Semantics and statistics for automated image annotation

Thesis

Semantics and Statistics for Automated Image Annotation

Ainhoa Llorente Coto

A Thesis presented for the degree of

Doctor of Philosophy

Knowledge Media Institute
The Open University
England

20 July 2010

In Memoriam

Lola Sagastibelza, my grandmother.

Semantics and Statistics for Automated Image Annotation

Ainhoa Llorente Coto

Submitted for the degree of Doctor of Philosophy

2010

Abstract

Automated image annotation consists of a number of techniques that aim to find the correlation between words and image features such as colour, shape, and texture to provide correct annotation words to images. In particular, approaches based on Bayesian theory use machine-learning techniques to learn statistical models from a training set of pre-annotated images and apply them to generate annotations for unseen images. The focus of this thesis lies in demonstrating that an approach, which goes beyond learning the statistical correlation between words and visual features and also exploits information about the actual semantics of the words used in the annotation process, is able to improve the performance of probabilistic annotation systems. Specifically, I present three experiments. Firstly, I introduce a novel approach that automatically refines the annotation words generated by a non-parametric density estimation model using semantic relatedness measures. Initially, I consider semantic measures based on co-occurrence of words in the training set. However, this approach can exhibit limitations, as its performance depends on the quality and coverage provided by the training data. For this reason, I devise an alternative solution that combines semantic measures based on knowledge sources, such as WordNet and Wikipedia, with word co-occurrence in the training set and on the web, to achieve statistically significant results over the

baseline. Secondly, I investigate the effect of using semantic measures inside an evaluation measure that computes the performance of an automated image annotation system, whose annotation words adopt the hierarchical structure of an ontology. This is the case of the ImageCLEF2009 collection. Finally, I propose a Markov Random Field that exploits the semantic context dependencies of the image. The best result obtains a mean average precision of 0.32, which is consistent with the state-of-the-art in automated image annotation for the Corel 5k dataset.

Declaration

The work in this thesis is based on research carried out at the Knowledge Media Institute, The Open University, England. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless explicitly stated in the text.

Acknowledgements

First, I would like to thank the Spanish organisations that financially supported this thesis. In particular, many thanks to Robotiker for funding me during the first two years of my PhD and Santander Universities for the last two.

Second, I would like to thank Asun Gómez-Pérez for her guidance before starting my PhD. Then, to my supervisors Stefan Rüger and Enrico Motta for their help. Special thanks to Manmatha, whose visit to KMi in the summer of 2009 marked a milestone in my PhD. Thanks for all the exhausting sessions at the whiteboard and for being so patient. Also many thanks to Stefanie Nowak for the good time spent together during our collaboration. Finally, thanks to my examiners Prof Anne De Roeck from the OU and Prof Joemon Jose from University of Glasgow for their positive attitude during my viva.

Third, from a personal perspective, this thesis would not have been possible without the support of all the nice people who were or still are in the Knowledge Media Institute (Jorge, Michele, Ivana, Vane, Carmencita and many others) and especially the people of my group. In particular, I would like to remember the following colleagues: Qiang Huang for his kindness, Sanyukta Shrestha for his enthusiasm, Rui Hu for her freshness, Anuj Panwar for his good heart, Serge Zagorac for being such a great colleague, and finally, Adam Rae for being the perfect English gentleman.

The story of this thesis goes back to Spain in 2006 when I was working in Robotiker and I started thinking about starting this crazy adventure. When I told my friends about the idea that was crossing my mind, many were unable to understand my motivations. However, I was very soon supported by all of them. Each one of them helped me so much during those difficult months in which I was getting ready to reduce my life to a shambles. Nerea, who was initially not very happy about the adventure, but came twice with Luis to check that I was doing fine. Sonia del Rio and Sonia Martinez who surprised me appearing out of nowhere when I was in the ferry queue, exactly at the moment when I was starting to feel very sad. Also, thanks to all those people who write to me periodically giving me strong support like Ana Ordóñez and Olga Navalón among others. Finally, it comes my family, I would like to thank my parents, sisters Paula and Marina, brothers-in-law, and Daniel and Cloe for supporting me during all these difficult years. My grandmother, Lola Sagastibelza, to whom this thesis is dedicated and who died in November 2008, at the age of 97 years. Thanks for being there all my life. Of course, I would not forget my two favourite cousins, Ana and Daniel, for giving me all the practical help for settling down in the UK, and sharing with me the secret "walk" alongside the south bank of the Thames. Special mention to my parents, Victor and Charo, who have always being there for me and whom I admire most for their capacity to enjoy life. Of course, Ringo, our dog, who died this summer and will always be special.

I arrived in Portsmouth after almost dying from seasickness in the "Pride of Bilbao" the last day of September in 2006. Like many of my fellow students, I had the usual initial shock after realising what Milton Keynes was like for real and had to deal with serious practical problems (like coping with unstable landlords) but very soon I found my way here. The first day in KMi I met Sofia whose arrival was the most similar

I have seen in my life to a real movie star appearance and I fell in love with her at first sight. Later on, Annalisa, who was like a ray of light, joined us and we created an extremely strong supporting therapy group that helped us during the difficult moments of the past four years. Girls, there is no need to remind you how important you have been to me all these years, and hopefully the following. Of course, Carlos Pedrinaci also occupies a special place for all his wise advices.

Last but not least, the story of this thesis has been the story of my relationship with Davide. He started being "yet another Italian in KMi" but soon he became a good friend with whom (alone with Nicola Gessa) I travelled around Perú and fulfilled one of the dreams of my life, visiting Machu Picchu!! Soon, he became my boyfriend, then my husband, and, in some months, he will be the father of my first child. Davide, you know already how much I love you but anyway, I would like to thank you for all the help provided during all these years, especially for your patience, common sense, maturity, and of course, love, which were invaluable in the numerous moments of crisis. Also, I would like to mention my new Italian family, who, despite the fact that the communication could be thought of as difficult, as I still do not talk Italian, have accepted me with great joy and open-mindedness since the very first day. A special kiss to my mother-in-law, Mimma.

As I have spent the last four years focussed on my PhD, the only moments in which I have been able to breath a bit of fresh air have been during my summer trips. The first summer, as mentioned before, was Perú!! The second, Davide and I lived a scary adventure crossing the American border in San Diego to attend a wedding in Tijuana. Unforgettable experience!!! We ended up living in the middle of a road movie visiting California, some southern States and San Francisco!!! The third summer, now as husband and wife, we visited the exotic and nice south of India. Thanks a lot

to Dnyanesh Rajpathak and Mugdha Karve for your hospitality and good moments shared!!!

There are lots and lots of people to mention such as Monia and Guido, my Oxley Park flatmates: Sofia, Manuela, and Carlo, all the rest of friends from Bilbao, Chiara, Rubén and his visits to Cranfield, Aneta and all the administrative people in KMi, my friends from school, the rest of my family, my cousins, friends from my first degree, friends from previous jobs, my Athens flatmates, and especially, all the people who once were close to me but due to life circumstances we have lost contact.

Finally, I would like to finish saying that as a whole the PhD was a great life experience and as such, very tough at times, but without any doubt deserved to be fully experienced.

Contents

List of Figures

List of Tables

List of Algorithms

Chapter 1

Introduction

The main objective of this chapter is to introduce the rationale behind this thesis. Section 1.1 provides a brief overview of the field of automatic image annotation. In particular, it analyses the terminology, the steps that an automated image annotation algorithm is made of, the strategies followed for producing the output, the different ways of approaching a model, how to classify these approaches, how to perform an evaluation, and which datasets are considered a benchmark in the field. Then, Section 1.2 discusses the motivation that originates this dissertation. This is followed by a review on classic probabilistic methods (Section 1.3) and by a failure analysis of an algorithm (Section 1.4), which belongs to this group. This helps to compile a set of requirements, which facilitates the formulation of the research questions in Section 1.5. Finally, the chapter concludes with a summary of the contributions provided by this thesis together with a description of how the thesis is structured in Section 1.6 and Section 1.7, respectively.

1.1 Overview of Automated Image Annotation

Automated image annotation aims to create a computational model able to assign terms to an image in order to describe its content. It provides an alternative to the time-consuming work of manually annotating large image collections. In the literature, it is referred, among others, as *image auto-annotation*, *automatic photo tagging*, *automatic image indexing*, *image auto-captioning*, and *multi-label image classification*. Initially, it was known as *object recognition* (Forsyth and Ponce 2003) and it was approached in the context of computer vision research area. However, it became very soon an independent area of study.

The starting point for most automated annotation algorithms is a training set of images that have already been annotated by a human annotator. The annotations are unstructured textual metadata made up of simple keywords that describe the content depicted in the image. In this thesis, the nomenclature used to refer to these keywords will be indistinctly annotation words, annotations, words, or terms. Alternatively, some authors speak of labels, tags, captions, footnotes, or semantic classes.

Most automated image annotation systems follow a simple process, which is characterised by three steps:

1. Image analysis techniques are used to extract features from the image pixels, such as colour, texture and shape. Features are obtained from either the whole image, *global* or *scene-oriented approach*, or from segmented parts, *segmentation approach*, such as *blobs*, which are irregularly shaped areas of connected pixels, or *tiles*, which are non-overlapping equally-sized rectangles.

2. Models that link the image features with the annotation terms are built.

3. The same feature information is extracted from unseen images in order to assess

the validity of the models generated at the previous step to produce a probability value associated to each image.

Several strategies can be adopted to produce the final output for these systems. One of them consists of an array of ones and zeros, with the same length as the number of terms in the vocabulary, which indicates the presence or absence of the different objects represented by the terms in the image. This is called *hard annotation* in contrast with *soft annotation*, which provides a probability score that gives some confidence for each word being present or absent in the image. Alternatively, other annotation frameworks consider *threshold-based annotations*; this strategy forces all the keywords with a probability value greater than the threshold to be considered as the final annotations. Authors such as Jin et al. (2004) propose an algorithm with *flexible annotation length*, which creates automatically annotations of different length for each image. However, most of automated image annotation frameworks implement a strategy that assumes a *fixed annotation length*. For instance, if the length of the annotation is k, the words with the top-k largest probability values are selected as annotations. In this case, it is essential to decide beforehand what annotation length is appropriate as the number of words in the annotation has a direct influence on the performance of the system. In general, shorter annotations would lead to higher precision (a measure of exactness) and lower recall (a measure of completeness). Consequently, short annotations might be more adequate for a casual user, more interested in quickly finding some relevant images without examining too much information. On the other hand, a professional user may be interested in higher recall and thus may need longer annotations. Nevertheless, most researchers have adopted a compromise that considers five as de facto annotation length.

Independently of the method used to define the annotations, automated image an-

notation systems generate a set of words that helps to understand the scene represented in the image.

From the theoretical point of view, the problem of annotating images can be formulated in two ways: one as a *direct retrieval* model and the other as *image annotation* model. Thus, the task of image retrieval is similar to the general ad-hoc retrieval problem. Given a text query $Q = \{w_1, \ldots, w_k\}$ and a collection of images denoted as T, the objective is to retrieve those images that contain objects described by the words $w_1, \ldots, w_k$, which implies ranking the images by the likelihood of their relevance to the query. On the other hand, the annotation model is formulated as follows: Given a collection of images denoted as T, the goal is to generate the set of words, $w_1, \ldots, w_k$, that best describe each one of them.

From a technical point of view, automated image annotation is formulated using machine learning techniques that, in turn, are based on statistical and probabilistic theories. There are different ways to categorise these techniques. Each categorisation represents a specific branch of machine learning methodologies that stem from different assumptions and philosophies and aim at different problems. However, these categorisations are not mutually exclusive. As a consequence, many machine learning applications fall into multiple categories simultaneously. According to Gong and Xu (2007), the following categories apply: supervised vs. unsupervised; generative models vs. discriminative models; models based on simple data vs. models based on complex data; models based on identification vs. models based on prediction. This classification will be partially followed when reviewing the classic probabilistic annotation methods (Section 1.3).

After completing the design of an automated image annotation algorithm, the next course of action is to measure its performance. Section 3.2 analyses several strategies

that accomplish the evaluation of the performance of annotation algorithms. All of them imply a comparison between the generated annotation words and the ground-truth provided by the human annotator. The main differences reside in the computation of the matches between these two sets. Additionally, some benchmark datasets are proposed in Section 3.7. The main benefit of these datasets is that they facilitate not only the comparison of results between algorithms but also the growth and expansion of the research undertaken in the field.

With respect to how to categorise automated image annotation methods, there is no common agreement in the literature as almost every author follows a different argumentation. The most widespread classification criteria are the following: Authors such as Srikanth et al. (2005), and Wang et al. (2006) divide them into classification and probabilistic-based methods. Others such as Liu et al. (2006) classify them into three categories: graph-based models, classification models, and probabilistic models.

1.2 Motivation

According to Manjunath et al. (2002), it has never been so easy to create multimedia content as it is nowadays. This is mainly due to the fact that digital cameras have become increasingly affordable, the new generation of mobile telephones integrate a digital camera, and the widespread use of personal computers with hundreds of gigabytes of storage space. This has converted almost each individual into a potential content producer, capable of producing content that can be easily distributed and published. The initial tendency of this digital content was to remain inaccessible as people kept their digital collections in their personal computers. However, Internet soon favoured the content to be fully accessible online. More recently, and with the advent of online collaborative communities, multimedia sharing through the Internet has become a com-

mon practice. Certainly, the value of this multimedia content resided in its capacity and ability of being discovered. Therefore, the content that could not be easily found, could not be used and had the same value as non-existing content. Thus, one of the most immediate requirements was to create information about this content, metadata, usually in the form of textual annotations. According to a social study conducted by Ames and Naaman (2007), there has been a shift in the practice of annotating images by individuals: from its being nearly avoided for personal off-line collections to its being enthusiastically embraced for online photo sharing communities. These annotations are inherently subjective and their usage is often confined to the application domain that the descriptions were created for. Contrastingly, the scenario in the commercial domain is significantly different. First, the number of commercial applications of annotating images are substantial. To mention a few: mobile applications related to cultural heritage, journalists searching for a specific picture to enrich their articles, digital libraries, medical applications, security applications, broadcast media selection, e-commerce, education, multimedia directory services, etc. Second, a correct image annotation has a direct influence on their revenues and on the efficiency in satisfying the consumers needs. This explains that these companies frequently employ teams of people to manually view each image and assign relevant annotation words to describe their content. However, the process of annotating visual content manually is not scalable to multi-million image libraries. In addition to that, the vast majority of images, particularly those residing on the Internet, have no associated keywords to describe their content. Hence, it is necessary to automatically and objectively describe, index and annotate multimedia information using tools that automatically extract visual features from the content to substitute or complement manual, text-based descriptions. The ultimate goal of annotating images is to allow for the retrieval of images based on natu-

ral language keywords as opposed to alternative content based image retrieval (CBIR) techniques, such as query by sketch or query by example. Finally, all these reasons fully justify the great interest amongst the computer vision and information retrieval community in the development of robust and efficient automated image annotation algorithms.

In particular, the motivation of this thesis originates in the observation of some limitations shared by some classic probabilistic annotation models as Section 1.4 will show. These limitations are the result of generating annotation words individually and independently, without considering that they share the same image context. These limitations may be addressed through the use of semantics of the relationships between annotation words although a precise way of approaching the problem raises several questions, which are formulated in the form of research questions in Section 1.5. Therefore, the focus of this thesis lies in demonstrating that the exploitation of the semantics of words combined with statistical models based on the correlation between words and visual features improve the performance of probabilistic automated image annotation systems.

Additionally, I would like to introduce some technical choices that this thesis makes. With respect to image features, the approach followed is that of global features; consequently, no partition strategy is needed. This is mainly due to two reasons. First, the success of segmented models are highly dependant on the accuracy of image segmentation algorithms. Second, approaches based on simple global features can achieve very good performance, such as demonstrated by Makadia et al. (2008).

Regarding the way the annotations are generated, this thesis follows the *fixed-length* annotation strategy. Specifically, the output of the annotation algorithm is a set of five words with the largest probability value. The reason behind this is because this is de

facto annotation length adopted by most researchers and this favours the comparison of results.

Other choices made, such as the use of the Corel 5k dataset and the use of mean average precision (MAP) as the preferred dataset and metric respectively in all the experiments conducted in this thesis, are as well the result of favouring the comparison of results. However, the Corel 5k dataset is used as a preliminary first evaluation set before doing deeper evaluation with large sets, as being aware of the limitations of the dataset (see Section 3.7.2). With respect to the mean average precision, not only MAP is widely used among the research community but also it has shown to have especially good discrimination and stability among evaluation measures.

1.3 Classic Probabilistic Models

The problem of modelling annotated images has been addressed from several directions in the literature. Initially, a set of generic algorithms were developed with the aim of exploiting the dependencies between image features and implicitly between words. In this thesis, they are denoted as *classic probabilistic models*. These probabilistic approaches use machine learning techniques to learn statistical models from a training set of pre-annotated images and apply them to generate annotations for unseen images using visual feature extracting technology. However, there exist different criteria about how to classify them. One proposed by Kamoi et al. (2007) makes reference to the way the feature extraction techniques treat the image either as a whole, in which case it is called *scene-orientated* or *global approach*, or as a set of regions, which is called *region-based* or *segmentation approach*. The latter implies the use of an image segmentation algorithm to divide images into a number of regions, which can be irregularly shaped *blobs* or equally rectangular *tiles*.

Table 1.1: Classic probabilistic methods for the Corel 5k dataset expressed in terms of number of recalled words (NZR), recall (R), precision (P), F-measure (F), and mean average precision (MAP). Subindices indicate the number of words used in the evaluation. Alternatively, figures with an asterisk indicate that 179 words were employed in the evaluation instead of all possible 260

Model	Author	NZR	R_{260}	P_{260}	F_{260}	F_{49}	MAP
Co-occurrence	Mori et al. (1999)	19	0.02	0.03	0.02	-	-
TM	Duygulu et al. (2002)	49	0.04	0.06	0.05	0.25	-
CMRM	Jeon et al. (2003)	66	0.09	0.10	0.09	0.44	0.17*
CRM	Lavrenko et al. (2003)	107	0.19	0.16	0.17	0.64	0.24*
CRM-Rect	Feng et al. (2004)	119	0.23	0.22	0.22	0.73	0.26
InfNet	Metzler and Manmatha (2004)	112	0.24	0.20	0.22	-	0.25*
InfNet-reg	Metzler and Manmatha (2004)	-	-	-	-	-	0.26*
MBRM	Feng et al. (2004)	122	0.25	0.24	0.24	0.76	0.30
Npde	Yavlinsky et al. (2005)	114	0.21	0.18	0.19	-	0.29*
Mix-Hier	Carneiro and Vasconcelos (2005)	137	0.29	0.23	0.26	-	0.31
LogRegL2	Magalhães and Rüger (2007)	-	-	-	-	-	0.28*
SML	Carneiro et al. (2007)	137	0.29	0.23	0.26	-	0.31
JEC	Makadia et al. (2008)	113	0.40	0.32	0.36	-	0.35
BHGMM	Stathopoulos and Jose (2009)	116	0.21	0.17	0.19	-	-

Others differentiate these algorithms into classification and probabilistic approaches. Classification approaches intend to associate words with images by learning classifiers. Probabilistic based methods attempt to infer the correlations or joint probabilities between images and words. Some examples of classification approaches for automated image annotation are *support vector machine methods*, such as those developed by Cusano et al. (2004), and *linguistic indexing of images*, as proposed by Li and Wang (2003).

However, this thesis only considers probabilistic models and categorises them with respect to the deployed machine learning technique. Thus, they can be divided into *co-occurrence models* (Mori et al. 1999); *generative hierarchical models* (Barnard and Forsyth 2001); *machine translation methods* (Duygulu et al. 2002); *probabilistic latent*

semantic analysis (Monay and Gatica-Perez 2004); *latent Dirichlet allocation* (Blei and Jordan 2003); *relevance models*: *continuous-space relevance model* (Lavrenko et al. 2003), *cross-media relevance model* (Jeon et al. 2003), and *multiple Bernoulli relevance model* (Feng et al. 2004); *inference networks* (Metzler and Manmatha 2004); *non-parametric density estimation* (Yavlinsky et al. 2005); *supervised learning models* (Carneiro and Vasconcelos 2005, Carneiro et al. 2007), and *information-theoretic semantic indexing* (Magalhães and Rüger 2007).

In what follows, the most relevant annotation algorithms will be analysed with a special emphasis on describing the employed methodology. Finally, Table 1.1 will establish a comparison among them in terms of their achieved performance.

The first automated image annotation model, called the *co-occurrence model*, was deployed by Mori et al. (1999), who exploited the co-occurrence information of low-level image features and words. The process first divides each training image into equal rectangular parts *tiles* ranging from 3×3 to 7×7. Features are extracted from all the parts. Each divided part inherits all the words from its original image and follows a clustering approach based on vector quantization. After that, they estimate the conditional probability for each word given a cluster as the equivalent of the number of times a word i appears in a cluster j by the total number of words in that cluster j. The process of assigning words to an unseen image is similar to one carried out on the training data. A new image is divided into parts, features are extracted, the nearest clusters are found for each part and an average of the conditional probabilities of the nearest clusters is calculated. Finally, words are selected based on the largest average value of conditional probability. They tested their approach using a Japanese multimedia encyclopaedia.

Barnard and Forsyth (2001) proposed a *generative hierarchical model*, which is a

hierarchical combination of the asymmetric clustering model that maps documents into clusters, and the symmetric clustering model that models the joint distribution of documents and features. The data is modelled as being generated by a fixed hierarchy of nodes, where the leaves correspond to clusters. Each node in the tree has associated some probability of generating each word, and each node has some probability of generating an image segment with given features. The documents belonging to a given cluster are modelled as being generated by the nodes along the path from the leaf corresponding to the cluster, up to the root node, with each node being weighted on a document and cluster basis. For their experimental procedure, they used different partitions of the Corel dataset (see Section 3.7.1). Later on, Barnard et al. (2001) extended their previous work incorporating statistical natural language processing in order to deal with free text and WordNet to provide semantic grouping information. In this case, they tested their results with a more difficult image collection, 10,000 images of work from the Fine Arts Museum of San Francisco.

Duygulu et al. (2002) improved the *co-occurrence method* of Mori et al. (1999) using a *machine translation model* that is applied in order to translate words into *blobs* in the same way as words from French might be translated into English using a parallel corpus. The dataset used by them, a subset of 5,000 images from the Corel dataset called the Corel 5k dataset, has become a popular benchmark of annotation systems in the literature as discussed in Section 3.7.2.

Monay and Gatica-Perez (2003) introduced *latent variables*[1] to link image features with words as a way to capture co-occurrence information. This is based on *latent semantic analysis* (LSA) (Landauer et al. 1998), which comes from natural language

[1]Latent variables are those that are not directly observed but are rather inferred (through a mathematical model) from other variables that are observed.

processing and analyses relationships between images and the terms that annotate them. The addition of a sounder probabilistic model to LSA resulted in the development of *probabilistic latent semantic analysis* (PLSA) (Monay and Gatica-Perez 2004). Comparison with other algorithms is impeded by their use of a non-standard set of 7,000 Corel images and a set of evaluation measures inherited from the computer vision world.

Blei and Jordan (2003) were one of the first authors to explore the dependence of annotation words on image regions. They generalised the problem of modelling annotated data to modelling data of different types where one type describes the other. For instance, image and their associated annotation words, papers and their bibliographies, genes and their functions. In order to overcome the limitations of the *generative probabilistic models* and *discriminative classification methods*, they proposed a framework that is a combination of both of them. This was culminated in the *latent Dirichlet allocation*, a model that follows the image segmentation approach and finds the conditional distribution of the annotation given the primary type. They used for their experiments a subset of the Corel dataset made up of 7,000 images.

Torralba and Oliva (2003) were the precursors of using global visual features rather than segmented ones. Their *scene-oriented approach* can be viewed as a generalisation of the previous one where there is only one region or partition which coincides with the whole image. It explores the hypothesis that objects and their containing scenes are not independent, learns global statistics of scenes in which objects appear and uses them to predict the presence or absence of objects in unseen images. Consequently, images can be described using basic keywords such as “street”, “buildings”, or “highways”, after using a selection of relevant low-level global filters.

Jeon et al. (2003) improved on the results of Duygulu et al. (2002) by recasting the

problem as cross-lingual information retrieval and applying the *cross-media relevance model* (CMRM) to the annotation task. They, too, utilised the Corel 5k dataset for their experiments as this allowed them to compare their results with other algorithms in a controlled manner. In particular, CMRM used the k-means algorithm to cluster the set of image features to form a visual codebook. However, the multinomial word smoothing mechanism applied by this model was demonstrated later on to be inadequate for image annotation and retrieval, given that many image collections have widely varying annotation lengths per image. A multinomial smoothing model focuses on the prominence of words rather than on the presence of words in the annotation. Clearly, this is highly undesirable. For example, the model will provide an image annotated with the words "person" and "tree" with a preference lower than an image only annotated with the word "person". The word "person" will have a probability of 1/2 in the first image while a probability of 1 in the second image.

To overcome this issue, Lavrenko et al. (2003) developed the *continuous-space relevance model* (CRM) to build continuous probability density functions that describe the process of generating blob features. They showed that CRM surpasses significantly the performance of the CMRM on the task of image annotation and retrieval for the Corel 5k dataset.

Metzler and Manmatha (2004) proposed an *inference network* approach to link regions and their annotations; unseen images can be annotated by propagating belief through the network to the nodes representing keywords.

Feng et al. (2004) used a *multiple Bernoulli relevance model* (MBRM), which outperforms CRM. MBRM differs from the latter in the image segmentation and in the distribution of annotation words. Thus, CRM segments images into *blobs* while MBRM imposes *tiles* on each image. The advantage of this tile approach is that it

reduces significantly the computational expense of a dedicated image segmentation algorithm and provides the model with a larger set of image regions for learning the association between regions and words. Additionally, CRM models annotation words using a multinomial distribution as opposed to MBRM that uses a multiple-Bernoulli distribution. This word smoothing model focuses on the presence or absence of words rather than their prominence as it does in the multinomial case. Finally, authors reported an increase in the performance of 38% over the CRM.

Yavlinsky et al. (2005)[2] followed the approach firstly introduced by Torralba and Oliva (2003) by using *simple global features* together with robust *non-parametric density estimation* and the technique of kernel smoothing. Their results are comparable with the inference network (Metzler and Manmatha 2004) and CRM (Lavrenko et al. 2003) approaches. Their major achievement lies in their demonstration that the Corel 5k dataset, proposed by Duygulu et al. (2002), could be annotated remarkably well just by using global colour information. The CRM model developed by Lavrenko et al. (2003) also utilises a kernel smoothing for image features. However, CRM uses kernel density estimators as part of a generative model that observes a set of blobs in a training image while Yavlinsky et al. (2005) used kernels for estimating densities of features conditional on each annotation word.

Carneiro and Vasconcelos (2005) presented a new method to automatically annotate and retrieve images using a vocabulary of image semantics under a *supervised learning formulation*. The novelty of their contribution resides in a discriminant formulation of the problem combined with a multiple instance learning solution together with a hierarchical description of the density of each image class that enables very efficient

[2]Chapter 4 shows an extension of this work as part of the experimental work undertaken in this thesis.

training. They compared their results with state-of-the-art approaches for the Corel 5k dataset and found their approach to outperform all the existing algorithms.

Magalhães and Rüger (2006) developed a clustering method, which they later integrated into an unique multimedia indexing model for heterogeneous data (Magalhães and Rüger 2007), which presents an *information-theoretic framework* for semantically indexing text, images, and multimedia information. As part of this framework, they proposed semi-parametric models such as Gaussian mixture as an alternative solution to robust methods based on kernel density estimators, such as those proposed by Lavrenko et al. (2003) and Yavlinsky et al. (2005), which are highly computationally expensive. To overcome the usual drawbacks derived from semi-parametric approaches, such as the difficulty of selecting a priori the number of components and of avoiding over-fitting in the training set, Magalhães and Rüger (2007) utilised the *expectation-maximization algorithm*, which allows to select automatically the number of components. Although their performance is slightly inferior to that obtained by Yavlinsky et al. (2005), they succeeded in deploying a solution with higher computational efficiency and greater flexibility. Finally, a point in common with Yavlinsky et al. (2005) is their use of *global image features*.

Makadia et al. (2008) showed that a proper selection of *global features* could lead to very good results for a k-nearest neighbours algorithm for the Corel 5k dataset. In their paper, they presented a complete state-of-art analysis of annotation algorithms for the Corel 5k image collection showing that their approach far outperforms all of them.

More recently, Stathopoulos and Jose (2009) proposed a novel Bayesian hierarchical method for estimating mixture models of Gaussian components, which was called *Bayesian mixture hierarchies model* (BHGMM). To validate their model, they

incorporated it in the supervised learning framework developed by Carneiro and Vasconcelos (2005) and tested it on the Corel 5k dataset. However, Stathopoulos and Jose (2009) did not achieve higher results. The reason behind this might be in the simple approach used by the authors when initialising the mixture models, which differs from the approach followed by Carneiro and Vasconcelos (2005).

To summarise this section, Table 1.1 shows a quantitative comparison of some probabilistic automated image annotation algorithms for the Corel 5k dataset in terms of performance. Only those algorithms tested with the Corel 5k dataset and using standard evaluation metrics are considered. Results are shown under three evaluation measures (see Section 3.2): the annotation metric expressed in terms of the number of recalled words (NZR), recall (R), and precision (P) using 260 words for the evaluation; the F-measure computed with 260 and 49 words, respectively; the rank retrieval metric expressed in terms of mean average precision (MAP) where figures with an asterisk indicate that 179 words were employed (all those that appear more than once in the test set) in the evaluation instead of all possible 260. Results are ordered according to the year that were released. A steady increment can be observed in the performance with respect to the F-measure. The best so far published performance on the Corel 5k dataset corresponds to Makadia et al. (2008) followed by the approaches presented by Carneiro and Vasconcelos (2005) and by Feng et al. (2004). This demonstrates that a careful selection of global features helps to significantly increase the final performance of an annotation algorithm.

1.4 Failure Analysis

The objective of this section is to study and identify categories of misclassification by analysing the output of a classic probabilistic annotation method, in particular, the

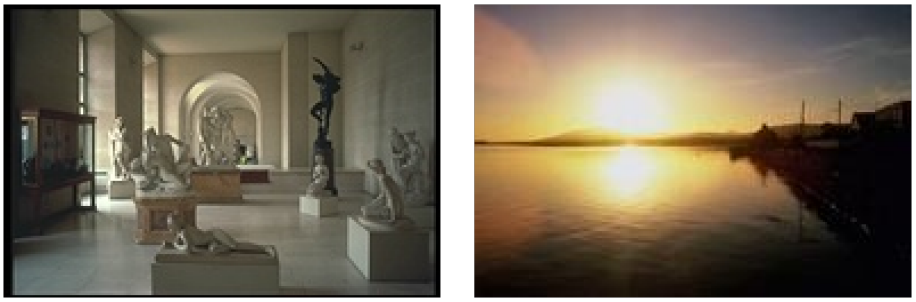

Figure 1.1: Examples of wrong automated annotations for the Corel 5k dataset

non-parametric density estimation algorithm developed by Yavlinsky et al. (2005). I systematically compared the annotations generated by this algorithm with the ground-truth annotations on the Corel 5k dataset (Duygulu et al. 2002), a usual benchmark in the field. I identified four categories of misclassification. For three of them, some possible solutions found in the literature are highlighted while for the fourth one, promising results may be obtained by using semantic relatedness measures.

The first group corresponds to *problems recognizing objects in a scene* as in the examples of Figure 1.1. The scene on the left-hand side represents a museum with some pieces of art in the background and the ground-truth is "art", "museum", and "statue". However, the machine-learning algorithm predicted "snow", "water", "ice", "bear", and "rocks". Note the similarity in terms of colour, shape, or texture existing in the image between the marble floor and a layer of ice; the bronze sculpture and a black bear; or the marble statue and white rocks. On the right-hand side, the scene depicts a sunset on a seashore with some houses in the background. In this case, the human annotator assigned "house", "shore", "sunset", and "water" to the image. However, the annotation algorithm detected "sky", "hills", "dunes", "sand", and "people". Once more, similar visual features might be shared between water and dunes as both of them present a surface with a wave-like texture; the houses in the background and the hills because of an analogous shape; and sunset and sand as they show an equivalent colour.

These problems are a direct consequence of the difficulty in distinguishing visually similar concepts. Duygulu et al. (2002) also identified this limitation although they consider it to be the result of working with vocabularies not suitable for research purposes. In their paper, they made the distinction between *concepts visually indistinguishable* such as "cat" and "tiger", or "train" and "locomotive" in opposition to *concepts visually distinguishable in principle* like "eagle" and "jet". However, the distinction between objects depends heavily on an adequate selection of visual features. Consequently, one way to overcome these limitations is to refine the image analysis parameters of the system. This is how Makadia et al. (2008) achieved very high results for the Corel 5k dataset with a careful selection of images features and using a simple k-nearest neighbour algorithm.

Other inaccuracies come from the *improper use of compound names* in some data collections, being usually handled as two independent words. For instance, in the Corel 5k dataset, the concept "lionfish", *a brightly striped fish of the tropical Pacific having elongated spiny fins*, is annotated with "lion" and "fish". As these words do not appear alone often enough in the learning set, the system is unable to disentangle them. Nevertheless, this problem may be overcome by applying methods for handling compound names such as those proposed by Melamed (1998).

Furthermore, the *over-annotation* problem arises when the ground-truth is made up of fewer words than the generated annotations. Note that many annotation algorithms return a fixed number of words, usually the five highest most probable words. One example is shown on the right-hand side image of Figure 1.2 where the ground-truth is "bear","black", "reflection" and "water", although the annotation system assigns additionally the word "cars". Over-annotation decreases the effectiveness of the image retrieval as it may introduce irrelevant words inside the annotations. However, Jin

Figure 1.2: Inconsistency and improbability appears when there is a lack of cohesion among annotation words

et al. (2004) proposed an algorithm with flexible annotation length in order to avoid this problem.

Finally, the last and most important group of inaccuracies corresponds to different levels of inconsistency among annotation words, which range from the improbability to the impossibility of some objects being together in the real world. This problem is the result of *each annotated word being generated individually and independently* without considering that they are part of the context represented in the image. Figure 1.2 shows examples of different lacks of cohesion among annotation words. For instance, the image on the right-hand side of Figure 1.2 shows the reflection of a black bear on the water. In this case, all generated annotation words ("bear","black", "reflection", "water", "cars") match the ground-truth except the word "cars". Clearly, there exists a certain unlikelihood between the words "cars" and "bear" as their co-occurrence is rare in a real life scenario. Imagine a North Pole scene depicting an animal surrounded by a snowed landscape. No matter how high is the probability value associated to the animal, it is unlikely to be a "camel".

