also an important step for an effective ensemble. According to the literature, the weighting voting scheme is one of the most used in text classification problems [2]. The weighting

[2]There are other important strategies to combine decisions. For example, the meta-learners (also known as *Staking*) where the idea is to focus in the decision pattern of the base classifiers, then train a meta-classifier over such

voting scheme is very useful when the members of the ensemble have similar performance, but difference confidence for the same task (Rokach, 2009). One example of this is the majority vote, also known as the *basic ensemble method* (BEM) that predicts the most voted class of the members. Typically BEM is used as a baseline in classification tasks. Equation 2.2 is the approaximation of (Rokach, 2009) of the latter ideas.

$$clase(x) = argmax_{c_i \in dom(y)} \left(\sum_k g(y_k(x), c_i) \right) \tag{2.2}$$

where $y_k(x)$ is the classification of the k-th classifier and $g(y, c)$ is an indicator function defined as:

$$g(y, c) = \begin{cases} 1 & y = c \\ 0 & y \neq c \end{cases} \tag{2.3}$$

predictions. Then, the meta-classifier perform the final decision (Wolpert, 1992).

CHAPTER 3

RELATED WORK ON CONTEXTUAL AND SEMANTIC VISUAL INFORMATION

In this chapter we present the most relevant related work on contextual and semantic visual information. The aim of this thesis is to propose novel and effective methods under the analogy visual-textual words. For this, it is hypothesized that some strategies in NLP exploiting contextual (spatial) and high level (semantic) information, might give rise to new methods and representations that improve the performance of approaches based on visual words. In this regard, in Section 3.1 and 3.2 of this chapter we present the most relevant computer vision works inspired by NLP, which consider the properties of the image domain in order to better exploit the contextual and high level information among visual words, especially those based on: n-grams (sequences of n elements), weighting schemes (weight functions for the visual elements), language models, and semantic distributional analysis. Finally, Section 4.1.3 explains some relevant concepts about information fusion that could help to better exploit the proposals in this thesis.

3.1 Contextual information under the analogy visual-textual words

The BoVW approach was introduced by Sivic and Zisserman for tackling the problem of video retrieval (Sivic and Zisserman., 2003). The outstanding performance and simplicity of BoVW quickly became popular in other computer vision tasks such as: image classification (Cruz-Roa et al., 2011a,b; Díaz and Romero, 2012), image retrieval (Csurka et al., 2004; Tirilly et al., 2009; Escalante et al., 2012), object recognition (Zhang et al., 2007), human-activity recognition (Wang et al., 2009), etc.

As already mentioned in the last chapter, a shortcoming of the BoVW like representations is the overlook of spatial relationships among words. In other words, histograms that account for visual word frequencies do not hold any spatial information about the occurrence of each visual word into the image. The spatial information, correctly exploited, has proven to be useful in several computer vision tasks(Tirilly et al., 2008). Given this scenario, a lot of work has been

devoted to capture such spatial relationships among visual words (Lazebnik et al., 2006; Zheng et al., 2008). For example, spatial relations have been represented using a graph of visual words in order to describe logos in sports photos (Jamieson et al., 2007). Other efforts have brought ideas from other areas such as NLP. For example in image retrieval, Zheng et al. (2006) 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 (2007, 2011), took advantage of the use of k-nearest neighbours 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 (Sivic and Zisserman, 2004). Quack et al. (2007) have explored local sets of visual words to detect frequent and distinctive 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 use LMs for computer vision tasks and perform several steps before training the LM (Wu et al., 2007), for example; the use of cocurrence and proximity information of neighbour visual words. The latter is because a LM needs to "read" the visual words in some direction. For example, Tirilly et al. (2008), used principal component analysis to project visual descriptors in a particular direction-axes, then induce a sequence of visual words (Tirilly et al., 2008). 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 most probable class. Another effective approach to capture spatial relationships among visual elements is the Spatial Pyramid Representation (SPR) proposed in (Lazebnik et al., 2006). The core idea of SPR relies in generating sub-windows of an image by using a sequence of increasingly coarser grids defined by a pyramid. For example, in a pyramid of three levels, there are three grids with sizes of 4x4, 2x2 and 1x1 cells. In this way, SPR computes a local BoVW in each cell of the image. The final representation of each image consists of an arrangement of its local histograms. Thus, the comparison of two images is done by using an intersection kernel computed between the representation vectors.

In all previous, works authors have proposed interesting extensions to the use of visual words, reporting improvements over the standard BoVW representation. However, these proposals do not necessarily correspond to the way in which sequences of words are processed in NLP tasks for boosting the performance (Wang and Manning, 2009). For instance, LMs are rarely used for text categorization, also it is know that LMC can have problems to handle imbalanced class problems (McCallum et al., 1998). Other previous works have proven that adding information

about occurrences of relative spatial information in visual words can enhance the performance, but at the cost of higher computational complexity, especially when relative distance and angles are considered (Tirilly et al., 2008; Khan et al., 2015).

3.2 Semantic information under the analogy visual-textual words

Regarding to the semantic information among visual words, probably the better known are the works using Latent Semantic Analysis (LSA) (Bosch et al., 2006) and Latent Dirichlet Allocation (LDA) (Bosch et al., 2006; Fei-Fei and Perona, 2005), where a new dimensional space has been computed by using the occurrence information of the visual words. Even though LSA and LDA are popular strategies to extract latent attributes in text classification tasks, they are different from distributional strategies for building the terms and instance representation in NLP (e.g., LSA builds a new feature space by means of singular value decomposition algorithms).

In this thesis, rather than define a set of latent attributes to represent images, we focus on studying the informative value of distributional features. More importantly, unlike other works using only standard representations like LSA, in this thesis besides adapting Distributional Term Representations (DTRs), we propose two novel approaches to better exploit distributional visual information. Our hypothesis is that by using distributional features in conjunction with visual words and visual n-grams, it is possible to obtain comparable results than other more elaborated strategies from the state of the art, such as LSA or LDA. DTRs are effective and efficient alternatives to model the semantic information. DTRs builds the representation of features by observing their occurrences and co-occurrences along the dataset, then in a second stage the instances are represented exploiting this distributional information (Lavelli et al., 2004). It is worth mentioning that most of the work related DTRs and visual elements has been devoted to the multi-modal distributional semantics for text classification, where authors intend to enrich the DTRs by using visual features (Feng and Lapata, 2010; Bruni et al., 2012, 2014). However, in this thesis we are interested in adapting and designing new methods based on DTRs to improve the image classification task. In this way, we are interested in observing the usefulness of such representations applied over pure visual features, in order to obtain new insights about the relation of techniques in the Language and Vision field. Finally, it is important to mention DTRs differ from other approaches, such as LSA and LDA, in the way of computing a low dimensional feature space. In DTRs the main idea is to exploit the distributional information in the datasets, by means of simple feature occurrences and co-occurrences. We consider that such way of computing distributional semantics features could be very useful to identify valuable patterns, that encompass complex visual elements that would be harder to model with other approaches.

3.3 Relevant Fusion Information Strategies

In this thesis we adopt several representations that has proven to be very helpful for text categorization. We focus in the idea of the visual word to demonstrate the usefulness and generality of NLP principles under the analogy of visual-textual words. In this work we are interested in at least two sets of visual features; structural and semantic. These visual features can be used to feed a wide range of different classification algorithms. Nonetheless, when instances can be represented under different sets of attributes, there exists several ways to take advantage of those different feature spaces (Rokach, 2009). In this context, we are interested in the following question: *How could features coming from Visual Words and Visual* n*-grams spaces be used together to enhance the performance of classification models?*. In this thesis we propose some interesting alternatives to deal with such question, especially for those handling contextual and semantic visual information. Thus, along this section we review two background topics in our proposals: i) early and late fusion information strategies, and ii) intermediate fusion strategies.

3.3.1 Early Fusion and Late Fusion Strategies

The extraction of several kinds of features (in our case contextual and semantic), raises new issues about the way to properly use them inside a classification system. In most previous works, authors have combined the extracted spatial-visual-features in order to improve the performance of their methods, then it is interesting to exploit this kind of information into the final representation. Nonetheless, the most used ways to combine heterogeneous attributes are simple fusion approaches; *early fusion* and *late fusion* (Bekkerman and Allan, 2004; Tan et al., 2002; Rokach, 2009). The main idea of *early fusion* is to concatenate the different feature spaces (e.g., words and n-grams) into single vectors, which are fed to a learning method (Rokach, 2009; Kuncheva, 2004). The Support Vector Machine (SVM) has shown to be effective using the *early* BoVW representation (Boiman et al., 2008; Cruz-Roa et al., 2011a; Díaz and Romero, 2012). On the other hand, *late fusion* strategies consider each feature space independently and build an ensemble learning system to combine the outputs of classifiers trained on different inputs (e.g., weighting vote ensemble classifier) (Breiman, 1996; Rokach, 2009). The underlying idea is to represent instances using vectors corresponding to each feature space in order to provide different perspectives/views of each instance. The problem of *early-late* fusion approaches is that they can be affected if the feature spaces are not diverse enough (Rokach, 2009; Kuncheva, 2004).

3.3.2 Intermediate Fusion Strategies

Multiple Kernel Learning (MKL) also know as *i*ntermediate fusion, is an attractive fusion scheme that has shown improvements over typical *early-late* fusion approaches (Gönen and Alpaydın, 2011), in part for performing the combination of information at a different level; at a *kernel* level. MKL methods build more accurate models using kernel functions that represent different similarity notions of the feature spaces (Gönen and Alpaydın, 2011). In the literature there are number of ways to perform the combination of kernel functions, for example; rule based operations (mean or product) over kernel matrices (Ben-Hur and Noble, 2005), alignment training techniques to weight the contribution of each kernel (Xu et al., 2010; Kloft et al., 2010), projected gradient updates (Rakotomamonjy et al., 2008), linear and conic analytical solutions for determining kernel weights (Cortes et al., 2010), etc. In this thesis, using different image collections, we evaluate the proposed visual π-grams representation and exploit different fusion alternatives based on Multiple Kernel Learning to combine features.

Part III

Contributions

CHAPTER 4

EXPLOITING THE CONTEXTUAL VISUAL INFORMATION

Image representation is one of the key procedures for building successful models in classification. In this regard, the BoVW is analogous to the well established Bag-of-Words (BoW) representation of text mining (see e.g., (Tan et al., 2002)). 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 computer vision 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. In this direction, contextual information among visual words could be captured exploiting the analogy visual-textual words using Natural Language Processing (NLP) approaches. To capture the spatial context, in this chapter we propose an effective feature inspired by one of the most used solutions in NLP for incorporating sequential information in documents representation: n-grams (sequences of n-words to capture compound word patterns). This type of representation can capture compound item-patterns; for example, in text mining; *united-states, very-good*, etc. In the case of visual imagery, we intend to capture frequent local coocurrences of visual elements. In this regard, we propose the jointly use of codebooks of visual words and visual n-grams (multidirectional sequences of visual words) to represent images under a bag of features formulation.

In other words, we propose the extension of the BoVW to the Bag of Visual n-grams (BoVN), which can be seen as representing images under different feature spaces (e.g., visual words and visual n-grams). These different feature spaces could be used together to enhance the performance of classification models, nonetheless this is not a trivial task (Kuncheva, 2004). In this work we propose to exploit two fusion strategies to combine information through Multiple Kernel Learning (MKL) methods. The first strategy consists in representing images under individual feature spaces (e.g., visual words and visual n-grams), then we use MKL strategies to exploit the information in the different spaces. The second strategy consists in representing images under the whole feature space (e.g., visual words and visual n-grams), but using different kernel functions to produce different notions of similarity that can be exploited by the proposed MKL strategy. MKL uses similarity kernel functions to delegate the construction of a new

combined kernel function to an algorithm (Gönen and Alpaydın, 2011; Alioscha-Pérez et al., 2012). Using the latter strategies to fuse information, we perform an extensive experimental work in order to establish a solid framework to exploit the proposed visual n-grams.

The main contributions of this chapter are twofold: i) to introduce the effective visual n-grams inspired in NLP, and ii) to propose MKL strategies to exploit the joint use of visual words and visual n-grams for Image Clasification (IC) tasks. This chapter is organized as follows. Section 4.1 introduces the proposed methodology to extract visual n-grams, whereas Section 4.1.3 presents the proposed strategies to take advantage of our visual n-grams for IC. Section 4.2 and 4.3 present the image collections and experimental settings, respectively. Section 4.4 reports the experimental results we obtained, and Section 4.5 outlines our conclusions and future avenues of inquiry.

4.1 Image Classification through Visual n-grams and MKL

In this section we describe the proposed Bag-of-Visual n-grams (BoVN) representation for image classification, as well as the proposed MKL fusion strategies. In Figure 4.1 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 a standard procedure outlined in Section 4.1.1. In the second step we extract visual n-grams to build a visual n-gram codebook (explained in Section 4.1.2). In the third and final step we merge the visual words codebook and the visual n-gram codebook in order to get a final codebook. We use our final codebook to build histograms, which are fed to the proposed strategies through MKL (Section 4.1.3). Each of these steps are described in the rest of this section.

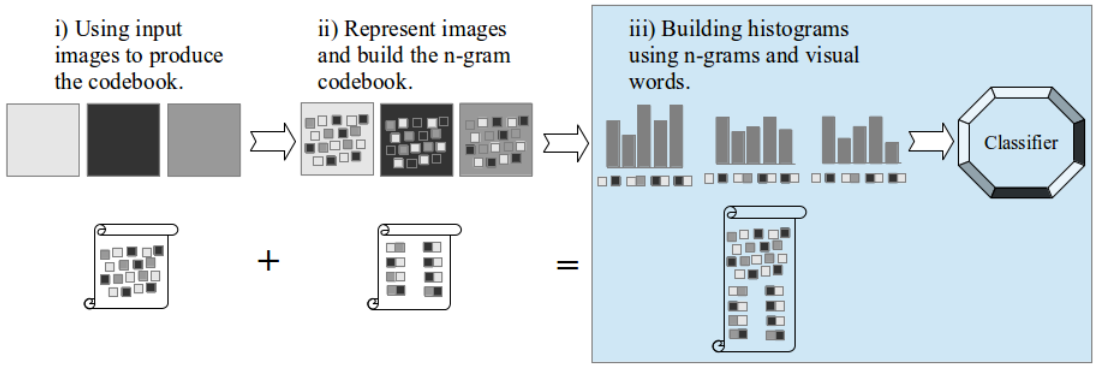

Figure 4.1: Image Representation through Bag-of-Visual-Ngrams.

4.1.1 Construction of the Visual Words Codebook

In this section we explain the first stage before building the Bag-of-Visual n-grams (BoVN) representation. In this context, we first need to build our Visual Words from the image collection. Such Visual Words will be the initial features used to generate the Visual n-grams. In Figure 4.2, 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 4.2. The next step consists in representing each extracted patch by a set of features (a visual descriptor[1]). The 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. In this process, all similar patch descriptors in the training set are grouped together independently of the source image. The k-means algorithm is used in this work to find a set of centroids which represent our visual words, which are labelled by an id and placed in the codebook.

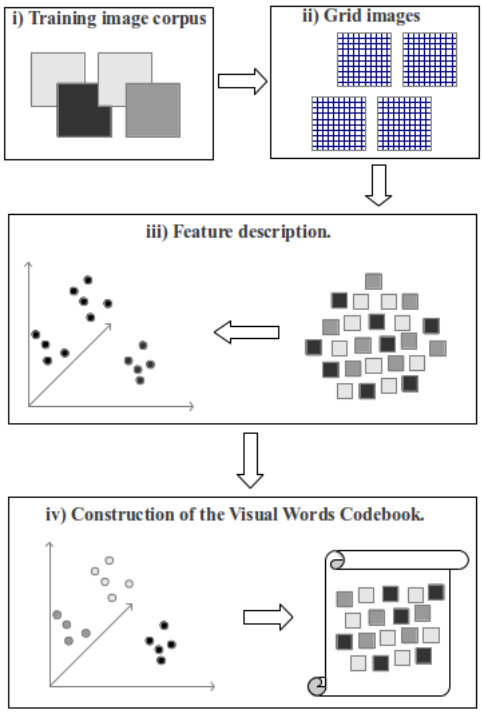

Figure 4.2: The process to build a visual word codebook.

To represent images using the codebook, each image is gridded and each image patch is replaced by its closest visual word in the codebook (see the detailed process in Figure 4.3). In

[1] In out case, we use DCT and SIFT descriptors since they are the reference point for the selected image collections.

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 constructed codebook in order to construct visual n-grams.

Figure 4.3: Example of a represented image using the Visual Word codebook. Left; Original image. Right; Visual Words representation.

4.1.2 Extracting visual n-grams

In this section we present the second stage to build our visual n-grams. For this we assume that there is a visual codebook which we will use to represent images. To capture spatial relationships among visual words, we inspired ourselves in the way word n-grams are used for text-classification. In NLP, n-grams are sequences of n consecutive words that help to maintain semantic relationships between words, which allows us to represent compound concepts like "bus stop" with a single attribute. In the image domain the extraction of visual n-grams have some additional issues. For example, a document can be read only in one direction, but sequences of image descriptors can be extracted horizontally, vertically, or diagonally). 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.