Another illustrative example occurs on the left-hand side image of Figure 1.2, which shows a boat in the water making waves. A human annotator produced the words "boats", "water", and "waves" as ground-truth annotations. However, the annotation algorithm generated the words "water", "desert", "valley", "people" and "street".

Some of them are inconsistent with the others as a "street" might not be found in the "desert", and "water" is not normally seen in a "desert". Moreover, depending on the context the incompatibilities between words vary. To reinforce the understanding of what should be typically seen in each scenario, the Oxford Dictionary of English (Simpson and Weiner 1989) is employed. For instance, "water" is defined as *a colourless, transparent, odourless, tasteless liquid that forms the seas, lakes, rivers, and rain.* According to this, objects that typically appear in a sea, lake, or river scenario are more likely to be found but not a "desert". If, on the other hand, the context corresponds to a "desert", *a dry, barren area of land, that is characteristically desolate, waterless, and without vegetation*, the words that might not belong to the context are "water", "valley", "people", and "street". If we were contemplating a "valley", *a low area of land between hills or mountains, typically with a river or stream flowing through it*, "water" would be perfectly plausible but not words like "desert", "people", and "street".

For the Corel 5k dataset, 17% of the misclassification corresponds to situations where the annotation algorithm under consideration is unable to interpret the image content, while the remaining 83%[3] correspond to images where there exist inconsistencies between annotation words. Note that the over-annotation and the improper use of compounds fall into this latter category as both of them present inconsistent annotations in spite of the different causes that originate each one of them.

As a result of this, a methodology able to overcome this final limitation is needed. Observations deducted from the failure analysis help to elaborate an initial set of requirements. First, an initial identification of the objects contained in the scene should be carried out by an annotation algorithm. Thus, the output of these algorithms are

[3]These figures were obtained from a rough analysis that compared the true annotations with those produced by the algorithm of Yavlinsky et al (2005) for the whole test set.

a set of words that represent the different objects depicted in the image. These words should provide an indication of the context depicted in the image. As these algorithms are prone to errors, it is highly probable that some words are incorrect. Consequently, mechanisms of detecting these errors should be defined and at the same time, ways of acting accordingly. Therefore, according to the image context different *degrees of cohesion or consistency* should be identified between the words as a way of detecting the mistakes created by the algorithm. As seen in the previous examples, one way of achieving this is by referring to the dictionary definition of the involved words or in another words, to their semantics. Additionally, the degree of "closeness" between the words should be measured. Finally, some decision rules that decide what to do when two words are inconsistent should be defined. Nevertheless, the most important requirement is that the algorithm should be able to accomplish the initial interpretation of the scene to a reasonable degree. Otherwise, the processing of nonsensical data will produce a nonsensical output. Thus, this solution depends on an initial annotation stage based on algorithms that exploit the correlation between image features and words. To summarise, in order to go from low-level (visual features) to the high-level features (semantics) of an image, the probability of each entity being present in a given scene should be first estimated and finally, semantic constraints such as relations among entities should be considered.

1.5 Research Questions

The main research question investigated in this thesis is:

How to combine statistical models based on the correlation between words and visual features with information coming from the actual semantics of the words used in the annotation process in order to increase the effectiveness of a probabilistic automated

image annotation system?

This research question is narrowed down to more specific research questions:

- *(i) How to successfully undertake the initial annotation of the scene?*
- *(ii) How to model semantic knowledge in an image collection?*
- *(iii) How to integrate semantic knowledge into the annotation framework?*

In particular, each question can be further explained as follows.

(i) How to successfully undertake the initial annotation of the scene?

An efficient algorithm able to interpret the context depicted by the image is needed. Such an algorithm should take into consideration the correlation between low-level features and annotation words in order to predict the objects that appear in the scene. The undertaking of this process in an effective way is crucial for the whole process as the processing of nonsensical data will produce a nonsensical output.

(ii) How to model semantic knowledge in an image collection? This question naturally evokes another one: *Which knowledge source is to be used?*

The main problem that this thesis attempts to overcome is a direct consequence of the *semantic gap* between low-level and high-level features (Santini and Jain 1998). To bridge the gap, the exploitation of the semantics between annotation words should be incorporated into the process.

The annotation words should provide a set of semantically related words. For instance, if an image is annotated with the word "jaguar" the rest of the annotation words should provide an indication of the context represented by the image: an animal or a car. If the accompanying words are "luxury", "sport", "sedan", clearly, one should be talking about a car. On the other hand, if the words are "cat", "tree", "forest", the animal jaguar is more probable. Apart from the knowledge that a jaguar is *a kind of cat*

found in a forest, one might be interested in obtaining additional information such as which kind of animal a jaguar is, or a physical description of the animal or information about its geographical location. In that case, external knowledge sources are required. Hence, one may refer to the definition provided by a dictionary and discover that a jaguar is *"a large, heavily built cat that has a yellowish-brown coat with black spots, found mainly in the dense forests of Central and South America"*. Then, additional words related to "jaguar", such as "large cat", "yellowish-brown coat", "black spots", "dense forests", "America", may be inferred from the definition. If, on the contrary, a lexical database that adopts a hierarchical structure like WordNet is employed, the outcome is the following: "jaguar" → "big cat" → "feline" → "carnivore" → "placental mammal" → "mammal" → "vertebrate" → "chordate" → "animal".

Consequently, the use of knowledge sources helps in providing semantically related words. Moreover, once some related words have been identified, it is crucial to define ways of assessing the degree of relatedness between them. As a result, it is essential to define a strategy able to model the semantics behind an image collection that has an adequate balance between information internal and external to the collection together with an adequate measure of semantic relatedness between words.

(iii) How to integrate semantic knowledge into the annotation framework?

Semantic knowledge is to be integrated either as part of the *annotation process* or as part of the *evaluation stage* that computes the performance of the model.

In the first case, the integration largely depends on how the annotation process is structured. Sometimes, there is an initial annotation of objects followed by a stage where semantically unrelated words are pruned. This is accomplished by adopting some decision rules once a pair of semantically inconsistent words are detected. The most delicate part of the process is how to integrate these decision rules in such as

way that it is ensured that the final performance is better than the initial. However, there exist other solutions that integrate the initial recognition of the objects with the semantic processing in the same stage of the process. This is usually the case of some hierarchical approaches where the semantic knowledge is modelled using a graph made up of annotation words. In this kind of approaches, the success of the whole approach resides in the correct modelling of the semantic graph together with an adequate strategy that selects the final annotation words. Examples of these approaches are Srikanth et al. (2005), Li and Sun (2006), Shi et al. (2006), Shi et al. (2007), Fan et al. (2007), or approaches based on Markov Random Fields, such as the one presented in Chapter 5.

The second case considers that the integration takes places in the *evaluation stage* of the process but the focus is on the special case in which the vocabulary of terms adopts the hierarchical structure of an ontology.

1.6 Thesis Contributions

In what follows, the main contributions of this thesis are ordered in terms of their importance and strength, starting from the most important to the less:

- The automated annotation model presented in Chapter 5, which demonstrates that Markov Random Fields provide a convenient framework for exploiting the semantic context dependencies of an image. With respect to the performance obtained, it is comparable to previous state-of-the-art algorithms. For the Corel 5k dataset, we obtained a MAP of 0.32 (higher than the popular models of Feng et al. (2004) and Carneiro et al. (2007)). For the more realistic dataset [4] used

[4] A subset of MIR Flickr dataset.

by the 2009 ImageCLEF competition, we were located in the position 21 out of 74 algorithms.

- The *semantic-enhanced annotation* model of Chapter 4, which achieves a MAP of 0.30 for the Corel 5k dataset and 0.31 for the ImageCLEF image collection. Both results were statistically significant over the baseline non-parametric density estimation algorithm. The strongest point of this model lies in the efficient combination of internal and external sources of knowledge.

- The experiments conducted in Chapter 6 that integrate semantics between annotation words with some evaluation measures to estimate the performance of annotation algorithms, when the annotation words adopt the hierarchical structure of an ontology. One of these novel evaluation measures has been successfully used to evaluate the performance of all submitted algorithms in the 2010 edition of the ImageCLEF evaluation campaign (Nowak and Huiskes 2010).

- The comprehensive review undertaken in Chapter 3 that shows the evolution of the evaluation metrics and how the Corel 5k dataset became a benchmark in the field. Additionally, the most important multimedia evaluation campaigns are revised, together with the most relevant research questions addressed by each one of them.

- An analysis of the limitations of *classic probabilistic* approaches, which helps in the identification of some gaps. Specifically, it identifies that probabilistic approaches are likely to have limited success as a result of the *semantic gap* that exists between the low-level features (image features) and the high-level features (semantics). The semantic gap term was firstly introduced by Santini and Jain (1998) to describe the inability to get high level features out of low level features

in multimedia retrieval. The only way of bridging this gap is by incorporating semantics into the process.

- The provided classification schema for automated image annotation algorithms. Thus, they are classified into three different groups: *classic probabilistic models*, *semantic-enhanced models*, and *fully semantic integrated models*.

- The identification of *semantic-enhanced models* as an independent group of annotation algorithms. I consider that they have been largely neglected in the research field and believe that they should be carefully taken into account. Chapter 2 presents an in depth analysis of these methods by reviewing previous approaches, analysing the methodology followed and exploring the limitations as well as the strong points of the proposed algorithms.

Finally, this thesis successfully proves that the exploitation of the semantics between words combined with statistical models based on the correlation between words and visual features definitively increases the effectiveness of probabilistic automated image annotation systems.

1.7 Thesis Structure

This thesis is structured in the following chapters.

Chapter 1 introduces the topic of automated image annotation and revises *classic probabilistic approaches* found in the literature. Then, there is a study of the limitations of previous approaches that helps to compile a list of requirements that should be taken into consideration. This set of requirements leads to a new kind of approaches, *semantic-enhanced models*, whose technical details are discussed in Chapter 2. Chapter 3 presents the methodology adopted as well as common evaluation

measures and benchmark datasets employed in the field. From Chapter 4-6, the experimental work undertaken in this thesis is introduced. In particular, Chapter 4 presents a *semantic-enhanced model*; Chapter 5 introduces a Markov Random Field model, which is part of the *fully integrated models*; and Chapter 6 shows an application of how semantics can be integrated with an evaluation measure to measure the performance of an annotation model, when the annotation words adopt the hierarchical structure of an ontology. Finally, Chapter 7 concludes by discussing the main achievements of this thesis as well as presenting future lines of work.

Several parts of this thesis gave rise to the following publications:

- Llorente, A., Manmatha, R., and Rüger, S. (2010). Image Retrieval using Markov Random Fields and Global Image Features. *Proceedings of the ACM International Conference on Image and Video Retrieval*, Xi'an, China, pp. 243-250. (Chapter 5, joint work with Manmatha of the University of Massachusetts at Amherst).

- Nowak, S., Llorente, A., Motta, E., and Rüger, S. (2010). The Effect of Semantic Relatedness Measures on Multi-label Classication Evaluation. *Proceedings of the ACM International Conference on Image and Video Retrieval*, Xi'an, China, pp. 303-310 (Chapter 6, joint work with Stefanie Nowak of Fraunhofer, Germany).

- Little, S., Llorente, A., and Rüger, S. (2010). An Overview of Evaluation Campaigns in Multimedia Retrieval, in eds. H. Mller; P. Clough; T. Deselaers & B. Caputo. *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*, pp. 507-522, Springer-Verlag. (Chapter 3).

- Llorente, A., Motta, E., and Rüger, S. (2010). Exploring the Semantics Behind a Collection to Improve Automated Image Annotation. *Proceedings of the 10th*

Workshop of the Cross-Language Evaluation Forum (CLEF 2009), LCNS 6242, Part II, Springer. (Chapter 4).

- Llorente, A., Motta, E., and Rüger, S. (2009). Image Annotation Refinement using Web-based Keyword Correlation. *Proceedings of the 4th International Conference on Semantic and Digital Media Technologies*, Graz, Austria, 5887, pp. 188-191. (Chapter 4).

- Llorente, A., and Rüger, S. (2009). Using Second Order Statistics to Enhance Automated Image Annotation. *Proceedings of the 31st European Conference on Information Retrieval*, Toulouse, France, 5478, pp. 570-577. (Chapter 4).

- Llorente, A., Overell, S., Liu, H., Hu, R., Rae, A., Zhu, J., Song, D., and Rüger, S. (2009). Exploiting Term Co-occurrence for Enhancing Automated Image Annotation. *9th Workshop of the Cross-Language Evaluation Forum*, LNCS 5706, pp. 632-639, Springer. (Chapter 4).

- Zagorac, S., Llorente, A., Little, S., Liu, H., and Rüger, S. (2009). Automated Content Based Video Retrieval. *TREC Video Retrieval Evaluation Notebook Papers*, NIST, 2009.

- Llorente, A., Little, S., and Rüger, S. (2009). MMIS at ImageCLEF 2009: Non-parametric Density Estimation Algorithms. *Working notes for the CLEF 2009 Workshop*, Corfu, Greece.

- Llorente, A., and Rüger, S. (2008). Can a Probabilistic Image Annotation System be Improved Using a Co-occurrence Approach?. *Proceedings of the Workshop on Cross-Media Information Analysis, Extraction and Management*, Koblenz, Germany, 437, pp. 33-42. (Chapter 4).

- Llorente, A., Zagorac, S., Little, S., Hu, R., Kumar, A., Shaik, S., Ma, X., and Rüger, S. (2008). Semantic Video Annotation using Background Knowledge and Similarity-based Video Retrieval. *TREC Video Retrieval Evaluation Notebook Papers*, Gaithersburg, Maryland, NIST. (Chapter 4).

- Overell, S., Llorente, A., Liu, H., Hu, R., Rae, A., Zhu, J., Song, D., and Rüger, S. (2008) MMIS at ImageCLEF 2008: Experiments combining Different Evidence Sources. *Working notes for the CLEF 2008 Workshop*, Aarhus, Denmark.

Chapter 2

Semantic Measures and Automated Image Annotation

The problem of modelling annotated images has been addressed from several directions in the literature. As seen in Section 1.3, a set of generic algorithms were initially developed with the aim of exploiting the implicit dependencies between image features and words. Recently, researchers have singled out limitations of these approaches (see Section 1.4), where individual words were generated independently without considering the occurrence of other words in the same image context. Addressing this limitation is the subject of this research and I believe that a solution can be obtained through the use of semantic relatedness measures. The main objective of this chapter is to review existing annotation algorithms in the literature, which make use of various semantic measures.

The rest of the chapter is organised as follows. Section 2.1 introduces the notions of semantic similarity and semantic relatedness. It is important to emphasise that although the distinction between them has created numerous debates in other fields, the distinction is not so important in the field of automated image annotation. In any case

the differentiation is reflected when introducing the measures in order to be rigourous. Then, a number of semantic relatedness measures will be analysed: Section 2.2 focuses on keyword correlation in a training set, Section 2.3 focuses on measures based on WordNet, Section 2.4 is devoted to web-based measures, Section 2.5 revises Wikipedia based approaches. Finally, Flickr-based measures are presented in Section 2.6. Each of these sections starts introducing the semantic measures and ends with a comparison of the annotation algorithms that use them. Section 5.6 summarises the main conclusions.

2.1 Semantic Similarity versus Semantic Relatedness

Meaningful visual information comes in the form of scenes. The intuition is that understanding how the human brain works in perceiving a scene will help to understand the process of assigning words to an image by a human annotator and consequently will help to model this process. Moreover, having a basic understanding of the scene represented in an image, or at least a certain knowledge of other objects contained there, can actually help to recognise an object. An attempt to identify the rules behind the human understanding of a scene was made by Biederman (1981). In his work, the author shows that perception and comprehension of a scene requires not only the identification of all the objects comprising it, but also the specification of the relations among these entities. These relations mark the difference between a well-formed scene and an array of unrelated objects. For example, the action of recognising a scene with "boat", "water" and "waves" (Figure 1.1) requires not only the identification of the objects, but also the knowledge that the "boat" is in the "water" and the "water" has got "waves". Thus, all the objects in the image are semantically related.

The distinction between semantic similarity and semantic relatedness has been a topic of continuous debate among researchers in the field of natural language processing

(NLP). Resnik (1999) introduces this difference by asserting that the semantic similarity between two concepts can be uniquely calculated using a concept inclusion relation (*is-a*) whereas semantic relatedness is the result of the aggregation of other semantic relations. Thus, semantic relatedness is a generalisation of semantic similarity. According to this, "car" and "vehicle" are semantically similar, as the only relation between them is the (*is-a*) relation; "steering wheel" and "car" are semantically related because the relationship between them is a meronym (*is-part-of*).

The application of this notion to the image domain is not straightforward. On the one hand, the relationship between concepts representing objects depicted in an image might be that of similarity, but also of relatedness. On the other, the antonym relationship (words with opposite meanings) is highly undesirable. Therefore, this thesis adopts the notion of semantic relatedness but excluding the antonym relation. Two words are semantically related if they refer to entities that are likely to co-occur such as "forest" and "tree", "sea" and "waves", "desert" and "dunes", etc.

In this work, the distinction between semantic similarity and relatedness is drawn only to introduce the different measures in the literature and in order to be consistent with the choice made by authors.

2.1.1 Introduction to Semantic Measures

According to Mohammad and Hirst (2005), humans are inherently able to assess whether two words are semantically related and even to estimate the degree of relatedness between them. However, a lot of work has been done in order to automate this process in the last fifteen years. In brief, automated systems assign a score of semantic relatedness to a pair of words calculated from a relatedness measure. Three kinds of approaches (Mohammad and Hirst 2005) have been adopted to evaluate computational

measures of semantic relatedness. The first one establishes a comparison with human similarity judgements; the second measures their performance in the framework of a particular application and the last one envisages the evaluation as a theoretical study where the measure is evaluated with respect to a set of mathematical properties that are considered desirable.

Finally, some datasets were proposed for accomplishing the evaluation of semantic measures. The first one was created by Rubenstein and Goodenough (1965) (R&G), who compiled 65 synonymous pairs of words that were assessed by 51 persons. Later on, Miller and Charles (1991) (M&C) extracted 30 pairs from the original 65 that were judged by 38 individuals. More recently, Finkelstein et al. (2002) (WS-353) proposed a collection of 353 word pairs.

Traditionally, proposed semantic relatedness measures relied either on *distributional* measures or on *semantic network* representations. The distributional similarity between two words occurs when they co-occur in similar contexts. The context considered may be a small or large window around the word, an entire document, or a corpus (a collection of documents). On the other hand, a *semantic network* is broadly described as *"any representation interlinking nodes with arcs, where the nodes are concepts and the links are various kinds of relationships between concepts"*, according to the definition provided by Lee et al. (1993). In particular, a taxonomy is a hierarchical representation of a semantic network with a partial ordering, typically given by the concept inclusion relation (*is-a*).

More recently, Gurevych (2005) proposed another classification schema for semantic similarity measures by making the distinction between intrinsic and extrinsic measures. The former denotes those measures that use only the information that are part of the data while the later refers to external information.

In the following sections, several semantic relatedness measures applied to enhance annotation algorithms will be analysed. Each section starts by summarising the measures proposed in the literature of automated image annotation. Then, it follows with an analysis of the performance of the annotation algorithms that incorporate them. Finally, the section concludes with a brief discussion about the benefits and drawbacks of the considered measures. In particular, keyword correlation in the training set, web-based measures, semantic network based measures using WordNet and Wikipedia, and Flickr-based measures will be revised.

2.2 Co-occurrence Models on the Training Set

A theoretical basis for the use of co-occurrence data was established by van Rijsbergen (1977) in text information retrieval. He argued that previous index terms used in automatic index term classification algorithms were considered independent because of mathematical convenience. Moreover, he considered this independence to be often an unrealistic assumption that is tolerated because it leads to a straightforward solution of a problem. Then, he set the foundations of a probabilistic model that incorporates dependence between index terms. The extent to which two index terms depend on one another is derived from the distribution of co-occurrence in the whole collection.

Later on, Hofmann and Puzicha (1998) defined the general setting described by the term co-occurrence data as follows. Let, $X = \{x_1, x_2, ..., x_n\}$ and $Y = \{y_1, y_2, ..., y_m\}$, be two finite sets of abstract objects with arbitrary labelling. As elementary observations pairs $(x_i, y_j) \in X \times Y$, that is, a joint occurrence of object x_i with object y_j. All data are numbered and collected in a sample set S =$\{(x_i(r),\ y_j(r),\ \mathrm{r})$ with $1 \leq \mathrm{r} \leq \mathrm{L}\}$. The information in S is completely characterised by its sufficient statistics $n_{ij} = |\ \{(x_i,\ y_j,\ \mathrm{r}) \in \mathrm{S}\}|$, which measures the frequency of co-occurrence of x_i and y_j.

However, different interpretations to the above formulation can be found depending on the displicine applied. In the case of image annotation, X corresponds to a collection of images and Y to a set of keywords. Hence n_{ij} denotes the number of occurrences of the word y_j in the image x_i.

The intrinsic problem of co-occurrence data is its sparseness. When the size of documents N and the size of keywords M are very large, a majority of pairs $(x_i,\ y_i)$ only have a small probability of occurring together in S. A typical solution consists in applying *smoothing techniques* to deal with zero frequencies of unobserved events[1], which uses a smoothing parameter, λ, different from zero. Additional approaches have been analysed by Jelinek and Mercer (1980), by Chen and Goodman (1996), and by Zhai and Lafferty (2001).

By applying *fuzzy set theory* to information retrieval (Baeza-Yates and Ribeiro-Neto 1999), *the degree of keyword co-occurrence* can be considered as a measure of semantic relatedness. The application to the image annotation field is then immediate. Thus, a co-occurrence matrix, C, can be constructed by defining the *normalised termed correlation index*, c_{ik}, between two words w_i and w_k as

$$c_{ik} = \frac{n_{ik}}{n_i + n_k - n_{ik}}, \tag{2.1}$$

where n_i and n_k, are the number of training set images that contains the words w_i and w_k respectively, while n_{ik} corresponds to the number of images containing both of

[1] Equation 2.9 shows an example of this technique.

them. c_{ik} oscillates between zero and one, then if

$$c_{ik} \begin{cases} = 0 \implies n_{ik} = 0, \text{ i.e., } w_i \text{ and } w_k \text{ do not co-occur (terms are mutually exclusive),} \\ > 0 \implies n_{ik} > 0, \text{ i.e., } w_i \text{ and } w_k \text{ co-occur (terms are non mutually exclusive),} \\ = 1 \implies n_{ik} = n_i = n_k, \text{ i.e., } w_i \text{ and } w_k \text{ co-occur whenever either term occurs.} \end{cases} \tag{2.2}$$

Jin et al. (2004) incorporate word-to-word correlation as part of their proposed model, which is called *coherent language model* (CLM). They define a language model, θ_w, as a set of word probabilities like $\{p_1(\theta_w), p_2(\theta_w), \ldots, p_n(\theta_w)\}$, where each probability, $p_j(\theta_w)$=p($w_j = 1|\theta_w$), determines how likely the j-th word will be used for annotation. They conclude that the estimation of word probability, $p_k(\theta_w)$, depends on the estimation of other word probabilities $p_j(\theta_w)$. Consequently, they demonstrate that the prediction of annotation words is no longer independent from each other. They evaluate their model with the Corel 5k dataset (Section 3.7.2) although their results cannot be compared to other works in the field as they compute precision and recall using 140 words, which are those that appear at least 20 times in the training set, instead of the usual 260 words that annotate the test set. They prove the validity of their approach by comparing their algorithm with their own implementation of a *relevance language model* (RLM) based on Jeon et al. (2003) and on Lavrenko et al. (2003), showing an improvement in the performance in terms of precision and recall.

Wang et al. (2006) propose an image annotation refinement algorithm (RWRM) using *random walks with restarts*. They revise the approach of Jin et al. (2005b)[2](see Section 2.3.6), who used WordNet to remove annotation words non-related to the oth-

[2]Jin et al (2005b) proposed a novel method that prunes the unrelated keywords generated by the translation model using WordNet semantic similarity measures. They reported a 56.87% improvement over the baseline translation method in terms of precision for the Corel 5k dataset.

ers. Wang et al. (2006) claim that the reasons why the performance of Jin et al. (2005b) decreases with respect to their baseline approach are twofold: One is due to the fact that WordNet is unable to reflect the characteristics of the image collection and the second is that many words of the vocabulary do not belong to WordNet. As a result, Wang et al. (2006) reformulate the image annotation process as a *graph ranking problem* using co-occurrence information from the training set. The graph is built as follows. First, a set of candidate annotations are produced using the *cross-media relevance model* (CMRM) by Jeon et al. (2003) as baseline algorithm. Each candidate annotation is considered a node of the graph; all nodes are fully connected with proper weights. The edge that connects two nodes has a weight given by the co-occurrence similarity value as

$$\text{sim}(w_i, w_k) = \frac{n_{ik}}{\min(n_i, n_k)}. \tag{2.3}$$

Then, an algorithm based on *random walks with restarts* (Page et al. 1999) re-ranks the candidate annotations producing more accurate ones. Wang et al. (2006) test their algorithm with two datasets, the Corel 5k and an image collection extracted from the web. For the Corel 5k dataset, they compare their results with the approach proposed by Jin et al. (2005b) and with CMRM. Their evaluation measures were averaged over the 49 words with best performance as in (Jeon et al. 2003). They outperform the previous refinement algorithm although they get comparable results to the baseline method in terms of F-measure (see Section 3.2.3).

Liu et al. (2006) propose a new image annotation method (AGAnn) based on *manifold ranking*, in which the visual and textual information are well integrated. They create a co-occurrence matrix, C, where each cell c_{ik} is represented by

$$c_{ik} = n_{ik} \cdot \log \frac{N}{n_i}, \tag{2.4}$$

where n_{ik} represents the number of images annotated by words w_i and w_k, and N is the total number of images in the training set. After normalising the matrix they combine it linearly with a measure based on WordNet. Finally, they prune irrelevant annotations for each unseen image using the semantic similarity information. For the Corel 5k dataset, they report better results than Jin et al. (2005b), who use solely WordNet as source of their semantic measures. However, comparison with the rest of literature is hindered by their use of precision and recall computed on the most frequent seven words of the Corel 5k collection.

Kang et al. (2006) present a novel framework called *correlated label propagation* (CLP) for multi-label learning that explicitly exploits high-order correlation between labels (annotation words). Unlike previous approaches that contemplate the propagation of a single class label between training examples and test examples, the proposed model considers the propagation of multiple labels simultaneously. The propagation from training examples to test examples is accomplished through their similarities. Thus, the similarity of the test label w_i to a training label w_k is defined as

$$\text{sim}(w_i, w_k) = \prod_j [p(j, \mathbf{x}_k)]^{x_{i,j}}, \tag{2.5}$$

where each $\mathbf{x}_k = (x_{k,1}, ..., x_{k,d})$ is an input image vector of d dimension. For the Corel 5k dataset, they report better results than the translation method (Duygulu 2003), and than a method based on support vector machines proposed by Joachims (1999).

Kamoi et al. (2007) present a new approach based on *visual cognitive theory* that improves the accuracy of image recognition by considering word-to-word correlation between words representing the context and the objects of the image. Although, they proved that their system has great potential for enhancing image recognition, a greater number of images and much larger amounts of knowledge are needed to make the system more practical. They assign a weight to every word calculated as the *term*

Table 2.1: Association or term-image matrix for the Corel 5k dataset

	J_1	J_2	J_3	...	$J_{\|\tau\|}$
city	0	0	0	...	1
mountain	1	0	0	...	1
sky	1	0	1	...	0
...	-	-	-	...	-
race	0	1	1	...	0
hawai	0	1	1	...	0

frequency (TF) $\mathrm{TF}(w_i) = n_i$, where n_i is the number of times the word w_i appears in the training set, or as the multiplication between the *term frequency* and the *inverse document frequency* (IDF), which is computed as

$$\mathrm{IDF}(w_i) = \log \frac{N}{n_i}, \tag{2.6}$$

being N the number of training images. They adopted the accuracy and the coverage of the collection as measures for the evaluation of results, which makes impossible their comparison with existing approaches in the field. However, they reach some interesting conclusions. The combination of TF and IDF is more suited for annotation of low frequency words, whereas TF performs better for high frequency words. Although TF accomplishes a better recognition, the combination of TF and IDF is able to generate more suitable annotations to the images. Therefore, they consider that the adequate treatment of the high and low frequencies may improve the accuracy of the final system.

Zhou et al. (2007) exploit the theory of *automatic local analysis* (Baeza-Yates and Ribeiro-Neto 1999) of text information retrieval to analyse the correlations between keywords on the training set. They build an association matrix, A, as seen in Table 2.1, a rectangular matrix where each cell a_{ij} represents whether or not the i-th word occurs in the annotation of the training set image J_j. The Jelinek-Mercer algorithm (Jelinek

and Mercer 1980) is used to avoid the sparseness problems in matrix A. The semantic correlation between words w_i and w_k is calculated as in Equation 2.1, where

$$n_{ik} = \sum_{J_j \in \tau} a_{ij} \cdot a_{kj} \tag{2.7}$$

and τ is the training set. However, the similarity between annotation words w_i and w_k follows the cosine distance given by

$$\text{sim}(n_i, n_k) = \frac{\vec{n_i} \cdot \vec{n_k}}{|\vec{n_i}| \cdot |\vec{n_k}|}, \tag{2.8}$$

being $\vec{n_i}$ and $\vec{n_k}$, a vector extracted, respectively, from the i-th row, and the k-th column of the co-occurrence matrix. Finally, the approach proposed in their paper (Anno-Iter) is compared to MBRM (Feng et al. 2004) obtaining a significant increment in recall and precision, which are 21% and 11% better than MBRM respectively.

Escalante et al. (2007a) use a interpolation smoothing technique in the form of

$$P(w_i|w_k) \approx \lambda \cdot \frac{n(w_i, w_k)}{n(w_k)} + (1 - \lambda) \cdot n(w_k) \tag{2.9}$$

in order to increase the accuracy of a k-nearest neighbour annotation method. λ is a smoothing factor. The correlation is computed off-line using an external image dataset. Experimental results of their method on three subsets of the benchmark Corel collection give evidence that the use of a naïve Bayes approach together with co-occurrence information results in signicant error reductions. Again, comparison with other approaches is not possible as they use evaluation methods inherited from the computer vision field.

Stathopoulos et al. (2008) propose a multi-modal graph based on *random walks with restarts* (RWR) (Lovász 1993) that exploits the co-occurrence of words in the training set. During the first run of the RWR algorithm the graph is built up without adding the edges between word nodes. In the second run, these edges are incorporated

Table 2.2: Co-occurrence Matrix

	city	**mountain**	**sky**	**...**	**race**	**hawaii**
city	2	0	1	-	1	0
mountain	0	1	1	-	0	0
sky	1	1	2	-	0	0
...	-	-	-	-	-	-
race	1	0	0	-	2	0
hawaii	0	0	0	-	0	0

after computing the similarity of words nodes given by the *automatic local analysis* theory (Baeza-Yates and Ribeiro-Neto 1999). Authors report results for the Corel 5k dataset using average accuracy, the normalised score, and average precision and recall, obtaining statistically significant improvement over the baseline RWR algorithm.

Tollari et al. (2008) also present a co-occurrence model in order to exploit the relationships among annotation keywords. The co-occurrence analysis was incorporated through some "resolution rules" that resolve the conflicting annotations. They considered two types of relations between concepts: exclusion and implication. By exclusion they mean concepts that never appear together and by implication they refer to relationships between concepts. Thus, their best performance is achieved when they use the exclusion and implication rules together. Results are solely presented for the ImageCLEF2008 collection.

Llorente and Rüger (2009a) propose a heuristic model able to prune the non-correlated keywords by means of computing statistical co-occurrence of pairs of keywords appearing together in the training set. This information is represented in the form of a co-occurrence matrix, which is estimated as follows. The starting point is the association matrix A exemplified in Table 2.1, where each row represents a word

of the vocabulary and each column an image of the training set. Each cell indicates the presence or absence of a keyword in the image. The co-occurrence matrix C (see Table 2.2) is obtained after multiplying the association matrix A by its transpose A^T. The resulting co-occurrence matrix ($C = A \cdot A^T$) is a symmetric matrix where each entry n_{jk} contains the number of times the keyword w_j co-occurs with the keyword w_k. For the Corel 5k dataset, they obtained statistically better results than the baseline approach, which is based on the probabilistic framework developed by Yavlinsky et al. (2005)[3], who used global features together with a non-parametric density estimation approach. However, they failed to obtain statistically significant results for the ImageCLEF2008 collection (Llorente et al. 2009c), and for TRECVid 2008 video collection (Llorente et al. 2008b). An explanation for this can be found in the small number of terms of the vocabulary for both collections that hinders the functioning of the algorithm. This makes sense as a big vocabulary allows us to exploit properly all the knowledge contained in the image context.

Garg (2009), more recently, proposes a simple correlation model based on naïve Bayes theory. Thus, he computes the annotation score, $S(w_j)$, for each annotation word, w_j, as

$$\log S(w_j) = \log P(\nu_1|w_j) + \ldots + \log P(\nu_k|w_j) + \log P(w_j), \tag{2.10}$$

where $\{\nu_1, \nu_2, \ldots, \nu_k\}$ are the set of *visterms* (quantised invariant local descriptors) that represent an image of the test set I, being the correlation model

$$P(\nu_i|w_j) = \frac{n(\nu_i, w_j)}{n(w_j)}, \tag{2.11}$$

where $n(\nu_i, w_j)$ denotes the number of training images with *visterm* ν_i and word w_j, and $n(w_j)$ is the number of images annotated by w_j. The comparison of results with

[3]Baseline approach will be explained in detail in Section 4.1.1.

other approaches is hindered by their use of non benchmark datasets.

2.2.1 Co-occurrence Discussion

Approaches based on statistical correlation on the training set may benefit[4] from the fact that they work with words instead of concepts so they do not need a prior disambiguation task as it happens in the case of thesaurus-based methods.

With respect to their limitations, all algorithms revised in Section 2.2 are based on the actual counts of co-occurrences in a corpus. However, the effectiveness of these methods relies heavily on the coverage and characteristics of the corpus used. In particular, some limitations in solutions based on the training set have been detected. The knowledge coming from the training set is internal to the collection and is limited to the scope of the topics represented. This information may not suffice to detect annotations that are not correlated with the others. For instance, if the collection contains a large amount of images of animals in a circus and just a few of animals in the wildlife, the co-occurrence approach will penalise combinations such as "lion" and "savannah" while promoting associations such as "lion" and "chair", given that the training set is dominated by images of lions in a circus (i.e. consider a lion tamer controlling a lion with a chair). In order to avoid being overly biased by the topics represented in the collection it is advisable to, additionally, incorporate the common-sense or world knowledge provided by external knowledge.