In order to construct visual n-grams we apply the following approach. First of all, we have each instance represented as the codeword matrix for each image (see Figure 4.4). Thus, let A be the $a \times b$ codebook matrix of a given image. The main idea is to produce n-grams ignoring

44	219	389	182	33	153	141	119
222	213	65	78	134	211	191	233
320	21	113	123	21	297	326	321
43	16	234	71	91	38	90	42
129	345	222	400	341	349	256	54
120	15	112	23	212	219	152	35
354	123	234	2	54	125	212	66
27	198	19	11	45	345	56	69

Figure 4.4: The process to build a Visual n-grams using a sliding window. Note that in the above image, each number represents the ID (codeword) of the closest visual word in the codebook. For the darkest item (65) the extracted n-grams are: 65-389, 65-219, 65-213, 65-21, 65-113, 65-123, 65-78, 65-182.

the orientation in which they appear. To construct n-grams we iterate over each element $a_{i,j}$ of the matrix A and we extract the neighbours in a straight fashion. 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 from the element $a_{i,j}$ only in horizontal, vertical and diagonal directions at angles of 45, 135, 225 and 315 degrees. In Figure 4.4 we illustrate the process to extract visual bigrams using an sliding window on each visual word to build its neighbours. The same process is applied to obtain $n > 2$ sequences, always producing straight fashion n-grams in horizontal, vertical and diagonal directions. The latter condition leaves us with eight possible n-grams for each position in the matrix. Finally, each n-gram is normalized to be read just in one way and consequently indexed as the same item in our new visual n-gram codebook. For example, in the visual n-gram codebook, a trigram 21-61-73 is indexed as the same item than 73-61-21. We use these normalized visual n-gram codebook to proceed with the image representation. For this, each image is represented using visual words and visual n-grams through histograms of the occurrence of visual n-grams found in the image.

4.1.3 Exploiting the jointly use of Visual words and Visual n-grams

According to previous sections, at this point there are at least two sets of visual features; visual words and the visual n-grams. These visual features can already be used to fed a wide range of different classification algorithms. Nonetheless, when instances can be represented under different sets of attributes, there exists several ways to take advantage of those different feature

spaces (Rokach, 2009). In this context, we are interested in the following question: *How can features coming from Visual Words and Visual n-grams spaces be used together to enhance the performance of classification models?*.

The appropriate use of several feature spaces to improve the discriminative power of a system is not a trivial task (Kuncheva, 2004). Two of the most popular strategies for combining information from different sources are *early-fusion* and *late-fusion* (Kuncheva, 2004; Breiman, 1996; Brown et al., 2005; Rokach, 2009). The former consists of merged attributes from two spaces into single space, then use standard supervised learning methods to build classification models. On the other hand, *late fusion* strategies use classifier ensembles to train individual models on each feature space, then perform a joint prediction using a voting decision or trained combiner. More specifically, the combination of features in early fusion consists in extending the space of visual words VW using the visual n-grams VN space to produce a new space with $|VW| + |VN|$ dimensions. The intuitive idea is that the learning algorithm (in our case SVM, which have shown high performance in similar situations) will be able to learn the important properties of the target problem in such space. On the other hand, in late fusion we build a SVM for each space VW and VN. For this purpose, we implemented each classifier to make a prediction using a vector of the probabilities of belonging to each class. For the final decision, we aggregate such vectors to determine the label as the i element with the maximum value. Nonetheless, as shown in experimental results of Section 4.4.2, sometimes several kinds of textual features could not be diverse enough to build accurately such ensemble models. For example, consider the following example keeping in mind our analogy visual-textual word and visual-textual n-gram. In text mining tasks, a wide variety of different kinds of textual features (e.g., words, word n-grams, frequent maximal sequences, collocations, etc.) are extracted just from one rigorous modality: the text (sequences of tokens). Thus, the space features could not be totally independent from others especially when one space was used as the base of a new one (e.g., n-grams are built from words). Thus, in some classification tasks, ensemble *early/late fusion* methods could not receive truly multi-modal features, which degrades the diversity in the feature space and makes difficult to build an accurate ensemble system (Kuncheva, 2004; Breiman, 1996; Brown et al., 2005; Rokach, 2009; Chávez et al., 2011). For this reason, we propose to use two alternative strategies to integrate visual n-grams through a more appropriate state-of-the-art scheme fusion. The *intermediate fusion* makes use of Multiple Kernel Learning (MKL) techniques to delegate the construction of a new combined kernel function to an algorithm (Gönen and Alpaydın, 2011). Using the latter strategy to fuse information, we establish a solid framework for the use of the proposed visual n-grams making it a good alternative to other approaches reported in the literature.

MKL Strategies to exploit Visual n-gram spaces

According to the literature, the most common/effective classifier under bag of features formulation is the Support Vector Machine (SVM) (Bekkerman and Allan, 2004; Sivic and Zisserman., 2003; Csurka et al., 2004; Boiman et al., 2008; Gönen and Alpaydın, 2011). SVM is a learning algorithm that aims to find an optimal separating hyperplane between instances belonging to two different classes (Gönen and Alpaydın, 2011). Let $\{\mathbf{x}_i, y_i\}$ be the training instance-class pairs examples, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y \in \{-1, +1\}$, with d dimensionality of the problem (say the size of the vocabulary). SVMs aim to determine a mapping from training examples to classes using the following linear function:

$$f(\mathbf{x}) = sign(\sum_i \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) - b) \tag{4.1}$$

where α_i and y_i are the weight and label of training example i. To map the $(\mathbf{x}_i, \mathbf{x}_j)$ input vectors into the feature space, the $k(\mathbf{x}_i, \mathbf{x})$ kernel function is applied. Intuitively, $k(\mathbf{x}_i, \mathbf{x})$ measures the similarity between instances x_i and x_j. [2]. Selecting the kernel function is an important issue in the training.

Multiple Kernel Learning (MKL) methods are popular solutions to face the problem of combining different feature spaces. The core idea relies in kernel functions; instead of choosing a single kernel function for a specific problem, it is better to have a set and let an algorithm to learn the best combination of them (Gönen and Alpaydın, 2011). To better explain this idea consider the following expression which represent combined kernels:

$$k_\eta(\mathbf{x}_i, \mathbf{x}_j) = f_\eta(\{k_m(\mathbf{x}_i^m, \mathbf{x}_j^m)\}_{m=1}^P | \eta) \tag{4.2}$$

the core is the combination function, $f_\eta : \mathbb{R}^P \rightarrow \mathbb{R}$, may be linear or not, the kernel functions, $\{k_m : \mathbb{R}^{D_m} \times \mathbb{R}^{D_m} \rightarrow \mathbb{R}\}_{m=1}^P$, take P feature representations of data instances: $\mathbf{x}_i = \{\mathbf{x}_i^m\}_{m=1}^P$ where $\mathbf{x}_j^m \in \mathbb{R}^{D_m}$, and D_m is the dimensionality of the m feature space. η parametrizes the combination function and usually are fixed parameters without any optimization during training.

In order to put the MKL framework in context of our extracted visual features, let $V_k = \{w_1, ..., w_d\}$ denote the d extracted features in space k (e.g., the codebook of visual words), $\Psi = \{V_1, ..., V_P\}$ the set of m considered feature spaces in the whole collection (e.g., codebooks for visual words and visual n-grams). In this thesis, we propose to fed the f_η function in Equation 4.2, using input instances represented through the following two strategies:

(a) **Strategy 1: Single Kernel - Several Spaces** : A fixed kernel function uses inputs coming from P different space representations (e.g., visual words and visual n-grams). The

[2] The parameters α and b, are learned using optimization techniques (Gönen and Alpaydın, 2011)

intuitive idea is to represent instances using vectors corresponding to each individual feature space in order to provide different perspectives/views of each image, then use MKL to learn a general perspective. Thus, we end up with $P = |\Psi|$ representations for each data instance: $\mathbf{x}_i = \{\mathbf{x}_i^m\}_{m=1}^P$ where $\mathbf{x}_j^m \in \mathbb{R}^{D_m}$, and D_m is the dimensionality of the m feature space. Having instances represented under P feature spaces is useful to generate diversity in the search space using the fixed kernel function. This allows to have P different representations that can be used by MKL to build the $k\eta$ general kernel.

(b) **Strategy 2: Several Kernels - Single Space**: A set of s different kernels functions correspond to different notions of similarity. The whole feature space (visual words and visual n-grams) are used to represent each instance using $P = s$ vector representations representations. The intuitive idea is that, instead of trying to find which is the best kernel function, a learning method do the picking or combination. In this strategy we represent data instances using vectors of $D = |\bigcup_{V_j \in \Psi} V_j|$ features. Thus, we end up with a vector for each data instance, but using the s kernel functions we produce the $P = s$ representations for each data instance: $\mathbf{x}_i = \{\mathbf{x}_i^m\}_{m=1}^P$ where $\mathbf{x}_j^m \in \mathbb{R}^D$, and D is the dimensionality of the whole feature space. We use those P representations in the $f\eta$ function and build the final kernel $k\eta$.

In order to solve the $f\eta$ for building the final kernel $k\eta$, we use the MKL methods outlined in Table 4.1. In this way, we analyze the performance of several MKL algorithms to produce a new kernel method that accurately describe each image collection exploiting our precalculated visual n-grams.

4.2 Image Collections

In order to perform the evaluation and demonstrate the generality of visual n-grams we used five different image collections that expose special particularities, which could be captured by visual sequences. Figure 4.5 and Table 4.2 briefly describes each collection. For example, the Histopathology image collection (Díaz and Romero, 2012; Cruz-Roa et al., 2011b) is class-imbalanced and contain complex visual patterns in tissues structures (healthy or pathological); the classification is related to pathological lesions and morphological-architectural features which can be captured by visual n-grams. Other collections like Birds, Butterflies and Scenes also have features of texture and structure not only in the target object (i.e., the bird), but also in the other surrounding visual elements like the grass, sky, water; which could play a role to determine or not the class label.

MKL algorithm	Description
1. SimpleMKL (Rakotomamonjy et al., 2008)	Iterative MKL algorithm that uses projected gradient updates and trains SVMs at each iteration to solve the optimization problem.
2. RBMKL (Ben-Hur and Noble, 2005)	Rule based MKL trains and SVM with the (mean or product) of the combined kernels.
3. NLMKL (Cortes et al., 2009)	A nonlinear MKL algorithm using an SVM as the base learner and a quadratic kernel.
4. LMKL (Gönen and Alpaydin, 2008)	Localized MKL algorithm using the softmax gating model using the concatenations of all feature representations in the gating model
5. GMKL (Varma and Babu, 2009)	The generalized MKL algorithm learn a kernel function instead of kernel matrix defining a kernel funtion in the space of kernels called *hyperkernel*, this use a convex combination of base kernels.
6. GLMKL (Xu et al., 2010; Kloft et al., 2010)	The group Lasso-based MKL algorithms, updates the kernel weights to learn a conic combination of the kernels.
7. CABMKL (Cortes et al., 2010)	Centered-alignment-based MKL algorithm. In a first step uses a linear analytical solution for determining the kernel weights. In a second step train a SVM with the kernel calculated with these weights.
8. ABMKL (Qiu and Lane, 2009)	Alignment-based MKL algorithms determine kernel weights using an heuristic, then train an SVM with the kernel calculated with these weights.

Table 4.1: Representative MKL algorithms (Gönen and Alpaydın, 2011).

Dataset	Classes	Distribution
1. Histopathology (Cruz-Roa et al., 2011a)	7	carcinoma (518), collgen (1238), epidermis (147), hair folicle (118), eccrine glands (126), sebaceous glands (136), inflammatory infiltrate (99).
2. Birds (Lazebnik et al., 2005)	6	egret (100), mandarin (100), owl (100), puffin (100), toucan (100), wood duck (100)
3. Butterflies (Lazebnik et al., 2004)	7	admiral (111), black-swallowtail (42), machaon (83), monarch-closed (74), monarch-open (84), peacock (134), zebra (91)
4. Scenes (Lazebnik et al., 2006)	15	bedroom (216), suburb (241), industrial (311), kitchen (210), livingroom (289), coast (360), forest (328), highway (260), insidecity (308), mountain (374), opencountry (410), street (292), tallbuilding (356), office (215), store (315)
5. 6-Caltech (Fei-Fei et al., 2007)	6	anchor (42), ant (42), camera (50), chair (62), crocodile (50), dollar-bill (52)

Table 4.2: Image collections used for the evaluation of visual n-grams. Histopathology collection is the only one with multi-label, which was approached as seven binary problems of the 1417 Histopathology image collection. The positive instances are images belonging to a target category. Basal cell carcinoma is the only one related with cancer diagnosis.

4.3 Experimental settings

We have performed several experiments for each dataset. In those experiments, we gridded images in patches of 8 pixels [3]. Among the wide variety of image descriptors in the literature, we use the Scale Invariant Feature Transform (SIFT) (Lowe, 2004) descriptor extracting edge points at two scales and eight orientations. We also 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. This strategy produces Visual Words that

[3] We experimentally test patches of size 8x8 and 16x16. The 8x8 size patch is an appropriated option, which have been also confirmed by other authors in the histopathology dataset (Díaz and Romero, 2012)

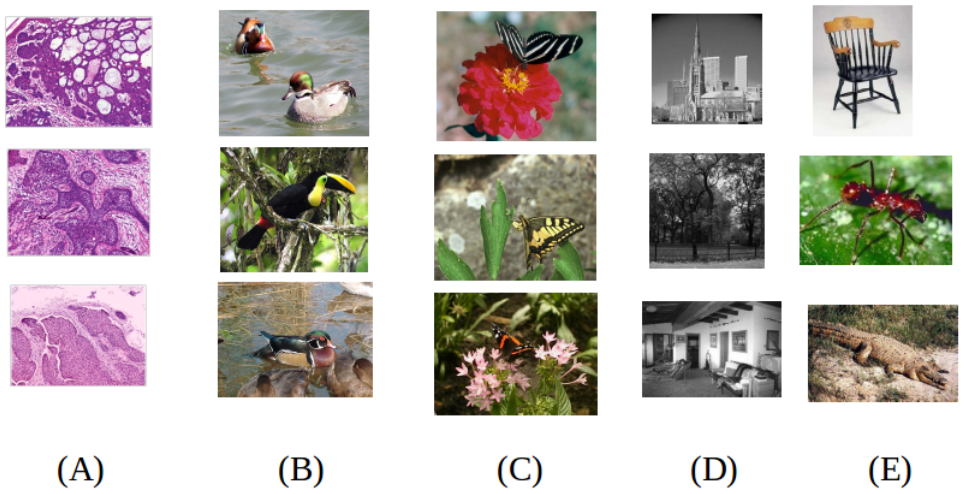

Figure 4.5: Image samples of the image collections (A) Histopathology, (B) Birds, (C) Butterflies, (D) Scenes y (E) 6-Caltech.

takes into account color and texture. We considered these features because in previous studies they have shown outstanding (e.g., DCT best descriptor found in (Cruz-Roa et al., 2011a,b; Díaz and Romero, 2012) for histopathology dataset) or at least competitive performance than other more complicated alternatives (Lowe, 2004). However, other types of feature-descriptors could be considered as well. 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 tasks (Bekkerman and Allan, 2004; Tan et al., 2002; Wang and Manning, 2009) (we also performed experiments with separated representations but we obtained worse results). Furthermore, we have 400 unigrams[4] and different number of n-grams for each different value of n (from 1 to 3). The latter means that, in an experiment of 3-grams ($1+2+3grams$) we have combined 400 unigrams plus x-top-frequent 2-grams and the x-top-frequent 3-grams features for our BoVN. Even though there is a number of ways to select generated n-grams (e.g., information gain), we are interested in observing if the simple top-frequent features can improve the performance, this is also a common practice in several text mining tasks (Bekkerman and Allan, 2004; Wang and Manning, 2009). Moreover, it is worth mentioning that we have normalized each space of attributes in an individual way (we represent each space as a probability distribution). In simple words, the total sum of values corresponding to each feature space (visual words and visual

[4]We chose 400 visual words as a fixed k value for all databases because of two reasons: for the histopathology dataset is a good configuration (Cruz-Roa et al., 2011a), and for the other datasets, the number of visual words was between 100 and 400, nonetheless 400 features did not present a severe impact in the performance (Lazebnik et al., 2005, 2006).

bigrams) is equal to one. In the evaluation we used stratified 10 fold cross validation (10FCV) for each dataset. For the histopathology dataset we report the average of the F-Measure obtained on each binary problem. For the rest of image collections we report the micro F-Measure (FM), which weights the F-measure performance in each class according to the number of instances in the class. The FM reflects the performance considering the precision and recall [5]. Equation 4.3 defines the F-Measure in terms of the precision and recall.

$$FMeasure = 2 \times \frac{precision \times recall}{precision + recall} \tag{4.3}$$

For the proposed MKL strategies we report the average time required to perform the learning and classification steps. In our experiments we used a computer with a CPU Intel-Corei7 3.6GHz and 64GB of RAM. Along Section 4.4 we will explain more details about each experiment such as: the way we measured the performance and other specific conditions for each experiment.