Lindsey et al. (2007) also observed this high dependency on the selected corpus: They detected statistically significant variations on the performance of two co-occurrence measures, the *normalized Google distance* defined by Cilibrasi and Vitanyi

[4]From a pragmatic point of view, these approaches do not need intensive use of search engines, which may be expensive, as it happens in web correlation approaches.

(2007) and the *pointwise mutual information* (PMI) defined by Turney (2001) using six different corpora: the World Wide Web (Google corpus), the Wikipedia corpus, the New York Times corpus, Project Gutenberg corpus, Google groups corpus and finally, the Enron email corpus. Contrary to their expectations, the best performance was achieved when using smaller corpora. In particular, the New York Times corpus outperformed Wikipedia, Enron, Google, Google gropus and Project Gutenberg corpus.

Finally, one problem that all approaches based on co-occurrence need to tackle is the sparseness of the data. Consequently, an adequate *smoothing technique* needs to be implemented. Jelinek and Mercer (1980), Chen and Goodman (1996), Hofmann and Puzicha (1998), and Zhai and Lafferty (2001), all have deployed smoothing techniques, which are widely used in the field of information retrieval.

2.3 WordNet-based Measures

The problem of assessing semantic similarity using semantic network representations has long been addressed by researchers in artificial intelligence and psychology. As a result, a number of semantic measures have been proposed in the literature. However, in what follows, I will only focus on those measures, which are used by annotation algorithms (see Section 2.3.6).

WordNet is a lexical database of English developed under the direction of Miller (1995). Thus, nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. In addition, each concept (or word sense) is described by a short definition or gloss. Synsets are interlinked with a variety of semantic relations. Table 2.3 summarises the semantic relations defined in one of the early versions of WordNet. Here, the subsumption (hypernymy/hyponymy) relationship constitutes the backbone of the noun subnetwork, accounting for 80% of

Table 2.3: Semantic Relations in WordNet where "n" stands for nouns, "v" for verbs,"a" for adjectives, and "r" for adverbs

Relation	Category	Definition	Example
Synonymy	n,v,a,r	"similar to"	"pipe" is synonym of "tube"
Antonymy	a, r,(n,v)	"opposite of"	"wet" is antonym of "dry"
Hyponymy	n	"is-a" or "is a kind of"	"tree" is a hyponym of "plant"
Hypernymy	n	"is a generalisation of"	"plant" is a hypernym of "tree"
Meronymy	n	"is-part-of"	"branch" is a meronym of "tree"
Meronymy	n	"is-member-of"	"tree" is a meronym of "forest"
Meronymy	n	"is-made-from"	"table" is a meronym of "wood"
Holonymy	n	"has-a", meaning object-component	"deer" is a holonym of "horn"
Holonymy	n	"has-a", meaning collection-member	"army" is a holonym of "soldier"
Holonymy	n	"is-a-substance-of"	"smoke" is a holonym of "fire"
Troponomy	v	"is a way to"	"to march" is a troponym of "to walk"
Entailment	v	"entails"	"to snore" is a troponym of "to sleep"

the links. Additionally, relations in WordNet do not cross part of speech boundaries, so the computation of semantic measures is limited to making judgements between noun pairs, adjective pairs, verb pairs, or adverb pairs. Another important feature is that WordNet works with concepts instead of words. Consequently, for a given pair of words, the first step consists in determining the appropriate sense according to a context. Moreover, the use of *word sense disambiguation* (WSD) techniques is an essential part of the process of estimating semantic measures between words. Finally, note that if the only semantic relation to be considered between concepts is an "is-a" relationship, the measure is called semantic similarity otherwise is called semantic relatedness.

Table 2.4 analyses several WordNet measures, which can be classified into three categories: path length measures, information content measures, and gloss based measures.

2.3.1 Path Length Measures

The first attempts to evaluate semantic measures in a taxonomy focused on measuring the length of the path established along the concepts whose semantic similarity or relatedness was being estimated. These methods treat the taxonomy as an undirected graph where the nodes (or vertices) corresponds to the concepts and the edges, arcs or links represent the relationship between them. Then, the shorter the distance between nodes, the higher the similarity. These measures are based on the observation that sibling terms deep in a tree are more closely related than siblings higher in the hierarchy.

Table 2.4: WordNet-based measures analysed in this thesis

Authors	Measure	Type	Bound
PATH	Similarity	Path Length	$(0, \infty)$
Wu and Palmer (1994) (WUP)	Similarity	Path Length	(0, 1]
Resnik (1995) (RES)	Similarity	Information Content	$[0, \infty)$
Jiang and Conrath (1997) (JCN)	Distance	Information Content	$[0, \infty)$
Leacock and Chodorow (1998) (LCH)	Similarity	Path Length	$[0, +\infty)$
Hirst and St-Onge (1998) (HSO)	Relatedness	Path Length	[0, 16]
Lin (1998) (LIN)	Similarity	Information Content	[0,1]
Banerjee and Pedersen (2003) (LESK)	Relatedness	Gloss	$[0, -\infty)$
Patwardhan et al. (2003) (VEC)	Relatedness	Gloss	[0,1]

Path (PATH)

This measure represents the simplest way of computing semantic similarity between two word senses as it counts the number of nodes along the shortest path in WordNet's noun and verb "is-a" hierarchies. Thus, it can be expressed as

$$\text{sim}(c_1, c_2) = \frac{1}{\text{length}(c_1, c_2)}, \tag{2.12}$$

where $\text{length}(c_1, c_2)$ is a function that belongs to the interval $[0, \infty)$ and that estimates the number the nodes along the shortest path defined by c_1 and c_2. The counts include the initial and end nodes. The length of the path between siblings nodes is three, for instance, the path length between "shrub" and "tree" yields three as the following dependencies can be found in WordNet hierarchy: "shrub" "is-a" "woody_plant" and "tree" "is-a" "woody_plant". The length of the path between members of the same

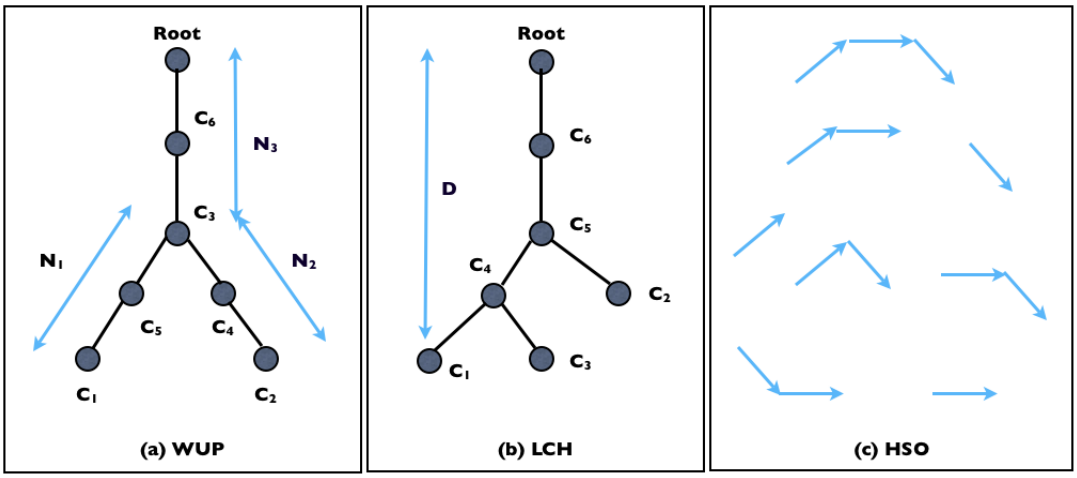

Figure 2.1: Examples of path-length WordNet measures

synset is one as this implies $c_1 = c_2$. Thus,

$$\text{length}(c_1, c_2) \begin{cases} > 0 & \text{if } c_1 \neq c_2 \implies \text{concepts belong to different synsets} \\ = 1 & \text{if } c_1 = c_2 \implies \text{concepts are members of the same synset} \\ = 0 & \text{if } c_1 \neq c_2 \implies \text{there is no path between concepts.} \end{cases} \tag{2.13}$$

From Eq. 2.12, it can be deducted that a longer path implies less relatedness between concepts. Then:

$$\text{sim}(c_1, c_2) \begin{cases} > 0 & \text{if } \text{length}(c_1, c_2) > 0 \implies c_1 \neq c_2 \\ = 1 & \text{if } \text{length}(c_1, c_2) = 1 \implies c_1 = c_2 \\ \to +\infty & \text{if } \text{length}(c_1, c_2) = 0 \implies \text{there is no path between } c_1 \textit{ and } c_2. \end{cases} \tag{2.14}$$

Wu and Palmer (WUP)

Wu and Palmer (1994) analysed the difficulties with the translation of English verbs into Mandarin Chinese. As part of their work, they defined the notion of "conceptual similarity" between a pair of concepts c_1 and c_2 in a hierarchical structure as:

$$\text{sim}(c_1, c_2) = \frac{2 \cdot N_3}{N_1 + N_2 + 2 \cdot N_3}, \tag{2.15}$$

where c_3=lso(c_1, c_2) and is the lowest super-ordinate or most specific common subsumer. This function calculates the most specific concept that subsumes both c_1 and c_2. For instance, the lowest super-ordinate of "tiger" and "cat" is "feline" but not "mammal", which is a concept higher in the hierarchy. Then, N_1 = length(c_1, c_3), N_2 = length(c_2, c_3), and N_3 = length$(c_3, root)$. An example of these parameters can be found in Figure 2.1(a), where $N_1 = N_2 = N_3 = 3$. As before, path lengths are estimated by counting nodes.

Resnik (1999) reformulated Eq. 2.15 as

$$\text{sim}(c_1, c_2) = \frac{2 \cdot \text{depth}(\text{lso}(c_1, c_2))}{\text{depth}(c_1) + \text{depth}(c_2)}, \tag{2.16}$$

where the function depth(c) measures the *distance* of a node c to the *root*. According to the previous example, depth(lso(c_1, c_2))=depth(c_1)=depth(c_2)=2.

This similarity is bounded in the interval (0, 1] as depth(lso(c_1, c_2))$\neq$ 0 and it has been established by convention that depth(root)= 1.

Leacock and Chodorow (LCH)

Leacock and Chodorow (1998) measured the semantic similarity between concepts c_1 and c_2 as the negative logarithm of the counting of nodes between them scaled by the depth of the hierarchy. Therefore, if length(c_1, c_2) is the number of nodes along the shortest path between c_1 and c_2 and D is the maximum depth of the taxonomy, the path length similarity between c_1 and c_2 is computed as:

$$\text{sim}(c_1, c_2) = \max\Big[-\log\big(\frac{\text{length}(c_1, c_2)}{2 \cdot D}\big)\Big]. \tag{2.17}$$

Figure 2.1(b) shows an example of this measure, where $\text{length}(c_1, c_2) = 4$ and $D = 5$. Then, as $D > 0$, the measure can take the following values:

$$\text{sim}(c_1, c_2) \begin{cases} = 0 & \text{length}(c_1, c_2) = 2 \cdot D \neq 0 \\ < 0 & \text{length}(c_1, c_2) > 0 \implies c_1 \neq c_2 \\ \to +\infty & \text{length}(c_1, c_2) = 0 \implies \text{there is no path between } c_1 \textit{ and } c_2. \end{cases} \tag{2.18}$$

Hirst and St-Onge (HSO)

As opposed to the other path-length approaches, Hirst and St-Onge (1998) described a method based on identifying *lexical chains* in text. In essence, a *lexical chain* is a cohesive chain in which the criterion for inclusion of a new word is that it bears some kind of cohesive relationship to another word that is already in the chain. Morris and Hirst (1991) claimed that the discourse structure of a text may be determined by finding lexical chains in it. Hirst and St-Onge (1998) showed how to construct these lexical chains by means of WordNet. The final goal of their research was to detect and correct automatically malapropism in a text. They defined a malapropism as *the confounding of an intended word with another word of similar sound or similar spelling that has a quite different meaning.* For instance, "word" and "world" constitute a clear example of this. The utility of their approach is evident as traditional spelling checkers cannot detect this kind of mistake.

They considered that links or relations in WordNet can be classified according to their direction and their weight. With respect to its direction, a link can be considered *horizontal* if the relation between the two concepts is of antonymy or synonymy; *upward* if the relation is hyponymy or meronymy; or *downward* when the relation is hypernymy or holonymy.

According to their weight, they can be considered *extra-strong*, *strong*, or *medium-strong*. An *extra-strong relation* holds between two instances of the same word; such relations have the highest weight of all relations but do not occur in this work. *Strong relations* occur in three cases. The first occurs when two words belong to the same synset such as "car" and "automobile". The second occurs when the words are connected through an horizontal link such as "cold" and "hot". The third occurs when one word is a compound word that includes the other, such as in the case of "school" and "boarding school". The value of a *strong relation* is $2 \cdot C$, where C is a constant. And, finally, a *medium-strong relation* or regular relation is defined as *a relation that occurs if there exists an allowable path connecting the synset associated with each word that is neither too long nor that changes direction too often.* The length of the path is defined as the number of links (edges) connecting the synsets; they may vary from two to five. However, a path is *allowable* if it corresponds to one of the eight patterns shown in Figure 2.1(c). Unlike *extra-strong* and *strong* relations, *medium-strong* relations have different weights, and their value is given by:

$$\text{weight} = C - \text{length}(c_1, c_2) - k \cdot \text{number of changes in direction}, \tag{2.19}$$

where C and k are constants set respectively to eight and one. Then, they calculate the semantic relatedness between two words w_1 and w_2 as:

$$\text{rel}(w_1, w_2) = \begin{cases} 2 \cdot C & \text{if words belong to the same synset, i.e. } c_1 = c_2 \\ 2 \cdot C & \text{if } c_1 \textit{ and } c_2 \text{ are connected by a horizontal link} \\ 2 \cdot C & \text{if one word is a compound word that contains the other} \\ \text{weight} & \text{otherwise.} \end{cases} \tag{2.20}$$

Consequently, the longer the path, and the more it changes direction, the lower the

weight. Note that, contrary to other WordNet based measures, they worked with words instead of with concepts and the length of the path is estimated by edge counting instead of by node counting. Finally, as they considered all the relations defined in WordNet their measure is that of semantic relatedness.

2.3.2 Information Content Measures

The problem of the previous approaches is that links in the taxonomy are considered to represent uniform distances and this is not usually the case. For example, in the plant (or flora) section of WordNet, the hierarchy is so dense that one parent node may have up to several hundred child nodes. As the degree of semantic similarity implied by a single link is not consistent, links between very general concepts may convey smaller amounts of similarity than links between very specific concepts do.

One way of addressing this limitation is to include additional information per concept in the form of *information content* (IC), which is computed counting the frequency of its occurrence in a large corpus.

Resnik (RES)

Resnik (1995) introduced an alternative way to evaluate the semantic similarity between the concept c_1 and c_2 in a taxonomy, based on the notion of *information content* (IC):

$$\text{sim}(c_1, c_2) = \max_{c \in \text{Super}(c_1,c_2)} |\text{IC}(c)|, \tag{2.21}$$

where $\text{Super}(c_1, c_2)$ represents the set of concepts that subsumes both c_1 and c_2, and c is the concept in $\text{Super}(c_1, c_2)$ with the highest value of IC.

He followed the standard argumentation of information theory that considers the information content of a given concept c in a taxonomy as the negative logarithm value

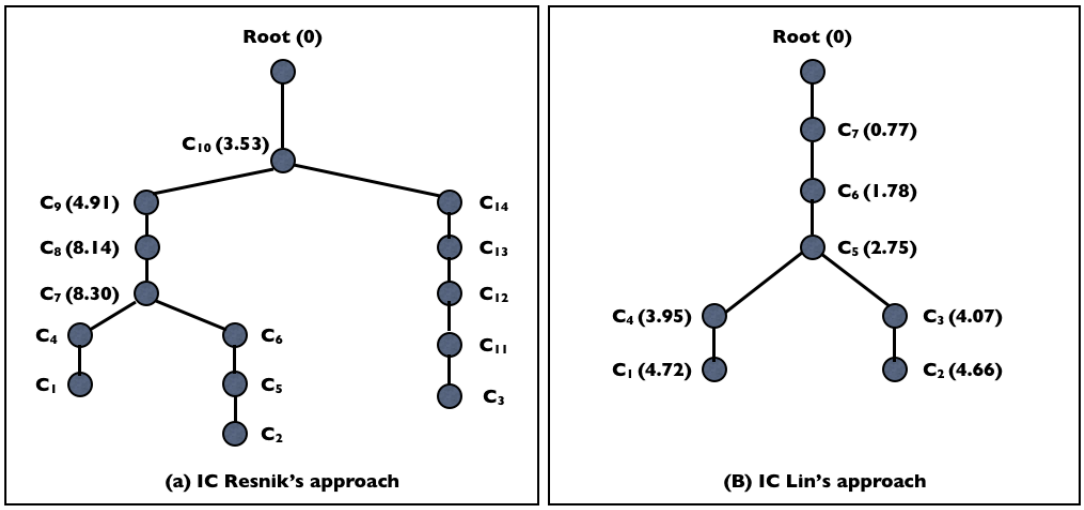

Figure 2.2: Example of some WordNet measures based on information content

of the probability of encountering an instance of itself:

$$\mathrm{IC}(c) = -\ \log\ \mathrm{p}(c), \tag{2.22}$$

ranging from zero to infinity. Additionally, when a node distinct from the root has a zero value of information content it implies a lack of data as there is no frequency count for the concept.

Therefore, when the probability increases, the informativeness decreases; in other words, the more abstract a concept, the lower its information content. The probability can be expressed as follows:

$$\mathrm{p}(c) = \frac{\mathrm{freq}(c)}{N} = \frac{\sum_n\ \mathrm{count}(n)}{N}, \tag{2.23}$$

where n represents each one of the words whose senses are subsumed by concept c, and N is the total number of nouns in the corpus. Thus, frequencies of concepts in the taxonomy were estimated gathering noun frequencies from a large external corpus such as the Brown Corpus of American English (Francis and Kučera 1982).

Figure 2.2(a) shows how to estimate Resnik's measure. Note that the number in brackets is the modulus of the *information content* of the associated node. Due to the

fact that, $\text{Super}(c_1, c_2) = \{c_7, c_8, c_9\}$ but c_7 is the one with the highest value of *information content*, $\text{sim}(c_1, c_2)$=$\text{sim}(c_1, c_5)$=$\text{sim}(c_1, c_6)$=$\text{sim}(c_4, c_2)$=$\text{sim}(c_4, c_5)$=$\text{sim}(c_4, c_6)$ = $\text{IC}(c_7)$. Likewise, $\text{sim}(c_1, c_3)$=$\text{sim}(c_5, c_{11})$=$\text{sim}(c_4, c_{13})$=$\text{sim}(c_8, c_{14})$=$\text{sim}(c_9, c_3)$=$\text{sim}(c_1, c_{12})$ = $\text{IC}(c_{10})$, as c_{10} contains the maximum *information content* value among all the concepts that subsume both classes.

This measure does not consider the *information content* of the concepts themselves, nor does it directly consider the path length. The potential limitation of this approach, as seen by the example, is that concepts having the same subsumers would have identical values of similarity assigned to them.

Jiang and Conrath (JCN)

Jiang and Conrath (1997) proposed a combined model derived from the edge-count approach by adding Resnik's information content as a decision factor. In particular, the model is based on estimating the link strength of an edge that links a parent node to a child node. Then, the link strength is simply the difference of the information content between a child and its parent node.

They defined the semantic distance between two nodes as the summation of edge weights (wt) along the shortest path between them:

$$\text{dist}(c_1, c_2) = \sum_{c \in \{\text{length}(c_1, c_2) - \text{lso}(c_1, c_2)\}} \text{wt}(c, p), \tag{2.24}$$

where $\text{length}(c_1, c_2)$ is the number the nodes along the shortest path defined by concepts c_1 and c_2, p is the parent node of c, and $\text{lso}(c_1, c_2)$ is the lowest super-ordinate function that calculates the most specific concept that subsumes both c_1 and c_2.

The overall edge weight for a child node c and its parent node p can be determined

considering factors such as local density, node depth, and link type as:

$$\mathrm{wt}(c,p) = \left(\beta + (1-\beta)\cdot\frac{\overline{\mathrm{E}}}{\mathrm{E}(p)}\right)\cdot\left(\frac{\mathrm{depth}(p)+1}{\mathrm{depth}(p)}\right)^{\alpha}\cdot\Big(\mathrm{IC}(c)-\mathrm{IC}(p)\Big)\cdot\mathrm{T}(c,p), \quad (2.25)$$

where depth(p) denotes the distance from the node p to the *root* in the hierarchy, $E(p)$ represents the number of edges in the child links and is called the local density, $\overline{E}$ is the average density in the whole hierarchy, IC(c) and IC(p) refer to the information concept defined in Equation 2.22 of the child node c and parent node p respectively, and $T(c,p)$ denotes the link relation or type factor. The parameters α ($\alpha \geq 0$) and β ($0 \leq \beta \leq 1$) control the degree of how much the node depth and density factors contribute to the edge weighting computation. These contributions become less significant when α approaches zero and β approaches one.

However, the distance function that is commonly adopted by researchers corresponds to the simplification of α=0, β=1, and $T(c,p)$=1, which leads to:

$$\mathrm{dist}(c_1,c_2) = \mathrm{IC}(c_1) + \mathrm{IC}(c_2) - 2\cdot\mathrm{IC}(c), \quad (2.26)$$

where c is the lowest super-ordinate of both concepts, i.e. c=lso(c_1,c_2). This formula results in a distance between the two concepts: concepts that are more related have a lower score than the less related ones. In order to maintain consistency among the measures, this measure of semantic distance can be converted into a measure of semantic relatedness as follows:

$$\mathrm{rel}(c_1,c_2) = \frac{1}{\mathrm{dist}(c_1,c_2)}. \quad (2.27)$$

According to the formula, the relatedness can be undefined when there is a zero in the denominator. This happens in the following situations:

1. $\mathrm{IC}(c_1) + \mathrm{IC}(c_2) = 2\cdot\mathrm{IC}(c)$:

   For example, when $\mathrm{IC}(c_1) = \mathrm{IC}(c_2) = \mathrm{IC}(c)$ and this happens when c_1, c_2, and c correspond to the same concept.

2. $\mathrm{IC}(c_1) = \mathrm{IC}(c_2) = \mathrm{IC}(c) = 0$:

   This implies that the value of the information content of the three nodes is zero. If $\mathrm{IC}(c) = 0$ is because the least common subsumer of c_1 and c_2 is the root node. The reason behind the zero value of the information concept of the nodes c_1 and c_2 is because of the lack of data.

Finally, the semantic measure has a lower bound of zero and no upper bound. Contrary to Resnik's measure, this semantic similarity depends on the value of the information content of the actual nodes so the fact of two pairs of nodes sharing the same least common subsumer does not imply that they share also the same similarity value.

Lin (LIN)

Lin (1998) presented a measure that scales the information content of the most specific concept that subsumes both c_1 and c_2 by the sum of the information content of the individual concepts. The similarity between concepts c_1 and c_2 is defined as:

$$\mathrm{sim}(c_1, c_2) = \frac{2 \cdot \mathrm{IC}(lso(c_1, c_2))}{\mathrm{IC}(c_1) + \mathrm{IC}(c_2)}. \tag{2.28}$$

Note that a zero in the denominator will give undefined relatedness. This will occur when the value of the information content of c_1 and c_2 is zero.

Figure 2.2(b) shows a fragment of WordNet's noun hierarchy where number in brackets represent the information content associated to each node. In the example, the node c corresponds to c_5.

The LIN measure is rather similar to JCN, as their semantic value depends on the IC of the actual nodes, contrary to RES. This measure is set to one as upper bound and it corresponds to the case when the information content of the three nodes coincides, the lower bound is set to zero.

2.3.3 Gloss-based Measures

In WordNet, each concept or word sense is defined by a short definition called gloss. For example, the gloss of the concept car is: *four wheel motor vehicle usually propelled by an internal combustion engine.*

The following measures use the text of that gloss as a unique representation for the underlying concept.

Adapted Lesk (LESK)

Lesk (1986) proposed that the relatedness of two words is proportional to the extent of overlaps of their dictionary definitions. For example, consider the WordNet glosses of car and tyre: *four wheel motor vehicle usually propelled by an internal combustion engine* and *hoop that covers a wheel, usually made of rubber and filled with compressed air.* The relationship between these concepts is given by their glosses sharing the word "wheel".

Banerjee and Pedersen (2003) extended this notion to use WordNet as the dictionary for the word definitions. The novelty of their proposed approach, the *extended gloss overlap* measure, resides in the way of finding and scoring the overlaps between two glosses. The original Lesk algorithm compared the glosses of a pair of concepts and computed a score by counting the number of words that are shared between them. Thus, this scoring mechanism does not differentiate between single word and phrasal overlaps and treats each gloss as a "bag of words". However, the *extended gloss overlap* algorithm considers multiple word matches, which are scored higher than single word matches

The score function provided by the *extended gloss overlap* measure is defined as follows. The final score for a given pair of glosses is computed by squaring and adding

together the sizes of the overlaps found. An overlap between two glosses is the longest sequence of one or more consecutive words that occurs in both glosses such that neither the first, nor the last word is a *function word*, a pronoun, preposition, article, or conjunction. If two or more such overlaps have the same longest length, then the overlap that occurs earliest in the first string being compared is reported. Given two strings, the longest overlap between them is detected, removed and in its place a unique marker is placed in each of the two input strings. The two strings obtained are, again, checked for overlaps, and this process continues until there is no more overlaps between them.

The *extended gloss overlap* measure computes the relatedness between two concepts c_1 and c_2 by comparing the glosses of all the concepts related to them through explicit relations provided by WordNet:

$$\text{rel}(c_1, c_2) = \sum \text{score}(R_1(c_1), R_2(c_2)),\ \forall (R_1, R_2) \in \text{relPairs}\,. \tag{2.29}$$

Here, the set relPairs guarantees that the final measure is reflexive, rel(c_1, c_2)=rel(c_2, c_1) and it is defined as follows:

$$\text{relPairs} = \{(R_1, R_2) \mid R_1, R_2 \in \text{RELS};\ \text{if } (R_1, R_2) \in \text{relPairs}, \text{then } (R_2, R_1) \in \text{relPairs}\}, \tag{2.30}$$

where RELS is a non-empty set of relations that consists of one or more WordNet relations as defined in Table 2.3:

$$\text{RELS} \subset \{r \mid r \text{ is a relation defined in WordNet}\}. \tag{2.31}$$

For instance, if RELS = {hypernymy}, $r(c_1)$ will return the associated gloss of the hypernymy synset of c_1.

Let's illustrate with an example how to compute this measure; given that RELS,

the set of relations of the concepts linked to concepts c_1 and c_2, is:

$$\text{RELS} = \{\text{gloss, hypernymy, hyponymy}\}, \tag{2.32}$$

then, the set relPairs is the following:

$$\begin{aligned}\text{relPairs} \;=\; & \big\{(\text{gloss, gloss}), (\text{hypernymy, hypernymy}), (\text{hyponymy, hyponymy}),\\ & (\text{hypernymy, gloss}), (\text{gloss, hypernymy})\big\}.\end{aligned} \tag{2.33}$$

Then, the semantic relatedness between concepts c_1 and c_2 can be defined as:

$$\begin{aligned}\text{rel}(c_1, c_2) \;=\; & \text{score}\big(\text{gloss}(c_1), \text{gloss}(c_2)\big) + \text{score}\big(\text{hypernymy}(c_1), \text{hypernymy}(c_2)\big) +\\ +\; & \text{score}\big(\text{hyponymy}(c_1), \text{hyponymy}(c_2)\big) + \text{score}\big(\text{hypernymy}(c_1), \text{gloss}(c_2)\big) +\\ +\; & \text{score}\big(\text{gloss}(c_1), \text{hypernymy}(c_2)\big).\end{aligned} \tag{2.34}$$

Vector measure (VEC)

Patwardhan (2003) observed that the *extended gloss overlap* measure depends on the exact match of words, missing overlaps between a term and its plural form or a semantically similar version. Thus, the presence of a word like "car" in two glosses would contribute to their overlap score. However, if one of the two glosses contained "car" and the other contained "cars", the overlap would be missed. Additionally, conceptual matches like "car" and "automobile" would not even be considered.

To overcome this limitation, they augmented the words in the glosses with data coming from external sources. They proposed a co-occurrence matrix from a corpus made up of WordNet glosses. Each word used in a WordNet gloss has an associated context vector. The word vector corresponding to a given word is calculated as a vector of integers; the integers are the frequencies of occurrence of each word from the word space in the context of the given word in a large corpus. The word space is determined

by a list of words used to form the vectors. The context of a given word is defined by the words that proceeds and follows it. Thus, each word in the word space represents a dimension of the vector space. Each gloss is represented by a gloss vector that is the average of all the context vectors of the words found in the gloss. Finally, the semantic relatedness between concepts c_1 and c_2 is measured by calculating the cosine of the angle between the corresponding gloss vectors $\vec{v_1}$ and $\vec{v_2}$:

$$\text{rel}(c_1, c_2) = \frac{\vec{v_1} \cdot \vec{v_2}}{|\vec{v_1}| \cdot |\vec{v_2}|}, \tag{2.35}$$

where $\vec{v_1}$ and $\vec{v_2}$ represent, respectively, the gloss vectors associated with concepts c_1 and c_2.

Finally, one of the advantages of this measure is that is not dependent on WordNet glosses and can be employed with any representation of concepts such as dictionary definitions using co-occurrence counts from any corpus.

2.3.4 Discussion on WordNet Measures

As seen in previous subsections, WordNet semantic measures can be divided in three types: *path length*, *information content*, and *gloss-based* measures.

The main problem of the *path length* measures is that WordNet is not equally balanced as some branches are more dense than others. Consequently, the estimation of the semantic relatedness may produce misleading results as these measures are based on computing distances.

Resnik's *information content* measure attempted to address this limitation by augmenting the information present in WordNet with statistical information from large corpora. However, it cannot differentiate the semantic similarity of those concepts that share the same least common subsumer concept.

Finally, some measures based on the definition associated to each concept called *gloss*

Table 2.5: Coefficient of the correlation between machine-assigned and human-judged scores for the best performing WordNet-based measures computed for the Miller and Charles (M&C) and for the Rubenstein and Goodenough (R&G) datasets

Measure	M&C	R&G
Jiang and Conrath (1997) (JCN)	0.850	0.781
Leacock and Chodorow (1998) (LCH)	0.816	0.838
Lin (1998) (LIN)	0.829	0.819
Resnik (1995) (RES)	0.774	0.779
Hirst and St-Onge (1998) (HSO)	0.744	0.786

were additionally proposed. Nevertheless, these *gloss-based* approaches presented as well a limitation as they cannot provide the semantic distance when there exists no shared word between the two glosses.

Because of these drawbacks, best results have been achieved when combining several semantic measures using appropriate fusion methods. Moreover and, as it can be seen in Section 2.3.6, the best performing results are obtained by combining the top performing measures. The performance of semantic relatedness measures is evaluated by establishing a comparison with human similarity judgements (see Section 2.1.1). Then, a coefficient that measures the correlation between them is computed. Table 2.5 shows a review by Budanitsky and Hirst (2006) on the best performing WordNet based semantic measures using two benchmark vocabularies: the Miller and Charles (1991) and the Rubenstein and Goodenough (1965) datasets. Figures refer to the absolute value of the coefficient of the correlation between machine-assigned and human-judged scores.

The application of these measures to automated annotation algorithms is not straightforward and it presents several shortcomings. The first is the necessity of handling some words that do not exist or do not have available relations with other words of

the thesaurus. For instance, the Corel 5k dataset is formed by 374 words but 63 of them correspond to plurals as shown by Table A.1 of the Appendix. Consequently, a morphological parser should be employed in order to detect the plurals and change them into their singular form. Additionally, there exist words such as "boeing", "f-16", "close-up","rockface", "white-tailed", and "dall" that are unknown to WordNet (version 3.0). However, none of the authors that use WordNet in their applications mentions how they deal with this problem. My intuition is that they use the common sense approach proposed by Agirre et al. (2009), which is based on replacing these words by others similar in meaning that actually belong to WordNet. For example, "boeing" could be replaced by "aircraft", "f-16" by "jet", etc.

Finally, in contrast with other methods, WordNet based measures work with concepts (word senses) instead of directly with words. Moreover, words are represented in the form of synsets, *a set of words that are interchangeable in some context without changing the truth value of the preposition in which they are embedded.* A synset presents the structure of (word#pos#lex_id), where "pos" stands for part of speech (noun "n", verb "v", adjective 'a", adjective satellite "s", and adverb "r") and "lex_id" is an integer number that ranges from one to 15, which is used to distinguish different senses of the same word. Therefore, another requirement is to find, for every word in the vocabulary, the sense attributed by the human annotator that provided the initial annotations to the image collection.

2.3.5 Word Sense Disambiguation Methods applied to WordNet

This section summarises some approaches that disambiguate words into concepts. Gale et al. (1992) experimentally proved the *one sense per discourse* hypothesis, which claims that well-written discourses tend to avoid multiple senses of a polysemous word.

Based on this assumption, Barnard et al. (2001) considered that a word has only one meaning within a context. Then, they proposed a disambiguation method applied to image collections, which consists in selecting the sense of a word whose hypernym is shared by its co-occurred words in the dataset according to WordNet. They employed this technique in two image collections, one the Corel dataset and the other a collection formed by images and their corresponding text extracted from the web. Additionally, they found that this disambiguation technique performs better for the Corel dataset. This is due to the fact that the Corel dataset has been annotated using one sense per word for the whole collection while this is not usually the case in a collection extracted from the web.

A very interesting approach was proposed by Weinberger et al. (2008), which is based on resolving ambiguity by proposing two words that are likely to co-occur with the ambiguous word but give rise to maximally different probability distributions. Their method was initially developed for Flickr but it can be applied to the case of the Corel 5k dataset or any other annotation benchmark.

Additionally, this thesis proposes in Chapter 4 a simple and straightforward approach that consists in assigning automatically to every word the first sense attributed in its synset, which is supposed to be the most frequent. Moreover, this approach is compliant with the one sense per discourse proposed by Gale et al. (1992). For example, the polysemic word "palm#n" presents the following senses according to WordNet:

1: the inner surface of the hand from the wrist to the base of the fingers;

2: a linear unit based on the length or width of the human hand;

3: palm tree, any plant of the family Palmae having an unbranched trunk crowned by large pinnate or palmate leaves;

4: an award for winning a championship or commemorating some other event.

This technique will select the sense of the first synset ("the inner surface of the hand..."), which might or might not match the sense of the collection. In the Corel 5k case, the sense attributed to images annotated with the word "palm" corresponds to "palm tree", the third sense in the synset. However, the accuracy in the disambiguation process remains around 81% for the Corel 5k dataset, which suggests that it is rather good; specially because of the simplicity of the approach.

Table A.2 of the Appendix shows all the cases (68) in which the sense of the first synset does not match the sense attributed to the Corel 5k dataset collection.

As a result, it is very important to adopt an effective disambiguation strategy as the inaccuracies in the disambiguation process might translate into inferior results for the resulting annotation method.

With respect to the influence of the breadth of the domain on the disambiguation strategy, when presented with a narrow technical domain, such as a set of images showing faults in aircraft fuselage, there may be reduced ambiguity and a high degree of similarity between the images. On the contrary, in case of a broader domain, such as a personal photo collection, the ambiguity between annotation words would be higher and an efficient WSD method should be implemented in order to obtain the correct annotations for the collection.

2.3.6 WordNet and Automated Image Annotation

The first attempt to improve the accuracy of the annotations by applying semantic similarity measures using WordNet was made by Jin et al. (2005b). They claimed that *whatever the statistical method employed*, the accuracy of the resulting annotations is quite low as a result of too many unrelated keywords. Their proposed *translation*

method with hybrid measures (TMHD) can be described as follows. Initial annotations are generated using a baseline annotation algorithm, the *translation method* (Duygulu et al. 2002). The next step consists in detecting annotation words unrelated to the others by applying semantic similarity measures based on WordNet. Finally, these unrelated words are discarded. They considered that two words are unrelated (noisy) when their semantic similarity value falls below a given threshold. They examined the following semantic similarity measures: Resnik measure (RES), Jiang and Conrath measure (JCN), Lin measure (LIN), Leacock and Chodorow measure (LCH), and Banerjee and Pedersen measure (LESK). They evaluated independently the performance of each measure over the baseline translation method. Finally, they proposed a combination of the best performing JCN, LIN, and LESK. After applying their model to the Corel 5k dataset, they reported a 56.87% improvement over the baseline method in terms of precision. Later on, the same authors (Jin et al. 2005a) presented an improved version of their previous paper, where they combined the semantic similarity measures selected JCN, LIN, and LESK using a *Dempster-Shafer evidence combination method* (Shafer 1976). They demonstrated that their system can overcome the majority of noisy words and provide the correct annotations for the image. They proposed the following pair of concepts, c_1 and c_2, as the disambiguated senses of words w_1 and w_2, as computed by the formula

$$(c_1, c_2) = \text{argmax}_{w_1,w_2} \, [\text{sim}(w_1, w_2)], \tag{2.36}$$

which means that for a given pair of words, w_1 and w_2, the corresponding concepts, c_1 and c_2, are those that provide the highest semantic similarity. However, one of the weakest point of this model is that the F-measure remains unchanged when compared to the baseline method so the benefit of this approach is unclear.

Srikanth et al. (2005) also built an automated image annotation framework using

the *translation method* (Duygulu et al. 2002) that makes use of WordNet in order to induce hierarchies on annotation words and then, improve the performance of the baseline model. Thus, the topology of the hierarchy is externally defined by WordNet and annotation words are attached to the nodes of the hierarchy based on their semantics. However, one of the initial steps in the process is to identify the sense attributed to each one of the annotation words in the collection. First, they made the assumption that a particular word is used with only one sense in the whole collection. Second, the sense is selected based on the number of times the *child term* and the *parent term* associated to that sense appear as annotation words in the collection. According to WordNet, the word "palm" has four senses: the inner surface of the hand, a palm tree, a linear unit, or an award for winning a championship. In this case, *parent terms* will be solely considered as some of them do not have associated *child terms*. The disambiguation process will be, then, a matter of counting the number of times their corresponding *parent terms*, "area", "tree", "linear measure", and "award", appear in the whole collection. Finally, the sense "palm tree" is selected as the Corel 5k dataset is populated by images of palm trees.

Then, they tested their approach using different configurations of the visual vocabulary. Finally, they successfully demonstrated that the hierarchical classification provides significant improvement over the translation method.

Li and Sun (2006) presented a new approach that incorporates lexical semantics into the image annotation process. The model works in two steps. There is an initial phase where annotation words are generated using k-means clustering combined with semantic constraints obtained from WordNet. The second phase refines these initial annotations through a *hierarchical ensemble* model composed of probabilistic *support vector machine* classifiers and a *co-occurrence language model*. With respect to the seman-

tic measure employed, they utilised the software developed by Pedersen et al. (2004), which provides nine WordNet semantic relatedness measures (WUP, LCH, PATH, HSO, RES, JCN, LIN, LESK, and VEC). However, the authors did not provide any indication about how they performed the disambiguation process. They reported results for the Corel 5k dataset, outperforming the *cross-media relevance model* (Jeon et al. 2003) and the *translation model* (Duygulu et al. 2002) in both average precision and recall.

Shi et al. (2006) also built a concept hierarchy using the terms of the vocabulary for the Corel dataset and WordNet. They applied the same disambiguation method as Barnard et al. (2001), who consider that a word has only one meaning within a context. With this assumption, the sense of a word, whose hypernym (*parent term*) is shared by its co-occurred words in the dataset, is selected during the disambiguation phase. They exemplified the process as follows. The term "path" is part of the Corel 5k vocabulary and has four senses in WordNet: way of life, a way especially designed for a particular use, an established line of travel or access, and a line or route along which something travels or moves. Their associated *parent terms* are "course of action", "way", "line" and "line", respectively. The word "path" appears 16 times in the training set of the Corel dataset and in seven times out of these 16, "path" is accompanied by the word "garden" while the rest of the time it is followed by "mountain", "grass", "tree", or "flower". This indicates that the right sense should be *a way especially designed for a particular use* as this sense is mostly shared by the rest of the words that co-occurs with it. Their final annotation algorithm, which follows a Bayesian learning approach, outperformed several techniques for automated image annotation such as Jeon et al. (2003) and Duygulu et al. (2002). They demonstrated the validity of their initial hypothesis that the use of a concept hierarchy facilitates the modelling of multi-level concept structures.

Liu et al. (2006) used WordNet semantic similarity measure linearly combined with statistical co-occurrence information (Eq. 2.4) among words to prune non-related annotation words. They employed the Jiang and Conrath Measure (JCN), which integrates the node-based and edge-based approach together, as it is proved to be one of the most effective WordNet measure. They used the same disambiguation strategy as Jin et al. (2005a), but applied to Equation 2.24. For the Corel 5k dataset, they reported better results than Jin et al. (2005b), who were the first to use WordNet and co-occurrence.

Shi et al. (2007) proposed a novel framework where they integrate a concept ontology derived from WordNet (Shi et al. 2006) with a text-based Bayesian learning model. They attempt to tackle the problem of expanding the training set for each annotation word without the need of additional human-annotators or using other training images from other collections. They demonstrated the validity of their approach for the Corel 5k datatet.

Fan et al. (2007) presented a *hierarchical classification* framework for automated image annotation that effectively bridges the semantic gap. They built a concept ontology using word co-occurrence and a semantic similarity measure based on Leacock and Chodorow (1998) WordNet measure. With respect to the disambiguation strategy, they applied the same approach as Srikanth et al. (2005), which means that for a given pair of words, w_1 and w_2, the corresponding concepts, c_1 and c_2, are those that provide the highest semantic similarity. Their experiments on large-scale image collections, such as Corel 5k, LabelMe, and Google Images, obtained very good results although they only provided precision and recall values for some words.

2.4 Web-based Correlation

The first proposed semantic web measures (Chen et al. 2006, Sahami and Heilman 2006) employ text *snippets*[5]. Bollegala et al. (2007) propose a robust semantic similarity measure that makes use of page count and text snippets returned by a web based search engine to compute the semantic similarity between two words. By using Miller and Charles dataset, they achieve better correlation with human judgement than previous web-based measures (Chen et al. 2006) and (Sahami and Heilman 2006) but lower than WordNet measures. Nevertheless, they achieve comparable results with the top performing WordNet measures such as (Lin 1998), (Li et al. 2003), and (Jiang and Conrath 1997).

More recently, a more successful measure, which relies on Google as search engine, was proposed by Cilibrasi and Vitanyi (2007). This new measure, the *normalized Google distance* (NGD), is based on information distance and Kolmogorov complexity and it counts the number of textual documents retrieved by Google, given the words w_1 and w_2. It is defined as

$$\mathrm{NGD}(w_1, w_2) = \frac{\max\{\log \mathrm{f}(w_1), \log \mathrm{f}(w_2)\} - \log \mathrm{f}(w_1, w_2)}{\log N - \min\{\log \mathrm{f}(w_1), \log \mathrm{f}(w_2)\}}, \tag{2.37}$$

where $\mathrm{f}(w_1)$ and $\mathrm{f}(w_2)$ are, respectively, the counts for search terms w_1 and w_2, and $\mathrm{f}(w_1, w_2)$ is the number of documents found on which both w_1 and w_2 occur. N is the total number of web pages searched by Google which, in 2007, was estimated to be more than 8bn pages. NGD ranges from zero to ∞ although most of the values fall between zero and one. Note that the smaller the value of NGD, the greater the semantic relation between the two words.

Later on, Cilibrasi and Vitany propose a generalisation of the previous by defining

[5]Snippet refers to a short piece of text.

the *normalized web distance*, $\mathrm{NWD}(w_1, w_2)$, the same NGD formula, but using any web search engine as source of frequencies. Despite the fact that NGD does not obey the triangular inequality property, it provides a good measure of how far two terms are semantically related. Moreover, experimental results demonstrate that the NGD is scale invariant as it stabilises with the growing Google dataset. However, a usual criticism is that it is neither a bounded measure nor it is normalised.

Gracia and Mena (2008) applied a transformation on Equation 2.37 proposing their web-based semantic relatedness measure between the words w_1 and w_2, as:

$$\mathrm{rel}(w_1, w_2) = e^{-2 \cdot \mathrm{NWD}(w_1, w_2)} \ . \tag{2.38}$$

This transformation was done in order to get a proper measure that is both a bounded value and in the range of $[0, 1]$ and at the same time increases inversely to distance. They considered the following web search engines in their experiments: Google, Yahoo!, Live Search, Altavista, Exalead, Ask, and Clusty. In order to validate their research, they compared their results with some WordNet measures. The best correlation with human judgment was obtained by Exalead-based measures, closely followed by Yahoo! and Altavista. On the contrary, WordNet-based measures such as Resnik, Adapted Lesk, Wu & Palmer, Hirst & St-Onge, Lin, and Leacock & Chodorow, obtained the lowest results.

Liu et al. (2007) proposed two types of word correlation measures with different characteristics. The first one is called *statistical correlation by search* and is a variation of Equation 2.37, which can be defined as

$$\mathrm{K_{SCS}}(w_1, w_2) = e^{-\gamma \cdot \mathrm{NGD}(w_1, w_2)}, \tag{2.39}$$

where NGD corresponds to the *normalized Google distance* but using an image search engine. Then, $\mathrm{f}(w_1)$ represents the number of images found in Google image search

engine using the word w_1 as a query and f(w_1, w_2) is the number of images indexed by both w_1 and w_2. N, the total number of indexed images by Google has not been published officially, is set to 4.5 billions. γ is an adjustable parameter. This formula was proposed in order to solve, again, the problem of NGD not being bounded and normalised. However, its novelty resides in the use of an image search engine instead of the usual textual search engine. The second type of word correlation is called *content-based correlation by search* and it is estimated by computing the visual consistence of the sets of images (top 20) retrieved by Google image search engine. Thus,

$$\mathrm{K_{CCS}}(w_1, w_2) = e^{-\sigma \cdot \frac{\mathrm{DPV}(S_{w1,w2})}{\min\{\mathrm{DPV}(S_{w1}),\ \mathrm{DPV}(S_{w2})\}}}, \tag{2.40}$$

where σ is a positive smoothing parameter, S_{w1} is the set of images retrieved by submitting the query w_1, S_{w2} is the set of images retrieved by submitting the query w_2, and $S_{w1,w2}$ is the set of images retrieved by submitting the query w_1 and w_2. The dynamic partial variance (DPV) is used to describe the visual consistence of a set of images:

$$\mathrm{DPV}(S) = \frac{1}{m} \cdot \sum_{i=1}^{m<d} \mathrm{var}_i(S), \tag{2.41}$$

d being the dimension of the visual feature and m the number of similar aspects activated in the measure. The variances of each dimensional feature among images in set S are ordered according to $\mathrm{var}_1(S) \leq \mathrm{var}_2(S) \leq ... \leq \mathrm{var}_d(S)$.

The underlying idea behind this measure is that two words that are semantically related should correspond to images whose visual features are visually consistent. Authors illustrate this idea with the example of the polysemous word “jaguar”. After submitting it as a query, the image search engine will return images according to all the senses of the word such as animal, car, or plane images. Then, those images not visually consistent with the others should be discarded. Finally, they propose the cor-

relation between words w_1 and w_2 as a linear combination of the previous Eq. 2.39 and Eq. 2.40 as shown in

$$\mathrm{rel}(w_1, w_2) = \lambda \cdot \mathrm{K_{SCS}} + (1 - \lambda) \cdot \mathrm{K_{CCS}}, \tag{2.42}$$

where $0 < \lambda < 1$.

2.4.1 WWW and Automated Image Annotation

Wang and Gong (2007) presented an approach that refines candidate annotations generated by a *relevance vector machine* approach using, for the first time, statistical correlation of words on the web. In particular, they modelled semantic relations between words using a *conditional random field* model where each node indicates the final decision (true/false) on a candidate annotation word. The statistical correlation is done using the *normalized Google distance* (Eq. 2.37). Their work is closely related to Jin et al. (2005b) and to Wang et al. (2006); although there are some differences. They adopted the same strategy as Wang et al. (2006) as both of them integrate the confidence score of the initial candidate annotations with the contextual knowledge in the refining stage. One difference is that each approach uses different knowledge source. Jin et al. (2005b) use WordNet, Wang et al. (2006) use statistical occurrence in the training set while Wang and Gong (2007) use the web through the NGD. Finally, they compared their results with Wang et al. (2006) as both of them use the same baseline annotation strategy. They used the Corel 5k image collection for their experiments demonstrating that their approach outperforms that proposed by Wang et al. (2006) in terms of precision and recall.

Liu et al. (2007) proposed a *dual cross-media relevance* model (DCMR), a new relevance model, which annotates images by maximizing the joint probability of images and words. Thus, they considered two types of relations: *image-to-word* and *word-to-*

word relations. The novelty of their approach lies in the fact that the estimation is based on the expectation over words in a given lexicon instead of over the training set as accomplished by traditional relevance methods such as (Jeon et al. 2003, Lavrenko et al. 2003, Feng et al. 2004). The *word-to-image* relation is obtained after submitting (textual) queries to an image search engine and estimating the similarity between the visual features extracted from the retrieved images and from the un-annotated image. The *word-to-word* relation combines a statistical correlation together with a content-based correlation by search as expressed in Equation 2.42. For the Corel 5k dataset, they outperformed state-of-the-art relevance approaches like the MBRM (Feng et al. 2004).

They conducted, additionally, another experiment in order to evaluate which of the semantic measures, WordNet (Jin et al. 2005b), training set correlation (Wang et al. 2006), or their proposed web correlation, achieve the highest performance. Thus, they replicated the MBRM model (Feng et al. 2004), which is used as baseline approach, and compared its performance with each one of the different measures. They reached the following conclusions. First, the statistical correlation using the training set, gains significant improvement over the baseline in terms of the number of recalled words (NZR) but it losses a little on the average precision. This shows that the correlation is capable of connecting more words through the statistical information, but the connections cannot ensure the relatedness on the semantic level. Second, the combined web correlation (Eq. 2.42) achieves overall improvement although, the *content-based* one (Eq. 2.41) seems to be better. This is because both web-based correlations are in the web context and accordingly provide the word correlations from a more general and reasonable level. Additionally, the *content-based correlation* is estimated using an adaptive measurement, i.e. DPV, which is more robust to web noise. Third, WordNet

shows the worst performance. One of the reasons behind this poor performance is that there are some words in the Corel dataset that do not exist in WordNet or have no available relations with other words in the thesaurus. Finally, the best performance is obtained when integrating the combined web correlation (Eq. 2.42) with the correlation in the training set (Wang et al. 2006). Both of them give a relatively precise and comprehensive representation of word semantic relatedness, which shares the advantages from each single correlation. However, they selected for their implementation the web correlation as they wished to make the correlation independent on a certain dataset.

Stathopoulos et al. (2008) proposed a multi-modal graph based on *random walks with restarts* (RWR) (Lovász 1993) that exploits the co-occurrence of words in the world wide web assuming a global meaning of words. They applied the *normalized Google distance* as seen in Eq. 2.37 for exploiting the correlation of words in the web. However, the average difference between the baseline method (RWR) and the semantic method (RWR+NGD) shows that they are not statistical significant for the Corel 5k dataset. Authors provide two possible explanations: First, NGD is not symmetric despite the fact that it is assumed that the search engine returns the same number of pages regardless of the order of the words in the query. Second, they doubt that the application of word correlation on the web can be beneficial for all the words in the vocabulary.

Llorente et al. (2009b) presented an algorithm that exploits the correlation between words computed using several web-based search engines such as Google and Yahoo. Specifically, they refined the annotations obtained from a *non-parametric density estimation* applying the semantic relatedness measure of Equation 2.38, which is a symmetric version of the *normalized Google distance*. Experiments were carried out with two datasets, the Corel 5k and the ImageCLEF 2008, achieving in both cases an

Table 2.6: Correlation between machine-assigned and human-judged scores for the Wikipedia-based measures using different datasets

Measure	M&C	R&G	WS-353
Gabrilovich and Markovitch (2007)	0.73	0.82	0.75
Milne and Witten (2008)	0.70	0.64	0.69
Ponzetto and Strube (2007a)	0.49	0.55	0.55

improvement in performance over the baseline in terms of mean average precision.

2.4.2 Web Correlation Discussion

Results confirm the expectation that distributional measures perform better than those based on lexical resources, such as WordNet. The reasons behind this is that they do not need to accomplish any disambiguation task and they do not need to deal with some vocabulary words not being part of the thesaurus. Additionally, the relations provided by WordNet are limited to *is-a* relations while the web allows the exploitation of additional relationships between words.

Web correlation methods also outperform statistical methods based on correlation on the training set as they are not biased by the topics represented in the collection.

2.5 Wikipedia-based Measures

Wikipedia is a free and on-line encyclopedia that was launched in 2001. Its structure is composed mainly of the following elements: *articles*, which are the basic unit of information; *redirects*, which are pages containing only a redirect link; *disambiguation pages*, which are special pages displaying several disambiguation options; and *categories*, which are merely nodes for organizing the articles they contain.

According to a review by Medelyan et al. (2009), the computation of semantic

relatedness using Wikipedia has been addressed from three different point of views; one that applies WordNet-based techniques to Wikipedia followed by Ponzetto and Strube (2007b); another that uses *vector model* techniques to compare similarity of Wikipedia articles proposed by Gabrilovich and Markovitch (2007); and, the final one, which explores the Wikipedia as a hyperlinked structure introduced by Milne and Witten (2008).

Table 2.6 shows the absolute values of the correlation coefficient between machine-assigned and human-judged scores obtained for the three semantic measures with the datasets M&C, R&G, and WS-353, which were introduced in Section 2.1.1.

In what follows, the Milne and Witten's measure is introduced, as it has been the only measure applied to automated image annotation. Thus, Milne and Witten (2008) proposed their *Wikipedia link-based measure* (WLM), which extracts semantic relatedness measure between two concepts using the hyperlink structure of Wikipedia. The semantic relatedness between concepts c_1 and c_2 is estimated by the angle between the vectors of the links found between the Wikipedia articles whose title matches each one of the concepts:

$$\text{rel}(c_1, c_2) = \frac{\vec{c_1} \cdot \vec{c_2}}{|\vec{c_1}| \cdot |\vec{c_2}|}, \tag{2.43}$$

where the vectors for article c_1 and c_2 are built using link counts weighted by the probability of each link occurring:

$$\vec{c_i} = (w(c_i \rightarrow l_1), w(c_i \rightarrow l_2), ..., w(c_i \rightarrow l_n)), \tag{2.44}$$

where $i = 1, 2$ and the operator $\rightarrow$ represents a link between two Wikipedia pages.

Thus, the weighted value w for the link $a \rightarrow b$ can be defined as:

$$w(a \rightarrow b) = |a \rightarrow b| \cdot \log \left(\sum_{c_1=l}^{t} \frac{t}{|c_1 \rightarrow b|} \right), \tag{2.45}$$

being t is the total number of articles within Wikipedia.

2.5.1 Wikipedia and Automated Image Annotation

Until now, the work that will be presented in Chapter 4 is the only one that uses semantic relatedness measures and Wikipedia to augment automated image annotation models. The approach adopted uses the measure introduced by Milne and Witten (2008), the WLM. Due to the fact that it works with the hyperlink structure of the Wikipedia, it is less computationally expensive than the ones that work with the whole content. In particular, the performance is comparable to WordNet results. However, Wikipedia is used combined with other semantic measures as its use alone does not provide any improvement over the baseline annotation method.

2.5.2 Wikipedia Discussion

Wikipedia needs to accomplish a previous disambiguation task as part of the computation of semantic relatedness. The disambiguation strategy adopted in the work proposed in Chapter 4 consists in automatically assigning for each word the most probable sense according to the content stored on the Wikipedia database. The accuracy in the disambiguation is similar but slightly higher than WordNet and it is around 90%. Again, these inaccuracies in the disambiguation process translate into inferior results for Wikipedia based annotation methods. Contrary to web-based methods, semantic relatedness measures using Wikipedia are computed off-line as Wikipedia allows to download dumps of its database that contains all the data. One of the advantages of using the WLM measure is that it works with the hyperlink structure of Wikipedia rather than with its whole content.

2.6 Flickr-based Measures

The use of Flickr as an external corpus to perform statistical correlation is a novel approach in automated image annotation.

Wu et al. (2008) propose a novel measure able to quantify semantic relationships between concepts in the visual domain. The process is defined as follows. First, authors select 1,000 semantic concepts from a pool of Flickr tags whose frequency range from 1k to 50k in order to eliminate the usual Flickr noise such as misspelling errors, combination of words, and affix variation. Then, they create a collection of images by selecting for each concept, the first 1,000 images retrieved by them. Finally, they define the *Flickr distance* (FD) between two concepts as the average square root of the Jensen-Shannon divergence between the two latent topic visual language models (VLM) associated to them. The general latent topic VLM algorithm is a modified version of the VLM model proposed by Wu et al. (2007) and is based on the assumption that images annotated by the same concept share not only similar features but also similar arrangement patterns. The rest of their paper aims to demonstrate that the *Flickr distance* is outperforming the Cilibrasi and Vitanyi's *normalized Google distance* (NGD) when applied to the multimedia field. The first point in their argumentation is that the NGD is not a symmetric measure quite contrary to the *Flickr distance*. Secondly, they argue that the NGD, as it counts the number of times two concepts co-occur in textual documents retrieved by Google, is unable to deal with meronymy and concurrence relationships. Later on, they demonstrate that their measure outperforms the NGD when 12 human evaluators score the relationship between two concepts. After this, they propose as more objective evaluation the estimation of precision and recall, considering as ground truth WordNet. Thus, they selected 497 concepts out of the initial 1,000 that belong both to WordNet and to Flickr. Their conclusion is that Flickr

outperforms Goggle in 17.2% in precision and 1.6 % in recall.

Jiang et al. (2009) exploit the context information associated with Flickr images to propose a new semantic measure. The resulting measurement, the *Flickr context distance* (FCS), is a variation of Eq. 2.37 but reflects the co-occurrence statistics of words in image context rather than in textual corpus. Thus, the NGD is converted to Flickr context similarity using a Gaussian kernel and it is expressed as

$$\mathrm{FCS}(x, y) = e^{-\mathrm{NGD}(x,y)/\rho}, \tag{2.46}$$

where the parameter ρ is estimated empirically as the average pairwise NGD among a randomly pooled set of words.

2.6.1 Flickr and Automated Image Annotation

Wu et al. (2008) replicate the *dual cross-media relevance* model proposed by Liu et al. (2007) but using the Flickr distance. Then, they compare the performance of the initial model that uses the normalized Google distance (NGD) in order to conclude that the Flickr distance outperforms the NGD.

Jiang et al. (2009) propose an annotation model called *semantic context transfer* and demonstrate that when combined with the FCS measure the performance notably increase. However, the comparison with the other Flickr approach is impeded by their use of video data in their experimental procedures.

2.6.2 Flickr Discussion

Measures based on Flickr, being distributional methods per se, inherit all the advantages of the statistical co-occurrence approaches. Their only difference is that the former rely on counting co-occurrences occurring in image collections while the latter do it on textual corpus. Despite the good performance obtained by the Flickr distance,

the main disadvantage is its lack of replicability as the experiments were carried out in a specific corpus, which was not made available to the rest of the scientific community.

2.7 Conclusions

The use of semantic relatedness measures augments automated image annotation algorithms by considering that annotation words should be semantically related to each other as they share the same context, which is that depicted in the image. I described several semantic measures, such as measures performing statistical correlation on a training set, the world wide web, or the Flickr image collection, as well as measures based on lexical resources, such as WordNet and Wikipedia. In general, the comparison of results in these algorithms confirm a priori expectations that measures based on statistical correlation perform better than those based on lexical resources, such as WordNet and Wikipedia. A plausible explanation might be that lexical resources deal with relations between concepts, where a concept refers to a particular sense of a given word, while distributional similarity is a corpus-dependent relation between words. Consequently, WordNet and Wikipedia perform a previous disambiguation task as part of their computation. Thus, the selection of an effective disambiguation strategy is essential to these methods to avoid inaccuracies coming from the disambiguation process translating into inferior results. Another observed limitation of measures based on lexical resources is that they are created manually and as a result they can exhibit some errors. In addition, relevant annotation words may not appear in them. On the other hand, distributional similarity depends on a corpus, which does not suffer from these limitations. However, it is affected by the data sparseness that can result in anomalous results. With respect to the corpora used by distributional measures, they can be divided into two categories: those using a textual corpus and those using Flickr

images, or other images retrieved by web search engines. Methods relying on statistical correlation on a training set are very important as they provide an indication of the nature of the collection. However, this information can be incomplete as it is limited to the topics represented in the collection. One way of overcoming this problem is by combining this information with external sources, such as the web, Flickr, WordNet, or Wikipedia.

Regarding the way these measures have been incorporated in automated image annotation algorithms, the following points should be considered. One strategy exploits these measures as part of a language model that is embedded in the annotation model. One example is when the measures are used to build a concept hierarchy, a graph, whose nodes are the annotation words. A second strategy is devoted to prune "noisy" (non-correlated) words from the annotations and is based on the assumption that highly correlated words should be kept while those non-correlated discarded. In this case, semantic relatedness measures are employed to compute the degree to which two words are correlated.

Table 2.7 shows a comparative performance of all the approaches analysed in this chapter for the Corel 5k dataset. In all cases, annotations are made up of five words. The symbol (-) indicates that the result was not provided. Results are shown under several evaluation metrics: the number of recalled words NZR, precision and recall evaluated using the 260 words that annotate the test set, P_{260} and R_{260}, respectively; precision and recall using 49 words in dataset, P_{49} and R_{49}, respectively; and finally, the F-measure, F_{260} and F_{49} computed with 260 and 49 words, respectively.

There exists a relatively large disparity in the experimental procedures adopted by researches in the field. Thus, many authors do not use the benchmark datasets, others use non-standard evaluation measures, others compute precision and recall using

Table 2.7: Semantic-enhanced models for the Corel 5k dataset expressed in terms of number of recalled words, precision, recall, and F-measure evaluated using 260 and 49 words, respectively. The symbol (-) indicates that the result was not provided

Model	Type	Author	NZR	R_{260}	P_{260}	F_{260}	F_{49}
CLM	Training	Jin et al. (2004)	-	-	-	-	-
KM-500	WordNet	Srikanth et al. (2005)	93	0.32	0.18	0.23	-
TMHD	WordNet	Jin et al. (2005b)	-	-	-	-	0.25
SCK+HE	WN+TS	Li and Sun (2006)	-	0.36	0.21	0.27	-
BHMMM	WordNet	Shi et al. (2006)	122	0.23	0.14	0.17	-
RWRM	Training	Wang et al. (2006)	-	-	-	-	0.43
AGAnn	WN+TS	Liu et al. (2006)	-	-	-	-	-
CLP	Training	Kang et al. (2006)	-	-	-	0.27	-
VisualCog	Training	Kamoi et al. (2007)	-	-	-	-	-
Anno-Iter	Training	Zhou et al. (2007)	-	0.18	0.21	0.19	-
TBM	WordNet	Shi et al. (2007)	153	0.34	0.16	0.22	-
HierarBoost	WN+TS	Fan et al. (2007)	-	-	-	-	-
RVM_CRF	Web	Wang and Gong (2007)	-	-	-	-	0.48
DCMRM	Web	Liu et al. (2007)	135	0.28	0.23	0.25	-
RWR+ALA	Training	Stathopoulos et al. (2008)	-	0.13	0.16	0.14	-
FD-DCMRM	Flickr	Wu et al. (2008)	-	-	-	-	-
Enhanced	Web	Llorente et al. (2009b)	-	-	-	-	-

non-standard number of words or just present their results in some selected words. Consequently, the comparison of results is rather difficult or even impossible to achieve by literature revisions alone. This is reflected in Table 2.7, where some interesting approaches are shown without any figures as they do not use benchmark evaluation measures.

Figure 2.3 shows a comparison with some of the traditional annotation algorithms according to the F-measure.

A suggestion that might improve research in the field is the use of large vocabularies together with a selection of annotation words that are to be found in any dictionary. Best performance is obtained when combining several semantic measures. Statistical correlation on the training set should be always used (with a proper smoothing strategy) as it provides first-hand knowledge about the collection. However, this should be used together with external information. Additionally, further research on combining these

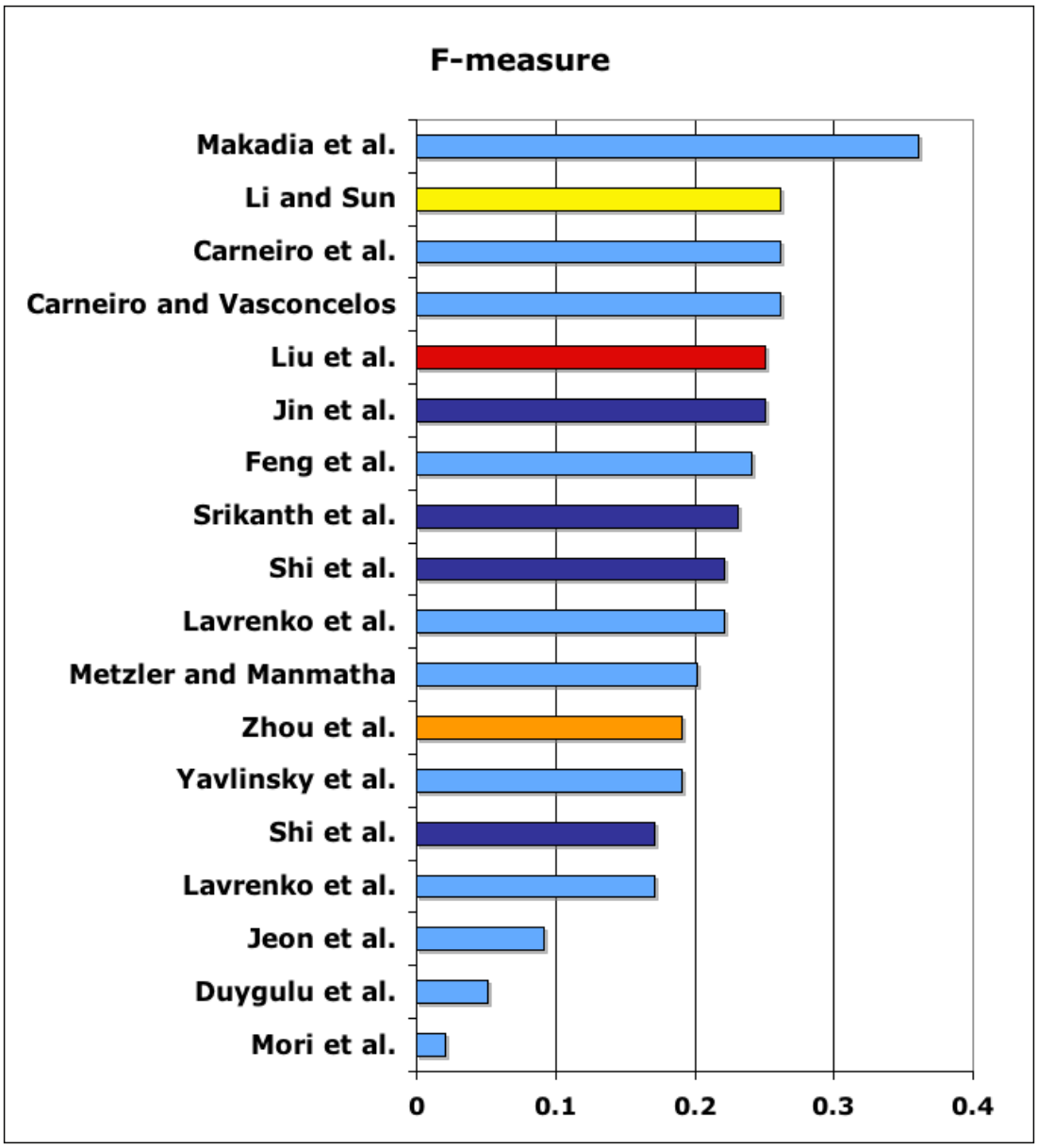

Figure 2.3: State of the art of traditional and semantic based methods in automated image annotation for the Corel 5k dataset. The horizontal axis represents the F-measure of the method represented in the vertical axis. The evaluation of the F-measure was accomplished using the 260 words that annotate the test set. Traditional methods are represented in pale blue, WordNet combined with training-based methods are in yellow, web-based methods in red, WordNet methods are in dark blue, and correlation methods on the training set are represented in orange. All methods correspond to annotation lengths of five words

methods is likely to increase the final performance. Some of these methods will be discussed in Chapter 4 together with the presentation of my method that enhances an automated image annotation solution with semantic measures. Before that, Chapter 3 will introduce the methodology used in this thesis.