4.3.1 Statistical significance of results

In this research, we are interested in observing if two approaches are statistically significantly different; these two approaches will be the proposed method (BoVN) and each of the proposed baselines (e.g., BoVW). For this reason we used the Wilcoxon signed-ranks (Wsr) (Wilcoxon, 1945) test for determining the statistical significance of differences in results. Wsr is the test recommended by J. Demšar for comparisons between two algorithms (Demšar, 2006). The Wsr is a non-parametric test, that makes no assumption that the differences between two random variables compared are distributed normally.

4.4 Experiments and Results

In this section, we explain the purpose and details of each experiment. We have chosen the most relevant experiments to show the different properties of the use of visual-n-grams for IC. The best result of each set of experiments has been set in bold.

4.4.1 Bag-of-Visual-Words versus Bag-of-Visual-Ngrams

The goal of this experiment is to show how the proposed visual n-grams could improve the classification performance. For this, we present experimental results comparing the performances of a traditional Bag of Visual Words (BoVW) and our proposed Bag-of-Visual n-grams (BoVN). In this first experiment, for the proposed BoVN, we extended the set of visual features by adding

[5] Precision is the fraction of instances that are relevant, this is the number of true positives over the number of true positives plus the number of false positives. Recall is the fraction of relevant instances that are classified, this is the number of true positives over the number of true positives plus the number of false negatives (Powers, 2011).

a set of 2-grams (multidirectional sequences of two visual words). Since the number of possible 2-grams are of hundreds of thousands we have fixed it to the top-frequent 2500 [6]. For this experiment we represented images under the BoVN through the *early fusion* scheme (called early BoVN). Experiments reported in Table 4.3 show the performance for *early* BoVN, and the traditional BoVW using SVM with a linear kernel (Chang and Lin, 2011).

Results BoVW vs BoVN						
		Averaged F-Measure per collection				
descriptor	model	Histopathology	Birds	Butterflies	Scenes	6-Caltech
DCT	early BoVN	**64.54**	47.91	**62.19**	**63.40**	**54.10**
DCT	BoVW	58.54	**52.90**	61.10	61.01	53.48
SIFT	early BoVN	**61.71**	**54.79**	52.31	**77.19**	**72.29**
SIFT	BoVW	53.41	53.12	**55.82**	74.10	70.51

Table 4.3: F-Measure results for visual words vs visual n-grams. For image preprocessing in these collections, settings from Section 4.3 where used.

The experimental results presented in Table 4.3 suggest that, independently of using the DCT or SIFT descriptor, the use of visual 2-grams outperforms the average classification performance of 1-grams in every image collection. The averaged better F-Measure obtained by the early BoVN, against the simple BoVW (which is the traditional BoVW using 1-grams), is in part due to the pairs of visual words representing structural visual patterns, which in some way reinforce some evidence in text mining (Bekkerman and Allan, 2004; Wang and Manning, 2009). It is worth noting that, using the DCT descriptor for the histopathology dataset produces better classification performance than SIFT descriptor. This is because the DCT descriptor considers important properties of texture and color, which are relevant for histopathology images (Cruz-Roa et al., 2011a). Furthermore, the histopathology images are captured in a more controlled environment, which makes possible to have images in the same scale and resolution. The DCT descriptor also obtained better results than SIFT for the butterflies dataset. This is also due to the color and texture properties of the images. Moreover, most images in the butterflies dataset have the object in similar positions, which alleviates problems related with rotation. On the other hand, for Birds, Scenes and 6-Caltech datasets, using SIFT descriptor leads to a better classification performance than DCT. This is mainly because natural images have some properties (e.g., different scales, resolutions, orientations) that SIFT descriptor can handle in a

[6] We analyzed how the dimensionality influences the performance of a Bag-of-Visual 2-grams (testing incrementally from one thousand to ten thousand of features), getting that the 2500 top frequent bigrams are a good balance (sightly better than experiments using less and more features) between the dimensionality and the performance of our approach.

more appropriated way (Lowe, 2004).

It is worth mentioning that under the same visual descriptor, computing the Wsr test over the outputs of the 10CFV in each dataset, we obtained more than 98% of statistical confidence in results comparing early BoVN and BoVW. In the following sections, we present more detailed experimental results. Given the evidence of the performance using DCT and SIFT descriptors in Table 4.3, for the **remainder of experiments** we used DCT descriptors for experiments in the Histopathology and Butterflies datasets, but SIFT descriptors for the rest of collections.

Longer Sequences of Visual Words

The purpose of these experiments is to expose whether considering n-grams of higher order than 2, could improve the performance of the classifier. Table 4.4 presents the results of the experiments of the BoVN approach for visual n-grams using unigrams (which are the traditional visual words and one of our baselines) to tetragrams [7]. From results in Table 4.4 we can figure out that the 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 (Tan et al., 2002). The second one is related with the high dimensionality: using longer sequences produces large vocabularies, which also produce sparse feature vectors (long sequences are more difficult to find (Wang and Manning, 2009)). According to the Wsr test, only the difference between using 1+2grams and 1+2+3grams is not statistical significantly. Nonetheless, using 1+2grams seems to be a better option given the compromise between the effectiveness and the required computational resources. For this reason, in our following experiments we used 1+2grams as visual features.

BoVN					
Averaged F-Measure per collection					
Config	Histopathology	Birds	Butterflies	Scenes	6-Caltech
1gram	58.54	53.12	61.10	74.10	70.51
$1+2$gram	**64.54**	**54.79**	**62.19**	**77.19**	**72.29**
$1+2+3$gram	62.69	53.31	62.09	76.12	71.11
$1+2+3+4$gram	61.34	51.11	61.05	74.32	69.31

Table 4.4: F-Measure results for visual words vs visual n-grams. For image preprocessing in these collections, settings from Section 4.3 where used.

[7] We selected the 2500 top frequent features for each n-gram space in the same way that in Section 4.4.1. Thus the experiment 1+2+3+4gram uses the information of the 400 visual words (1grams) and 7500 sequences of visual words (n-grams)

4.4.2 Strategies to Exploit Visual n-grams

Visual words and visual n-grams can be seen as two different sets of visual features. These visual features can be already used together (e.g., *early or late fusions*) to fed a wide range of different classification algorithms. Nevertheless, as explained in Section 4.1.3 it is possible to exploit these feature spaces using more appropriated fusion methods. The purpose of experiments in this section is to show that the MKL strategies can improve the classification performance taking advantage of the joint use of Visual Words and Visual n-grams. For this we analyze the proposed visual n-grams under a MKL formulation, solving the kernel combination using a wide variety of MKL methods outlined in Table 4.1. We present the general obtained results by Strategy 1 and 2 under the following specific kernel combinations [8]:

1. Linear $k_{LIN}(\mathbf{x}_i, \mathbf{x}_j) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$
2. Intersection $k_{INT}(\mathbf{x}_i, \mathbf{x}_j) = \sum_{h}^{d} min(x_{i,h}, x_{j,h})$
3. The fusion of k_{LIN} and k_{INT} under MKL schemes.

In Tables 4.5 and 4.6 we report experimental results per collection. We also report the time required by each method to perform the 10CFV over all datasets. The time required to build the kernel matrix is what varies from one MKL strategy to another. Once the matrix kernel is learned/built, it is fed into an standard SVM. For these results, the proposed BoVN using *early* or *late* fusion strategies outperforms the traditional BoVW in each dataset. For the sake of comparison, in Table 4.5 we also evaluate other approach in the literature that also takes advantage of contextual information; a Language Model Classifier (LMC). As explained in Chapter 3, language models have been used in previous works (Tirilly et al., 2008; Wu et al., 2007) for building classifiers. Thus, we have implemented a Language Model Classifier (LMC) as the one used in (Tirilly et al., 2008), which is based on the *CMU-Cambridge Statistical Language Modeling Toolkit v2* (Clarkson and Rosenfeld, 1997). 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). [9] From Table 4.5 it can be seen that MKL Strategy 1 (specially RBMKL) is a better option than BoVN

[8] We study other basic kernel functions, as polinomial and gaussian, and their combination. Nevertheless, obtained results are much lower than the performance of linear and intersection kernel. This is due in part to that linear and intersection kernels, are more appropriated for working with data represented using histograms.

[9] The language model classifier works as follows: i) For each binary problem, it takes the training documents and builds two model languages (one for positive class and one for the negative), and ii) 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.

Strategy 1: Single Kernel (Linear | Intersection) - Several Spaces						
	F-Measure performance by collection					
	dataset					
	Histopathology	Birds	Butterflies	Scenes	6-Caltech	hours
	k_{LIN}|k_{INT}	k_{LIN}|k_{INT}	k_{LIN}|k_{INT}	k_{LIN}|k_{INT}	k_{LIN}|k_{INT}	
SimpleMKL	65.12 | 66.11	55.67 | 55.18	62.82 | 62.14	77.22 | 77.62	72.16 | 71.07	7.91
RBMKL	**66.53** | 66.43	55.01 | **55.98**	62.53 | **63.48**	76.12 | **77.79**	73.22 | **73.82**	3.71
NLMKL	62.41 | 61.11	54.33 | 55.43	62.31 | 62.24	75.11 | 75.53	71.87 | 70.52	12.81
LMKL	61.17 | 60.54	54.74 | 52.23	60.31 | 62.29	76.24 | 77.71	72.91 | 71.21	9.21
GMKL	65.20 | 67.42	54.54 | 55.33	61.28 | 58.22	73.32 | 74.03	72.34 | 70.93	6.56
GLMKL	65.69 | 65.21	54.73 | 54.37	61.44 | 60.51	77.01 | 76.44	73.12 | 72.17	6.58
CABMKL	60.55 | 61.31	53.53 | 53.02	62.28 | 62.89	76.98 | 75.38	72.12 | 72.44	5.97
ABMKL	60.91 | 59.88	53.50 | 54.73	62.55 | 62.15	73.91 | 72.21	69.13 | 67.12	4.21
early BoVN	64.31 | 63.71	54.79 | 54.51	62.19 | 61.13	77.19 | 76.62	72.29 | 72.94	2.21
late BoVN	60.31 | 61.31	53.05 | 54.34	61.51 | 61.28	76.04 | 75.25	71.34 | 72.01	2.36
simple BoVW	58.59 | 57.21	53.12 | 54.10	61.10 | 62.66	74.10 | 73.11	70.51 | 69.38	1.16
LMC	53.00	54.76	61.14	62.12	68.41	5.24

Table 4.5: Strategy 1: Single Kernel (Linear | Intersection) - Several Spaces. This Table shows experiments using sequences of Visual Words (Uni-Bi-grams) early and late fusion. For these experiments we compute the F-Measure for the positive class in each category on 10-fold cross validation using unigrams and bigrams in each of the problems. Simple BoVW is the only experiment using just the 400 visual words.

Strategy 2: Several Kernels (Linear + Intersection) - Single Space						
	F-Measure performance by collection dataset					
model	Histopathology	Birds	Butterflies	Scenes	6-Caltech	hours
SimpleMKL	67.12	56.12	63.82	78.01	72.34	17.54
RBMKL	**68.53**	**56.41**	**64.00**	**78.22**	**74.12**	7.23
NLMKL	62.41	55.31	63.89	78.10	73.21	28.71
LMKL	61.17	53.71	63.83	77.21	73.84	23.34
GMKL	66.20	55.76	64.15	77.13	72.41	14.21
GLMKL	66.69	53.30	63.05	78.02	73.92	14.56
CABMKL	60.55	54.33	63.29	76.72	73.01	11.10
ABMKL	60.91	54.96	63.25	77.34	72.74	9.92
	k_{LIN}|k_{INT}	k_{LIN}|k_{INT}	k_{LIN}|k_{INT}	k_{LIN}|k_{INT}	k_{LIN}|k_{INT}	k_{LIN}|k_{INT}
early BoVN	64.31 | 63.71	54.79 | 54.51	62.19 | 61.13	77.19 | 76.62	72.29 | 72.94	6.74
late BoVN	60.31 | 61.31	53.05 | 54.34	61.51 | 61.28	76.04 | 75.25	71.34 | 72.01	6.98
simple BoVW	58.59 | 57.21	53.12 | 54.10	61.10 | 62.66	74.10 | 73.11	70.51 | 69.38	2.24

Table 4.6: Strategy 2: Several Kernels (Linear + Intersection) - Single Space. This Table shows experiments using sequences of Visual Words (Uni-Bi-grams) early and late fusion. For these experiments we compute the F-Measure for the positive class in each category on 10-fold cross validation using unigrams and bigrams in each of the problems. Simple BoVW is the only experiment using just the 400 visual words.

and LMC in most datasets, except for Caltech dataset where an *early* BoVN seems to obtain competitive results. Nevertheless, from the considered strategies and kernels to evaluate MKL using visual n-grams, the most competitive seems to be results in Table 4.6, which corresponds to Strategy 2: "Several (Linear + Intersection) Kernels - Single Space)". In this strategy, linear and intersection kernels correspond to different notions of similarity of the whole space (visual words plus 2grams). Then, instead of trying to find which kernel is the best, MKL method performs combination. Experimental results show improvements when using several MKL strategies, but the bests results in Table 4.6 were obtained by RBMKL. We individually validate RBMKL-Strategy-2 using the Wilcoxon Signed Rank test against: simple BoVW, early-BoVN, late-BoVN and RBMKL-Strategy-1. The output obtained by this test is above of 98% of statistical confidence. The best outcomes of the RBMKL method confirms what is reported in (Gönen and Alpaydın, 2011), which in similar domains obtains competitive performance or overcomes other approaches. The RBMKL method combines kernels performing a linear operation using the similarities matrices of each kernel, in our case the mean of the linear and intersection kernel matrices. RBMKL methods, through a simple-effective kernel operation can derive a kernel that better reflects similarities among instances represented under histograms of visual words and visual bigrams. This suggest that without using elaborated kernel learning techniques (note the time required by RBMKL), and under the proposed visual n-grams space, it is possible to compute useful similarity kernel matrices. We hypothesize that the main reason of this result is that instances are under a space of high dimensionality (400 visual words + 2500 2-grams), which allows to obtain more appropriate similarity measures. The LMC does not provide better performance than BoVN. This can be due in part to the highly unbalanced data in some collections, 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.

Specific Detailed Results by Class for MKL

In this section, we present some of the most relevant results by class obtained by MKL strategies. The purpose is to expose the performance of visual n-grams in some specific interesting classes.

Histopathology dataset: Results in Table 4.7 show that most methods using 1+2grams overcome 1grams methods in most classes. This is more visible 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 (see Figure 4.6). Visual words (1-grams) are competitive 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).

Experiments for Histopathology dataset using Strategy 2								
	Detailed F-Measure by class							
	class							
model	1	2	3	4	5	6	7	avg
RBMKL	**96.46**	**99.08**	**84.68**	**55.28**	**52.41**	**56.30**	**35.52**	**68.53**
BoVW	86.10	94.80	74.40	36.80	35.80	48.00	34.20	58.59

Table 4.7: Strategy 2: Several (Linear + Intersection) Kernels - Single Space. Table shows detailed experiments in the Histopathology dataset using sequences of Visual Words (Uni-Bi-grams) under RBMKL and the traditional BoVW. The class 1 is the only one related with cancer diagnosis.

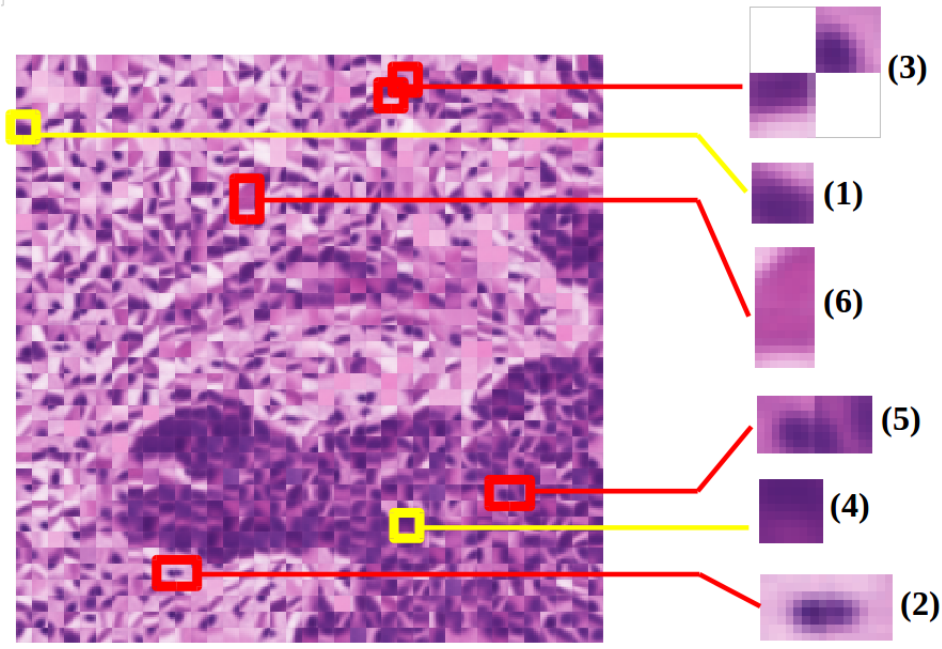

Figure 4.6: Example of an image related with cancer diagnosis (class 1). The image is represented under the computed visual words codebook using DCT descriptor. According to information gain implemented in (Hall et al., 2009), we rank the 6 most discriminative visual features found in this image. We highlight in yellow and red, the most discriminative visual words and visual n-grams respectively.

Birds and Butterflies dataset: Table 4.8 presents experimental results using the Birds dataset. Methods using 1+2grams outperform simple BoVW (1grams) in some specific classes. This is more visible in classes Mandarin, Puffin, Toucan, and Wood duck. In a similar way that the Histopathology dataset, we think those results are in part due to the complexity of each class image. Figure 4.7 shows one instance of the Egret class (left) and one of the mandarin class (right). Results suggest that, simple BoVW, in some way can solve the Egret class because the target object contains low variety of visual words and there are less structural local visual patterns (captured by the n-grams). On the other hand, the image belonging to the mandarin class, expose more visual spatial patterns that could be extracted (see example in Figure 4.8). [10]. A similar situation is presented for results in Table 4.9 for the Butterflies image collection (see the example in Figure 4.9).

Experiments for Birds dataset using Strategy 2							
Detailed F-Measure performance by class							
classes							
model	Egret	Mandarin	Owl	Puffin	Toucan	Wood duck	avg
RBMKL	52.32	**47.71**	**68.03**	**55.53**	**64.32**	**50.55**	**56.41**
BoVW	**52.62**	41.54	66.75	54.37	60.21	49.15	54.10

Table 4.8: Strategy 2: Several (Linear + Intersection) Kernels - Single Space. Table shows detailed experiments in the Birds dataset using sequences of Visual Words (Uni-Bi-grams) under early, late fusion and RBMKL (MKL intermediate fusion). The F-Measure value of each class for BoVW correspond to the best kernel configuration we found (linear or intersection). The F-Measure value of each class for BoVW correspond to the best kernel configuration we found (linear or intersection).

Finally, experimental results using visual n-grams for the Scenes datasets, also showed similar properties for specific classes. For example, in this collection there are classes where results of experiments using visual n-grams are closer to the pure use of visual words (1-grams), having low gain/lost performance. Some classes with more differences in performance are: kitchen, living room, bedroom and store. These kind of indoor classes appear to be the more complicated given the high variety of objects that could be found (other interesting classes are street and suburb). In those classes, visual n-grams provided an improvement in the performance. On the other hand, simple visual words get better results in natural scenes like mountain, forest, open country and coast have more plain unstructured visual elements like sky, grass, water, etc. We think such images are better classified by a simple BoVW because there are structural visual

[10] Analogous characteristics present other classes like Owl (a mostly white bird) and toucan (a bird with more contrast and structural characteristics)

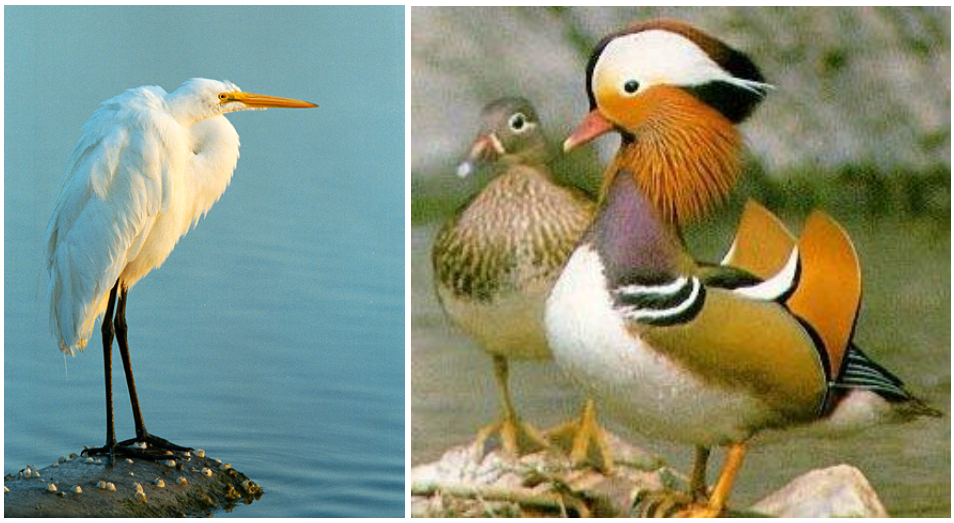

Figure 4.7: Left *Egret* class and Right *Mandarin* of the Birds dataset. Sample instances to expose the image characteristics of those two classes and their performance when using visual n-grams to capture the context.

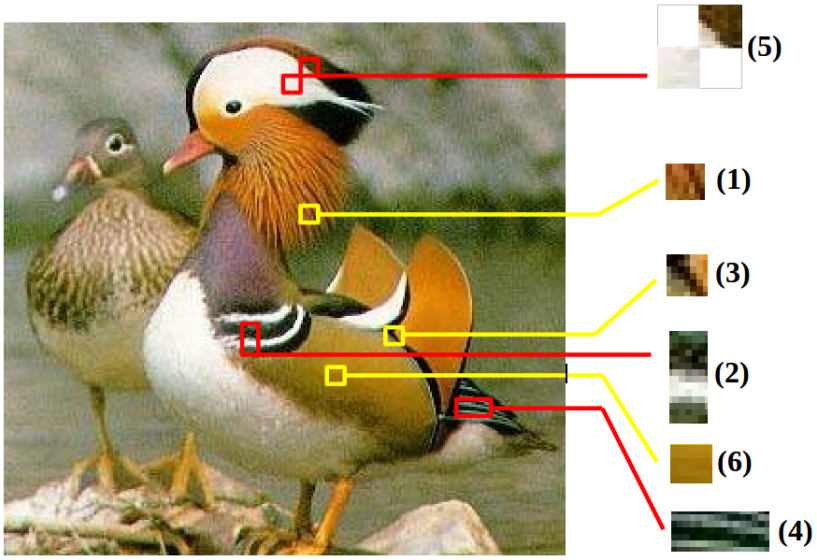

Figure 4.8: Example of a Mandarin duck image. According to information gain implemented in (Hall et al., 2009), we rank the 6 visual regions that produced visual features with most discriminative information. We highlight in yellow and red, the regions that using SIFT descriptors produced discriminative visual words and visual 2grams respectively.

elements that need to be captured in a more global way (visual n-grams capture local visual patterns).

Experiments for Butterflies dataset using Strategy 2								
Detailed F-Measure performance by class								
clases								
model	Admiral	Swallow tail	Machaon	Monarch 1	Monarch 2	Peacock	Zebra	avg
RBMKL	57.31	**44.73**	**70.61**	**57.61**	**67.14**	**72.32**	**65.11**	**62.11**
BoVW	**58.82**	42.05	68.02	54.95	65.15	70.72	64.10	60.45

Table 4.9: Strategy 2: Several (Linear + Intersection) Kernels - Single Space in the Butterflies dataset. Table shows detailed experiments in the Butterflies dataset using sequences of Visual Words (Uni-Bi-grams) under early, late fusion and RBMKL (MKL intermediate fusion). The F-Measure value of each class for BoVW correspond to the best kernel configuration we found (linear or intersection).

Figure 4.9: Left *Balck Swallowtail* class and Right *Peacock* of the Butterflies dataset. Sample instances to expose the image characteristics of those two classes and their performance when using visual n-grams to capture the context.

4.4.3 Bag-of-Visual n-grams and the Spatial Pyramid Representation (SPR)

In this section we present an experimental evaluation to deepen the analysis of the proposed visual n-grams and the BoVW. In spite of the simplicity and effectiveness of the BoVW approach, there have been several efforts to incorporate spatial information to it. In addition to the language models used in Section 4.4.2, there have been other approaches to capture spatial information at different levels. The Spatial Pyramid Representation (SPR) (Lazebnik et al., 2006) described in Section 3.1, is one of the most notable works to improve the performance of BoVW. In the following subsections we evaluate and compare the performance of the proposed BoVN and the SPR. We explain the key differences between the two approaches and discuss some properties of each dataset that makes possible the outstanding performance of each approach. Furthermore, we

found that both approaches can be easily integrated to obtain an improvement in the classification performance.

BoVN vs SPR

The purpose of this first experiment is to compare the classification performance of the proposed strategy and the Spatial Pyramid Representation (SPR) (Lazebnik et al., 2006). SPR is a method proposed in the literature which exploits the use of spatial information in other ways (see Section 3.1 for more details). For this purpose we evaluate our best strategy using the visual 1+2grams and RBMKL. We also evaluate the performance of the SPR representation on each dataset using the original implementation provided by (Lazebnik et al., 2006) [11]. Some interesting findings can be highlighted from results in Table 4.10. For example, it is interesting to see that, according to the Wsr test, the proposed RBMKL approach significantly outperforms the SPR in the Birds and the Butterflies datasets. Regarding to the 6-Clatech dataset, RBMKL also outperforms SPR, however, the difference is not statistically significant. Finally for the 15-scenes dataset, SPR significantly outperforms our approach. We think such results on each dataset are due to specific characteristics of each domain. More specifically, the proposed visual n-grams extract very *local* visual patterns. For example, visual bigrams are useful to capture some characteristic lines of the mandarin duck (see Figure 4.8), but may fail to capture more global visual patterns (e.g., the mandarin ducks are usually surrounded by visual elements similar to watter or grass). On the other hand, SPR captures more *global* and *absolute* visual patterns, which can be matched by the intersection kernel according to each region in corresponding levels of the pyramid. We believe that this is one of the reasons of the high performance of SPR in natural scenes dataset, where images belonging to the same class, share more visual words in each pyramid level (e.g., in buildings images the top part usually is the sky and clouds). Nonetheless, SPR may fail to account for very local and relative patterns, especially when an image presents the target object in a wide range of positions and rotations variants (the case of Birds and Butterflies dataset). The latter scenario hinder to match several levels of the pyramid (except for the coarser level), which produce a more noisy representation.

[11] We use the following experimental settings: 400 visual words and 8x8 size patches, besides our best descriptor for each collection; DCT descriptor for histology and Butterflies datasets, while SIFT descriptor for the rest. An intersection kernel as defined in (Lazebnik et al., 2006) is used into an SVM. By using these fixed experimental settings, we experimental determine to 3 the number of the pyramid levels in SPR by exploring values between 2 and 5

Results RBMKL vs SPR					
Averaged F-Measure per collection					
model	Histopathology	Birds	Butterflies	Scenes	6-Caltech
RBMKL	**68.53***	**56.41***	**64.00***	78.2	**74.12**
SPR	67.32	53.38	61.97	**80.10***	72.13

Table 4.10: F-Measure results for RBMKL vs SPR. For image preprocessing in these collections, settings from Section 4.3 where used.

Extending SPR using visual n-grams and MKL

The purpose of this second experiment is to evaluate the classification performance of visual n-grams when they are integrated into the SPR. For this, we evaluate SPR using the experimental settings of Section 4.4.3. In Table 4.11, there are three SPR strategies. The first one (SPR), corresponds to the standard implementation as described in (Lazebnik et al., 2006). The second one (SPR+VN), is a simple extension where local histograms of visual 1+2grams are computed from each cell in the pyramidal representation. Thus, the final pyramidal representation of an image is the arrangement of histograms of 1+2grams corresponding to each cell. Finally, SPR+VN+RBMKL is an approach, that under the Strategy 2, uses the representation vectors of SPR+VN to learn a new kernel [12]. From results in Table 4.11, we can observe that integrating visual 2grams into the SPR results in an improvement of the classification performance. We can also note that by using RBMKL to learn the kernel, the impact in the performance is positive. It is worth mentioning that SPR+VN and SPR+VN+RBMKL, according to the Wsr, are significantly better than SPR. We think that the improvement in the results are due to the complementary *local* and *global* information carried by each of the combined methods.

Results of integrating visual n-grams and MKL into SPR					
Averaged F-Measure per collection					
model	Histopathology	Birds	Butterflies	Scenes	6-Caltech
SPR	67.32	53.38	61.97	80.10	72.13
SPR+VN	68.90	57.71	63.15	80.78	74.83
SPR+VN+RBMKL	**70.02**	**58.19**	**65.31**	**81.29**	**76.17**

Table 4.11: F-Measure results for SPR extended with visual n-grams and MKL. For image preprocessing in these collections, settings from Section 4.3 where used.

[12] A linear and an intersection kernel is built using image representation of SPR+VN, then RBMKL is used to learn a new kernel function

4.5 Final Remarks

The underlying motivation of this chapter was to improve the state-of-the-art in BoVW like approaches through fusion strategies that integrates the visual n-grams (multi-directional sequences of visual words) as attributes. Thus, we took the analogy visual-textual words into a new higher level combining contextual (visual n-grams) and non-contextual (visual words) information through alternative fusion strategies. Motivated by the analogy visual-textual, we considered the fusion of the contextual and non-contextual information in NLP tasks. Thus, we evaluated visual n-grams in order to consider visual spatial information from a NLP perspective. Regarding to typical fusion strategies, simple *early fusion* strategy showed better/similar performance than *late fusion* approach. This is due in part to that in text classification most of textual feature spaces are derived from one rigorous modality: the text. This condition degrades the diversity among the search space, which is one of the most important aspects for building ensembles. The results show evidence of the usefulness of integrating visual n-grams under the proposed MKL strategies, showing that, every experiment using visual bigrams outperforms unigrams and other methodologies.

CHAPTER 5

EXPLOITING THE SEMANTIC VISUAL INFORMATION

One of the main motivations of the BoVW representations is similar to the Bag-of-Words (BoW) used in text mining tasks: to build word histograms that represent documents. In this regard there are many alternative ways of improving the BoW representation within the text mining community that can be applied in computer vision as well. This chapter proposes the adaptation of Distributional Term Representations (DTRs) for image classification. DTRs represent images by exploiting statistics of feature occurrences and co-occurrences along the dataset. We focus on the suitability and effectiveness of adapting well-known DTRs in different image collections. Furthermore, we devise two novel distributional strategies that learn appropriated groups of images to compute better suited distributional features.

In summary, we introduce the idea of exploiting distributional visual information among high level visual features, and expose its benefits for image classification tasks. Thus, the contributions of this chapter are threefold: *i)* the proposal of adapting DTRs for image classification by exploiting the analogy visual-textual features, *ii)* two novel distributional visual feature strategies, that automatically obtain better suited representations according to each image domain, and *iii)* The evaluation of DTRs in different image domains using features based on visual words. We evaluate DTRs using two different high level features, namely, visual words and sequences of visual words. Experimental results over the five proposed datasets suggest evidence of the usefulness of the distributional information in DTRs over BoVW and other methods in the literature.

5.1 Distributional Term Representations, from text to images

This section describes the general framework for popular DTRs. By exploiting DTRs, we aim to represent images in a low dimensional and non-sparse space that captures more relevant visual information. Our goal is to overcome, to some extent, the issues naturally inherited by the BoVW, and therefore to improve the classification performance. For this purpose, this section shows the adaptation of traditional DTRs by taking advantage of the *visual-textual word* analogy and the distributional hypothesis. The idea is that visual words that occur in similar visual contexts

should have similar representations. Thus, the main goal is to build enriched distributional visual-word and image representations, which capture contextual information by means of visual word occurrences and co-occurrences.

DTRs comprise two main stages. In a first stage, visual-words are represented in a new distributional space (see step 1 in Figure 5.1). In a second stage images are represented under that new *distributional* space (see step 2 in Figure 5.1). In the following sections, we consider the visual words in the codebook as the terms to build the DTRs. More formally, let $\mathcal{I} = \{(I_1, y_1), \ldots, (I_n, y_n)\}$ be a training set of labeled images, that is, $\mathcal{I}$ is a collection of $n-$pairs of images (I_i) and labels (y_i); where the latter indicates the category associated with the image, with $y_i \in \mathcal{C} = \{C_1, \ldots, C_q\}$. Also let $\mathcal{V} = \{v_1, \ldots, v_m\}$ denote the vocabulary of terms, which in our case is the codebook of visual words precomputed in the collection under analysis. Given the latter context, DTRs attempt to capture in different ways valuable distributional information of the $\mathcal{V}$ features along the dataset. DTRs begin associating each visual word $v_i \in \mathcal{V}$ with a vector $\mathbf{t_i} \in \mathbb{R}^r$, i.e., $\mathbf{t_i} = \langle t_{i,1}, \ldots, t_{i,r} \rangle$, where r is the number of distributional features according to each specific DTR, and $t_{i,j}$ indicates the contribution of distributional feature j to the representation of visual word v_i.