Chapter 3

Methodology

The objective of this chapter is to describe the specific techniques or set of procedures used in conducting research in the field of automated image annotation.

At the beginning of the experimentation in the field, there existed a lack of a common experimental procedure. Section 3.1 provides a comprehensive overview of how experimental procedures evolved in the field until one final set-up was adopted as a common standard by researchers. This study is complementary to the overview provided in Section 1.3, although the focus here is on describing the methodology. In particular, the description of which image collections are used, how experiments were conducted, parameter estimation done, and how the annotation results are evaluated are the main focus of this chapter. Then, Section 3.2 revises the evaluation metrics mostly used in the field. Section 3.5 follows with a description of the most popular evaluation campaign initiatives, whose main objective is to develop common infrastructures for evaluating information retrieval systems. Section 3.6 reviews some past evaluation campaigns. Finally, Section 3.7 summarises the most common benchmark collections in the field, with a particular focus on the datasets used in this thesis. Section 3.8 concludes with an analysis of the main points discussed in the chapter.

3.1 A Review on Experimental Procedures

The first work done on automated image annotation (Mori et al. 1999) used a collection of 9,681 images that were accompanied by an average of 32 words each. This dataset was extracted from a Japanese multimedia encyclopedia. The annotation words were obtained from the documents cited by the images in the encyclopedia after undergoing a processing based on natural language techniques. A two-fold cross validation technique is used in the experiment. The collection is divided randomly into two groups of 4,841 and 4,840 images, one is used as a training set for learning purposes and the other as a test set for producing the annotations and vice versa. The final annotation words were the three words with largest average value of likelihood. The evaluation of results is accomplished by dividing the number of hits per image by the number of annotation words and then, averaging them for all the images of the test set in the two-fold validation. A hit is counted each time a word is correctly predicted by the annotation algorithm. Different results are shown for different values of the parameters of the system. This evaluation measure is called *hit rate* and it can be considered as the precursor of the precision in an annotation system.

Later on, Barnard and Forsyth (2001) approached image annotation as a form of *object recognition* estimating joint probability distributions for images and words. They used for their annotation experiments different subsets extracted from the Corel Stock Photo CDs (Section 3.7.1). After that, they evaluated their results applying three scoring methods. The first one attempted to average the predicted probability of each observed keyword, after scaling it by the probability of occurrence, assuming uniform statistics. A second score was proposed by normalising each prediction by the overall probability of the word given the model. For the third measure, they looked at the 15 top ranked words, scoring each inversely to its rank. Thus, if a document word

is predicted at rank five, it will give a score of 1/5. They considered that, if a word does not occur in the top 15 results, it is not worth looking at it.

Barnard et al proposed two extensions (Barnard et al. 2001) and (Barnard et al. 2002) of the method explained in Barnard and Forsyth (2001) but using a more complicated dataset: 10,000 images from the Fine Arts Museum of San Francisco. They worked with 3,319 words as vocabulary for the annotations, some of them coming from the associated text accompanying the images and others being extracted from WordNet. They defined two tasks, one of image retrieval called auto-illustration and the second was purely an annotation task called auto-annotation. Unfortunately, they did not accompany their annotation results with any evaluation measure.

Duygulu et al. (2002) were the first research team who started using a particular subset of the Corel Stock Photo CDs made up of 5,000 images (4,500 training and 500 test images), and a vocabulary of 371 words. This dataset became a benchmark known in the literature as the Corel 5k dataset (Section 3.7.2).

Barnard et al. (2003) presented an overview of different automated image annotation algorithms on a subset of 80 Corel Stock Photo CDs, from which ten different training and test sets were sampled. The average number of images used was 7,000, from which 5,200 corresponded to training images while 1,800 images were used as test set. They proposed the *Kullback-Leibler divergence* and the *normalized classification score* as evaluation metrics for algorithms based on language models. However, for methods inherited from object recognition that were based on predicting a specific correspondence between regions and words they proposed the *prediction score* and the *manual correspondence scoring*, which is introduced in order to corroborate the prediction score measure and it consists of checking the correspondence between regions and words by hand.

In her thesis, Duygulu (2003) claims that annotation performance is measured by comparing the predicted words with the words that are actually present as an annotation keyword in the image. She proposed three measures for comparing the predictions with the actual data: a measure for computing the difference between the actual and the desired probability distributions, which is called the *Kullback-Leibler divergence* (Kullback and Leibler 1951); the *normalized classification score* (Barnard et al. 2003); and the *prediction score* (Barnard et al. 2003).

Blei and Jordan (2003) introduced the *correspondence latent Dirichlet allocation* and conducted their experiments following the same experimental procedure as Barnard et al. (2003) using the same subset of the Corel Stock Photo CDs made up of 7,000 images. However, they proposed a different evaluation measure, the *perplexity* of the given caption for each image of the test set. This measure is an inherited metric from the language modelling community and is equivalent algebraically to the inverse of the geometric mean-per-word likelihood.

It was not until late 2003, when researches from the University of Massachusetts proposed an evaluation metric (precision, recall, and number of recalled words) that was later on adopted as a standard metric in the field together with the Corel 5k dataset, firstly introduced by Duygulu et al. (2002). This standard metric is explained in detail in Section 3.2. Jeon et al. (2003) undertook a comparison of previous models in automated image annotation, the *co-occurrence model* (Mori et al. 1999), *translation model* (Duygulu et al. 2002) with their own *cross-media relevance model* (CMRM). They did this comparison using the Corel 5k dataset. In order to evaluate the annotation task, they computed *precision*, *recall* and *F-measure* on 70 vocabulary words[1].

[1]These 70 words are the result of the union of the words with a recall greater than zero for the co-occurrence model, translation model, and CMRM.

To estimate the performance of the two proposed retrieval models, they used *non-interpolated average precision* over queries made up of one, two, three and four words. Thus, they only considered as queries words or combination of words that occur more than once in the test set: 179 queries of one word, 386 queries of two words, 178 queries of three words, and 24 queries of four words.

Lavrenko et al. (2003) presented a model called *continuous relevance model* (CRM) and established a comparison with previous models, the *co-occurrence model* (Mori et al. 1999), *translation model* (Duygulu et al. 2002), and *CMRM* model (Jeon et al. 2003). They employed the Corel 5k dataset and evaluated their results using *precision and recall* on 49 best words[2] and all the 260 words in the test set for the annotation task. For the retrieval task, they computed *mean average precision* on 179 queries of one word, 386 queries of two words, 178 queries of three words, and 24 queries of four words. They showed that their CRM model outperformed significantly all others known models.

However, some other researches continued using the same experimental procedure, dataset, and evaluation measures suggested by (Barnard et al. 2003). For instance, Monay and Gatica-Perez (2003) applied two latent space models namely *latent semantic analysis* (LSA) and *probabilistic LSA* (PLSA) to the problem of automated image annotation. They used a similar dataset than (Barnard et al. 2003) extracted from a subset of the Corel Stock Photo CDs. Therefore, they performed the comparison between the two models computing the *normalized score* measure. A year later, Monay and Gatica-Perez (2004) proposed a variation to the probabilistic latent semantic analysis algorithm. They proved their method by employing the same dataset as before and as evaluation metrics the *annotation accuracy* and the *normalized score*.

[2]These are the words whose recall is greater than zero for the translation method.

By 2005, many researchers such as Carneiro and Vasconcelos (2005), Jin et al. (2005a), Jin et al. (2005b), Srikanth et al. (2005), and Yavlinsky et al. (2005) and many others, started adopting the Corel 5k dataset together with the evaluation metrics proposed by the researchers of the University of Massachusetts. Despite the critics (Müller et al. 2002) and (Tang and Lewis 2007) received by the Corel 5k dataset along the years, its adoption as a benchmark dataset facilitated the necessary comparison of results among researchers that contributed undoubtedly to the establishment of automated image annotation as a solid area of research.

3.2 Standard Evaluation Metrics

Evaluation metrics measure the effectiveness of a system. In order to accomplish this, three things are needed: a collection of images, a test set of information needs expressed as queries, and a set of relevance judgements, a binary assessment of either relevant or non-relevant for each query-image pair.

Jeon et al. (2003), Lavrenko et al. (2003), and then, Feng et al. (2004) proposed the evaluation scenarios described as annotation and retrieval tasks, which are exemplified as follows.

3.2.1 Annotation Task

This task consists in computing the probability, $\mathrm{p}(w|I)$, of an image I being annotated with the word w, for all the words of the vocabulary. This results in the top five words with the highest probability. The evaluation is done by comparing the generated automatic annotations with the human annotations or ground-truth. Note that in this task it is not necessary to perform any actual ranking. Finally, the following set-based evaluation measures are computed:

Mean per-word recall (R) For each word w of the test set, recall is computed as the number of images correctly annotated with w, divided by the number of images that have that word w in the human annotations or ground-truth. *Mean per-word recall* is obtained after averaging *recall* for all the words in the test set.

Mean per-word precision (P) For each word w of the test set, precision is estimated as the number of images correctly annotated with w, divided by the total number of images automatically annotated with that particular word w, correctly or not. *Mean per-word precision* is obtained after averaging *precision* for all the words in the test set.

Mean per-word recall and *mean per-word recall* are often denoted as recall and precision, respectively.

Non-zero recall (NZR) The number of words with recall greater than zero provides an estimation of the learning capabilities of the system. It is also called the number of recalled words of the system.

For the Corel 5k dataset, these measures are usually computed for the 260 words of the test set. However, Jeon et al. (2003), Lavrenko et al. (2003), and Feng et al. (2004) report results additionally for the 49 query words that retrieve at least one relevant image in the model proposed by Duygulu et al. (2002). The number of queries that retrieve at least one relevant image vary depending on the models. For example, the co-occurrence model (Mori et al. 1999) has 19, the translation model (Duygulu et al. 2002) has 49, and the CMRM (Jeon et al. 2003) has 66 queries.

For a given annotation word w, the measures of precision and recall concentrate the evaluation on the return of the number of images correctly annotated with w, asking what percentage of images that contain w in the ground-truth have been found and how

many images have been incorrectly annotated with w. Nevertheless, the two quantities clearly trade off against one another. In general, the desired scenario is to get some amount of recall while tolerating only a certain percentage of incorrectly annotated images.

3.2.2 Retrieval Task

The description of this task is as follows. Given an image of the test set I, p($query|I$) is computed as before but instead of a word, now it is a query. All the images are ranked according to their value p($query|I$). A query is a combination of one or several words that occurs in the test set. Given a query and the n top images matches retrieved the following measure are proposed:

Mean average precision (MAP). Average precision is the average of precision values at the ranks where relevant items occur. This is averaged over the entire query set in order to obtain the mean average precision. Mean average precision provides a single-figure measure of quality across recall levels. Among evaluation measures, MAP has shown to have especially good discrimination and stability. This measure is more appropriate when the user wants to find a large proportion of relevant items.

Precision @ n. For some applications, such as web search engines, what matters is how many good results there are in the first one or two pages. This leads to measuring precision at fixed low levels of retrieved results, such as 10 or 30 images. This is referred to precision after n retrieved images. Precision is the proportion of top n images that are relevant, where relevant means that the ground-truth annotation of this image contains the query word. It has the advantage of not requiring any estimate of the size of the set of relevant images. However, it is the

least stable of the commonly used evaluation measures and it does not average well, since the total number of relevant images for a query has a strong influence on precision at n.

Discussion on Queries used for Ranked Retrieval Task

There exist two different trends in the literature about the use of queries in the *mean average precision* computation.

Jeon et al. (2003) and Lavrenko et al. (2003) worked with multiple word queries. For the Corel 5k dataset, they proposed four set of queries constructed from the combination of one, two, three and four words that occur *at least twice* in the test set: 179 queries of one word, 386 queries of two words, 178 queries of three words, and 24 queries of four words. The reason behind selecting queries that occur more than twice in the test set is to reduce the number of combinations of three and four words that otherwise could have resulted in a prohibitively large number of queries. This approach is also taken by Metzler and Manmatha (2004), Yavlinsky et al. (2005), and Magalhães (2008).

However, Feng et al. (2004) designed the retrieval task as made up of single word queries. They considered as queries all the words that occur *at least once* in the test set. For the Corel 5k dataset, they used 260 word queries. This approach is followed by Carneiro and Vasconcelos (2005), Carneiro et al. (2007), and many others.

3.2.3 Other Common Metrics

A single measure that trades off precision versus recall is the F-measure (Manning et al. 2009), which is the weighted harmonic mean of precision and recall:

$$\mathrm{F} = \frac{2 \cdot \mathrm{P} \cdot \mathrm{R}}{\mathrm{P} + \mathrm{R}}. \tag{3.1}$$

A question that may arise is why use a harmonic mean rather than a simple arithmetic mean. Note that a 100% recall can always be obtained by just returning all images, and therefore by applying the same process a 50% arithmetic mean can also be achieved. This strongly suggests that the arithmetic mean is an unsuitable measure to use. The harmonic mean is always less than or equal to the arithmetic mean and the geometric mean. In the case of the values of two numbers differing greatly, the harmonic mean is closer to their minimum than to their arithmetic mean.

Another evaluation metric followed is based on ROC curves (Fawcett 2006). ROC stands for "receiver operating characteristics". Initially, a ROC curve was used in signal detection theory to plot the sensitivity versus (1 - specificity) for a binary classifier as its discrimination threshold is varied. Later on, ROC curves were applied to information retrieval (Manning et al. 2009) in order to represent the fraction of true positives (TP) against the fraction of false positives (FP) in a binary classifier. The equal error rate (EER) is the error rate at the threshold where FP=FN, being FN the number of false negatives. A ROC curve always goes from the bottom left to the top right of the graph. For a good system, the graph should climb steeply on the left side. For unranked result sets, specificity, which is given by $\frac{\mathrm{TN}}{\mathrm{FP+TN}}$, where TN represents the number of true negatives, may not be seen as a very useful notion. Because the set of true negatives is always so large, its value would be almost one for all information needs. As a result of this, a common aggregate measure to report is the area under the ROC curve, AUC, which is the ROC analogue of mean average precision. AUC is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. This metric was used in ImageCLEF evaluation campaign during the 2008 and 2009 editions.

3.3 Multi-label Classification Measures

Traditionally, research in machine learning and neural networks was focused on defining single-label classifiers where labels were mutually exclusive by definition. Those classifiers were learnt from a set of examples associated with a single label coming from a set of disjoint labels. However, many real world situations require classes that are not mutually exclusive. This happens, for instance, with document classification where a document can belong to several classes at the same time. In the same way as standard single-label classifier approaches cannot be applied straightaway to multi-label classification, multi-label classification requires the definition of its own evaluation measures.

In the past, researchers, such as Godbole and Sarawagi (2004), Shen et al. (2004), and

Tsoumakas and Vlahavas (2007), proposed several measures that can be classified into two main categories. The first category is called *concept-based* evaluation measures, which groups any known measure, such as *equal error rate* or *area under curve* for binary evaluation. In this case, binary evaluation measures are computed for every concept and the evaluation procedure is the same as for single-label classification evaluation. The second category stands for *example-based* evaluation measures as they compare for each document the actual set of labels (ground-truth) with the predicted set using set-theoretic operations. The α-evaluation (Shen et al. 2004) and the *ontology-based score* (OS) (Nowak and Lukashevich 2009) fall into the latter category.

Depending on the averaging process, traditional information retrieval measures such as precision, recall, F-measure, accuracy, and mean average precision can be computed *concept-based* or *example-based*. An extensive comparison of the characteristics of these *concept-based* and *example-based* evaluation measures for multi-label evaluation can be found in Nowak et al. (2010b).

One drawback of these measures, and contrary to the OS, is that they do not differentiate between semantically false annotated labels and labels that have a similar meaning to the correct label. For example, the annotation of *plant* instead of *flower* would be regarded as incorrect although the concepts are semantically similar and *flower* is a specialisation of *plant*.

A practical application of multi-label classification occurs in automated image annotation whose main purpose is to generate a set of labels that best characterises the scene depicted in the image. Some recent applications, such as Fan et al. (2008), Marszałek and Schmid (2007), Schreiber et al. (2001), and Srikanth et al. (2005), use vocabularies adopting the form of taxonomies or ontologies as a natural way to classify objects. This is the case of the ImageCLEF benchmark where a task was posed about automated multi-label image annotation using ontology knowledge (Nowak and Dunker 2009b). In this task, the Consumer Photo Tagging Ontology defined by Nowak and Dunker (2009a) was provided and could be integrated in the classification procedure. Besides the OS, neither of the above mentioned evaluation measures addresses the problem of multi-label evaluation when the vocabulary adopts the hierarchical form of an ontology.

The OS uses ontology information to detect violations against real-world knowledge in the annotations and calculates costs between misclassified labels. However, there exist some cases in which the OS does not work adequately. This occurs as the OS bases its costs computation on measuring the path between concepts in the ontology. Therefore it assumes that the number of links between two concepts is determined by their mutual similarity. But links in an ontology do not usually represent uniform distances.

These limitations are illustrated on the example of the Consumer Photo Tagging Ontology as seen in Figure A.2 of the Appendix. For example, the cost obtained

between the concepts "Landscape" and "Outdoor" is 0.86, which is the same between "Landscape" and "Indoor". However, taking into consideration real-world knowledge, it is more likely that a scene depicting a landscape is an outdoor scene rather than an indoor scene. Another example arises when the cost between "Landscape" and "Trees" yields 0.93. This high value implies that they are quite distant in the ontology so this should mean that the likelihood of them appearing together should be low. However, this assumption contracts sharply with real-world expectations. Therefore, the computation of this cost is heavily influenced by the structure adopted by the ontology, how dense it is, the fact that it is well-balanced or not, rather than a real semantic distance between terms.

To overcome these limitations, Chapter 6 extends the example-based OS measure and investigates the behaviour of different *cost functions* in the evaluation and ranking of multi-label classification systems. These *cost functions* estimate the semantic relatedness between visual concepts considering several knowledge bases such as Wikipedia, WordNet, Flickr, and the World Wide Web as discussed in Chapter 2.

3.4 A Note on Statistical Testing

The main goal of statistical testing is to evaluate the probability that the observed results could have occurred by chance.

Hull (1993) summarises the statistical tests applicable in *information retrieval* as well as analysing their benefits and limitations. In the case of the comparison being held between two retrieval systems, the following test apply: the sign test, the Wilcoxon signed rank test and the Student's t-test. There has been intense debate (van Rijsbergen 1979), (Hull 1993), and (Sanderson and Zobel 2005) about which is the appropriate test to utilise in the field of information retrieval. The first two mentioned tests are non-

parametric whilst the last assumes a normal distribution. Additionally, the Wilcoxon signed rank test assumes a continuous distribution. This thesis utilises the sign-test as it makes no specific assumptions and does not require the data to have particular characteristics.

The statistical tests undertaken in this thesis are concerned with testing whether or not the difference between two methods, expressed in terms of mean average precision, is statistically significant. A statistical test is given by the values of the *statistical level* and the *p-value*. The *statistical level* α is defined as the probability of making a decision to reject the *null hypothesis* when the *null hypothesis* is actually true. A *null hypothesis* implies that there is no difference between the results. The decision is made using the p-value: if the p-value is less than the significance level, the null hypothesis is rejected, and the results are statistically significant. Usual levels of significance are 5%, 1%, and 0.1%. In all the experiments conducted in this thesis, α has always been set to 5%.

3.5 Benchmarking Evaluation Campaigns

According to Smeaton et al. (2006), evaluation campaigns in information retrieval have become very popular in recent years for diverse reasons. First, they allow researches to compare their work with others in an open, metric-based environment. Second, they provide shared data, common evaluation measures, and often offer collaboration and sharing of resources. Finally, they have the potential of attracting funding agencies and outsiders due to their capacity of acting as a showcase for research results.

With respect to the benefits and shortcomings of these benchmarking evaluation campaigns, the most important points can be summarised as follows. The most obvious positive point is that they provide datasets, which are at the disposal of the participants. This, together with the fact that they use the same evaluation metrics

for evaluation and the same ground truth for measurement, facilitates the comparison of research results among research groups. Moreover, participants are invited to complete their tasks simultaneously following the same common guidelines to ensure a fair competition. As a result, researchers can be confident that their algorithms are sound and well-jugded. Additionally, the participation in these campaigns may facilitate the mobility of research groups into a new area of research. Finally, they enable the collaboration between participants and also the donation of datasets, which undoubtedly enrich research in the field.

Within the negative points, a valid criticism of evaluation campaigns is that datasets can both define and restrict the problems to be evaluated.

Table 3.1 summarises some evaluation campaigns related to multimedia data collections. The first block represents current conferences while the second refers to past campaigns.

The only reason for including text retrieval based conferences such as TREC and CLEF is because, as they are the precursors of TRECVID and ImageCLEF respectively, they represent their role model with respect to the way they are organised. Additionally, within each conference the emphasis is placed on the description of those tasks similar to the focus of this thesis: the automatic annotation of multimedia content.

3.5.1 TREC

The pioneer in evaluation conferences is the Text REtrieval Conference (TREC), which started in 1992. The main objective of TREC is to promote research in information retrieval by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. It is co-sponsored by the U.S. National Institute of Standards and Technology (NIST) and U.S. Department of Defence.

Table 3.1: Summary of the most relevant evaluation campaigns. The second block refers to past campaigns

Name	Data	URL
TREC	Text	http://trec.nist.gov
TRECVID	Video	http://trecvid.nist.gov
CLEF	Text	http://www.clef-campaign.org
CLEF 2010	Text	http://clef2010.org
MediaEval's VideoCLEF	Video	http://www.multimediaeval.org
PASCAL VOC	Images	http://pascallin.ecs.soton.ac.uk/challenges/VOC
PETS	Video	http://pets2010.net
FRVT	Images	http://www.frvt.org
ImagEVAL	Images	http://www.imageval.org
Benchathlon	Images	http://www.benchathlon.net
CLEAR	Video	http://www.clear-evaluation.org

A TREC edition consists of a series of *tracks*, which are areas of focus where particular retrieval tasks are defined. *Tracks* often become the test bed for new research areas. The running of a new track for the first time helps to define better the research problem that attempts to address. Additionally, each track creates the necessary infrastructure (test collections, evaluation methodology, etc.) to support research on its task.

TREC is supervised by a program committee consisting of representatives from government, industry, and academia. Within each campaign, NIST provides a test set of documents and questions. Participants run their own retrieval systems on the data, and return a list of the retrieved top-ranked documents. NIST pools the individual results, judges the retrieved documents for correctness, and evaluates the results. The cycle finalises with a workshop held at NIST's facilities in Gaithersburg, MD, where

participants share their results and experiences.

Among its main achievements, it is worth mentioning the immense growth that has experienced in terms of number of systems, of tasks, of research groups, and of countries participating every year. The TREC test collections and evaluation software are usually available to the retrieval research community. Additionally, the TREC conference has successfully met its goals of not only improving the state-of-the-art in information retrieval but also of facilitating the transfer of technology. Moreover, the effectiveness of retrieval systems practically doubled in the first six years of the campaign. A final accomplishment is that most of the current commercial search engines include technology that was first developed in TREC.

3.5.2 TRECVID

In 2001 and 2002 the TREC campaign supported a new track devoted to research in automatic segmentation, indexing, and content-based retrieval of digital video. At the beginning of 2003, this track became an independent evaluation campaign by itself with a workshop taking place just before TREC at the NIST's facilities in Gaithersburg. It was called TRECVID, which stands for TREC Video Retrieval Evaluation.

The annual TRECVID cycle is defined as follows. It starts more than a year before the November workshop, as NIST works with the sponsors to secure the data and defines the tasks and measures to be used, which are presented for discussion at the November workshop a year before. A set of guidelines is created and a call for participation is sent out by early February. In the spring and early summer, the data is distributed to the participants. Then, researchers develop and test their systems, run them on the test data, and submit the output to NIST. This happens from August to early September, depending on the task. Results of the evaluations are returned to the participants in

September and October. After that, participants summarise their work together with preliminary conclusions in the working notes, which are discussed at the workshop in mid-November. In the months following the workshop, the analysis and description of the work is finalised, which completes the cycle.

The data collections employed have varied greatly throughout the years. Initially, TREVID used video data from broadcast news organisations, TV program producers, and surveillance systems. These organisations imposed limits on programme style, content, production qualities, language, etc. From 2003 to 2006, TRECVID supported experiments in automatic segmentation, indexing, and content-based retrieval of digital video using broadcast news in English, Arabic, and Chinese. They also completed two years of pilot studies on exploitation of unedited video rushes provided by the BBC. From 2007 to 2009, the focus was on cultural, news magazine, documentary, and education programming supplied by the Netherlands Institute for Sound and Vision. The tasks associated to this data were video segmentation, search, feature extraction, and copy detection. The surveillance event detection task was accomplished using airport surveillance video provided by the UK Home Office.

The 2010 edition will challenge participants with a new set of videos characterised by a high degree of diversity in creator, content, style, production qualities, encoding, language, etc. Furthermore, the collection has associated keywords and descriptions provided by the video donor. The videos are available under creative commons licence from the Internet Archive. The only selection criteria imposed by TRECVID is that the video duration should be less than 4 minutes. In addition to this dataset, NIST is developing an Internet multimedia test collection (HAVIC) with the Linguistic Data Consortium and plans to use it as an exploratory pilot task during this edition.

The tasks outlined for the 2010 edition are the following: known-item search task

(interactive, manual, automatic); semantic indexing; content-based multimedia copy detection; event detection in airport surveillance video; and instance search.

Semantic Indexing

This task corresponds to the usual video annotation task called *high-level feature extraction* that was part of TRECVID competition since 2002. The description of the task is as follows. Given the test collection, master shot reference, and feature definitions, return for each feature a list of, at most, 2000 shot IDs from the test collection, ranked according to the possibility of detecting the high-level features. Note that in this context the term "feature" refers to semantic concept.

However, there are several differences with respect to previous editions. First, the collection of concepts or high-level features has increased from 20 (last year) to 130. These include all the previous concepts used in the *high level feature task* from 2005 to 2009 plus a selection of *large scale concept ontology for multimedia* (LSCOM) (Naphade et al. 2006) concepts. As a result, the goals have changed. Currently, the focus is on promoting research for indexing collections with large number of concepts, and, at the same time, on investigating the benefits of using ontology relations to improve the detection.

Another difference with last year's edition lies in the complexity in the definition of the concepts. For instance, concept number 18 is described as: "a structure carrying a pathway or roadway over a depression or obstacle. Label as positive any shots that contain a structure containing a pathway or roadway over a depression or obstacle and as negative those shots that do not contain such a structure. Shots containing structures over non-water bodies such as an overpass or a catwalk were also labelled as positive, includes model bridges". Whereas, concepts of the past editions were simple keywords

accompanied by a short description. For example, last year's concept number 2 was "chair" - "a seat with four legs and a back for one person". A final positive point of this year's edition is that cross-domain evaluations are encouraged, especially with other evaluation campaigns such as Pascal VOC challenge (Section 3.5.6) or ImageCLEF (Section 3.5.4).

With respect to the evaluation measures used, this year introduces two new measures, the *mean extended inferred average precision* (Yilmaz et al. 2008) and the *delta (M)AP* (Yang and Hauptmann 2008). Additionally, the evaluations will be performed with the usual metrics based on recall and precision, which is provided by the "trec_eval" software[3] developed by Chris Buckley.

Known-item Search Task

This task is a variation of the usual search task. The search task used to work with video shots while the emphasis is currently placed on the whole video. The *known-item search* task models the situation in which someone knows of a video, has seen it before, believes it is contained in a collection, but does not know where to look. To begin the search process, the searcher formulates a text-only description, which captures what the searcher remembers about the target video. The task can be formulated as follows: Given a topic (a text-only description of the desired video desired) and a test collection of video with associated metadata automatically return a list of 100 video IDs ranked by their probability. Alternatively, return the ID of the sought video and elapsed time to find it.

[3]Trec_eval can be downloaded from http://trec.nist.gov/trec_eval

Content-based Copy Detection

This task was presented for the first time as a pilot task in the 2008 edition of TRECVID. A *copy* is a segment of video derived from another video, usually by means of various transformations, such as addition, deletion, modification of aspect, colour, contrast, encoding, video recording, etc. Detecting copies is important for copyright control, business intelligence and advertisement tracking, law enforcement investigations, etc. *Content-based copy detection* offers an alternative to watermarking[4]. The TRECVID 2010 copy detection task will be will be based on the framework tested in TRECVID 2008, as past years.

Surveillance Event Detection

This task started in 2008. The rationale behind it is to detect human behaviours efficiently in vast amounts of surveillance video, both retrospectively and in realtime. This technology is fundamental for a variety of higher-level applications of critical importance to public safety and security.

Instance Search

This is a pilot task first introduced in the 2010 edition. It attempts to fulfil a real need in many situations involving video collections where there is a necessity to find more video segments of a certain specific person, object, or place, given a visual example. Due to its nature of pilot task, its main objective is to explore the definition of the task and evaluation issues using data and an evaluation framework in hand.

[4]Watermarking is the process of embedding information into a video.

Event Detection in Internet Multimedia

This is also a new task in the 2010 edition that will be developed as a pilot task. The objective is, given a collection of test videos and a list of test events, to indicate whether each of the test events is present anywhere in each of the test videos and give the strength of evidence for each such judgement.

3.5.3 Cross-Language Evaluation Forum

The Cross-Language Evaluation Forum (CLEF) is an annual system evaluation campaign whose goal is to develop an infrastructure for evaluating information retrieval systems operating on European languages, in both monolingual and cross-language contexts.

The first CLEF evaluation campaign appeared in early 2000 (Peters and Braschler 2001) and was divided into different tasks, called *tracks*, devoted to different research objectives. Since then, each edition has published a collection of working notes with descriptions of the experiments conducted within the campaign. The results of the experiments were presented and discussed in the CLEF workshop. Finally, the final papers were published by Springer in their Lecture Notes for Computer Science series as CLEF Proceedings. CLEF has been mainly sponsored by different programmes of the European Union.

The 2009 edition marked the end of the CLEF series and also the participation of Carol Peters, after ten years of work as main organiser. However, there exists a follow-up, the CLEF 2010 conference, which is the continuation of the CLEF campaigns and will cover a broad range of issues from the fields of multilingual and multimodal information access evaluation.

CLEF 2010 will consist of two parts, of two days each. One part will be devoted to

presentations of papers on all aspects of the evaluation of information access systems while the other two-day part will be devoted to a series of "labs". Two different kinds of labs will be offered: labs can either be run "campaign-style" during the twelve month period preceding the conference, or adopt a more "workshop"-style format that can explore issues of information access evaluation and related fields. The labs will culminate in sessions of a half-day, one full day or two days at the CLEF 2010 conference.

3.5.4 ImageCLEF

ImageCLEF is an evaluation campaign part of the Cross-Language Evaluation Forum (CLEF) initiative. The main objective of ImageCLEF is to advance the field of image retrieval and offer evaluation in various fields of image information retrieval. The evaluation procedure is accomplished in a manner similar to the way TREC's results are evaluated by NIST.

It started as a new track for the CLEF 2003 edition (Clough and Sanderson 2004) led by University of Sheffield. The initial cross language image retrieval proposed task was: Given a user need expressed in a language other than English, find as many relevant images as possible. In order to facilitate the task, textual captions were provided. Every year new tasks were added.

In 2004, a medical retrieval task started, in which an example image was used to perform a search against a medical image database consisting of images such as scans and x-rays.

In 2005, an automatic image annotation task appeared but on medical images and it was not until the 2006 edition when this task was extended to include a normal photographic collection.

The tasks proposed for the 2010 edition were the following: medical retrieval, photo

annotation, robot vision, and Wikipedia retrieval.

Visual Concept Detection and Annotation Task

This task corresponds to the *image annotation task* that started in 2006. The task is defined in the same way as last year's where participants were asked to annotate the images of the test set with a predefined set of keywords. This defined set of keywords allowed for an automatic evaluation and comparison of the different approaches. However, this year's task can be solved following three different approaches: annotation using only visual information; annotation using Flickr user tags (tag enrichment); and a multi-modal approach that considers visual information, and/or Flickr user tags, and/or EXIF information. The image collection is a subset of MIR Flickr 25,000 image dataset (Huiskes and Lew 2008).

The evaluation follows both the *concept-based* and *example-based* evaluation paradigm. For the *concept-based* evaluation, the mean average precision (MAP) will be utilised for the first time. This measure showed better characteristics than the usual EER and AUC employed in previous editions. For *example-based* evaluation, the F-measure will be applied. Additionally, the *ontology-based score* (OS) proposed by Nowak and Lukashevich (2009) will be used but with a different cost map based on Flickr metadata (Nowak et al. 2010a). The final goal is to investigate whether or not this adaption can cope with the limitations of the *ontology-based score*.

With respect to the annotations, they are a collection of 93 keywords that contain additionally the 53 words proposed in the previous edition.

Medical Retrieval Task

The medical retrieval task of ImageCLEF 2010 will use a similar database to 2008 and 2009 but with a larger number of images. The dataset contains all images from articles published in radiology and radiographics including the text of the captions and a link to the web page of the full text articles. Over 77,000 images are currently available. This task is divided into three subtasks: the modality classification, the ad-hoc retrieval, and the case-based retrieval tasks.

Robot Vision Task

The third edition of this challenge will focus on the problem of visual place classification, with a special focus on generalisation. Participants will be asked to classify rooms and functional areas on the basis of image sequences, captured by a stereo camera mounted on a mobile robot within an office environment. The test sequence will be acquired within the same building but at a different floor than the training sequence. It will contain rooms of the same category such as "corridor", "office", "bathroom". Additionally, it will also contain room categories not seen in the training sequence such as "meeting room", or "library". The system built by participants should be able to answer the question "where are you?" when presented with a test sequence imaging a room category seen during training, and it should be able to answer "I do not know this category" when presented with a new room category.

Wikipedia Retrieval Task

This is an ad-hoc image retrieval task. The evaluation scenario is thereby similar to the classic TREC ad-hoc retrieval task and the previous ImageCLEF photo retrieval task. The aim is to investigate retrieval approaches in the context of a large and

heterogeneous collection of images searched by users with diverse information needs. In 2010, the task will use a new collection of over 237,000 Wikipedia images that cover diverse topics of interest. These images are associated with unstructured and noisy textual annotations in English, French, and German.

3.5.5 MediaEval's VideoCLEF

MediaEval is a benchmarking initiative that was launched by the PetaMedia Network of Excellence in late 2009. It serves as an umbrella organization to run multimedia benchmarking evaluations. It is a continuation and extension of VideoCLEF, which ran as a track in the CLEF campaign in 2008 and 2009.