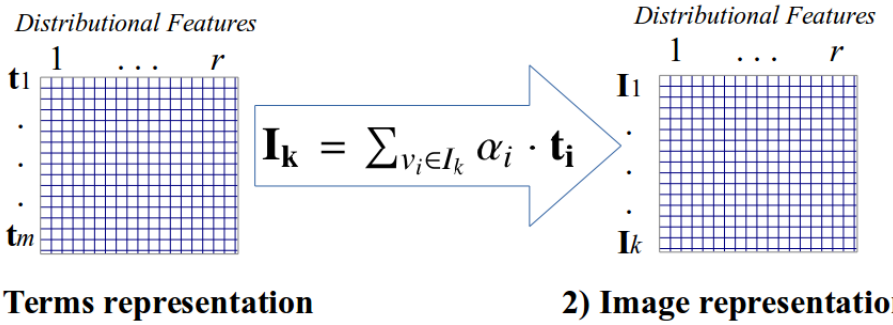

Figure 5.1: DTRs build visual words representations (term vectors) that describe its distributional properties (Step 1), then use such term vectors to build the image representation (Step 2).

Image representation is obtained by aggregating the representation of terms that occur in the image (see Figure 5.1). This is, let $\mathbf{t_i}$ denote the DTR of the visual word v_i, then the image representation of I_k is computed as $\mathbf{I_k} = \sum_{v_i \in I_k} \alpha_i \cdot \mathbf{t_i}$. Where the scalar α_i weights the relevance of visual word v_i in image I_k. In this way, the representation of each image is the weighted aggregation of the DTRs of the visual words occurring in the image [1]. The resultant DTRs are low-dimensional, nonsparse and capture more useful contextual term information. In the test phase, test images are represented by combining DTRs from the training term vectors

[1] Note that α_i does not need to be optimized, and could be any term weighting scheme in text mining to capture the feature contribution (e.g., frequency, boolean, information gain, etc.). In our evaluation we use the simplest one: the visual word occurrence frequency.

as well. In the following sections we describe in detail how each of the DTRs computes the $\mathbf{t_i}$ representation taking visual words as terms.

5.1.1 Image Occurence Representation (IOR)

The main idea of Image Occurrence Representation (IOR) consists in capturing the semantics of a visual word by observing the distribution of occurrence statistics over the images in the dataset (Lavelli et al., 2004). More formally, each visual word v_i is represented as a vector $\mathbf{t_i} = \langle t_{i,1}, \ldots, t_{i,|\mathcal{I}|} \rangle$, where $|\mathcal{I}|$ is the number of images in the training collection, and $t_{i,k}$ indicates the relevance of the image I_k to characterize v_i. Equation 5.1 presents the above ideas.

$$t_{i,k} = df(v_i, I_k) \cdot \log \frac{|\mathcal{V}|}{|\mathcal{N}_k|} \tag{5.1}$$

where $\mathcal{N}_k \subseteq \mathcal{V}$ is the set of different terms in the image I_k, and $df(v_i, I_k)$ is defined in Equation 5.2.

$$df(v_i, I_k) = \begin{cases} 1 + \log(\#(v_i, I_k)) & \text{if } \#(v_i, I_k) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.2}$$

where $\#(v_i, I_k)$ indicates the frequency of term v_i in I_k. The intuitive idea is that, the importance of an image I_k to characterize v_i is given by the frequency of the term v_i in I_k. Also note that the number of different terms contained in I_k is inversely proportional to its contribution to represent v_i. Finally, the DTR of each term $\mathbf{t_i}$ is normalized so that $\|\mathbf{t_i}\|_2 = 1$

5.1.2 Visual-Feature Co-occurence Representation (VCOR)

The principle behind the Visual-Feature Co-occurrence Representation (VCOR) is that, the semantics of a visual word can be captured by observing its co-occurrences with other visual words across images in the dataset (Lavelli et al., 2004). Thus, each visual word v_i is associated to a vector $\mathbf{t_i} = \langle t_{i,1}, \ldots, t_{i,|\mathcal{V}|} \rangle$, where $|\mathcal{V}|$ indicates the codebook size, and $t_{i,k}$ denotes the contribution of the visual word v_k to the semantic description of v_i. Equation 5.3 presents the above ideas.

$$t_{i,k} = tff(v_i, v_k) \cdot \log \frac{|\mathcal{V}|}{|\mathcal{T}_k|} \tag{5.3}$$

where $\mathcal{T}_k \subseteq \mathcal{V}$ is the set of different visual words co-occurring with t_i in at least one document. *tff* is defined in Equation 5.4.

$$tff(v_i, v_k) = \begin{cases} 1 + \log(\#(v_i, v_k)) & \text{if } \#(v_i, v_k) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{5.4}$$

where $\#(v_i, v_k)$ indicates the number of images in which visual words v_i and v_k co-occur. Finally $\mathbf{t_i}$ vector is normalized so that $\|\mathbf{t_i}\|_2 = 1$.

5.1.3 Class Occurrence Representation (COR)

The intuitive idea of a Class Occurrence Representation consists in representing the terms by their relation with each target class (Li et al., 2011; López-Monroy et al., 2015). This can be done by exploiting occurrence-statistics over the set of documents in each of the target classes. In this way, we represent each visual word $v_i \in V$ with a vector $\mathbf{t_i} = \langle t_{i,1}, \ldots, t_{i,q} \rangle$, where the $t_{i,k}$ is the degree of association between visual word v_i and class C_k. Under this DTR, the weight $t_{i,k}$ is related with the occurrence of term v_i in images that are labelled with class C_k. The relationship between the i^{th} visual word and the k^{th} class can be defined according to Equation 5.5.

$$w_{i,k} = \sum_{\forall I_j : y_j == C_k} \log_2 \left(1 + \frac{tf(v_i, I_j)}{len(I_j)} \right) \tag{5.5}$$

where $tf(v_i, I_j)$ is the visual word-occurence frequency of the visual word v_i in the image I_j, and $len(I_j)$ indicates the number of visual words in I_j. The $\log_2$ function aims to soften the relevance of highly frequent visual words.

Finally, in order to produce the final $\mathbf{t_i} = \langle t_{i,1}, \ldots, t_{i,q} \rangle$ representation, two normalizations are performed. Equation 5.6.1 shows the first one in order to consider the proportion of the $|\mathcal{V}|$ terms in each class, whereas, Equation 5.6.2 shows the second one in order to take into account the weights computed for the $|\mathcal{C}|$ classes, making weights $w_{i,\cdot}$ comparable among classes.

$$(5.6.1)\ \hat{t}_{i,k} = \frac{w_{i,k}}{\sum_{i=1}^{|\mathcal{V}|} w_{i,k}} \qquad (5.6.2)\ t_{i,k} = \frac{\hat{t}_{i,k}}{\sum_{k=1}^{|\mathcal{C}|} w_{i,k}} \tag{5.6}$$

5.2 New Distributional Visual-Feature Representations

DTRs presented in Section 5.1 are useful methodologies to harness distributional information of visual words. Notwithstanding the classification performance that traditional DTRs bring to the standard BoVW, each of these proposals is somewhat limited because they only capture very specific distributional visual patterns. For image classification there are a number of complex scenarios (i.e., high intra-class visual content, wide and narrow domains, etc.) where modeling the visual content, might require the analysis of multiple levels of the contextual information. In this regard, if we carefully observe the DTRs presented in Section 5.1, one can notice that each DTR focuses at different levels of the contextual information. For example, VCOR aims to model the term relevance in a very detailed way by observing the term co-occurrences. In the case of IOR, it models the term importance by using a more general principle; the occurrences in images. Finally, COR pushes to the limit the generalization principle, modeling the term relevance by exploiting the occurrences in the categories. According to the characteristics of the domain in which a DTR is used, the latter principles of generalization can become in a strength or in a weakness to achieve better accuracy rates. In this regard, it would be promising to propose alternative strategies for building better distributional visual representations for images; that is precisely the purpose of this section.

5.2.1 Subclass Occurrence Representation (SOR)

The first proposed strategy is the Subclass Occurrence Representation (SOR), which can be seen as an improvement of COR. The idea of SOR is to alleviate the idea of class *homogeneity* (e.g., a whole class can be represented in one dimension) exploited by COR. Thus, the idea is to generate a new space of subclasses by means of a class-driven procedure that allow us to model the *heterogeneity* inside each target class. In this way, SOR exploits *intra-class* information by means of a clustering procedure in each target class. Thus, in a new stage the visual features and images could be represented in a new discovered subclass-space, which still being low dimensional, non-sparse and captures more fine grained class-specific information (see Figure 5.2). Algorithm 1 shows the main steps of this approach. More precisely, a clustering procedure is separately applied in training images belonging to each of the categories $\mathcal{C} = \{C_1, \ldots, C_q\}$ (lines 3 and 4). Thus, for each class C_i, a set $\mathcal{R}_i = \{S^i_1, \ldots, S^i_{r_i}\}$ of r_i subclasses are generated (line 4). The final set of subclasses is the combination of all generated clusters: $\mathcal{S} = \{S^1_1, \ldots, S^1_{r_1}, \ldots, \ldots, S^q_1, \ldots, S^q_{r_q}\}$, where $|\mathcal{S}| = \sum_{j=1}^{q} |\mathcal{R}_j|$ (line 5). Once that subclasses are generated, images are represented under this subclass space by using COR as described in Section 5.1.3, this time making $\mathcal{C} = \mathcal{S}$ (line 8).

Algorithm 1 Subclass Occurrence Representation (SOR)

Require: $\mathcal{I}_{\text{Train}} = \{(I_1, y_1), \ldots, (I_n, y_n) \mid y_i \in \mathcal{C}$ is the class label of $I_i\}$

Ensure: $\mathbf{I}^{\text{SOR}}_{\text{Train}}$ (Train matrix of n images by $|\mathcal{S}|$ distributional features)

1: $\mathbf{I}^{\text{BoVW}}, \mathbf{y} = \text{BoVW}(\mathcal{I}_{\text{Train}})$

2: $\mathbf{I}^{\text{COR}} = \text{COR}(\mathbf{I}^{\text{BoVW}}, \mathbf{y})$

3: **for** C_i **in** $\mathcal{C} = \{C_1, \ldots, C_q\}$ **do**

4: Perform clustering on $\mathbf{I}^{\text{COR}}_{\{k:y_k = C_i\}}$ for obtaining intra-classes in S_i

5: $S^i = \{S^i_{r_i} \mid S^i_{r_i}$ is the r th intra-class label of the class $i\}$

6: New intra-class labels are given to the images $I_k : \hat{y}_k = S^i_{ji} \in S^i$

7: **end for**

8: $\mathbf{I}^{\text{SOR}}_{\text{Train}} = \text{COR}(\mathbf{I}^{\text{BoVW}}, \hat{\mathbf{y}})$

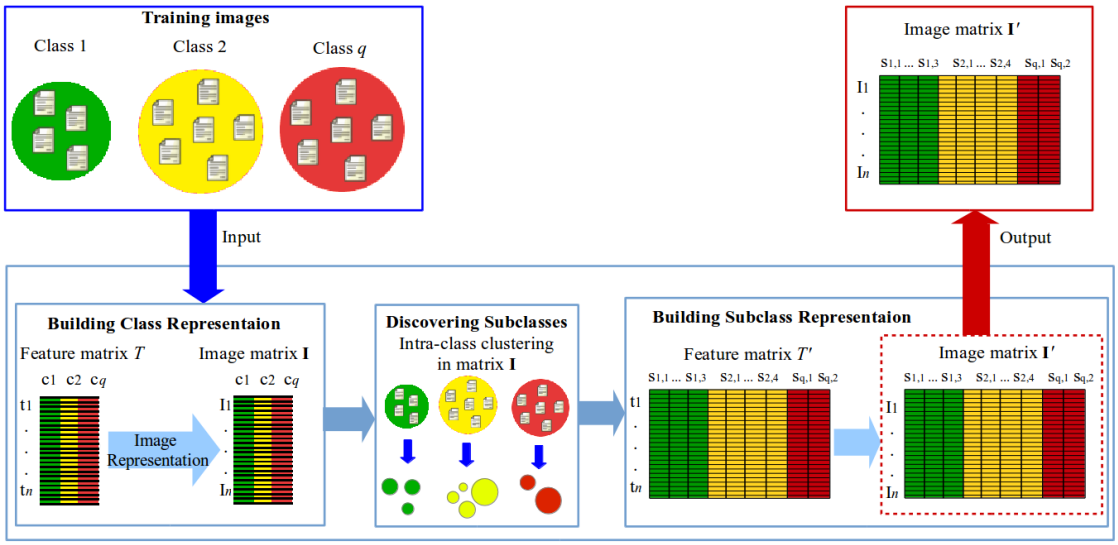

Figure 5.2: Subclass Occurrence Representation (SOR). The purpose of this representation is to model the *intra-class* information by means of a class driven procedure.

5.2.2 Group Occurrence Representation (GOR)

The latter approach automatically learns more fine-grained distributional information of the features. The strategy allows to discover the relevant subcategories of images in which the DTR should focus. Notwithstanding that SOR can capture automatically finer grained information, the strategy still being in some way forced to focus in each of the target categories. Thus, we take this weakly supervised SOR into an unsupervised version named Group Occurrence Representation (GOR)[2]. GOR is an extension of COR, adapted for automatically discovering visual patterns in the training dataset by discovering *inter-group* information by means of clustering. Thus, the visual features and images are represented under a new discovered group-space. The intuitive idea is that, the generated groups expose more relevant information about the images in the dataset (see Figure 5.3). GOR keeps COR properties, namely, low dimensional and non-sparse by capturing information of natural clusters into the dataset. The Algorithm 1 shows the main steps of this approach. To build GOR a clustering procedure is applied to the whole training images (line 3). Thus, GOR generates a final set of subclasses $\mathcal{S} = \{S_1, \ldots, S_r\}$, where $|\mathcal{S}| = r$ (line 4). After creating the new subclass space, the images are represented by using COR as described in Section 5.1.3 making $\mathcal{C} = \mathcal{S}$ (line 6).

Algorithm 2 Group Occurrence Representation (GOR)

Require: $\mathcal{I}_{\text{Train}} = \{(I_1, y_1), \ldots, (I_n, y_n) \mid y_i \in \mathcal{C}$ is the class label of $I_i\}$
Ensure: $\mathbf{I}^{\text{SOR}}_{\text{Train}}$ (Train matrix of n images by $|\mathcal{S}|$ distributional features)
1: $\mathbf{I}^{\text{BoVW}}, \mathbf{y} = \text{BoVW}(\mathcal{I}_{\text{Train}})$
2: $\mathbf{I}^{\text{COR}} = \text{COR}(\mathbf{I}^{\text{BoVW}}, \mathbf{y})$
3: Perform clustering on $\mathbf{I}^{\text{COR}}$ for obtaining inter-classes in S
4: $S = \{S_r \mid S_r$ is the r th inter-class label in the corpora $\}$
5: New inter-class labels are given to the images $I_k : \hat{y}_k = S_j \in S$
6: $\mathbf{I}^{\text{SOR}}_{\text{Train}} = \text{COR}(\mathbf{I}^{\text{BoVW}}, \hat{\mathbf{y}})$

[2]Please note that the process to build the representation is totally unsupervised, but it is used in image classification which is a supervised problem.

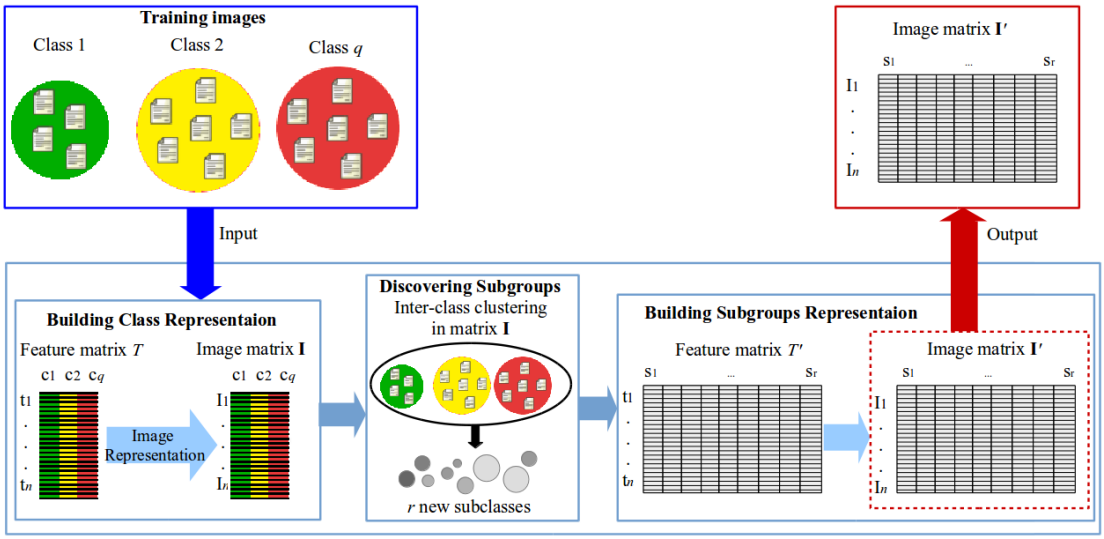

Figure 5.3: Group Occurrence Representation (GOR). The purpose of this representation is to model the *inter-class* information by means of clustering the entire collection of images.