The 2010 cycle for MediaEval started with the data release in June and will conclude with a workshop in October.

The initiative is divided into several tasks. In the 2010 edition, there are two annotation tasks called the *tagging* task but designed with two variations, the *professional* version and the *wild wild web* version. The *professional version tagging* task requires participants to assign semantic theme labels from a fixed list of subject labels to videos. The task uses the TRECVID data collection from the Netherlands Institute for Sound and Vision. However, the tagging task is completely different than the original TRECVID task since the relevance of the tags to the videos is not necessarily dependent on what is depicted in the visual channel. The *wild wild web* task requires participants to automatically assign tags to videos using features derived from speech, audio, visual content or associated textual or social information. Participants can choose which features they wish to use and are not obliged to use all features. The dataset provided is a collection of Internet videos.

Additional tasks for the 2010 initiative are the *placing* task or *geo-tagging* where

participants are required to automatically guess the location of the video by assigning geo-coordinates (latitude and longitude) to videos using one or more of: video metadata, such as tags or titles, visual content, audio content, social information. Any use of open resources, such as gazetteers[5], or geo-tagged articles in Wikipedia is encouraged. The goal of the task is to come as close to possible to the geo-coordinates of the videos as provided by users or their GPS devices. Other tasks are the *affect* task whose main goal is to detect videos with high and low levels of dramatic tension; the *passage* task where, given a set of queries and a video collection, participants are required to automatically identify relevant jump-in points into the video based on the combination of modalities such as audio, speech, visual, or metadata; and the *linking* task, where participants are asked to link the multimedia anchor of a video to a relevant article from the English language Wikipedia.

The video collection used belong to the creative commons licence or data from the Netherlands Institute of Sound and Vision.

One of the strongest points of this competition is that it attempts at complementing rather than duplicating the tasks assigned by TRECVID evaluation campaign. Traditionally, TRECVID tasks are mainly focused on finding objects and entities depicted in the visual channel whereas MediaEval concentrates on what a video is about as a whole.

3.5.6 PASCAL Visual Object Classes Challenge

The PASCAL Visual Object Classes (VOC) challenge (Everingham et al. 2009) started in 2005 supported by the EU-funded Network of Excellence on "Pattern Analysis, Statistical Modelling and Computational Learning" (PASCAL). The focus of this challenge

[5] A gazetteer is a geographical index or dictionary.

is on object recognition. Additionally, it addresses the following objectives: to compile a standardised collection of *object recognition* databases; to provide standardised ground-truth object annotations across all databases; to provide a common set of tools for accessing and managing the database annotations; and to run a challenge evaluating performance on object class recognition.

The goal of the 2010 challenge is to recognise objects from a number of visual object classes in realistic scenes. It is fundamentally a supervised learning learning problem because a training set of labelled images is provided. The twenty object classes that have been selected belong to the following categories: person, animal, vehicle, and objects typically found in an indoor scene. There is an annotation task called ImageNet *large scale visual recognition taster* competition. Its main goal is to estimate the content of images using a subset of the large hand-labeled ImageNet dataset (Deng et al. 2009) as training. Test images will be presented with no initial annotations and algorithms will need to assign labels that will specify what objects are present in the images. In this initial version of the challenge, the goal is only to identify the main objects present in images, not to specify the location of objects.

Another tasks of the 2010 edition are the following: the *classification* task, which predicts the presence or absence of an instance of the class in the test image; the *detection* task, which determines the bounding box and label of each object in the test image; the *segmentation* task, which generates pixel-wise segmentation giving the class of the object visible at each pixel; the *person layout taster* competition, which consists of predicting the bounding box and label of each part of a person such as "head", "hands", and "feet"; and the *action classification taster* competition, which deals with predicting the actions being performed by a person in a still image.

The challenges culminates every year with a workshop where participants are invited

to show their results. The workshop is generally allocated with a relevant conference in computer vision such as the International or European Conference on Computer Vision, ICCV or ECCV, respectively.

3.5.7 PETS

The Performance Evaluation of Tracking Systems (PETS) initiative was initially funded by the European Union through the FP6 project called "Integrated Surveillance of Crowded Areas for Public Security" (ISCAPS). The main goal of the project was to reinforce security for the European citizens and to downsize the terrorist threat by reducing the risks of malicious events. This is to be undertaken by providing efficient, real-time, user-friendly, highly automated surveillance of crowded areas that may be exposed to terrorist attacks. Therefore, the main objective of the PETS initiative is to perform crowd image analysis, crowd count, density estimation, tracking of individual(s) within a crowd, and detection of separate flows and specific crowd events. The 2010 edition of PETS workshop was held in conjunction with the 2010 edition of the IEEE International Conference on Advanced Video and Signal-Based Surveillance.

3.6 Past Benchmarking Evaluation Campaigns

This section summarises other relevant benchmarking evaluation campaigns. Note that none of them are currently operative. However, it is worth mentioning them as they keep on-line their research questions, objectives, results, and the used datasets.

The Face Recognition Vendor Test (FRVT) 2006 was the latest in a series of large scale independent evaluations for face recognition systems organised by the U.S. National Institute of Standards and Technology. Previous evaluations in the series were the FERET, FRVT 2000, and FRVT 2002. The primary goal of the FRVT 2006 was

to measure progress of prototype systems and commercial face recognition systems since FRVT 2002. Additionally, FRVT 2006 evaluated the performance on high resolution still imagery, 3D facial scans, multi-sample still facial imagery, and re-processing algorithms that compensate for pose and illumination.

ImagEVAL evaluation conference was launched in France in 2006. It attempted to bring some answers to the question posed by Carol Peters, in the CLEF workshop of 2005, where she wondered why systems that show very good results in the CLEF campaigns have not achieved commercial success. The point of view of ImagEVAL is that the evaluation criteria "do not reflect the real use of the systems". Although, the initiative was fairly concentrated on the French research domain, it was accessible to other researchers as well. They divided the campaign into several tasks related to content based image retrieval including recognition of image transformations like rotation or projection; image retrieval based on combining text and image; detection and extraction of text regions from images; detection of certain types of objects in images such as cars, planes, flowers, cats, churches, the Eiffel tower, table, PC or TV or the US flag; and semantic feature detection like indoor, outdoor, people, night, day, etc. Unfortunately, the campaign only lasted one edition.

During the 2000 Internet Imaging Conference, a suggestion to hold a public contest to assess the merits of various image retrieval algorithms was proposed. Since the contest would require a uniform treatment of image retrieval systems, the concept of a benchmark quickly entered into the scenario. The contest would exercise one or more such content based image retrieval (CBIR) benchmarks. The contest itself became known as the Benchathlon and was finally held at the Internet Imaging Conference in January 2001. Despite their initial objectives, no real evaluation ever took place, although many papers were published in this context and also a database created.

The Classification of Events, Events, Activities and Relationships (CLEAR) evaluation conference was an international effort to evaluate systems that are designed to recognise events, activities, and their relationships in interaction scenarios. Its main goal was to bring together projects and researchers working on related technologies in order to establish a common international evaluation in this field. It was divided into the following tasks: person tracking (2D and 3D, audio-only, video-only, multimodal); face tracking; vehicle tracking; person identification (audio-only, video-only, multimodal); head pose estimation (2D and 3D); and acoustic event detection and classification. The latest edition, which was held in 2007, was supported by the European Integrated project "Computers In the Human Interaction Loop" (CHIL) and NIST.

3.7 Benchmark Datasets

This section provides a detailed description of the Corel dataset, a popular collection in the field together with other datasets used in the experiments undertaken in this thesis, such as the collection used in the annotation task in the 2008 edition of TRECVID and the collection used in the 2008 and 2009 editions of ImageCLEF.

3.7.1 Corel Stock Photo CDs

The Corel Stock Photo CDs is a large collection of stock photographs compiled by the Corel corporation. The dataset is commercially available as a set of libraries. There are currently four libraries available on the Internet. Each library is composed of 200 CDs and each CD contains 100 images about a specific topic. In total, there are 800 Photo CDs.

Initially, researches created different datasets extracting images from various CDs as noted in Section 3.1. Müller et al. (2002) claimed that many research groups compared

their results by using different subsets of the Corel Stock Photo CDs. They demonstrated how dependent is the performance of a content-based retrieval (CBIR) system on the dataset used as training, the selected queries and even, the relevant judgements. They highlighted the need for standard evaluation measures and specially the need for a standard image database.

Westerveld and de Vries (2003) argued that the Corel dataset is a relatively easy set and that researchers should be careful when carrying over results from this collection to other datasets.

3.7.2 Corel 5k Dataset

This image collection has been extensively employed as a standard benchmark in the field of automated image annotation. It was firstly proposed by Duygulu et al. (2002). The data that they used for their experiments can still be found on-line: http://kobus.ca/research/data/eccv_2002/index.html.

The collection is made up of 5,000 images that were selected from 50 out of 200 CDs of the Corel Stock Photo collection. The dataset is divided into 4,500 images that correspond to the training set, and 500 to the test set. The vocabulary is made up of 374 words, out of which 371 appear in the training set, while 260 in the test set.

Images are annotated with words that range from one to five, most of them contains four annotation words, while a few have one, two, three, or five. For example, the first five images of the training set are annotated as follows:

1000	*city mountain sky sun*
1003	*bay lake sun tree*
1004	*sea sun*

1005	*beach sea sky sun*
1006	*clouds sea sky sun*

The collection represents topics such as: sunrises and sunsets; air shows; bears; Fiji; tigers; foxes and coyotes; Greek isles, etc. The rest of the list together with the complete list of terms of the vocabulary are represented in Section A.1 of the Appendix.

Despite its popularity, the dataset has received numerous critics. Tang and Lewis (2007) argued that this collection can be easily annotated when training on the Corel 5k training set, because the training and test sets contain many very globally similar images. Additionally, other limitation of the Corel 5k dataset is that it suffers from generalisation errors and over-fitting due to its small size. They proposed a very simple algorithm based on *support vector machine* (SVM) and applied to a global feature vector, the MPEG-7 *colour structure descriptor* (CSD), which was able to get comparable results with the best performing methods of that time, MBRM (Feng et al. 2004) and Mix-Hier (Carneiro and Vasconcelos 2005).

Viitaniemi and Laaksonen (2007) reviewed the influence of the metric used together with the annotation length on the performance of annotation algorithms for the Corel 5k dataset. As the previous authors, they proposed an algorithm that combines three classifiers and that by using global MPEG-7 descriptors achieves a performance higher than that obtained by MBRM and Mix-Hier.

Athanasakos et al. (2010) claimed that the reason why some annotation algorithms perform so high for the Corel 5k dataset is due to some collection-specific properties of the collection and not to the actual models. In addition to that, they observed that the evaluation settings under which the annotation algorithms have been compared differ from algorithm to algorithm in terms of different descriptors, collections or part of the

Table 3.2: Highest performing algorithms for the Corel 5k dataset ordered according to their F-measure value

Model	Author	NZR	R_{260}	P_{260}	F_{260}
MBRM	Feng et al. (2004)	122	0.25	0.24	0.24
Mix-Hier	Carneiro and Vasconcelos (2005)	137	0.29	0.23	0.26
SML	Carneiro et al. (2007)	137	0.29	0.23	0.26
CSD-SVM	Tang and Lewis (2007)	-	0.28	0.25	0.26
PicSOM	Viitaniemi and Laaksonen (2007)	-	0.35	0.35	0.35
JEC	Makadia et al. (2008)	113	0.40	0.32	0.36

collections used, or "easy" settings used. They consider that this makes their results non-comparable. Consequently, they proposed a framework for evaluating automated image annotation algorithms: a set of test collections, a sampling method that extracts normalised and self-contained samples, a variable-size block segmentation technique, and a set of multimedia content descriptors. Finally, they demonstrated that a simple SVM approach with global features (MPEG-7) achieves better results than MBRM and Mix-Hier.

Being conscious of the limitations of the Corel 5k dataset, the approach followed in this thesis consists in using it as a preliminary first evaluation set before doing deeper evaluation with additional sets. This allows the methods deployed in this thesis to have a preliminary estimation of their performance before being tested on more difficult datasets. The reason for selecting the datasets coming from ImageCLEF over TRECVID's as the additional evaluation set is because the description of the annotation task and the underlying research questions addressed by ImageCLEF adjust completely to the research goals approached by this thesis. TRECVID's focus is on video research and as such, it is more concerned with some research problems that are not applicable to the case of image research, such as movement detection.

By taking into consideration these new generation of algorithms based on simple classifying approaches and MPEG-7 global features, the actual scenario for the highest performing algorithms[6] for the Corel 5k dataset is depicted in Table 3.2. The algorithms in the second block correspond to global features. Due to the fact that highest performing algorithm, which was provided by Makadia et al. (2008) was tested on two additional datasets other than the Corel 5k and that they were able to maintain the same good performing results, I consider that the solution for automated image annotation algorithms would come from research done on global images features.

3.7.3 TRECVID 2008 Video Collection

Participants of the high-level feature or annotation task of the 2008 edition of TRECVID were provided with 100 hours of video as training set, and an additional 100 hours as test set. The objective of the task was to provide semantic annotations for the test video. The video was in format MPEG-1 and was provided by the Netherlands Institute for Sound and Vision. The topics covered were news magazine, science news, news reports, documentaries, educational programming, and archival video. The set of videos used for training and development purposes were annotated with a collection of 20 words. Section A.2 of the Appendix contains the complete list. The difficulty of the task lied in the annotation words that were rather specific compared to initiatives like ImageCLEF where the vocabularies are more general. TRECVID vocabulary contained words such as "two people", which forced the algorithm to be able to count people, or "emergency vehicle", which made the algorithm to be able to differentiate between different kinds of vehicles, or "singing", which implied that the algorithm should be

[6] According to Table 2.7 there are algorithms that beat Mix-Hier and MBRM, such as Li and Sun (2006) and Kang et al (2006), both with a F value of 0.27.

able to detect actions.

3.7.4 ImageCLEF 2008 Image Dataset

The dataset provided for the annotation or visual concept detection task was a subset of the IAPR TC-12 image collection (Grubinger et al. 2006). It was made up of a training set of 1,800 images and a test set of 1,000. The annotation words were 17 and were rather general, with terms such as "indoor", "outdoor", "person", "animal", etc. This together with the fact that the size of the collection (2,800 images) was quite small, made participants to achieve performance superior to that obtained for the Corel 5k dataset. The vocabulary was presented adopting a hierarchical structure but the organisers did not provide any indication about how to benefit from it. Section A.3 of the Appendix provides the complete list of terms.

Finally, the evaluation measures adopted were the equal error rate (EER) and the area under the ROC curve (AUC) as discussed in Section 3.2.3.

3.7.5 ImageCLEF 2009 Image Dataset

The *large scale visual concept detection and annotation* (Nowak and Dunker 2009b) task of the 2009 edition presented a list of research questions that the participants should be able to address. In particular, they were interested in learning whether or not image classifiers could scale to the large amount of concepts and data, and whether an ontology could help in large scale annotations. The novelty of this edition lay in the increment of the size of the image collection (18,000 images in total), and also in the number of vocabulary terms (53 visual concepts), together with the provision of an ontology, which structured the hierarchy between the visual concepts.

The image collection used was a subset of the MIR Flickr 25k image dataset (Huiskes

and Lew 2008). It was divided into a training set of 5,000 images and a test set of 13,000 images.

The difficulty of this edition was in handling a large collection of images together with the nature of the provided vocabulary. Specially as many of them did not correspond to visual terms. Thus, there were some words that were rather subjective such as "fancy", "overall_quality", and "aesthetic_impression". Others represented negations such as "no_visual_season","no_visual_place", "no_visual_time", "no_persons". While others made reference to the number of people such as "single_person", "small_group", and "big_group".

With respect to the evaluation measures, they proposed the usual EER and AUC, adopted by this initiative, and additionally, the *ontology-based score* proposed by Nowak and Lukashevich (2009), which is designed for the case when the vocabulary adopts the hierarchical structure of an ontology.

3.8 Conclusions

This chapter has revised the methodology most commonly adopted in the field in terms of evaluation measures and benchmark datasets used. Moreover, an especial emphasis has been placed on the evaluation campaigns as they have a great influence on the evolution of the field. In particular, the most relevant initiatives related to automated image annotation are ImageCLEF, TREVID, PASCAL VOC challenge, and MedieEval's VideoCLEF. Despite the fact that initially it may seem that they address the same research problem, a more detailed look will reveal that each one of them has not only defined the task differently but also they have placed the focus on different aspects. Moreover, they explore different evaluation measures. Consequently, they are rather complementary.

With respect to the experimental work conducted in this thesis, the evaluation metric adopted has been the mean average precision as it is rather stable and widely used, which facilitates the comparison of results. Being aware of the limitations of the Corel 5k dataset, I have utilised it in my experiments as a preliminary first evaluation set and with the solely intention of establishing a comparison of results with other algorithms. Additionally, I have always used another dataset, usually provided by the evaluation campaigns ImageCLEF2008 and ImageCLEF2009.

Finally, the reason for selecting ImageCLEF (Llorente et al. 2009c) and (Llorente et al. 2010b) as the preferred evaluation conference is because the description of the annotation task and the underlying research questions that they address adjust completely to my research goals. I also participated in the 2008 edition of TRECVID (Llorente et al. 2008b) with limited success as their focus was on aspects of the video research that were not considered in my work, such as movement detection.

Subsequent chapters will introduce the experimental work undertaken in this thesis.

Chapter 4

A Semantic-Enhanced Annotation Model

This chapter presents an automated image annotation algorithm enhanced by a combination of several semantic relatedness measures. The hypothesis to be tested is whether background knowledge coming from various sources together with semantic relatedness measures can increase the effectiveness of a baseline image annotation system. The process is based on re-ranking the baseline annotations following a heuristic algorithm that attempts to prune those annotations that do not belong to the same context. Several knowledge sources are employed, one internal and others external to the collection. The training set has been used as internal knowledge base and Wikipedia, WordNet and the World Wide Web as external. In all cases, several semantic relatedness measures have been applied. The performance of the proposed approaches is calculated in order to make an analysis of their benefits and limitations. Finally, the effectiveness of the final combined approach is illustrated by showing very good results obtained using two datasets, Corel 5k, and ImageCLEF2009. In both cases, statistically significant better results are obtained over the baseline annotation method.

4.1 Model Description

The hypothesis to be tested is whether or not semantic relatedness measures can substantially increase the performance of a baseline image annotation system by pruning unrelated keywords. This is based on the observation (see Section 1.4) that annotation words are generated independently without taking into consideration that they should be consistent with each other as they share the same image context. Thus, two words are considered to be related or not related based on their semantic relatedness value falling below or above a given threshold.

In particular, given the vocabulary $V = \{w_1, ..., w_n\}$, the image training set $\tau = \{J_1, ..., J_m\}$ and test set $T = \{I_1, ..., I_p\}$, I propose a heuristic algorithm that automatically refines the image annotation keywords generated by a baseline non-parametric density estimation algorithm (NPDE). The model detects unrelated words with the help of the semantic measures, discards them and finally, re-ranks the baseline annotations generating a set of more accurate annotations.

This model is divided into several parts. The algorithms that constitute each one of them are detailed in the following. Algorithm 1 describes the baseline algorithm that generates the initial annotations. Algorithm 2 accomplishes the computation of semantic relatedness values for every pair of words in the vocabulary. Finally, Algorithm 3 describes the heuristic algorithm that prunes the baseline annotations.

4.1.1 Baseline NPDE Algorithm

The baseline algorithm is a variation of the probabilistic framework developed by Yavlinsky et al. (2005), who used global features together with a non-parametric density estimation (NPDE). This approach is based on the Bayes formulation, being the ultimate goal to model the conditional probability density f($x|w$) for each annotation keyword

Algorithm 1 NPDE(norm, scale)

Input: Vocabulary: $V = \{w_1, ..., w_n\}$;

Images of the Test Set: $T = \{I_1, ..., I_p\}$;

Test set feature vector $x = (x_1, ..., x_d)$, $x \in \mathbb{R}$;

Images of the Training Set: $\tau = \{J_1, ..., J_m\}$;

TS: a text file with annotation words for images.

Output: Probability Matrix: $P \in \mathbb{R}^p \times \mathbb{R}^n$ where each cell is $\mathrm{p}(w_j|I_i)$ and it is estimated by applying Bayes rule.

foreach $I_i \in T$ **do**
  foreach $w_j \in V$ **do**
    Create T_{w_j}, set of training images annotated by w_j
    //Starting kernel estimation:
    $h_l \leftarrow \sigma \cdot scale$ *//σ: standard deviation of feature l*
    //Laplacian kernel:
    if $norm = 1$ **then**
      $\mathrm{k} \leftarrow \prod_{l=1}^{d} \frac{1}{2h_l} \cdot e^{\frac{-|x_i - x_{w_j}^{(l)}|}{h_l}}$
    end if
    //Gaussian kernel:
    if $norm = 2$ **then**
      $\mathrm{k} \leftarrow \prod_{l=1}^{d} \frac{1}{\sqrt{2\pi h_l^2}} \cdot e^{-\frac{1}{2}\left(\frac{x_i - x_{w_j}^{(l)}}{h_l}\right)^2}$
    end if
    //Modelling x_i upon the assignment of w_j:
    $\mathrm{f}(x_i|w_j) \leftarrow \frac{1}{|T_{w_j}|} \cdot \sum_{t=1}^{n} \mathrm{k}(x_i - x_{w_j}^{(t)}; h)$
    //Modelling the prior probability of word w_j:
    $\mathrm{p}(w_j) \leftarrow \frac{|T_{w_j}|}{\sum_{w_j} |T_{w_j}|}$
    //Approximating the probability density of x_i:
    $\mathrm{f}(x_i) \leftarrow \sum_{w_j} \mathrm{f}(x_i|w_j) \cdot \mathrm{p}(w_j)$
    //By applying Bayes formula:
    $\mathrm{p}(w_j|x_i) \leftarrow \frac{\mathrm{f}(x_i|w_j) \cdot \mathrm{p}(w_j)}{\mathrm{f}(x_i)}$
  end for
end for
return P

w, where x is a d-dimensional feature vector of real values representing a test image. The non-parametric approach is employed because the distributions of image features have irregular shapes that do not resemble a priori any simple parametric form.

Thus, the density of function f($x|w$) is estimated placing a kernel function over each point, which represents an image of the training set. Two kernel functions are investigated, the Laplacian and the Gaussian kernel. Both can be formulated under the generalized Gaussian distribution as:

$$\mathrm{k_{general}} = \prod_{l=1}^{d} \frac{1}{2\Gamma(1+1/norm)A(norm,h_l)} \cdot e^{\left(\frac{-|x_i - x_{w_j}^{(l)}|}{A(norm,h_l)}\right)^{norm}}, \tag{4.1}$$

where Γ is the gamma function, $A(norm, h_l) = \left[\frac{h_l^2 \cdot \Gamma(1/norm)}{\Gamma(3/norm)}\right]^{\frac{1}{2}}$ is a scaling factor, and $norm$ is a positive real number that determines the shape of the curve, a Laplacian when it is set to one and a Gaussian when set to two. Note that $\Gamma(1) = \Gamma(2) = 1$, $\Gamma(3) = 2$, $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, and $\Gamma(\frac{3}{2}) = \frac{\sqrt{\pi}}{2}$. The kernel bandwidth h_l, another positive real number, is set by scaling the sample standard deviation of feature component l of the image by the same constant. This constant is called *scale*. Then, the Laplacian kernel is described as

$$\mathrm{k_L} = \prod_{l=1}^{d} \frac{1}{2h_l} \cdot e^{\frac{-|x_i - x_{w_j}^{(l)}|}{h_l}}, \tag{4.2}$$

while the Gaussian kernel is formulated as:

$$\mathrm{k_G} = \prod_{l=1}^{d} \frac{1}{\sqrt{2\pi h_l^2}} \cdot e^{-\frac{1}{2}\left(\frac{x_i - x_{w_j}^{(l)}}{h_l}\right)^2}. \tag{4.3}$$

The baseline annotation algorithm yields a probability value, p($w_j|I_i$), for every word w_j being present in an image I_i of the test set, which is used as a confidence score. The final annotations are generated after selecting the five words with the highest confidence scores. Algorithm 1 summarises the computation of the baseline algorithm.

Algorithm 2 SemanticComputation(measure)

Input: Vocabulary: $V = \{w_1, ..., w_n\}$

TS: a text file with annotation words for images.

Output: Similarity matrix: $S \in \mathbb{R}^n \times \mathbb{R}^n$.

if *measure* = "TrainingSet" **then**
  Read TS
  Initialisation of matrix $M \in \mathbb{R}^m \times \mathbb{R}^n$
  foreach $image \in TS$ **do**
    Read annotation word w_i
    while $w_i \neq$ NULL **do**
      $M(image, i) \leftarrow 1$
    end while
  end for
  $M^T \leftarrow$ Transpose(M)
  $S \leftarrow M^T \cdot M$
end if
if *measure* = "WebCorrelation" **then**
  foreach $(w_i, w_j) \in V^2$ **do**
    S$(w_i, w_j) \leftarrow$ rel (w_i, w_j) as defined by Eq. 2.38.
  end for
end if
if *measure* = "WordNet" **then**
  foreach $(w_i, w_j) \in V^2$ **do**
    $c_i \leftarrow$ WNDisambiguation(w_i)
    $c_j \leftarrow$ WNDisambiguation(w_j)
    $S(w_i, w_j) \leftarrow$ rel (c_i, c_j) as defined in (Pedersen et al. 2004).
  end for
end if
if *measure* = "Wikipedia" **then**
  foreach $(w_i, w_j) \in V^2$ **do**
    $c_i \leftarrow$ WikiDisambiguation(w_i)
    $c_j \leftarrow$ WikiDisambiguation(w_j)
    $S(w_i, w_j) \leftarrow$ rel (c_i, c_j) as defined in Eq.2.43.
  end for
end if
return S

4.1.2 Semantic Relatedness Computation

The main objective of this section is to compute the semantic relatedness between all pairs of words coming from the vocabulary of the collection. This data is stored in the form of a similarity matrix.

Sections 2.2 to 2.6 were devoted to the revision of several semantic measures and their performance using human similarity judgement. This performance should not be considered in any case determinant as it can change in the framework of the image annotation application although it is very useful as a preliminary estimation. In total, 14 semantic relatedness measures are explored: four distributional measures and ten that uses semantic network representations like WordNet and Wikipedia.

The first measure is based on training set correlation as seen in Section 2.2. In particular, the context of the images is computed using statistical co-occurrence of pairs of words appearing together in the training set. This information is represented in the form of a co-occurrence matrix as described in Algorithm 2. The resulting co-occurrence matrix S is a symmetric matrix where each entry s_{ij} contains the number of times the annotation word w_i co-occurs with w_j.

The use of the World Wide Web as an external corpus to perform statistical correlation of words is a recent approach in automated image annotation. Thus, the correlation between words is computed using several web-based search engines such as Google, Yahoo, and Exalead and it is based on the normalized Google distance (NGD) defined in Equation 2.37. However, this work uses a variation of the previous NGD, which was accomplished by Gracia and Mena (2008) in order to get a proper relatedness measure that is a bounded value and, at the same time, increases with decreasing distance. Their web-based semantic relatedness measure between words w_i and w_j is defined in Equation 2.38.

The problem of assessing semantic similarity using semantic network representations has long been addressed by researches in artificial intelligence and psychology. As a result of this, a sheer number of semantic measures have been proposed in the literature. Some of them were initially developed for generic semantic representations other than WordNet. However, Pedersen et al. (2004) adapted them to WordNet (Miller 1995) and released a very useful Perl implementation. This work adopts Pedersen's nine measures: Jiang and Conrath (1997) (JCN), Hirst and St-Onge (1998) (HSO), Leacock and Chodorow (1998) (LCH), PATH (Pedersen et al. 2004), (Wu and Palmer 1994) (WUP), Resnik (1995) (RES), Patwardhan (2003) (VEC), Lin (1998) (LIN), and Adapted Lesk (Banerjee and Pedersen 2003) (LESK). These measures have been reviewed in Section 2.3.

Finally, the semantic relatedness measure applied to Wikipedia (WIKI) used in this research was developed by Milne and Witten (2008) as seen in Equation 2.43.

Note that measures based on WordNet and Wikipedia work with concepts rather than with words so a previous disambiguation process is required. This is achieved by the functions WNDisambiguation(w_i) and WikiDisambiguation(w_i). In both cases, the disambiguation task is accomplished by assigning automatically to every word the most usual sense. In the case of WordNet, this sense corresponds to the first sense in the synset (word#n#1) as explained in Section 2.3.5 while in Wikipedia corresponds to the sense of the word more probable according to the content stored on the Wikipedia database as seen in Section 2.5.2. Although this naïve disambiguation approach works reasonably well for the collections tested in this research, it might have a negative impact on the performance of the algorithm in other domains. Consequently, more sophisticated disambiguation approaches will be implemented in the future as future lines of work.

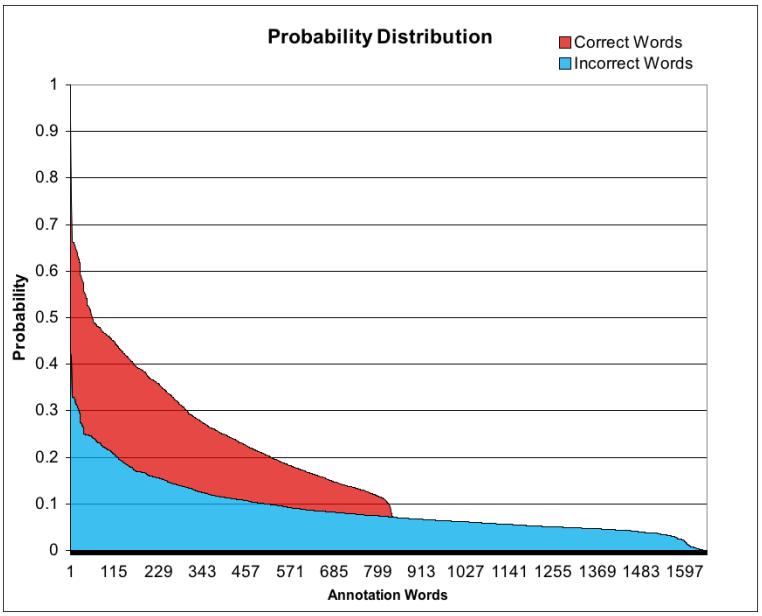

Figure 4.1: Probability distribution for the Corel 5k dataset

Independently of the semantic measure employed, the process followed in this research is summarised in Algorithm 2. Basically, it consists in creating a similarity matrix, which will be later used by Algorithm 3.

4.1.3 Pruning Algorithm

The main goal of the pruning algorithm is to detect which words are not semantically related with the others in a given image. In this thesis, these words are denoted as "noisy" annotation words. Consequently, the starting point of the algorithm is the set of five *candidate annotations* per image generated by the baseline algorithm, which is denoted by Anno in Algorithm 3.

However, Section 1.4 shows that sometimes the baseline annotation system is unable to interpret adequately the content depicted in the image and, the only way to generate the correct annotations is by refining the image visual parameters. The pruning algorithm contemplates this issue by being it solely applicable to images where at least some objects have been successfully detected. Otherwise, a nonsensical output

Algorithm 3 Pruning(threshold$_\alpha$, threshold$_\beta$)

Input: Similarity matrix: $S \in \mathbb{R}^n \times \mathbb{R}^n$;

Probability Matrix: $P \in \mathbb{R}^p \times \mathbb{R}^n$

Output: Modified Probability Matrix: $P' \in \mathbb{R}^p \times \mathbb{R}^n$

FinalAnno(I_i): set of 5 annotation words

foreach $I_i \in T$ **do**
  //Select the initial annotation words:
  Anno(I_i) $\leftarrow$ $\{(w_t, P(I_i, w_t))$ with $t = 1...5$ and $\{w_t\}$ sorted according to probability$\}$
  FinalAnno(I_i) $\leftarrow \emptyset$
  //Criteria for selecting an image:
  if $P(I_i, w_1) > threshold_\alpha$ **then**
    foreach (w_j, w_k) with $j \neq k \in$ Anno(I_i) **do**
      //If words are dissimilar:
      if S$(w_j, w_k) < threshold_\beta$ **then**
        if $w_k \notin$ FinalAnno(I_i) **then**
          FinalAnno(I_i) $\leftarrow \{w_k\} \cup$ FinalAnno(I_i)
          $P'(I_i, w_k) \leftarrow$ lowerProbability (I_i, w_k)
        end if
      end if
    end for
    foreach $w_t \in$ FinalAnno(I_i) **do**
      //Lowering probabilities of related terms:
      lowerRelatedProbabilities(I_i, w_t)
    end for
  else
    FinalAnno(I_i) $\leftarrow$ Anno(I_i)
  end if
end for
return P'

may be obtained.

Figure 4.1 represents all the annotation words generated by the baseline method for the Corel 5k dataset. The horizontal axis shows all the words, while the vertical axis displays the probability value of the words, which is denoted as *confidence score*. The graph attempts to study whether or not there exists a correlation between the probability value of some words and their correctness. According to this graphic, the probability value of most annotation words, which have been incorrectly generated, is very low. Additionally, the probability value of the correctly guessed annotation words is within a range.

By considering this, only those images whose confidence score is greater than a threshold (α) have been considered. The threshold is set after performing cross-validation on the training set. Then, the semantic relatedness measure is computed for each pair of candidate annotations until candidates that are not semantically related to the others are detected. Two words are unrelated when their semantic relatedness value falls below a given threshold (β). Once detected, the confidence score of the "noisy" candidates is reduced. Additionally, related candidates to the noisy ones are detected and their score is reduced accordingly. The pruning process concludes after re-ranking and selecting, again, the five with the highest confidence scores (FinalAnno).

4.2 Experimental Work

The Corel 5k dataset (Section 3.7.2) has been used as a preliminary dataset for the experiments. Additionally, I have employed the collection provided by the Photo Annotation Task of the 2009 edition of ImageCLEF campaign (Section 3.7.5). To establish a proper comparison of results between the two datasets, I have defined a common experimental ground: I utilise the same baseline algorithm, extract the same image features,

perform the same parameter estimation, and apply the same evaluation measures.

4.2.1 Image Features

The features used in the model correspond to those that achieve the highest performance, among the proposed features for the baseline approach by Yavlinsky et al. (2005). In particular, the features used are a combination of the colour CIELAB and texture Tamura descriptors. As they were extracted globally from the whole image without segmenting or performing object recognition the approach followed is considered as global features. However, the image is tiled in order to capture a better description of the distribution of features across the image. Afterwards, the features are combined to maintain the difference between images with, for instance, similar colour palettes but different spatial distribution across the image. The process for extracting each of these features is as follows: each image is divided into nine equal rectangular tiles and the mean and standard deviation feature per channel are calculated in each tile. The resulting feature vector is obtained after concatenating all the vectors extracted in each tile.