5.3 Visual features used as terms

Terms can be any set of patterns taken as features in a target domain. Regarding to text mining, one of the most important set of textual features are specific lexical units, for example; words. Also in this direction, maybe the second most popular textual feature, are the sequences of n words (n-grams) that helps to capture spatial relationships among words (i.e., representing as one attribute concepts like "pattern recognition"). Regarding to image classification, in this thesis we use the analogous versions of those two terms: visual words and visual bigrams.

5.3.1 Visual words

We compute the visual words using a traditional BoVW formulation. This framework can be easily explained in the following four steps (see Figure 2.1 in Chapter 2):

1. **Griding images**: We use a grid to extract image patches [3].
2. **Descripting patches**: We represent each extracted patch using a visual descriptor (i.e., SIFT or DCT).
3. **Building the dictionary**: A codebook of visual words is learned by using a clusterer (i.e., k-means) on the training descriptors. Cluster centroids are taken as visual words.

[3] Albeit there are many alternative ways to extract image patches (i.e., dense or regular grid-based, keypoint-based, etc.), in this work we use regular-grid-based to simplify the explanation of the visual features that we use as terms.

44	219	389	182	33	153	141	119
222	213	65	78	134	211	191	233
320	21	113	123	21	297	326	321
43	16	234	71	91	38	90	42
129	345	222	400	341	349	256	54
120	15	112	23	212	219	152	35
354	123	234	2	54	125	212	66
27	198	19	11	45	345	56	69

Figure 5.4: Example of an image representation using the codebook (López-Monroy et al., 2013). For the current-target item (65) the generated bigrams are: 65-389, 65-219, 65-213, 65-21, 65-113, 65-123, 65-78, 65-182.

4. **Indexing images**: Each grid-patch descriptor in image is replaced with the code of the nearest learned visual word.

5.3.2 Visual bigrams

There are several alternatives to model the spatial relationships among visual regions. In order to extract visual features for DTRs, we are interested in those related with analogies of *visual-textual* words. In this regard, sequences of visual words is one of the next logical steps that several authors have proposed (López-Monroy et al., 2013; Zheng et al., 2006; Yuan et al., 2007; Tirilly et al., 2008; Yuan et al., 2011). In this thesis, we exploit the idea of visual n-grams proposed in (López-Monroy et al., 2015) and explained in Chapter 4. We used this strategy because it produces sequences of neighbouring items in a very similar way to word n-grams (see Figure 5.4). The dictionary of those normalized visual n-grams are then used to proceed with the image representation through histograms that account occurrences of the visual n-grams found in the image.

5.4 Experimental Settings

For evaluating the effectiveness of the proposed DTRs, we use the five datasets proposed in Chapter 5[4]. For the experiments in each image collection we also follow a similar experimental setting to the one presented in Chapter 5. Thus, we have built a dictionary of 400 visual words by griding images in patches of 8x8 pixels. To represent each patch we perform experiments using two image descriptors, the first one is the Scale Invariant Feature Transform (SIFT) (Lowe, 2004). The Second one is the discrete cosine transform (DCT) computed merging 64 coefficients on each channel of the RGB color space. Finally, the experiments using the Bag-of-Visual bi-grams (BoVB) baseline consist in the concatenation of the histograms of visual words and the histogram of the occurrences of sequences of two visual words. The number of possible sequences generated from a dictionary of size 400 is 400^2. Thus, as determined in (López-Monroy et al., 2015) we only use a subset of the most 2500 top frequent bi-sequences, which for our datasets, usually are sequences with at least 10 to 50 occurrences according to each dataset. For the evaluation, we use a 10 cross fold validation framework (10CFV) on each dataset. The classifier used for experiments in this chapter is a linear SVM in (Fan et al., 2008), which have achieved outstanding performances in similar classification problems. We report experimental results using the accuracy, which represents the percentage of images that were correctly classified. In this chapter we use the accuracy to directly compare the new baselines (LSA and LDA), which have reported results in some of the target collections. Nonetheless, in Appendix A we also report full F-measure results for all the experiments of this chapter. For automatically creating the $\mathcal{S}$ number of clusters (subclasses) in the proposed SOR and GOR, we have used the Expectation Maximization (EM) clustering algorithm provided by the Weka framework (Hall et al., 2009). The EM algorithm assumes that the set of images I is a set of objects generated by a probability distribution, which is a combination of n different Normal distributions. According to (Hall et al., 2009), the EM clustering algorithm works as follows: i) n is set to one, ii) 10 folds are created in the training data, iii) EM clusterer is performed in a cross validation way, iv) the likelihood is averaged over all runs, and v) if likelihood is greater than the test $n-1$, then n is increased and continues in step 2. For comparing the performance of the DTRs, under the same experimental configuration, we have used three of the alternatives mentioned in Chapter 3: LSA (Bosch et al., 2006), LDA (Fei-Fei and Perona, 2005), and SPR (Lazebnik et al., 2006) [5]. Given that context, we also estimated if the proposed DTRs and each

[4] It is worth mentioning that, in this chapter, for the histopathology dataset, we delimit the analysis to the most important class: *basal cell carcinoma*.

[5] The experimental settings for LSA and LDA were experimentally determined between 200 and 60 concepts for each dataset in the same way that in (Lazebnik et al., 2006). For the SPR, the parameter to build the pyramid representation was fixed to three as suggested in (Lazebnik et al., 2006).

of the proposed baselines (e.g., IOR, BoVW) are statistically significantly different (using a confidence value above 95%). Thus, we used the Wilcoxon signed-ranks (Wsr) test, which is the recommended test by J. Demšar for comparisons between two methods (Demšar, 2006).

5.5 Experiments and Results

In this section we explain the purpose and details of each experiment to evaluate the DTRs. In Section 5.5.1, we evaluated and compared the DTRs using visual words as terms, versus the BoVW and other typical baselines in the literature. Finally in Section 5.5.2, we evaluated the DTRs jointly exploiting visual words and visual bigrams. Furthermore, to get more insights about the robustness of DTRs, we also present experimental results where the visual words are built based on two different visual descriptors: SIFT and DCT.

5.5.1 Distributional representations on visual words

The main goal of this experiment is to evaluate the usefulness of the DTRs using visual words as terms, since they are the most representative and basic visual feature in the *visual-textual* words analogy. From experimental results in Table 5.1, we can observe that independently of using DCT or SIFT descriptors, most of DTRs (except COR) outperformed the standard BoVW. This is in part due to the captured distributional information of visual words along the datasets, similar to the evidence exposed in text mining tasks (Lavelli et al., 2004). More specifically, in VCOR we can notice some clues of the improvement capturing the distribution of local pattern co-occurrences, whereas in IOR we can see outstanding performances especially in the 6-Caltech dataset. The best result of IOR in the 6-Caltech dataset is somehow expected since DOR is the dual in TFIDF weighting scheme of text mining tasks. In other words, IOR allows to capture the importance of features in documents with respect to the collection, which is very useful to discriminate in wide domains with different objects like chairs or ants. Also, regarding to the other DTRs, it is not surprising that COR did not outperform the BoVW, this is because COR is designed to extract very few attributes (one attribute per class) that have proven to be effective only in collections where the number of images and the dictionary size usually are much larger (López-Monroy et al., 2015). This is similar to what happens in scenarios where simple methods overcome complex ones when large amounts of data are available.

From Table 5.1 we can observe that most of the best results were obtained by the proposed GOR and SOR. We infer that the outstanding results are because such representations were built in a softer and less hand-crafted way than the others. For example, SOR is built in a weakly directed way creating subclasses from the target classes, and GOR is totally unsupervised on the entire dataset. In this way, SOR seems to be more useful in the dataset of Scenes, which presents

high intraclass variability (there could be a lot of types of objects inside the scenes). On the other hand, the GOR representation tends to obtain better results in the Histopathology, Butterflies and Birds datasets, which besides of having high intraclass variability, are a narrower domain since the target classification object belongs to a more general category (e.g., a butterfly or a bird). Regarding to visual descriptors, although DCT is sensible to scale and rotation of the image, it considers properties of texture and color, which produces better results for histopathology and butterflies datasets. Finally, just as expected, SIFT descriptor worked better for the other collections, which are natural images with different scales, resolutions and orientations. Besides the usefulness of DTRs over BoVW, the DTRs also obtained better results than LSA and LDA, which are two traditional feature extraction methods in the literature. Finally, it is worth mentioning that, computing the Wsr test over the outputs of the 10CFV of GOR and each of the proposed baselines, we obtained more than 98% of statistical confidence in results.

Accuracy Performances using visual words						
			SIFT			
Method	Hist.	6-Caltech	Butterflies	Birds	Scenes	avg.
BoVW	75.04	64.76	50.72	50.00	69.65	62.03
LDA	56.21	59.39	30.04	36.83	61.56	48.80
LSA	77.23	65.77	48.78	51.50	73.11	63.27
VCOR	72.13	67.47	50.2	51.22	69.11	62.02
IOR	79.18	**69.43**	51.08	53.11	70.21	64.60
COR	68.88	61.74	43.45	42.16	69.16	57.07
SOR	79.03	68.11	52.24	54.50	**74.02**	65.58
GOR	**81.20**	68.01	**54.40**	**56.66**	71.10	**66.27**
			DCT			
Method	Hist.	6-Caltech	Butterflies	Birds	Scenes	avg.
BoVW	86.22	54.02	61.55	50.83	52.15	60.95
LDA	76.12	40.93	28.91	32.00	37.03	42.99
LSA	87.77	40.26	41.68	39.00	53.77	52.49
VCOR	81.43	55.51	63.48	53.21	52.62	61.25
IOR	88.11	57.86	65.03	57.06	52.10	64.03
COR	75.55	45.97	41.84	38.83	49.98	50.43
SOR	89.23	**58.88**	67.21	58.47	**53.42**	65.44
GOR	**91.12**	58.11	**68.01**	**60.10**	52.82	**66.03**

Table 5.1: Performances for the evaluated strategies using a dictionary of visual words generated using SIFT-based and DCT-based visual words.

5.5.2 Distributional representations on Visual Words and Visual n-grams

The purpose of these experiments is to expose whether considering visual bigrams[6] (which are local structural visual patterns) in DTRs, LSA, LDA and SPR, could improve the performance of the classifier. The way we integrate such information is separately building two image representations for each dataset. The first representation is built using visual words as terms, whereas the second representation uses the visual bigrams as described in Section 5.3.2. The combination is done by an early fusion strategy, which is a merely concatenation of both matrices (Kuncheva, 2004). For example, in the case of the Bag-of-Visual Bigrams (BoVB) we concatenate the histograms of visual words and visual bigrams.

Table 5.2 shows high improvements in the experimental results, in part due the extra information of the local visual patterns captured by visual bigrams, but also by the inclusion of its distributional information along the dataset. Furthermore, the Spatial Pyramid Representation (SPR) (Lazebnik et al., 2006) briefly described in Chapter 3 is also evaluated. In this regard, both BoVB and SPR representations took advantage of the spatial information, BoVB considers local visual patterns, whereas SPR considers local and global visual patterns in the image. The evaluation showed better performance for the proposed GOR and SOR than the rest of methods in most of the datasets. Notwithstanding that SPR can capture local and global visual patterns at different levels, it does not consider the distribution of visual patterns along the entire collection in the same way that DTRs. Thus, we infer that this extra distributional visual information is what allows GOR and SOR to obtain better performance than SPR in most of the datasets. We think that SPR is better than GOR and SOR in the Scenes dataset because scenes images appear in horizontal positions most of the time, which represent an advantage for *absolute* spatial information captured by SPR from one image to another. In conclusion, the proposed SOR and GOR are expected to work better in datasets that have high intra-inter class variability. For example, Table 3 shows more performance improvements over a typical BoVW for histology, birds and Scenes datasets. The kind of images presented in such datasets are much more complex than 6-Caltech and Birds. Finally, we separately used the Wsr test over the outputs of GOR and SOR for comparing them with SPR. The test produces more than 98% of statistical confidence in results, showing that GOR and SOR significantly outperform SPR except for the Scenes dataset.

[6]This version of visual bigrams as features (López-Monroy et al., 2015) requires the use of a building strategy in the order of $O(n^2)$ for time and space.

Accuracy performances using visual words + visual n-grams						
			SIFT			
Method	Hist.	6-Caltech	Butterflies	Birds	Scenes	avg.
BoVW	75.04	64.76	50.72	50.00	69.65	62.03
LDA	69.23	58.05	30.69	34.16	66.80	51.78
LSA	76.23	65.77	49.11	52.33	75.42	63.77
BoVB	80.71	71.23	52.34	57.21	77.04	67.70
SPR	75.55	72.25	52.22	56.88	**81.44**	67.66
VCOR	79.94	75.55	51.21	56.14	72.11	66.99
IOR	85.21	**76.84**	52.58	58.16	73.44	69.24
COR	79.03	61.40	44.42	46.33	72.53	60.74
SOR	84.23	76.11	53.27	60.27	81.15	71.00
GOR	**86.52**	76.32	**54.11**	**61.71**	79.21	**71.57**
			DCT			
Method	Hist.	6-Caltech	Butterflies	Birds	Scenes	avg.
BoVW	86.22	54.02	61.55	50.83	52.15	60.95
LDA	75.21	40.92	29.24	32.66	40.53	43.71
LSA	89.41	39.93	43.13	39.50	58.66	54.12
BoVB	91.02	56.81	62.13	51.33	54.14	63.08
SPR	80.11	57.22	63.05	55.55	59.44	63.07
VCOR	86.25	54.20	60.41	59.10	53.20	62.63
IOR	92.25	**59.72**	69.46	68.33	56.90	69.33
COR	87.21	43.28	42.48	36.33	54.82	52.82
SOR	93.45	59.12	69.32	67.21	**60.31**	69.88
GOR	**95.57**	58.88	**71.11**	**69.07**	56.11	**70.14**

Table 5.2: Performances for the evaluated strategies using a dictionary of visual words + visual bi-grams generated using SIFT-based and DCT-based visual words.

5.6 Final Remarks

In this chapter the main motivation was to improve the popular BoVW through DTRs applied on the top of analogous visual-textual features. This took the analogy visual-textual words into a new higher level by exploiting distributional and contextual (visual n-grams) relevant information. The evaluation suggested evidence of the usefulness of building DTRs on the top of visual words and visual n-grams. Experimental results showed that in general, DTRs outperform BoVW and other methodologies in the literature. In this way, the results suggested evidence of the usefulness of DTRs in different image collections and two different visual descriptors. We think this is because DTRs are finding better notions of similarity for each space, which could be difficult to obtain with other typical approaches. To the best of our knowledge, the usefulness of DTRs had never been evaluated for different image classification domains. Moreover, the proposed SOR and GOR seem to be better alternatives that allows to build suitable representations according to each domain. Future research paths include the use of different weighting schemes inside DTRs that help improve the image classification task.

Part IV

General Conclusion

CHAPTER 6

GENERAL CONCLUSIONS

The interest of this research lied in the field of Vision and Language. In this regard, we showed evidence that the contextual and semantic visual information, are important elements to improve state-of-the-art visual words approaches (such as the BoVW). We did that by exploiting the analogy between visual and textual words by means of NLP strategies. The two main contributions to the typical BoVW framework were the following; i) the proposal of capturing the contextual information by means of visual n-grams and MKL strategies, and ii) the proposal of capturing distributional semantics by means of novel Distributional Terms Representations. The experimental evaluation showed the usefulness of the proposals in several image collections. Below we detail the main conclusions of this research work:

- In chapter 4 the Bag-of-Visual n-grams representation successfully improves the traditional BoVW by capturing the spatial relations among visual words. This is because the proposed algorithm extracts the multi-directional sequences of visual words, which encompass valuable discriminative visual patterns for each collection. In this regard, we also concluded that the most useful size of visual n-grams are 2 (bigrams). The reason of this is, to some extent, to the number of images in the collection and the generated representations, since longer visual n-grams would require more training instances. In this way, if longer sequences of visual words are harder to find, then sparse and high dimensional representations are generated. In this chapter, we also showed that the Multiple Kernel Learning (MKL) strategies (in particular RBMKL) are helpful to separately represent the spaces of visual words and visual n-grams, which could be harder to achieve with other typical approaches like early and late fusion. In general, we showed the usefulness of visual n-grams under different fusion strategies in different image domains.