CIE L*a*b* (CIELAB) (Hanbury and Serra 2002) is the most perceptually accurate colour space specified by the International Commission on Illumination (CIE). Its three coordinates represent the lightness of the colour (L*), its position between red/magenta and green (a*) and its position between yellow and blue (b*). The histogram was calculated over two bins for each coordinate.

The Tamura texture feature is computed using three main texture features called "contrast", "coarseness", and "directionality". Contrast aims to capture the dynamic range of grey levels in an image. Coarseness has a direct relationship to scale and repetition rates and it was considered by Tamura et al. (1978) as the most fundamen-

Table 4.1: Parameters used for baseline runs for Corel 5k and ImageCLEF09 collection

Task	Corel 5k	ImageCLEF09
MAP	0.2861	0.3079
Queries	179	53
Scale	1.78	4.7
Norm	2	1

tal texture feature and finally, directionality is a global property over a region. The histogram was calculated over two bins for each feature.

4.2.2 Evaluation Measures

As discussed in Section 3.2, the evaluation of the performance of an automated image annotation algorithm can be accomplished following two different metrics, the image annotation and the ranked retrieval. In this research, results are shown under the rank retrieval metric that consists in ranking the images according to their probability of annotation. Retrieval performance is evaluated using the mean average precision (MAP), which is the average precision, over all queries, at the ranks where recall changes as relevant items occur. I proposed as queries those keywords that annotate more than two images in the test set. For the Corel 5k dataset this makes 179 single-word queries, and 53 for ImageCLEF09. Single-word queries are used instead of multi-word queries in order to follow the approach of Feng et al. (2004), who designed the retrieval task as made up of single word queries. For compatibility purposes with the baseline approach, queries are selected based on their occurrence in the test set more than twice.

4.2.3 Parameter Estimation

I performed a 10-fold cross validation on the training set in order to tune the parameters of the system, the kernel bandwidth scaling factor called *scale* and the kernel shape as given by the *norm* (Section 4.1.1).

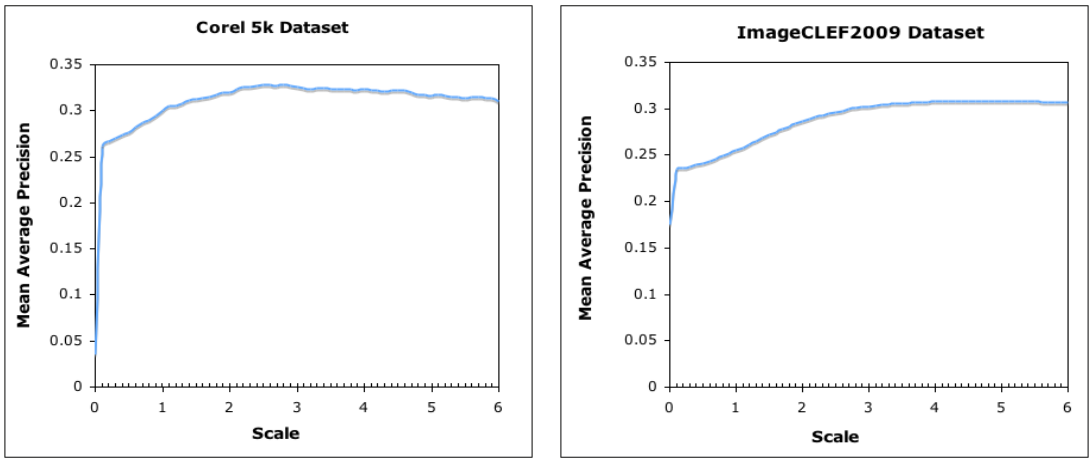

Figure 4.2: Parameter optimisation for the Corel 5k and ImageCLEF09

Thus, the dataset was divided into three parts: a training set, a validation set, and a test set. The validation test is used to find the parameters of the model. After that, the training and validation set are merged to form a new training set. Figure 4.2 represents the dependency of the mean average precision with the kernel bandwidth scaling factor for both datasets. As observed, the larger the scale, the higher the MAP.

4.3 Analysis of Results

The baseline results for the two datasets under the rank retrieval metric are presented in Table 4.1. The performance of the ImageCLEF09 collection is slightly higher than that obtained by the Corel 5k. This is the result of the vocabulary of the ImageCLEF09 being composed of more general terms than the Corel 5k dataset as reflected in Appendix A.1 and Appendix A.4. Consider, for example, the animal category in the ImageCLEF09 and Corel 5k datasets. In the first case, the term "animals" refers to all kinds of animals whereas in the second there exist much more specific terms such as "polar bear", "grizzly", "black bear" that forces the annotation system to identify different types of bears with the subsequent loss of precision. However, the increment produced over the Corel 5k is not very significant because of the difficulty in detect-

Table 4.2: Results obtained for the two datasets expressed in terms of mean average precision. Bold figures indicate that values are statistically significant over the baseline according to the sign test. The significant level α is 5% and p-value < 0.001

Algorithm	Corel5k	ImageCLEF09
Baseline	0.2861	0.3079
Training set correlation	**0.2922**	0.3081
Web Corpus (Google)	**0.2882**	**0.3095**
Web Corpus (Yahoo)	**0.2901**	0.3090
Web Corpus (Exalead)	**0.2900**	0.3091
Jiang and Conrath (JCN)	**0.2870**	0.3081
Hirst and St-Onge (HSO)	**0.2870**	0.3081
Leacock&Chodorow (LCH)	0.2861	0.3078
PATH	**0.2870**	0.3081
Wu and Palmer (WUP)	**0.2870**	0.3081
Resnik (RES)	0.2868	0.3078
Patwardhan (VEC)	**0.2870**	0.3081
Lin (LIN)	**0.2870**	0.3081
Adapted Lesk (LESK)	0.2868	0.3079
Milne and Witten (WLM)	**0.2870**	0.3119

ing some words of the ImageCLEF09 vocabulary. Some terms are quite subjective such as "Aesthetic_Impression", Overall_Quality", "Fancy". Others are negations like "No_Persons", "No_Visual_Season, etc. Consequently, an algorithm generating words based solely on their visual properties might have difficulties in achieving good performance.

Table 4.2 shows the resulting MAP for each run using different knowledge bases and semantic relatedness measures. The results, which are statistically significant over the baseline according to the sign test (see Section 3.4), are represented in bold characters. The significant level α is set to 5% and p-value < 0.001. In total, there are 15 runs: one baseline, one statistical correlation, three applying NGD to Google, Yahoo, and Exalead, nine applied to WordNet, and, finally, WLM applied to Wikipedia.

Results confirm previous expectations that the use of semantic measures increase the performance of a baseline probabilistic method. However, increments differ from measure to measure.

4.3.1 Discussion

For the Corel dataset, the best improvement (2%) corresponds to keyword correlation in the training set, closely followed by correlation using an external web corpus like Yahoo (1.40%). Yahoo together with Exalead beat Google. Regarding the WordNet relatedness measures, the best performing were JCN, HSO, LIN, WUP, PATH, and VEC. However, the improvement achieved by measures based on Wikipedia and WordNet are not dramatic in spite of both being statistically significant.

For the ImageCLEF09 dataset, although the best result corresponds to Wikipedia it is discarded in benefit of the web correlation using Google with 0.52% of improvement, which is statistically significant over the baseline method. Measures based on WordNet achieved a rather poor performance, being the best performing, again, JCN, HSO, LIN, WUP, PATH, and VEC. Regarding the low values obtained with ImageCLEF09 in contrast to Corel 5k, previous experiences (Llorente et al. 2008b, Llorente et al. 2009c,Llorente and Rüger 2009a) have confirmed that vocabularies with a small number of terms hinders the functioning of this algorithm.

In general, these results confirm a priori expectations that measures based on word correlation perform better than those based on lexical resources such as WordNet and Wikipedia. A plausible explanation might be that they do not need to perform a previous disambiguation task as part of their computation, as it happens in the case of WordNet and Wikipedia. Although both methods present similar disambiguation capabilities: around 70% of accuracy for the ImageCLEF09 and a bit higher, 90%, for

the Corel 5k dataset, WordNet performs slightly worse. Table 4.3 shows some examples where the most popular sense of a word does not match the sense attributed in the collection. As these inaccuracies in the disambiguation process may have translated into inferior results for WordNet and Wikipedia based methods, a more sophisticated disambiguation strategy will be implemented as future line of work.

The most important limitation, affecting approaches that rely on a training set, is that they are limited to the scope of the topics represented in the collection. Additionally, the sparseness of the data could affect the performance of the final model. The smoothing strategy adopted here is very simple and consists in replacing all zeros by a small number different from zero.

4.3.2 Combination of Results

The motivation behind the combination of results comes from the graphical observation that there exist huge variations per word depending on the selected method. Figure 4.3 represents the average precision per word of a given method divided by the average precision of the baseline run for all the words in the vocabulary. In particular, the following approaches were considered: training set correlation, Yahoo correlation, WordNet (HSO), and Wikipedia. The peaks observed in Figure 4.3 show that for many words the performance of one method is clearly better than the performance of the others. This observation is confirmed by Table 4.4, which shows the best ten performing words for the Corel 5k dataset and with which measure they achieve the highest performance. Moreover, the same behaviour is observed for the ImageCLEF09 image collection. This suggests that an increment in the performance could be gained after combining appropriately the outputs of the annotation algorithms.

The problem of *rank aggregation* or *rank fusion* has been addressed in the field of

Table 4.3: Word sense disambiguation (WSD) for the Corel 5k and for ImageCLEF09 dataset performed by Wikipedia and WordNet. Senses wrongly disambiguated by measures based on WordNet and Wikipedia are marked with an asterisk

Term	Wikipedia Sense	WordNet Sense	Dataset Sense
pillar	*pillar (band)	*a fundamental principle	column
kit	*body kit	*a case for containing a set of articles	a young cat or fox
range	*range (mathematics)	*an area in which something acts or operates	a range of mountains
palm	*hand	*the inner surface of the hand	palm tree
model	*model (abstract)	*a hypothetical description of a process	pattern
prop	*rugby league positions	*a support beneath something to keep it from falling	propeller
run	*execution (computers)	*a score in baseball	to move at fast speed
outdoor	*Outside(magazine)	the region outside of something	situated out of doors
canvas	canvas	*the setting for a fictional account	canvas for painting
macro	*macro(computer science)	*a single computer instruction	macro lens
plants	plant	*building for carrying on industrial labour	a living organism
small group	*group (auto-racing)	any number of entities considered as a unit	group

Table 4.4: Semantic measure that performs better for the top ten best performing words of the Corel 5k dataset. The third column shows the % improvement of the semantic combination method (SC) over the baseline for every word

Word	Best Method	Δ
herd	Web Corpus (Yahoo)	56%
lawn	Training set Correlation	78%
shadows	Web Corpus (Yahoo)	96%
nest	Web Corpus (Yahoo)	50%
light	Wikipedia	54%
forest	Training set Correlation	35%
frozen	Training set Correlation	59%
reefs	Web Corpus(Yahoo)	22%
meadow	Training set Correlation	20%
locomotive	Training set Correlation	17%

information retrieval by many researchers, such as Shaw and Fox (1994), Bartell et al. (1994), Aslam and Montague (2001), and Wilkins et al. (2006). In particular, a *rank fusion* task is described as follows: Given a set of rankings, the task consists in combining these lists in a way that the performance of the combination is optimised. The ranks to be combined should be compliant with the following requirements. All ranks should have outputs on the same scale; all ranks should produce accurate estimates of relevance, and they should be independent from one to the other.

Fusion strategies based on the *Borda-fuse*, and *weighted Borda-fuse* methods were attempted. In particular, the *Borda-fuse* method, which is based on a voting model, works as follows. Each voter ranks a fixed set of c candidates in order of preference, the top ranked candidate is given c points, the second $c-1$ and so on. Then, the candidates are ranked according to the total number of points. Finally, the candidate with the greatest number of points wins the election. The *weighted Borda fusion* is a variation of the previous where each score is multiplied by a weight α_i. This weight can be an assessment of the performance of the system such as the average precision.

Table 4.5: Final results expressed in terms of MAP. Both results are statistically significant over the baseline, with a significant level α of 5%

Algorithm	**Corel5k**	**ImageCLEF09**
Baseline	0.2861	0.3079
Semantic Combination	**0.3007**	**0.3134**
P-value	< 0.001	< 0.001

The previous methods were initially applied as an *annotation aggregation* strategy. This strategy is defined as follows. Candidates are the set of five annotation words generated by each algorithm and the score is assigned according to the order provided by the probability value. However, due to the poor performance of the combination, the strategy was discarded and the *rank aggregation* strategy considered. Finally, a new method was finally proposed as Borda-based approaches did not increase the final performance of the system. Specifically, the *semantic combination* (SC) method consists in recording during the training phase the best performing method based on the highest average precision per word. Thus, the algorithm applies in each step and for every word the best recorded semantic measure and this translates into substantial increments in the results for both collections.

Final results are represented in Table 4.5, where the p-value shows that the performance improvement over the baseline is statistically significant according to the sign test for the two collections as the calculated p-value is lower than 0.001. I have considered a significant level α of 5%. Besides that, a 5% and 2% improvement is obtained for the Corel 5k and ImageCLEF09, respectively.

Figure 4.4 demonstrates the efficiency of the presented method as it shows large improvements for the ten best performing words for the Corel 5k and the ImageCLEF09, respectively.

From the point of view of the semantic measures, the fact of combining several

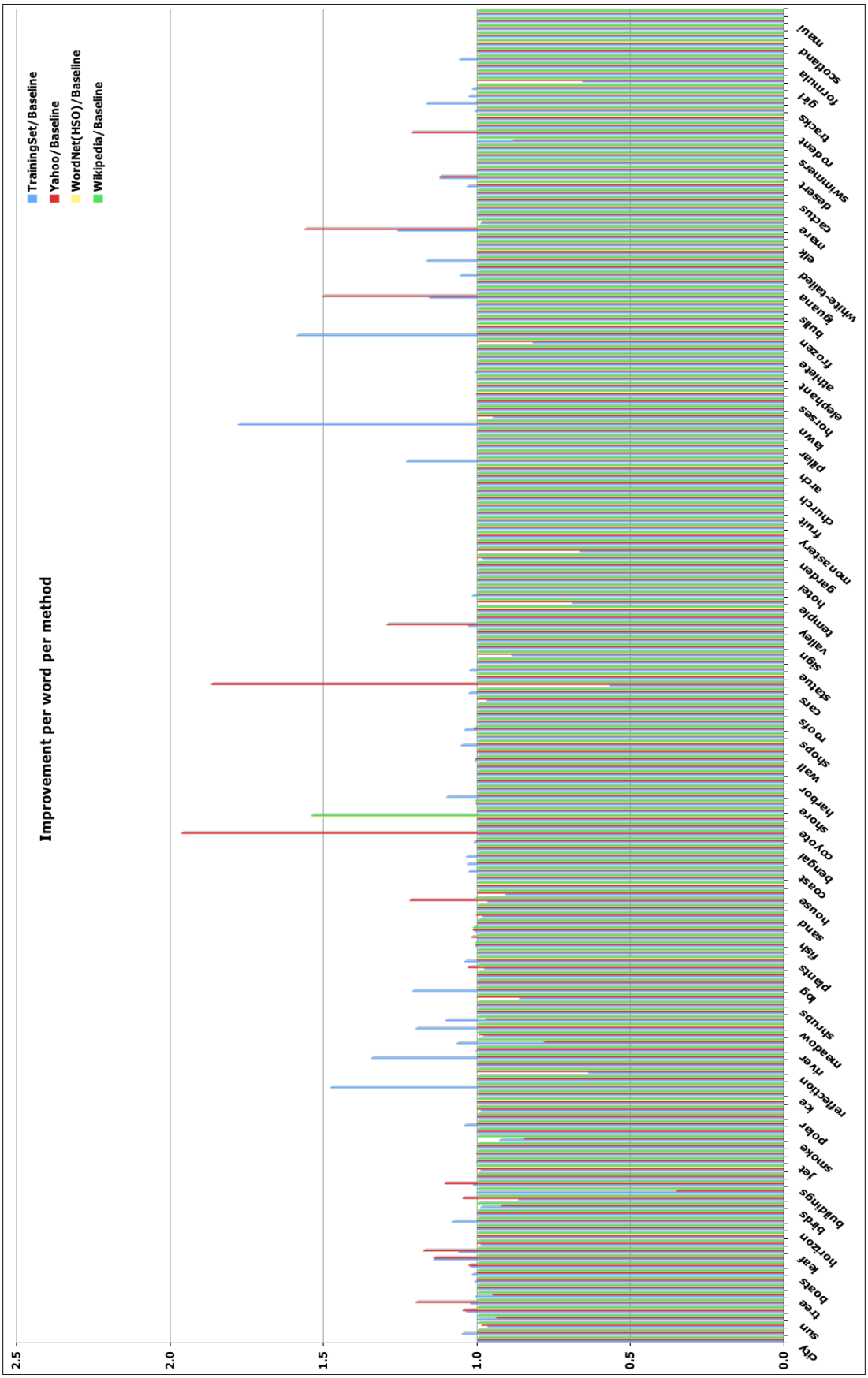

Figure 4.3: Improvement of each method over the baseline in terms of precision per word for the Corel5k dataset

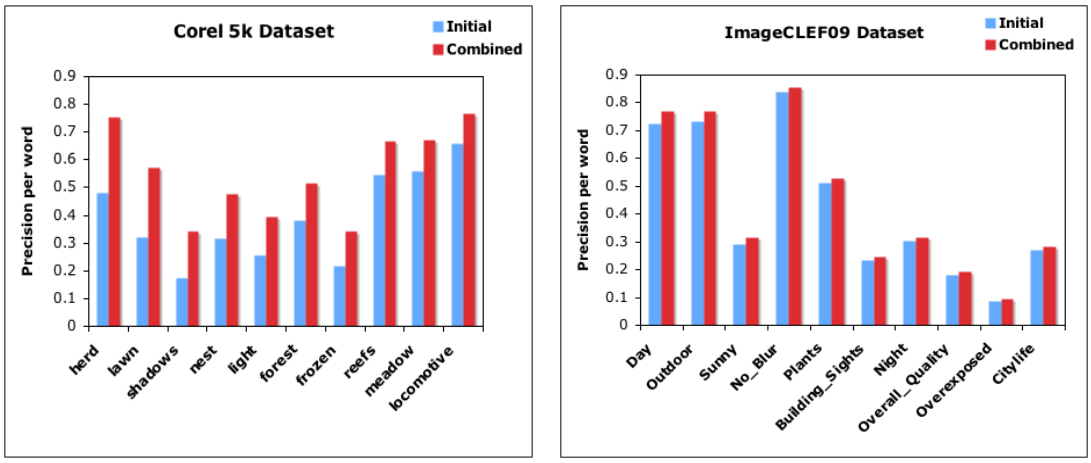

Figure 4.4: Ten best performing words for the Corel 5k and ImageCLEF09 datasets expressed in terms of precision per word

measures is perfectly justifiable. On the one hand, a semantic measure based solely on correlation on the training set is limited to the topics represented in the collection but on the other, it is absolutely necessary in order to get a sense of what the collection is about. However, if this information is combined with world knowledge coming from external sources such as correlation in the web, WordNet and Wikipedia, it is clear that the information generated is much more valuable than in the first case. Consequently, this appropriate combination leads to the significant increment in the performance observed in the results.

4.4 Conclusions

The main goal of this experimental work is to improve the accuracy of a traditional automated image annotation system based on a machine learning method.

I have demonstrated that building a system that models the context of the image on top of another that is able to accomplish the initial annotation of the objects increases significantly the mean average precision of the final annotation system. The context of the image is defined by the objects represented there together with the semantic

relations that connect them. Thus, unrelated objects are to be discarded from the annotations.

For this purpose, 14 semantic relatedness measures have been implemented, two distributional measures: statistical correlation in the training set and in a web corpus and 10 based on semantic networks representations like WordNet and Wikipedia. The performance of the proposed approaches is computed in order to analyse their benefits and limitations. Finally, I propose a combination method that achieves statistically significant better results over the baseline. Experiments have been carried out with two datasets, Corel 5k and ImageCLEF09.

As future lines of work, I intend to enhance the encouraging results shown in this research by introducing Semantic Web technologies in order to further improve the algorithm. I plan to use ontologies to model generic knowledge (i.e. that can be used with different datasets) about images, and then exploiting them to additionally prune incoherent words and representing the relationships among objects contained in the scene.

Chapter 5

A Fully Semantic Integrated Annotation Model

In this chapter, we[1] propose an automated image annotation algorithm that constitutes an example of the *fully semantic integrated models*. In particular, it is a direct image retrieval framework based on Markov Random Fields (MRFs).

The novelty of our approach lies in the use of different kernels in our non-parametric density estimation together with the utilisation of configurations that explore semantic relationships among concepts at the same time as low-level features, instead of just focusing on correlation between image features like in previous formulations. Hence, we introduce several configurations and study which one achieves the best performance.

Results are presented for two datasets, the usual benchmark Corel 5k (Section 3.7.2) and the collection proposed by the 2009 edition of the ImageCLEF campaign (Section 3.7.5). We observe that, using MRFs, performance increases significantly depending on the kernel used in the density estimation for the two datasets. With respect to

[1] In this chapter, "we" refers to the joint work done with Manmatha of the University of Massachusetts at Amherst.

the language model, best results are obtained for the configuration that exploits dependencies between words together with dependencies between words and visual features. For the Corel 5k dataset, our best result corresponds to a mean average precision of 0.32, which compares favourably with the highest value ever obtained, 0.35, achieved by Makadia et al. (2008) albeit with different features. For the ImageCLEF09 collection, we obtained 0.32, as mean average precision.

5.1 Background

As discussed in Chapter 1, the problem of modelling annotated images has been addressed from several directions in the literature. Initially, a set of generic algorithms were developed with the aim of exploiting the dependencies between image features and implicitly between words. However, many algorithms do not explicitly exploit the correlation between words. These set of algorithms correspond to *classic probabilistic models*.

The human understanding of a scene was a topic confronted by many researchers in the past. Authors like Biederman (1981), and then, Torralba and Oliva (2003) supported the hypothesis that objects and their containing scenes were not independent. For example, the prediction of the concept "beach" is usually followed by the presence of "water" and "sand". On the other hand, a "polar bear" should never appear in a "desert" scenario, no matter how high the probability of the prediction. As a result, a new collection of algorithms, devoted to exploring word-to-word correlations, shortly emerged. Chapter 2 revises in detail these algorithms that correspond to *semantic-enhanced models*. These methods relied on either filtering the results obtained by a previous baseline annotation method or on creating adequate language models as a way to boost the efficiency of previous approaches. However, as seen before, the

Table 5.1: Best performing automated image annotation algorithms expressed in terms of number of recalled words (NZR), recall (R), precision (P), and F-measure for the Corel 5k dataset. The first block represents the *classic probabilistic models*, the second is devoted to the *semantic-enhanced models*, and the third depicts *fully integrated semantic models*. The evaluation is done using 260 words that annotate the test data. (-) means numbers not available

Model	Author	NZR	R_{260}	P_{260}	F_{260}
CRM	Lavrenko et al. (2003)	107	0.19	0.16	0.17
Npde	Yavlinsky et al. (2005)	114	0.21	0.18	0.19
InfNet	Metzler and Manmatha (2004)	112	0.24	0.20	0.22
CRM-Rectangles	Feng et al. (2004)	119	0.23	0.22	0.22
MBRM	Feng et al. (2004)	122	0.25	0.24	0.24
SML	Carneiro et al. (2007)	137	0.29	0.23	0.26
JEC	Makadia et al. (2008)	113	0.40	0.32	0.36
BHMMM	Shi et al. (2006)	122	0.23	0.14	0.17
Anno-Iter	Zhou et al. (2007)	-	0.18	0.21	0.19
TBM	Shi et al. (2007)	153	0.34	0.16	0.22
KM-500	Srikanth et al. (2005)	93	0.32	0.18	0.23
DCMRM	Liu et al. (2007)	135	0.28	0.23	0.25
SCK+HE	Li and Sun (2006)	-	0.36	0.21	0.27
MRFA-region	Xiang et al. (2009)	124	0.23	0.27	0.25
MRFA-grid	Xiang et al. (2009)	172	0.36	0.31	0.33

improvement in the performance might be hindered by error propagation of the baseline classifiers and by the lack of sufficient data, which can lead to over-fitting.

Nevertheless, Markov Random Fields (MRFs) provide a convenient way of modelling context-dependent entities like image content. This is achieved through characterizing mutual influences among such entities using conditional MRF distributions. The main benefit of using a MRF comes from the fact that we can model correlations between words explicitly. Models based on MRF are called *fully integrated semantic models*. Additionally, we observe from Table 5.1 that these models present higher performance than the *classic probabilistic* and *semantic-enhanced models*. Therefore, Table 5.1 shows an updated version of Table 1.1 and Table 2.7 for the Corel 5k dataset. The table is

divided into three blocks. The first block represents the *classic probabilistic models*, the second is devoted to the *semantic-enhanced models*, and the third corresponds to *fully integrated semantic models*. The evaluation measures considered are the number of recalled words (NZR), precision (P), recall (R), and F-measure; all computed for the 260 words that annotate the test set. Note that results are ordered in each block according to the increasing value of the F-measure.

Specifically, this chapter presents a direct image retrieval framework that makes use of different configurations to model the image content. Besides that, the application of MRF theory allows us to easily formulate the joint distribution of the graph. The novelty of our approach lies in the use of different kernels, in our non-parametric density estimation, together with the utilisation of configurations that explore semantic relationships among concepts and low-level features instead of just focusing on correlation between image features like in previous formulations. The emphasis of this work is placed on the model and on obtaining a better kernel estimation. As Makadia et al. (2008) show, a good choice of features can give very good results. Here our focus is not on the features. We use simple global features.

The rest of the chapter is structured as follows. Section 5.2 discusses state-of-the-art automated image annotation algorithms. Section 5.3 introduces our Markov Random Field model. Section 5.4 explains the experiments undertaken, while Section 5.5 analyses our results. Finally, Section 5.6 explains our conclusions.

5.2 Related Work

Markov Random Fields have been widely used in computer vision applications to model spatial relationships between pixels.

Escalante et al. (2007b) proposed a MRF model as part of their image annotation

framework, which additionally uses word-to-word correlation. Hernández-Gracidas and Sucar (2007) carried out another variation of the previous approach placing emphasis on the spatial information relation among objects. Both works are based on the MRF model proposed by Carbonetto et al. (2004), whose approach is considered to be out of the scope of this work as it is more aligned with the approaches usually adopted in the field of computer vision.

Qi et al. (2007) proposed a model based on Gibbs Random Fields applied to video annotation. Their method, the correlative multi-label (CML) framework, simultaneously classifies concepts while modelling the correlations between them in a single step. They conduct their experiments on TRECVID 2005 dataset outperforming several algorithms.

Feng and Manmatha (2008) were the first to do direct retrieval (without an intermediate annotation step) using a MRF model. By ranking while maximising average precision the model is simplified due to the fact that the normaliser does not need to be calculated. They used discrete image features and obtained comparable results to the state-of-the arts algorithms. Later on, Feng (2008) presented a similar model but applied it to the case of continuous image features. He achieved better performance with the continuous model than with the discrete model although the latter was more efficient in terms of speed. Both models were based on the Markov Random Field framework developed by Metzler and Croft (2005), who modelled term dependencies in text retrieval. The novelty of their approach lies in training the model that maximises directly the mean average precision instead of maximising the likelihood of the training data.

More recently, Xiang et al. (2009) presented a new approach able to perform directly automated image annotation. They adopt a MRF to model the context relationships

among semantic concepts with keyword subgraphs generated from training sample for each keyword. Thus, they defined two potential functions in cliques up to order two: the site potential and the edge potential. The former models the joint probability of an image feature and a word and was modelled using the multiple Bernoulli relevance model (MBRM) (Feng et al. 2004). The edge potential approximates the joint probability of an image feature and a correlated word. The parameter estimation is done adopting a pseudo-likelihood scheme in order to avoid the evaluation of the partition function. Finally, they showed significant improvement over six previous approaches for the Corel 5k dataset.

5.3 Markov Random Fields

For the basic Markov Random Field model we followed the approach and the notation used by Feng (2008). However, our graph configurations are different, and consequently, the two models differ. The only similarity is that both of us explored the dependencies between words and image regions. Nevertheless, his focus is on exploring the dependencies between image regions, while ours is on the relationships between words.

Let G be an undirected graph whose nodes are called I and Q. A Markov Random Field (MRF) is an undirected graph G which allows the joint distribution between its two nodes to be modelled in terms of:

$$P_\Lambda(I,Q) = \frac{1}{Z_\Lambda} \prod_{c \in C(G)} \psi(c;\Lambda), \tag{5.1}$$

where $C(G)$ is the set of cliques defined in the graph G, $\psi(c;\Lambda)$ is a non-negative potential function over clique configurations parametrized by Λ, and Z_Λ is the value that normalised the distribution. When applied to the image retrieval case, the nodes of the graph, I and Q, represent respectively a image of the test set and a query. The

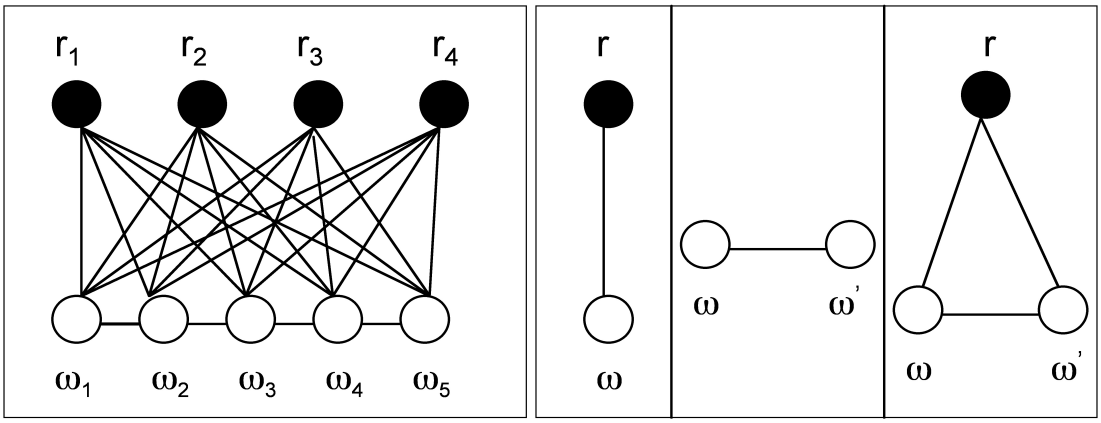

Figure 5.1: Markov Random Fields graph model. On the right-hand side, we illustrate the configurations explored in this chapter: one representing the dependencies between image features and words (r-w), another between two words (w-w'), and the final one shows dependencies among image features and two words (r-w-w').

image is represented by a set of feature vectors r and the query by a set of words w. Following the same reasoning as Feng (2008) in his continuous model developed, we approximate the joint distribution using the following exponential form:

$$\psi(c;\Lambda) = e^{\lambda_c f(c)}. \tag{5.2}$$

Therefore, we arrive at the following model where images are ranked according to their posterior probability:

$$P_\Lambda(I|Q) \stackrel{rank}{=} \sum_{c \in C(G)} \lambda_c f(c), \tag{5.3}$$

where $f(c)$ is a real-valued feature function defined over the clique c weighed by λ_c.

Figure 5.1 shows a graph representing the dependencies explored in our model. The left side of the image illustrates the clique configurations considered in this research which contemplates cliques of up to third order. A 2-clique (r-w) consisting of a query node w and a feature vector r, followed by a 2-clique (w-w') representing the dependencies between words w and w', and, finally a 3-clique (r-w-w') capturing the relation between a feature vector r and two word nodes w and w'.

According to the graph, the posterior probability is expressed as:

$$P_\Lambda(I|Q) \stackrel{rank}{=} \sum_{c \in T} \lambda_T f_T(c) + \sum_{c \in U} \lambda_U f_U(c) + \sum_{c \in V} \lambda_V f_V(c), \tag{5.4}$$

where T is the set of 2-cliques containing a feature vector r and a query term w, U is the set of 2-clique (w-w') representing the dependencies between two words w and w' and V is the set of 3-cliques (r-w-w') capturing the relation between a feature vector r and two word nodes w and w'. Finally, and for simplicity, we make the assumption that all image features are independent of each other given some query Q.

The differences between this work, Feng (2008), and Feng and Manmatha (2008) reside mainly in the divergent associations defined in our respective graphs. Both approaches investigate the dependencies between image regions and words (configuration r-w). However, their focus is on exploring the dependencies between various image regions while ours relies on the relationships between words. Thus, the rest of the configurations presented in this chapter are new. Another differing point is that we work with feature vectors extracted from the entire image instead of with image regions. Additionally, both works differ on their selection of visual features. Finally, Feng (2008) and Feng and Manmatha (2008) employ a Gaussian kernel in their density estimation while our strongest point is exploring additional kernels such as the "square-root" or the Laplacian kernel.

In what follows, we explain in detail the different configurations followed in this research.

5.3.1 Image-to-Word Dependencies

This configuration is formed by the set of 2-cliques r-w and it corresponds to the *Full Independence Model* developed by Feng (2008). The potential function associated to this clique expresses the probability of generating the word w, for a given image feature,

scaled by the prominence of the feature vector r in the test set image I, as shown in:

$$f_T(c) = P(w|r)P(r|I), \tag{5.5}$$

where $P(r|I)$ is set to be the inverse of number of features vector per image, as we make the assumption that the distribution is uniform. $P(w|r)$ is estimated applying Bayes' rule:

$$P(w|r) = \frac{P(w,r)}{\sum_w P(w|r)}, \tag{5.6}$$

where $P(w,r)$ is computed in a similar way to the continuous relevance model (CRM) developed by Lavrenko et al. (2004):

$$P(w,r) = \sum_{J \in \tau} P(J)P(w|J)P(r|J), \tag{5.7}$$

where τ represents the training set and J, a training image, and $P(J) \approx \frac{1}{|J|}$.