- In chapter 5 the proposals inspired in Distributional Term Representations (DTRs) captured the distributional semantic information. First of all, the adapted traditional DTRs were helpful to capture a variety of distributional patterns according to each DTRs (e.g., visual feature occurrences and co-occurrences). Furthermore, the experimental results showed the contribution and usefulness of distributional visual information independently of using

DCT or SIFT descriptors. Regarding to the two novel DTRs, both approaches built highly discriminative representations by building distributional representations of visual words and visual n-grams contained in the images. The first proposed DTR captures the *intra-class* information, which is very useful if high intra-class variability is present in the images (e.g., target classes in the Scenes dataset could have a wide variety of different objects). The second approach captures the *inter-class* information, which is better for narrow domains where the target instances in the dataset belongs to a more general category (e.g., butterfly or bird). Although DTRs have been previously used in several text mining problems, to the best of our knowledge, this is the first time that DTRs are analyzed and found useful, in the context of image classification.

Those two chapters encompass the main contributions of this thesis, which expose the strengths of capturing contextual and semantic visual information. We evaluated them using several image collections used in several pattern recognition studies and comparing with methods reported in the literature. Furthermore, the results have been supported with statistical tests, which provide evidence of the effectiveness of the proposals and that the goals stated in this research have been reached.

In this Vision and Language research, we showed that several ideas from NLP can be exploited in a visual context, nonetheless we are aware that there is a large path to explore. As part of this future work, we consider the following:

- **Fine grained specific classification problems:** This consists in taking advantage of the previously proposed methods in more specific classification problems, such as narrow domains. This is an interesting research path of this work, since we found that some of the more significant differences in the classification performance were achieved in problems with similar target classes (e.g., Birds, Butterflies, and Histopathology). Thus, we infer that many of the core research done in this thesis can be useful in fined grain image classification problems (see Appendix B).

- **Exploit the analogy visual-textual words in multimodal-problems:** In this work we have showed that, to some extent, visual features can be used in similar ways that textual features. Having evidence of this fact, it is promising to push to the limit the analogy of visual-textual words, by means of merging both features into unified multi-modal representations. Multi-modal representations take raw features coming from different sources or domains (e.g., raw text and images) in order to compute new multi-modal features to represent the target instances (e.g., web pages containing text and images). This research path can be very useful in data mining problems where there are images

associated with text descriptions or labels (e.g., text illustration tasks, image annotation, etc.).

- **Semantic and Contextual Visual Information by means of Representation Learning:** Representation Learning strategies consist in using the learned parameters of specific models (e.g., neural networks or support vector machines) to represent key elements of the target problem. For example, deep-learning uses the hidden neurons of Convolutional Neural Networks in order to represent images. In our context, it could be possible to model the semantic information by means of representation learning strategies (e.g., Recurrent Neural Networks and Word2Vec) to improve the discriminative power in more complex images. This can be achieved by exploring neural network based techniques in text mining that automatically learns the contextual information in a more global way among visual features.

6.1 Scientific Publications

The following list shows the research papers derived from this thesis, or those where ideas from this work have been used:

- **Journals papers:**

  1. **López-Monroy, A. P.**, Montes-y-Gómez, M, Escalante, H. J., Cruz-Roa, A. and González, F. A. Improving the BoVW via discriminative visual n-grams and MKL strategies. Neurocomputing, 175, Part A (2016), pp. 768 - 781.
  2. **López-Monroy, A. P.**, Montes-y-Gómez, M., Escalante, H. J. and González, F. A. Distributional Visual-Feature Representations for Image Classification. **Under review (September 2016)**.

- **International conferences:**

  3. **López-Monroy A. P.**, Montes-y-Gómez, M, Escalante, H. J., Cruz-Roa, A., and González, F. A. Bag-of-visual-ngrams for histopathology image classification. Proc. SPIE 8922, IX International Seminar on Medical Information Processing and Analysis, 89220P (November 19, 2013).
  4. **López-Monroy, A. P.**, Montes-y-Gómez, M., Escalante, H. J., González, F. A. Image Classification through Text Mining techniques: a Research Proposal. Postgraduate Students' Meeting. 6th Conference on Pattern Recognition, MCPR 2014. *Best poster award*.

5. Pellegrin, L. and **López-Monroy, A. P.** and Escalante, H. J. and Montes-y-Gómez, M. INAOE's participation at ImageCLEF 2016: Text Illustration Task. Notebook for ImageCLEF at CLEF 2016. Évora, Portugal, September 2016.

- **Supplementary work:** The following works do not involve any treatment of visual information, but rather only text analysis. Although evaluated only in text mining tasks[1], such ideas were developed in parallel with this thesis and culminated in the core basis of the *intra-inter class* methods presented in Chapter 6 and the journal paper *"Distributional Visual-Feature Representations for Image Classification"*. This supplementary work generated the following journal (JCR) and conference papers:

6. **López-Monroy, A.P.**, Montes-y-Gómez, M., Escalante, H.J., Villaseñor-Pineda, L., Stamatatos, E. Discriminative subprofile-specific representations for author profiling in social media. Knowledge-Based Systems. 89 (2015), pp. 134 - 147.
7. Álvarez-Carmona, M. A., **López-Monroy, A. P.**, Montes-y-Gómez, M., Villaseñor-Pineda, L., and Escalante, H. J. INAOE's participation at PAN'15: Author Profiling task. Notebook for PAN at CLEF 2015. Toulouse, France, on September 8-11, 2015. CEUR-WS.org. ISSN 1613-0073. *Best overall performance of 22 participating universities.*
8. **López-Monroy, A. P.**, Montes-y-Gómez, M., Escalante, H. J., and Villaseñor-Pineda, L. Using Intra-Profile Information for Author Profiling. Notebook for PAN at CLEF 2014. Sheffield, UK, on September 15-18, 2014. *Best overall performance of 11 participating universities.*
9. **López-Monroy, A. P.**, Montes-y-Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., and Villatoro-Tello, E. INAOE's participation at PAN'13: Author Profiling task. Notebook for PAN at CLEF 2013. Valencia, España, September 2013. *Best overall performance of 21 participating universities.*

[1] In particular, for the Author Profiling task. A document supervised classification task, where each instance corresponds to a document belonging to an author (e.g., a blog), and target classes are, for example, gender and age intervals.

BIBLIOGRAPHY

Abbasi, A., Chen, H., 2008. Writeprints: a stylometric approach to identity-level identification and similarity detection in cyberspace. ACM Transactions on Information Systems 26, Article 7, 29p.

Alioscha-Pérez, M., Sahli, H., González, I., Taboada-Crispi, A., 2012. Sparse and non-sparse multiple kernel learning for recognition. Computación y Sistemas 16 (2), 167–174.

Bekkerman, R., Allan, J., 2004. Using bigrams in text categorization. Tech. rep., Department of Computer Science, University of Massachusetts, Amherst.

Ben-Hur, A., Noble, W. S., 2005. Kernel methods for predicting protein–protein interactions. Bioinformatics 21 (suppl 1), i38–i46.

Boiman, O., Shechtman, E., 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. pp. 1–8.

Bosch, A., Zisserman, A., Muñoz, X., 2006. Scene classification via plsa. In: Computer Vision–ECCV 2006. Springer, pp. 517–530.

Breiman, L., 1996. Bagging predictors. Mach Learn 24, 123–140.

Brown, G., Wyatt, J., Harris, R., Yao, X., 2005. Diversity creation methods: a survey and categorisation. Inf. Fusion 6, 5–20.

Bruni, E., Boleda, G., Baroni, M., Tran, N.-K., 2012. Distributional semantics in technicolor. In: ACL. ACL, pp. 136–145.

Bruni, E., Tran, N.-K., Baroni, M., 2014. Multimodal distributional semantics. J. Artif. Intell. Res.(JAIR) 49, 1–47.

Cao, Y., Wang, C., Li, Z., Zhang, L., Zhang, L., 2010. Spatial-bag-of-features. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, pp. 3352–3359.

Chang, C.-C., Lin, C.-J., 2011. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1–27:27.

Chávez, R. O., Montes, M., Sucar, L. E., 2011. Using a markov random field for image re-ranking based on visual and textual features. Computación y Sistemas 14 (4), 393–404.

Chen, G., Yang, J., Jin, H., Shechtman, E., Brandt, J., Han, T. X., 2015. Selective pooling vector for fine-grained recognition. In: 2015 IEEE Winter Conference on Applications of Computer Vision. IEEE, pp. 860–867.

Clarkson, P., Rosenfeld, R., 1997. Statistical language modeling using the cmu-cambridge toolkit. In: Proceedings of EUROSPEECH. Vol. 97. International Speech Communication Association Rhodes,, Greece, pp. 2707–2710.

Cortes, C., Mohri, M., Rostamizadeh, A., 2009. Learning non-linear combinations of kernels. In: Advances in neural information processing systems. pp. 396–404.

Cortes, C., Mohri, M., Rostamizadeh, A., 2010. Two-stage learning kernel algorithms. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10). pp. 239–246.

Cruz-Roa, A., Caicedo, J. C., González, F. A., 2011a. Visual pattern mining in histology image collections using bag of features. Artificial intelligence in medicine 52, 91–106.

Cruz-Roa, A., Díaz, G., Romero, E., 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.

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., Bray, C., 2004. Visual categorization with bags of keypoints. In: International Workshop on Statistical Learning in Computer Vision, ECCV. Vol. 1. p. 22.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R., 1990. Indexing by latent semantic analysis. Journal of the American society for information science 41 (6), 391.

Demšar, J., 2006. Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research 7, 1–30.

Díaz, G., Romero, E., 2012. Micro-structural tissue analysis for automatic histopathological image annotation. Microscopy Research and Technique 75, 343–358.

Dumais, S. T., 2004. Latent semantic analysis. Annual review of information science and technology 38 (1), 188–230.

Escalante, H. J., Sucar, L. E., Montes-y Gómez, M., 2012. Semantic cohesion for image annotation and retrieval. Computación y Sistemas 16 (1), 121–126.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J., 2008. LIBLINEAR: A library for large linear classification. Journal of Machine Learning Research 9, 1871–1874.

Fei-Fei, L., Fergus, R., Perona, P., 2007. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Computer Vision and Image Understanding 106 (1), 59–70.

Fei-Fei, L., Perona, P., 2005. A bayesian hierarchical model for learning natural scene categories. In: CVPR. Vol. 2. IEEE, pp. 524–531.

Feng, Y., Lapata, M., 2010. Topic models for image annotation and text illustration. In: NAACL. ACL, pp. 831–839.

Ferraro, F., Mostafazadeh, N., Huang, T.-H., Vanderwende, L., Devlin, J., Galley, M., Mitchell, M., September 2015. A survey of current datasets for vision and language research. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Lisbon, Portugal, pp. 207–213.
URL http://aclweb.org/anthology/D15-1021

Freund, Y., Schapire, R. E., 1996. Experiments with a new boosting algorithm. Machine learning: proceedings of the thirteenth international conference, 325–332.

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

Gavves, E., Fernando, B., Snoek, C. G., Smeulders, A. W., Tuytelaars, T., 2013. Fine-grained categorization by alignments. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1713–1720.

Gavves, E., Fernando, B., Snoek, C. G., Smeulders, A. W., Tuytelaars, T., 2015. Local alignments for fine-grained categorization. International Journal of Computer Vision 111 (2), 191–212.

Gönen, M., Alpaydin, E., 2008. Localized multiple kernel learning. In: Proceedings of the 25th international conference on Machine learning. ACM, pp. 352–359.

Gönen, M., Alpaydın, E., 2011. Multiple kernel learning algorithms. The Journal of Machine Learning Research 12, 2211–2268.

Gosselin, P.-H., Murray, N., Jégou, H., Perronnin, F., 2014a. Revisiting the fisher vector for fine-grained classification. Pattern Recognition Letters 49, 92–98.

Gosselin, P.-H., Murray, N., Jégou, H., Perronnin, F., 2014b. Revisiting the fisher vector for fine-grained classification. Pattern Recognition Letters 49, 92 – 98.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I., 2009. The weka data mining software: An update. SIGKDD Explorations 11.

Jamieson, M., Fazly, A., Dickinson, S., Stevenson, S., Wachsmuth, S., 2007. Learning structured appearance models from captioned images of cluttered scenes. In: ICCV. pp. 1–8.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093.

Joachims, T., 1998. Text categorization with support vector machines: Learning with many relevant features. In: Nédellec, C., Rouveirol, C. (Eds.), Machine Learning: ECML-98. Vol. 1398 of Lecture Notes in Computer Science. Springer Berlin Heidelberg, pp. 137–142.

Joachims, T., 1999. Advances in kernel methods. MIT Press, Cambridge, MA, USA, Ch. Making Large-scale Support Vector Machine Learning Practical, pp. 169–184.

Kanan, C., 2014. Fine-grained object recognition with gnostic fields. In: IEEE Winter Conference on Applications of Computer Vision. IEEE, pp. 23–30.

Khan, R., Barat, C., Muselet, D., Ducottet, C., 2015. Spatial histograms of soft pairwise similar patches to improve the bag-of-visual-words model. Computer Vision and Image Understanding 132 (0), 102 – 112.

Khosla, A., Jayadevaprakash, N., Yao, B., Fei-Fei, L., June 2011. Novel dataset for fine-grained image categorization. In: First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition. Colorado Springs, CO.

Kloft, M., Brefeld, U., Sonnenburg, S., Zien, A., 2010. Non-sparse regularization and efficient training with multiple kernels. Arxiv preprint arXiv 1003 (0079), 186.

Krapac, J., Verbeek, J., Jurie, F., 2011. Modeling spatial layout with fisher vectors for image categorization. In: Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, pp. 1487–1494.

Krause, J., Stark, M., Deng, J., Fei-Fei, L., 2013. 3d object representations for fine-grained categorization. In: Proceedings of the IEEE International Conference on Computer Vision Workshops. pp. 554–561.

Krizhevsky, A., Sutskever, I., Hinton, G. E., 2012. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems. pp. 1097–1105.

Krogh, A., Vedelsby, J., 1995. Neural network ensembles, cross validation and active learning. Adv. Neural Inf. Process Syst. 7, 231–238.

Kuncheva, L. I., 2004. Combining Pattern Classifiers: Methods and Algorithms. Wiley-Interscience.

Landauer, T. K., Foltz, P. W., Laham, D., 1998. An introduction to latent semantic analysis. Discourse processes 25 (2-3), 259–284.

Landauer, T. K., McNamara, D. S., Dennis, S., Kintsch, W., 2013. Handbook of latent semantic analysis. Psychology Press.

Lavelli, A., Sebastiani, F., Zanoli, R., 2004. Distributional term representations: an experimental comparison. In: CIKM. ACM, pp. 615–624.

Lazebnik, S., Schmid, C., Ponce, J., 2005. A maximum entropy framework for part-based texture and object recognition. In: ICCV. Vol. 1. IEEE, pp. 832–838.

Lazebnik, S., Schmid, C., Ponce, J., 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR. Vol. 2. IEEE, pp. 2169–2178.

Lazebnik, S., Schmid, C., Ponce, J., et al., 2004. Semi-local affine parts for object recognition. In: BMVC. pp. 779–788.

Li, Z., Xiong, Z., Zhang, Y., Liu, C., Li, K., 2011. Fast text categorization using concise semantic analysis. Pattern Recognition Letters 32 (3), 441 – 448.

Lin, T.-Y., RoyChowdhury, A., Maji, S., 2015. Bilinear cnn models for fine-grained visual recognition. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1449–1457.

López-Monroy, A. P., Montes-y Gómez, M., Escalante, H. J., Cruz-Roa, A., González, F. A., 2013. Bag-of-visual-ngrams for histopathology image classification. In: IX International Seminar on Medical Information Processing and Analysis. Vol. 8922. SPIE, p. 89220P.

López-Monroy, A. P., Montes-y Gómez, M., Escalante, H. J., Villaseñor-Pineda, L., Stamatatos, E., 2015. Discriminative subprofile-specific representations for author profiling in social media. Knowledge-Based Systems 89, 134 – 147.

Lowe, D. G., 2004. Distinctive image features from scale-invariant keypoints. International journal of computer vision 60 (2), 91–110.

López-Monroy, A. P., y Gómez, M. M., Escalante, H. J., Cruz-Roa, A., González, F. A., 2015. Improving the bovw via discriminative visual n-grams and mkl strategies. Neurocomputing, Accepted.

Maimon, O., Rokach, L., 2002. Improving supervised learning by feature decomposition. Proceedings of foundations of information and knowledge systems, Salzan Castle, Germany, 178–196.

Maji, S., Kannala, J., Rahtu, E., Blaschko, M., Vedaldi, A., 2013. Fine-grained visual classification of aircraft. Tech. rep.

McCallum, A., Nigam, K., et al., 1998. A comparison of event models for naive bayes text classification. In: AAAI-98 workshop on learning for text categorization. Vol. 752. Citeseer, pp. 41–48.

Miranda-García, A., Calle-Martín, J., 2005. Yule's k characteristic k revisited. Language Resources and Evaluation 39, 287–294.

Nguyen, D., Gravel, R., Trieschnigg, D., Meder, T., 2013. How old do you think i am?: A study of language and age in twitter. In: Seventh International AAAI Conference on Weblogs and Social Media.

Pang, B., Lee, L., 2008. Opinion mining and sentiment analysis. Foundations and trends in information retrieval 2 (1-2), 1–135.

Pavelec, D., Justino, E., Batista, L. V., Oliveira, L. S., 2008. Author identification using writer-dependent and writer-independent strategies. In Proceedings of the 2008 ACM Symposium on Applied Computing - SAC08, 414–418.

Phan, X.-H., Nguyen, L.-M., Horiguchi, S., 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In: WWW. WWW '08. pp. 91–100.

Plakias, S., Stamatatos, E., 2008. Tensor space models for authorship attribution. In Proc. of the 5th Hellenic Conference on Artificial Intelligence (SETN'08), LNCS 5138, 239–249.

Powers, D. M., 2011. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation.

Provost, F. J., Kolluri, V., 1999. A survey of methods for scaling up inductive learning algorithms. Proceeding of 3rd international conference on knowledge discovery and data mining 3, 131–139.

Qiu, S., Lane, T., 2009. A framework for multiple kernel support vector regression and its applications to sirna efficacy prediction. Computational Biology and Bioinformatics, IEEE/ACM Transactions on 6 (2), 190–199.

Quack, T., Ferrari, V., Leibe, B., Van Gool, L., 2007. Efficient mining of frequent and distinctive feature configurations. In: ICCV. IEEE, pp. 1–8.

Rakotomamonjy, A., Bach, F., Canu, S., Grandvalet, Y., 2008. Simplemkl. Journal of Machine Learning Research 9, 2491–2521.

Rokach, L., 2009. Ensemble-based classifiers. Artificial Intelligence Review 33, 1–39.

Sahlgren, M., 2008. The distributional hypothesis. Italian Journal of Linguistics 20 (1), 33–54.

Schler, J., Koppel, M., Argamon, S., Pennebaker, J., 2006. Effects of age and gender on blogging. In: Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs. pp. 199–205.

Sebastiani, F., 2008. Machine learning in automated text categorization. ACM Computer Surveys 34 (1), 1–47.

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

Sivic, J., Zisserman, A., 2004. Video data mining using configurations of viewpoint invariant regions. In: CVPR. Vol. 1. IEEE, pp. I–488.

Solorio, T., Pillay, S., Raghavan, S., Montes-y Gómez, M., 2011. Modality specific meta features for authorship attribution in web forum posts. In Proceedings of the 5th International Joint Conference on Natural Language Processing, 156–164.

Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H., Demirbas, M., 2010. Short text classification in twitter to improve information filtering. In: SIGIR. SIGIR '10. pp. 841–842.

Stamatatos, E., 2008. Author identification: Using text sampling to handle the class imbalance problem. Information Processing and Management 44, 790–799.

Stamatatos, E., 2009. A survey on modern authorship attribution methods. Journal of the American Society for Information Science and Technology 60, 538–556.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A., 2015. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–9.

Tan, C. M., Wang, Y. F., Lee, C. D., 2002. The use of bigrams to enhance text categorization. Information processing and management 38, 529–546.

Tirilly, P., Claveau, V., Gros, P., 2008. Language modeling for bag-of-visual words image categorization. In: CIVR. pp. 249–258.

Tirilly, P., Claveau, V., Gros, P., 2009. A review of weighting schemes for bag of visual words image retrieval. Tech. rep., TEXMEX - INRIA - IRISA.

Tommasi, T., Orabona, F., Caputo, B., 2007. Image annotation task: an svm-based cue integration approach. In: 2007 CLEF Workshop.

Tumer, K., Ghosh, J., 1996. Error correlation and error reduction in ensemble classifiers. Connection science, special issue on combining artificial neural networks: ensemble approaches 8, 385–404.

Turney, P., P., P., 2010. From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research 37, 141–188.

Varma, M., Babu, B. R., 2009. More generality in efficient multiple kernel learning. In: Proceedings of the 26th Annual International Conference on Machine Learning. ACM, pp. 1065–1072.

Wang, H., Ullah, M. M., Klaser, A., Laptev, I., 2009. Evaluation of local spatio-temporal features for action recognition. In: BMVC. pp. 1–11.

Wang, S., 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. pp. 90–94.

Wilcoxon, F., 1945. Individual comparisons by ranking methods. Biometrics Bulletin 1 (6), pp. 80–83.

Wolpert, D. H., 1992. Stacked generalization. Neural Networks 5, 241–259.

Wu, L., Li, M., Li, Z., Ma, W. Y., Yu, N., 2007. Visual language modeling for image classification. In: ACM Proceedings of the international workshop on Workshop on multimedia information retrieval. pp. 115–124.

Xu, Z., Jin, R., Yang, H., King, I., Lyu, M. R., 2010. Simple and efficient multiple kernel learning by group lasso. In: Proceedings of the 27th international conference on machine learning (ICML-10). pp. 1175–1182.

Yang, S., Bo, L., Wang, J., Shapiro, L. G., 2012. Unsupervised template learning for fine-grained object recognition. In: Advances in Neural Information Processing Systems. pp. 3122–3130.

Yuan, J., Wu, Y., Yang, M., 2007. Discovery of collocation patterns: from visual words to visual phrases. In: CVPR. pp. 1–8.

Yuan, J., Yang, M., Wu, Y., 2011. Mining discriminative co-occurrence patterns for visual recognition. In: CVPR. IEEE, pp. 2777–2784.

Zhang, J., Marszalek, M., Lazebnik, S., 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.

Zheng, Q. F., Wang, W., Gao, W., 2006. Effective and efficient object-based image retrieval using visual phrases. In: ACMMM. pp. 77–80.

Zheng, Y.-T., Zhao, M., Neo, S.-Y., Chua, T.-S., Tian, Q., 2008. Visual synset: towards a higher-level visual representation. In: CVPR. IEEE, pp. 1–8.

Appendices

APPENDIX A

SUPPLEMENTARY RESULTS

The aim of this appendix is to present the experimental results in Chapter 5 using the F-Measure. In that chapter we reported experimental results using the accuracy to directly compare the LSA and LDA baselines, which have reported results in some of the target collections. The experiments in Chapter 5 showed the usefulness of the DTRs exploiting visual words and visual n-grams as terms, by using the classifier in the LibLINEAR framework (Fan et al., 2008). In this regard, in Tables A.1 and A.2 we present F-Measure results that confirm the main conclusions obtained in Chapter 5. First, in Table A.1 we can observe that independently of using DCT or SIFT descriptors, most of distributional representations achieve better performances than the standard BoVW. Second, in in Table A.2 we can also observe that visual bigrams (which are local structural visual patterns), under the DTR formulation, improve the performance of the classifier, particularly the proposed GOR and SOR.

F-Measure Performances using visual words						
			SIFT			
Method	Hist.	6-Caltech	Butterflies	Birds	Scenes	avg.
BoVW	74.18	64.61	50.62	50.00	69.51	61.78
LDA	55.74	56.31	29.03	36.20	60.23	47.50
LSA	76.66	63.42	48.41	51.19	72.59	62.45
VCOR	71.25	66.57	48.88	50.03	68.87	61.12
IOR	78.63	**68.91**	50.27	52.04	69.19	63.80
COR	67.49	60.91	41.97	40.21	68.49	55.81
SOR	78.21	67.52	51.47	53.41	**73.31**	64.78
GOR	**80.15**	67.09	**53.78**	**55.64**	70.28	**65.38**
			DCT			
Method	Hist.	6-Caltech	Butterflies	Birds	Scenes	avg.
BoVW	85.81	53.62	61.33	50.43	51.72	60.58
LDA	75.31	39.80	27.65	31.21	34.99	41.79
LSA	86.82	39.01	40.65	38.62	52.57	51.53
VCOR	80.28	54.67	62.51	52.47	51.19	60.22
IOR	86.71	56.54	64.40	56.00	51.14	62.95
COR	74.41	44.70	40.09	37.14	48.25	48.91
SOR	88.15	**57.21**	66.64	57.19	**52.21**	64.28
GOR	**90.08**	56.98	**67.37**	**59.71**	51.78	**65.18**

Table A.1: F-Measure: Performances for the evaluated strategies using a dictionary of visual words generated using SIFT-based and DCT-based visual words.

F-Measure performances using visual words + visual n-grams						
			SIFT			
Method	Hist.	6-Caltech	Butterflies	Birds	Scenes	avg.
BoVW	74.18	64.61	50.62	50.00	69.51	61.78
LDA	68.64	57.10	30.51	33.63	66.02	51.18
LSA	74.12	63.74	48.90	52.21	75.11	62.81
BoVB	79.81	70.02	51.47	56.37	76.40	66.81
SPR	74.34	71.54	51.48	55.71	**80.80**	66.77
VCOR	78.57	73.81	48.97	55.41	71.47	65.64
IOR	85.04	**76.01**	52.51	58.01	73.33	68.98
COR	78.01	59.67	42.52	45.66	71.95	59.56
SOR	83.25	75.41	52.47	59.74	80.51	70.27
GOR	**85.79**	75.49	**53.17**	**60.05**	78.41	**70.58**
			DCT			
Method	Hist.	6-Caltech	Butterflies	Birds	Scenes	avg.
BoVW	85.81	53.62	61.33	50.43	51.72	60.58
LDA	74.13	40.44	28.10	32.22	38.79	42.73
LSA	88.55	38.84	42.33	39.23	57.81	53.35
BoVB	90.55	55.84	61.09	50.42	53.26	62.23
SPR	79.05	56.14	62.00	54.37	58.43	61.99
VCOR	85.47	53.18	59.57	58.16	52.17	61.71
IOR	91.81	**58.92**	69.00	68.10	56.60	68.88
COR	84.95	41.84	40.43	35.14	53.72	51.21
SOR	92.18	58.27	68.54	66.07	**59.24**	68.86
GOR	**94.61**	57.47	**70.05**	**68.24**	55.09	**69.09**

Table A.2: F-Measure: Performances for the evaluated strategies using a dictionary of visual words + visual bi-grams generated using SIFT-based and DCT-based visual words.

APPENDIX B

FINE GRAIN CLASSIFICATION RESULTS

The aim of this appendix is to enrich the evidence about the usefulness of the proposed strategies in specific domains. In Chapter 4 and Chapter 5, some of the most important improvements in results where observed in narrow domains, where the target classes belongs to a more general category (e.g., Birds, Butterflies and Histopathology datasets). Thus, it is promising to find similar problems and applications with this narrow particularity. Recently, the computer vision community have been interested in a similar problem: the *Fine Grained Classification* (FGC). From a general perspective, the main goal of FGC is to classify at a finer level of granularity. In this regard, there are a number of image classification problems that meet this property, for example the identification of fish species, dog breeds, car models, aircraft manufacturers, etc. The FGC is a challenging problem because the target classes often share a similar appearance, which can only be discriminated based on subtle specific details. For example, the discrimination between the different versions of Boeing 747 aircraft can only be possible by counting the windows (Gosselin et al., 2014b).

In this research work we begin the study of the proposed methods in FGC tasks. For this purpose, we perform an experimental evaluation by using the following three fine-grain datasets in the literature:

- **FGVC-Aircraft (Maji et al., 2013):** This dataset contains 10,000 images of aircraft. There are 100 images for each of the 100 different aircraft model variants, most of which are airplanes. We show some sample images in Figure B.1.

- **Stanford BMW-10 (Krause et al., 2013):** This dataset is a small, ultra-fine-grained set of 10 BMW sedans (512 images) hand-collected by the authors. This dataset provides the specific train and test images. We show some sample images in Figure B.2.

- **Stanford Dogs (Khosla et al., 2011):** This is a challenging and large-scale collection specially suited for fine-grained image categorization. This dataset comprises more than 22,000 annotated images of dogs belonging to 120 species. We show some sample images in Figure B.3.

Figure B.1: Three images of different aircraft classes in the FGVC-Aircraft: (a) Boeing 737-200, (b) Boeing 737-300 and (c) Boeing 737-400.

Figure B.2: Three images of different bmw cars in the BMW-10 dataset.

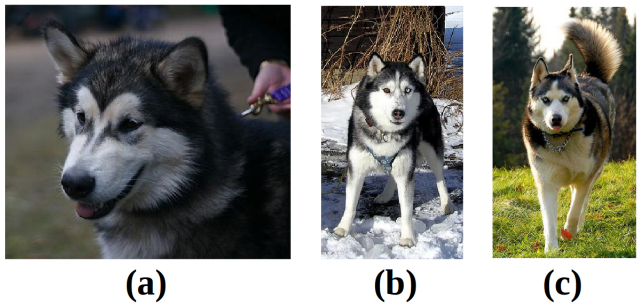

Figure B.3: Three images of different dog breeds in the Stanford Dogs dataset: (a) Malamute, (b) Husky and (c) Eskimo.

The experimental settings for the proposed methods and baselines of this thesis are the same than those presented in Section 5.4. Furthermore, we also introduce two new reference methods based in Convolutional Neural Networks (CNN). Recently, CNNs have shown outstanding performances in a number of image classification problems, specially those involving object recognition. In this regard, we use two of the most successfully models referenced in the literature: AlexNet (Krizhevsky et al., 2012) and GoogLeNet (Szegedy et al., 2015). For this purpose, we use the implementations in the popular Caffe Deep Learning Framework (Jia et al., 2014). In our experimental evaluation for CNN we maintain the structural elements of the algorithms (e.g., layers) provided in Caffe (Jia et al., 2014), but in validation folds we explore

some low level parameter of each model (e.g., number of epochs, learning rate, weight decay, momentum, etc.). We present experimental results for the following variants/strategies in deep learning:

- **relu-7:** This strategy consists in representing the target images by using as features the *relu-7* activation layer of the proposed model in (Krizhevsky et al., 2012). Then, it is possible to use a regular classifier, which in our case is the SVM in LibLINEAR framework (Fan et al., 2008).
- **scratch:** This strategy consists in the random initialisation of all elements in the CNN, then learning from scratch the appropriated model for the task.
- **fine-tuned:** This strategy consists in two steps. First, we pre-initialize the CNN with ImageNet-trained weights (Krizhevsky et al., 2012). Second, the learning rate of the CNN for the last layer (target classes in our FGC) is normal, but the rest of the layers learn at a diminished rate.
- **full-train:** This strategy consists in two steps. First, we pre-initialize the CNN with ImageNet-trained weights (Krizhevsky et al., 2012). Second, we train the entire network at a normal rate.

The experimental results show interesting evidence of the proposed methods in this research work. In Table B.1 we can observe that most of the main conclusions of this work are also valid for this fine-grained domains. For example, we can conclude that visual bigrams (BoVB) are a better option than only visual words (BoVW). We can also observe that, the Image Occurrence Representation (IOR) outperforms traditional strategies such as BoVW and Latent Semantic Analysis (LSA). More importantly, the best results were obtained by the two proposed distributional strategies SOR and GOR. Regarding to the deep learning strategies, it can be seen that none of the variants achieve important results. In fact, in some cases CNN were outperformed by the typical Bag-of-Visual Words. Our experimental results, reinforce the evidence showed by a number of works in the literature, which expose the problems of CNN in fine grained domains (Gosselin et al., 2014a; Gavves et al., 2015; Lin et al., 2015). It is worth noting that most of the methods in the state-of-the-art for FGC relies on SIFT features, HOG based features, Fisher Vectors and Spatial Pyramid Representation (Chen et al., 2015; Kanan, 2014; Gavves et al., 2015, 2013; Yang et al., 2012).

Accuracy performances in FGC - DTRs exploiting visual ɲ-grams				
	SIFT			
Method	Aircraft	Cars	Dogs	avg.
Rand	1.43	0.52	0.81	.92
BoVW	45.21	30.25	33.10	36.18
LDA	32.27	21.81	20.09	24.72
LSA	49.21	41.19	34.45	41.61
BoVB	50.01	42.23	**35.54**	42.59
SPR	**55.41**	**44.80**	35.21	**45.14**
CaffeNet relu-7	27.23	20.11	17.10	21.48
CaffeNet scratch	1.75	0.61	1.82	1.39
CaffeNet fine-tuned	41.01	24.73	34.02	33.25
CaffeNet full-train	45.69	24.56	33.58	34.61
GoogLeNet scratch	27.21	11.54	9.18	15.97
GoogLeNet fine-tuned	**48.94**	33.91	**35.21**	39.35
GoogLeNet full-train	47.64	**36.56**	34.43	**39.54**
VCOR	39.66	33.45	21.21	31.44
IOR	**53.31**	**51.24**	**36.51**	**47.02**
COR	44.41	30.21	33.21	35.94
SOR	58.76	54.23	**40.54**	51.17
GOR	**59.41**	**55.14**	39.15	**51.23**

Table B.1: Performances for the evaluated strategies using a dictionary of visual words + visual bi-grams generated using SIFT-based visual words.