The function $P(r|J)$ is estimated using a non-parametric density estimation approach as represented in:

$$P(r|J) = \frac{1}{m} \sum_{t=1}^{m} k\left(\frac{|r - r_t|}{h}\right), \tag{5.8}$$

where r is a real-valued image feature vector of dimension d, m is the number of feature vectors representing the image J, t is an index over the set of biagrams in J. We propose as kernel function a Generalized Gaussian Distribution (Domínguez-Molina et al. 2003) whose probability density function (pdf) is defined as:

$$\text{pdf}(x; \mu, \sigma, p) = \frac{1}{2\Gamma(1 + 1/p)A(p,\sigma)} e^{-\frac{|x-\mu|}{A(p,\sigma)}^p}, \tag{5.9}$$

where $x, \mu \in \mathbb{R}$, p, $\sigma > 0$ and $A(p,\sigma) = \left[\frac{\sigma^2\Gamma(1/p)}{\Gamma(3/p)}\right]^{\frac{1}{2}}$. The parameter μ is the mean, the function $A(p,\sigma)$ is a scaling factor that allows the variance of x to take the value of σ^2, and p is the shape parameter that we will call norm. When $p = 1$, the pdf corresponds to a Laplacian or double exponential function and to a Gaussian when $p = 2$. Note that p can take any real value in $(0, \infty)$.

However, in this work, we will experiment with three types of kernels: a d-dimensional Laplacian kernel, which after simplification of Equation 5.9 yields

$$k_L(t;h) = \prod_{l=1}^{d} \frac{1}{2h_l} e^{-\left|\frac{t_l}{h_l}\right|}, \tag{5.10}$$

a Gaussian kernel, expressed as

$$k_G(t;h) = \prod_{l=1}^{d} \frac{1}{\sqrt{2\pi h_l^2}} e^{-\frac{1}{2}(\frac{t_l}{h_l})^2}, \tag{5.11}$$

and the "square-root" kernel (p=0.5)

$$k_{SQ}(t;h) = \prod_{l=1}^{d} \frac{1}{2h_l} e^{-\left|\frac{2t_l}{h_l}\right|^{\frac{1}{2}}}, \tag{5.12}$$

where $t = r - r_t$, and h_l is the bandwidth of the kernel, which is set by scaling the sample standard deviation of feature component l by the same constant *scale* (sc).

Finally, $P(w|J)$ is modelled using the same multinomial distribution as Lavrenko et al. (2004):

$$P(w|J) = \lambda_1 \frac{N_{w,J}}{N_J} + (1 - \lambda_1)\frac{N_w}{N}. \tag{5.13}$$

$N_{w,J}$ represents the number of times w appears in the annotation of J, N_J is the length of the annotation, N_w is the number of times w occurs in the training set and N is the aggregate length of all training annotations. λ_1 is the smoothing parameter and together with the coefficient that scales the kernel bandwidth represents the two parameters that are estimated empirically using a held-out portion of the training set.

5.3.2 Word-to-Word Dependencies

The 2-clique w-w' models word-to-word correlation and is approximated by the following potential function:

$$f_U(c) = \gamma f(w, w') = \gamma \sum_{w'} P(w|w'), \tag{5.14}$$

$$P(w|w') = \frac{P(w,w')}{P(w')} = \frac{\#(w,w')}{\sum_w \#(w,w')}, \tag{5.15}$$

where $\#(w,w')$ denotes the number of times the word w co-occurs together with the word w' annotating an image of the training set. To avoid the problem of the sparseness of the data, we follow a smoothing approach:

$$P(w|w') = \beta \frac{\#(w,w')}{\sum_w \#(w,w')} + (1-\beta)\frac{\sum_w \#(w,w')}{\sum_J \sum_w \#(w,w')}, \tag{5.16}$$

where β is the smoothing parameter.

5.3.3 Word-to-Word-to-Image Dependencies

The model consists of 3-cliques formed by the words, w and w' and the feature vector r, and captures the dependencies among them. The underlying idea behind this model is that a feature vector representing two visual concepts should imply a degree of compatibility between the visual information and the concepts, and between the concepts themselves. This compatibility is measured by the potential function. For instance, assume that we have a marine scene representing a portion of the sea and a boat, the visual features should reflect the visual properties of the boat and the sea regarding colour and texture and, at the same time, the concepts "sea" and "boat" should pose a degree of semantic relatedness as both represent objects that share the same image context. Thus, the potential function over the 3-clique r-w-w' can be expressed as:

$$\lambda_V f_V(c) = \delta f((w,w'),r), \tag{5.17}$$

where δ is the weight of the potential function. This can be formulated as the possibility of predicting the pair of words (w,w') given the feature vector r, weighted by the importance of the vector in the image I:

$$f((w,w'),r) = P((w,w')|r)P(r|I), \tag{5.18}$$

where $P(r|I) \approx \frac{1}{|I|}$, and $|I|$ is set to the number of feature vectors that represent a test image. By applying Bayes formula and the continuous relevance model (CRM) developed by Lavrenko et al. (2004) but adapted to $P((w, w'), r)$, we have the following:

$$P((w, w')|r) = \frac{\sum_{J \in \tau} P(J)P((w, w')|J)P(r|J)}{\sum_{(w,w')} P((w, w'), r)}, \tag{5.19}$$

where J refers to a training image, and τ to the training set. The rest of the terms are computed as follows. $P(J)$ is aproximated by $\frac{1}{|J|}$. $P((w, w')|J)$ is estimated following a generalisation of a multinomial distribution (Lavrenko et al. 2004) as seen in Section 5.3.3. Finally, $P(r|J)$ is calculated following a Generalized Gaussian kernel estimation as in Equation 5.10, 5.11, and 5.12. In this model, we have three parameters: the smoothing parameter λ_2 of the multinomial distribution and two additional ones derived from the kernel estimation (scale sc, and γ) that are estimated during the training phase.

Multinomial Distribution of Pairs of Words

The multinomial distribution of pairs of words is modelled using the formula:

$$P((w, w')|J) = \sum_{w'} \lambda_2 \frac{N_{(w,w'),J}}{N_J} + (1 - \lambda_2)\frac{N_{(w,w')}}{N}. \tag{5.20}$$

The distribution measures the probability of generating the pair w and w', as annotation words, for the image J based on their relative frequency in the training set. Therefore, the first term reflects the preponderance of the pair of words (w, w') in the image J whereas the second is added as smoothing factor and registers the behaviour of the pair in the whole training set. Thus, $N_{(w,w'),J}$ represents the number of times, zero or one, (w, w') appears in the image J, N_J is the number of pairs that could be formed in the image J, $N_{(w,w')}$ is the number of times (w, w') occurs in the whole training set, $N = \sum_J N_J$, and λ_2 is the smoothing parameter.

For instance, when estimating the distribution of pairs of words formed by the term "tree" in an image annotated with the words "palm", "sky", "sun", "tree" and, "water", we should consider the weight of all pairs appearing in the image as well as in the rest of the training set:

$$\begin{aligned} P((\text{“}tree\text{”}, w')|J) &= \left[\lambda_2 \frac{1}{10} + (1-\lambda_2)\frac{221}{20,972}\right]_{tree-sky} + \\ &+ \left[\lambda_2 \frac{1}{10} + (1-\lambda_2)\frac{23}{20,972}\right]_{tree-sun} + \\ &+ \left[\lambda_2 \frac{1}{10} + (1-\lambda_2)\frac{143}{20,972}\right]_{tree-water} + \\ &+ \left[\lambda_2 \frac{1}{10} + (1-\lambda_2)\frac{22}{20,972}\right]_{tree-palm} + \\ &+ \sum_{w'} \left((1-\lambda_2)\frac{N_{(tree,w')}}{20,972}\right)_{tree-w'}, \end{aligned}$$

where w' represents the rest of vocabulary words that co-occur with "tree" in the rest of the training set, but not in J. Additionally, m is an integer value that represents the number of words annotating an image J, N_J is equal to $\binom{m}{2}$, and N is a constant for a given collection, and it is set to $20,972$ for the Corel 5k dataset. Even if there are images annotated by one single word or without annotations, $P((w, w')|J)$ might be different from zero due to the contribution of the second factor in Equation 5.20.

5.4 Experimental Work

For our experiments, we have adopted a standard annotation database, the Corel 5k dataset, which is a considered benchmark in the field (Section 3.7.2). Additionally, we use the collection provided by the Photo Annotation Task of the 2009 edition of ImageCLEF campaign (Section 3.7.5). Note that, although we did not participate in that edition, we compare our results with the other participants in order to provide an estimation of our performance.

5.4.1 Visual Features

Prior to our modelling phase, we undertake a *feature selection task*. Its objective is to select from a set of available features an optimal subset that achieves the highest efficiency in terms of performance. Our initial set features are the top-performing global image descriptors proposed by Little and Rüger (2009). Our ultimate goal is to find an adequate combination where the redundancy between individual features is minimised and where, at the same time, combinations either highly correlated or suffering from multivariate prediction are discarded. This is achieved through training our baseline model r-w as described in Section 5.4.3.

Our final selection corresponds to a combination of global features. In particular, a 3x3 tiled marginal histogram of global CIELAB colour space computed across 2+2+2 bins, with a 3x3 tiled marginal histogram of Tamura texture across 2+2+2 bins with coherence of 6 and coarseness of 3, with a 3x3 tiled marginal histogram of global HSV colour space computed across 2+2+2 bins, and with a Gabor texture feature using six scales and four orientations.

The CIELAB colour and Tamura texture were introduced in Section 4.2.1.

HSV is a cylindrical colour space with H (hue) being the angular, S (saturation) the radial and V (brightness) the height component. The H, S and V axes are subdivided linearly (rather than by geometric volume) into two bins each.

The final feature extracted is a texture descriptor produced by applying a Gabor filter to enable filtering in the frequency and spatial domain. We applied to each image a bank of four orientation and six scale sensitive filters that map each image point to a point in the frequency domain.

The image is tiled in order to capture a better description of the distribution of features across the image. Afterwards, the features are combined to maintain the

difference between images with, for instance, similar colour palettes but different spatial distribution across the image.

5.4.2 Evaluation Measures

In this research, we present our results under the rank retrieval metric which consists in ranking the images according to the posterior probability value $P_\Lambda(I|Q)$ as estimated in Equation 5.3. Then, retrieval performance is evaluated with the mean average precision (MAP), which is the average precision, over all queries, at the ranks where recall changes where relevant items occur. For a given query, an image is considered relevant if its ground-truth annotation contains the query. For simplicity, we employ as queries single words. For the Corel 5k dataset we use 260 single word queries and 53 for the ImageCLEF09; in both cases we use all the words that appear in the test set.

5.4.3 Model Training

The training was done by dividing the training set into two parts: the training set and the validation or held-out set. The validation test is used to find the parameters of the model. After that, the training and validation set were merged to form a new training set that helps us to predict the annotations in the test set. For the Corel 5k dataset, we partitioned the training set into 4,000 and 500 images. The ImageCLEF09 was divided into 4,000 as training set, and 1,000 as held-out data.

Metzler and Croft (2005) argued that, for text retrieval, maximising average precision rather than likelihood was more appropriate. Feng and Manmatha (2008) showed that this approach worked for image retrieval and we also maximised average precision. We followed a hill-climbing mean average precision optimisation as explained by Morgan et al. (2004).

Table 5.2: State-of-the-art of algorithms in direct image retrieval expressed in terms of mean average precision (MAP) for the Corel 5k dataset. Results with an asterisk show that the number of words used for the evaluation are 179, instead of the usual 260. The first block corresponds to the *classic probabilistic models*, the second illustrates models based on Markov Random Fields, and the last shows our best performing results

Model	Author	MAP
CMRM	Jeon et al. (2003)	0.17*
CRM	Lavrenko et al. (2003)	0.24*
CRM-Rectangles	Feng et al. (2004)	0.26
LogRegL2	Magalhães and Rüger (2007)	0.28*
Npde	Yavlinsky et al. (2005)	0.29*
MBRM	Feng et al. (2004)	0.30
SML	Carneiro et al. (2007)	0.31
JEC	Makadia et al. (2008)	0.35
Discrete MRF	Feng and Manmatha (2008)	0.28
MRF-F1	Feng (2008)	0.30
MRF-NRD-Exp1	Feng (2008)	0.31
MRF-NRD-Exp2	Feng (2008)	0.34
MRF-Lplcn-rw	sc=7.4, λ_1=0.3	0.26
MRF-Lplcn-rw-ww'	sc=7.1,λ_1=0.9,γ=0.1,β=0.1	0.27
MRF-Lplcn-rww'	sc=7.1, λ_2=0.7	0.27
MRF-SqRt-rw-ww'	sc=9.6,λ_1=0.8,γ=0.1,β=0.9	0.29
MRF-SqRt-rw	sc=2, λ_1=0.3	0.32
MRF-SqRt-rw-ww'	sc=2.0,λ_1=0.3,γ=0.1,β=0.1	0.32
MRF-SqRt-rww'	sc=1.8, λ_2=0.3	0.32

5.5 Results and Discussion

We analyse the behaviour of three models obtained by combining the clique configurations shown in Figure 5.1 for the two datasets. In particular, we join the image-to-word with the word-to-word model and investigate whether its performance is higher than the image-to-word and the word-to-word-to-image separately. We also explore the effect of using different kernels in the non-parametric density estimation. Finally, we study which combination of parameters achieves the best performance. The parameters under consideration depend on the selected language model.

The name assigned to each of our models is made up of three parts. The first

Table 5.3: Top 20 best performing words in Corel 5k dataset ordered according to the columns

Word	Word
land	runway
flight	tails
crafts	festival
sails	relief
albatross	lizard
white-tailed	mule
mosque	sphinx
whales	man
outside	formula
calf	oahu

refers to the fact that it is a MRF model. The second applies to the kind of kernel considered: "Lplcn" for Laplacian, "Gssn" for Gaussian, and "SqRt" for the "square-root" kernel. The third part corresponds to the language model used: [-rw] refers to the image-to-word model, [-ww'] to the word-to-word model, and [-rww'] to the word-to-word-to-image model.

Our top results are represented in Table 5.2 for the Corel 5k dataset, and in Table 5.4 for the ImageCLEF09 collection. In Table 5.2, we have included other state-of-the-art algorithms for comparison purposes.

The "square root" kernel provides the best results in any configuration modelled for the two datasets. These results are followed by the Laplacian kernel whereas the Gaussian produces the lowest performance.

For the Corel 5k dataset, the best result corresponds to the word-to-word-to-image configuration, with a MAP of 0.32, closely followed by the image-to-word model, and by the combined image-to-word and word-to-word configuration. The kernel used in the three cases corresponds to the "square root". This result outperforms previous probabilistic methods, with the exception of the continuous MRF-NRD-Exp2 model

Table 5.4: Top performing results for the ImageCLEF09 dataset expressed in terms of mean average precision using 53 words as queries

Model	Parameters	MAP
MRF-Gssn-rw	sc=4.2, λ_1=0.1	0.2981
MRF-Gssn-rw-ww'	sc=2.5,λ_1=0.2,γ=0.1,β=0.1	0.3027
MRF-Lplcn-rw-ww'	sc=4.2,λ_1=0.4,γ=0.1,β=0.1	0.3195
MRF-Lplcn-rw	sc=4.2, λ_1=0.1	0.3197
MRF-SqRt-rww'	sc=4.6,λ_2=0.001	0.3205
MRF-SqRt-rw-ww'	sc=5.3,λ_1=0.3,γ=0.1,β=0.1	0.3217
MRF-SqRt-rw	sc=5.9, λ_1=0.1	0.3220

of Feng (2008), and the JEC system proposed by Makadia et al. (2008). It is worth mentioning that the good results obtained by Makadia et al. are due to their careful use of visual features. Note that the top 20 best performing words, which are represented in Table 5.3, have an average precision value of one. This means that the system is able to annotate these words perfectly.

For the ImageCLEF09 collection, the best performance is achieved by the image-to-word configuration. We consider that this behaviour is very revealing as the correlation between concepts is very rare in the collection, because of the nature of its vocabulary. Thus, as the correlation between words does not provide any added value to the model, the best performing is the image-to-word model, which detects concepts only based on low-level features. The corresponding MAP is of 0.32, which translated into the evaluation measures followed by ImageCLEF competition yields EER of 0.31 and AUC of 0.74. After comparing our results with the rest of the algorithms submitted to the competition, we are located in the position 21 (out of 74 algorithms). Again, best results were obtained using a "square root" kernel.

Finally, we represent in Table 5.5, the top ten best performing words for the image-to-word model. Not surprisingly, the best performing words correspond to visual con-

Table 5.5: Average Precision per Word for the top ten best performing words in ImageCLEF09

Word	Avg. Precision
Neutral_Illumination	0.97
No_Visual_Season	0.94
No_Blur	0.86
No_Persons	0.81
Sky	0.77
Outdoor	0.76
Day	0.74
No_Visual_Time	0.74
Clouds	0.61
Landscape_Nature	0.60

cepts, while the worst performing correspond to the most subjective concepts.

5.6 Conclusions

We have demonstrated that Markov Random Fields provide a convenient framework for exploiting the semantic context dependencies of an image. In particular, we have formulated the problem of modelling image annotation as that of direct image retrieval. The novelty of our approach lies in the use of different kernels in our non-parametric density estimation together with the utilisation of configurations that explore semantic relationships among concepts at the same time as low-level features, instead of just focusing on correlation between image features like in previous formulations.

Experiments have been conducted on two datasets, the usual benchmark Corel 5k and the collection proposed by the 2009 edition of the ImageCLEF campaign.

Our performance is comparable to previous state-of-the-art algorithms for both datasets. We observed that the kernel estimation has a significant influence on the performance of our model. In particular, the "square root" kernel provides the best performance for both collections. With respect to the language model, the best result

corresponds to the configuration that exploits dependencies between words at the same time as dependencies between words and visual features. This makes sense as it is the configuration that makes use of the maximum amount of information from the image. However, the ImageCLEF achieves the best performance with the word-to-image configuration although closely followed by word-to-word-to-image model. We consider that this behaviour is very revealing as the correlation between concepts is very rare in the collection, as a result of the nature of its vocabulary. Thus, as the correlation between words does not provide any added value to the model, the best performing is the image-to-word model, which detects concepts only based on low-level features.

As for future work, we intend to consider other kernels to see whether we can improve our results even more. Additionally, we will study whether a better choice of features as in Makadia et al. (2008) might improve our performance.

Chapter 6

The Effect of Semantic Relatedness Measures on Multi-label Classification Evaluation

In this chapter, we[1] explore different ways of formulating new evaluation measures for multi-label image classification when the vocabulary of the collection adopts the hierarchical structure of an ontology. Automated image annotation constitutes a practical application of multi-label image classification.

We apply several semantic relatedness measures based on web-search engines, WordNet, Wikipedia, and Flickr to the ontology-based score (OS) proposed by Nowak and Lukashevich (2009). The novelty of this measure, as seen in Section 3.3, lies in being the only measure formulated for working with ontologies. In particular, the OS uses ontology information to detect violations against real-world knowledge in the annotations

[1] In this chapter, "we" refers to the joint work done with Stefanie Nowak, of Fraunhofer, Germany.

and calculates costs between misclassified labels.

The final objective of this chapter is to assess the benefit of integrating semantic distances to the OS measure. Hence, we have evaluated them in a real case scenario: the results (73 runs) provided by 19 research teams during their participation in the ImageCLEF 2009 Photo Annotation Task (see Section 3.7.5).

Two experiments were conducted with a view to understand what aspect of the annotation behaviour is more effectively captured by each measure. First, we establish a comparison of system rankings brought about by different evaluation measures. This is done by computing both the Kendall and the Kolmogorov-Smirnov correlation coefficient between the ranking of pairs of them. Second, we investigate how stable the different measures react to artificially introduced noise in the ground-truth. The *example-based* F-measure is utilised as baseline to compare the results of the experiments. We conclude that the distributional measures based on image information sources show a promising behaviour in terms of ranking and stability.

The rest of the chapter is organised as follows. Section 6.1 introduces the OS measure. Section 6.2 lists techniques to estimate the semantic relatedness between concepts. The evaluation framework is introduced in Section 6.3, and Section 6.4 explains the experimental setup. The results of the experiments are analysed and discussed in Section 6.5. Finally, Section 6.6 draws our conclusions.

6.1 Ontology-based Score (OS)

This section explains the definition of the ontology-based score (OS) by Nowak et al. (2010b), which serves as the basis for the experiments with the semantic relatedness measures. The OS is an *example-based* evaluation measure for the evaluation of multi-label annotations.

This evaluation measure belongs to the group of *example-based* evaluation measures. Consequently, it is able to assess partial matches between the predicted set of labels and the ground-truth and to provide an evaluation score per image. Nowak et al. (2010b) identified three requirements that an *example-based* multi-label evaluation measure should fulfil. The first two are a direct consequence of the concepts being part of an ontology while the third one is related to the subjectivity process of providing ground-truth.

First, the evaluation measure should return an appropriate score when a related label to the ground-truth was mistakenly assigned. This requirement is addressed by incorporating the *depth-dependent distance-based misclassification costs* (DDMC), a hierarchical measure inherited from the single-label classification world (Freitas and de Carvalho 2007). This hierarchical measure computes the shortest path in the hierarchy between the predicted and the ground-truth label by counting the number of edges between them and assigning different costs depending on the depth of the link in the hierarchy.

Second, the relationships between concepts in an ontology is taken into account by this evaluation measure. Thus, when two predicted labels are disjoint to each other or when pre-conditions for relationships are ignored, the annotation system is penalised accordingly.

Third, the manual process of assigning labels to an image is highly subjective. Thus, the degree of agreement among annotators for each concept is computed over a reference set of images and serves as a weighting factor for the costs for each misclassified label. Consequently, the more subjective concepts are weighted less than the objective ones in case of misclassification.

The crucial point in *example-based* multi-label evaluation is the way the predicted

label set $\mathcal{P}$ is mapped to the ground-truth label set $\mathcal{G}$. In most cases these sets are partly consistent. The OS defines a matching procedure which calculates costs between the sets $\mathcal{P}$ and $\mathcal{G}$ for each example X. First, the false positive labels $\mathcal{P}' = \mathcal{P} \setminus (\mathcal{P} \cap \mathcal{G})$ and the missed labels $\mathcal{G}' = \mathcal{G} \setminus (\mathcal{P} \cap \mathcal{G})$ are computed, as only a matching between these labels is necessary. If $\mathcal{P} = \emptyset$, the matching costs for all labels of $\mathcal{G}' = \mathcal{G}$ are set to the maximum. A crosscheck on the predicted label set $\mathcal{P}$ is performed. If labels in $\mathcal{P}$ violate relationships from the ontology, these labels get the maximum costs of one as penalty assigned and are removed from $\mathcal{P}'$, $\mathcal{G}$ and $\mathcal{G}'$ if contained. This ensures that the measure does not assign costs twice. Then for each label l_i from $\mathcal{P}'$ a match to a label l_j from $\mathcal{G}$ is calculated and for each label l_m from $\mathcal{G}'$ a mapping to a label l_n from $\mathcal{P}$ is performed in an optimization procedure that determines the lowest costs between two labels:

$$\begin{aligned} \text{match}(\mathcal{P}, \mathcal{G}) = &\sum_{l_i \in \mathcal{P}'} \left((\min_{l_j \in \mathcal{G}} \text{cost}(l_i, l_j)) \cdot a({l_j}^*) \right) \\ &+ \sum_{l_m \in \mathcal{G}'} \left((\min_{l_n \in \mathcal{P}} \text{cost}(l_n, l_m)) \cdot a(l_m) \right), \end{aligned} \tag{6.1}$$

with ${l_j}^* = \text{argmin}_{l_j \in \mathcal{G}}(\text{cost}(l_i, l_j))$ and $a(l)$ as annotator agreement factor (see Sec. 6.3.3).

The function $\text{cost}(l_i, l_j)$ depends on the shortest path in the hierarchy between two mapped labels l_i and l_j in the original proposed OS measure. Each link in the hierarchy is associated with a cost that is cut in halves for each deeper level of the tree and that is at most one for a path between two leaf nodes of the deepest level. The costs for a link at level l of the hierarchy are calculated as follows:

$$\text{cost link}_l = \frac{2^{(l-1)}}{2^{(L+1)} - 2}, \tag{6.2}$$

L being the number of links from the leaf node to the root. Finally, the costs $\text{cost}(l_i, l_j)$ are calculated by summing up all link costs at the shortest path between these concepts.

The overall score for the OS is then determined as follows:

$$\text{score}(X) = 1 - \frac{\text{match}(\mathcal{P}, \mathcal{G})}{|\mathcal{P} \cup \mathcal{G}|}. \tag{6.3}$$

where X is the multimedia document to evaluate, $\mathcal{P}$ the set of labels predicted by the system and $\mathcal{G}$ the ground-truth. Thus, the score is one if all the predicted labels are correct and goes to zero if no concept was found.

6.2 Semantic Relatedness Measures

This section introduces the various semantic relatedness measures employed. For a more detailed description, one may refer to Chapter 2. In this chapter, the measures are grouped into thesaurus-based, document-based, and image-based information sources, according to the information source used.

6.2.1 Thesaurus-based Relatedness Measures

Thesaurus-based methods rely on a hierarchical representation of concepts and relations as nodes and links, respectively. A fair amount of thesaurus-based semantic relatedness measures were proposed and investigated on the WordNet hierarchy of nouns, see Budanitsky and Hirst (2006) for a detailed review. Specifically, we employ the following WordNet measures: WUP (Wu and Palmer 1994); LCH (Leacock and Chodorow 1998); PATH; RES (Resnik 1995); JCN (Jiang and Conrath 1997); LIN (Lin 1998); HSO (Hirst and St-Onge 1998); LESK (Banerjee and Pedersen 2003); and VEC (Patwardhan 2003).

Additionally, a semantic relatedness measure (WIKI) (Milne and Witten 2008) based on Wikipedia is used. In this case, the Wikipedia's hyperlink structure is considered.

6.2.2 Distributional Methods

More recently several semantic relatedness measures based on search engines have been proposed. Often they are referred to as *distributional methods*, as they define the semantic relatedness between two terms as their co-occurrence in similar contexts. In this work, we differentiate between *distributional methods* that rely on text documents to extract the knowledge and distributional methods that gain the knowledge from images and associated metadata. The former are called *methods relying on document-based information*, the latter ones are called *methods based on image resources*.

Document-based Relatedness Measures

These measures use the World Wide Web as a corpus for distributional semantic relatedness estimation. The correlation between terms is computed by crawling them with web-search engines and by weighting the results based on some distance criterion. In particular, we use the transformation (Equation 2.38) applied to the *normalized Google distance* (Cilibrasi and Vitanyi 2007) as proposed by Gracia and Mena (2008). In the following, we denote www_G as the measure that employs Google to find the correlation between terms, whereas www_Y is the measure that utilises the Yahoo web search engine.

Image-based Relatedness Measures

The distributional measures of the previous section rely on the number of hits in textual documents retrieved by web search engines. It is rather debatable whether these textual documents can represent the co-occurrence relationship of visual concepts adequately. Subsequent research investigates the utilisation of information from photo communities, such as Flickr for the definition of semantic relatedness between concepts.

Jiang et al. (2009) proposed the *Flickr context similarity* (FCS). They used the Flickr search functionality to search for concepts in image tags, descriptions, and comments and apply the same formula as Equation 2.38 to estimate a relatedness value. In their work, they utilised the FCS to automatically select video concept detectors from a pool of detectors to answer a user query. Additionally, they performed a small experiment between the *Flickr tag similarity* (FTS)[2] and the FCS and concluded that FCS has a better word coverage.

We incorporated the FTS and FCS relatedness measures as part of our experimental work. The number of photos on Flickr recently crossed the 4.3 billion threshold, which was used as N in our computation.

The *Flickr distance* (FD) was proposed by Wu et al. (2008) to quantify semantic relationships between concepts in the visual domain. For each concept, 1000 images are downloaded, visual features are extracted and a latent topic visual language model is computed. Finally, they defined the *Flickr distance* between two concepts as the average square root of the Jensen-Shannon divergence between the two latent topic visual language models associated to them. This method, although promising in revealing visual co-occurrence, is computationally expensive and relies on the visual features and the language model as additional parameters. It is unclear whether a study on evaluation could benefit from this model, as the visual features are already incorporated in the annotation process of the participants. For these reasons, it is not used in our experiments.

[2]Same as FCS but only searching for concepts on the image tags.

6.3 Evaluation Framework

The experiments are conducted in an evaluation framework that is schematically illustrated in Figure 6.1. In this framework, the OS evaluation measure is implemented as introduced in Section 6.1.

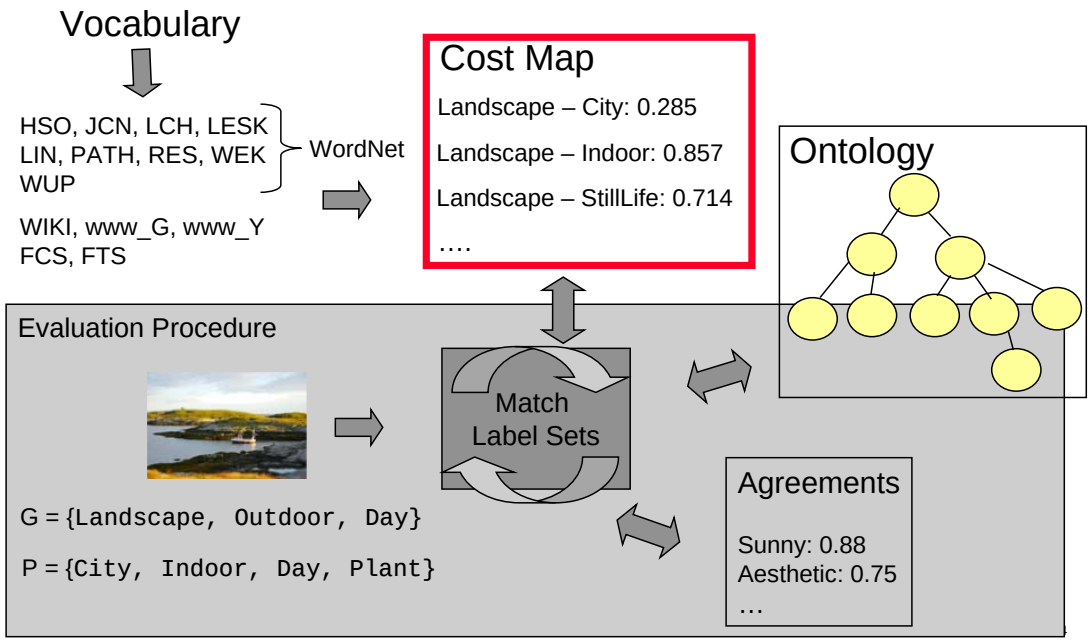

Figure 6.1: Schematic representation of the evaluation framework

The core of the evaluation framework is the *matching procedure*, that matches labels of the predicted set $\mathcal{P}$ to the ground-truth set $\mathcal{G}$ and vice versa according to Equation 6.1. The matching procedure takes as pluggable information resources a *costmap*, an *ontology*, and an *agreement map* into account. These information resources are further denoted as *plug-ins* and explained in detail in the following subsections. The predicted labelset $\mathcal{P}$ and the ground-truth labelset $\mathcal{G}$ serve as input for each image and the framework outputs a score that describes the annotation quality by applying Equation 6.3.

	www_G	HSO	JCN	LCH	LESK	LIN	HS/OS	PATH	RES	VEC	WIKI	WUP	www_Y	FCS	FTS	F
www_G		.908	.859	.925	.859	.931	.860	.879	.941	.904	.906	.943	.933	.910	.893	.615
HSO	.855		.946	.965	.925	.952	.826	.964	.953	.973	.958	.898	.935	.931	.918	.674
JCN	.813	.910		.917	.932	.904	.775	.976	.902	.929	.930	.847	.893	.893	.889	.690
LCH	.889	.956	.890		.901	.959	.831	.941	.959	.963	.966	.927	.958	.932	.919	.648
LESK	.933	.869	.832	.886		.899	.803	.922	.899	.913	.914	.844	.900	.922	.930	.746
LIN	.882	.906	.836	.935	.877		.858	.923	.987	.963	.954	.941	.944	.935	.916	.645
HS/OS	.662	.610	.539	.629	.648	.685		.796	.859	.841	.831	.887	.853	.864	.850	.599
PATH	.835	.937	.968	.922	.850	.865	.565		.922	.950	.950	.870	.915	.911	.902	.675
RES	.885	.903	.835	.930	.881	.975	.673	.859		.956	.952	.944	.945	.942	.922	.644
VEC	.884	.930	.866	.957	.880	.936	.651	.891	.921		.959	.908	.931	.927	.917	.661
WIKI	.876	.882	.859	.913	.873	.893	.639	.877	.876	.905		.910	.940	.930	.920	.664
WUP	.856	.828	.757	.864	.827	.913	.721	.787	.913	.870	.824		.922	.904	.887	.599
www_Y	.902	.868	.809	.881	.910	.873	.699	.835	.870	.881	.859	.833		.934	.922	.650
FCS	.929	.839	.779	.860	.934	.876	.698	.801	.885	.861	.847	.859	.916		.978	.688
FTS	.916	.819	.767	.839	.922	.850	.708	.786	.857	.845	.836	.833	.910	.969		.706
F	.921	.867	.847	.883	.975	.870	.629	.859	.874	.873	.861	.817	.899	.918	.906	

Table 6.1: Kendall τ correlation coefficient between ranking of runs evaluated with different semantic relatedness evaluation measures. Upper triangle shows results for the complete measures while the lower depicts results for the costmap measures. As baseline for comparison, F-measure (F) is illustrated in light grey. Cells in gray illustrate the combinations where the Kolmogorov-Smirnov test showed concordance in the rankings

6.3.1 Plug-in: Costmap

The costmap plug-in is the most important part of the evaluation framework for our study. It describes the costs between pairs of concepts in case of misclassification and can be determined in various ways. Originally, the OS used as costmap the *depth-dependent distance-based misclassification costs* (DDMC) as explained in Section 6.1. As the experiments deal with a classification task, the vocabulary of concepts is fixed from the beginning and consists of 53 concepts in this study. The vocabulary is used to build a costmap for each pair of concepts. The costmap is represented as a *confusion matrix* and it is symmetric. The costs are defined in the range of $[0, 1]$, where one determines the highest cost and zero indicates no cost or equality of concepts. In our experiments, we investigate the 14 semantic relatedness measures introduced in Section 6.2 (www_G, HSO, JCN, LCH, LESK, LIN, PATH, RES, VEC, WUP, WIKI, www_Y, FCS and FTS) and turn them into a costmap. The semantic relatedness measures were normalised and the relatedness value is converted into a cost by subtracting it from one. These semantic costmaps are compared to the original proposed costmap of the OS, which realises the cost function of Equation 6.2, the annotator agreements, and the ontology knowledge. If the ontology and the annotator agreement factors are not used, the measure is called *hierarchical score* (HS).

6.3.2 Plug-in: Ontology

The ontology structures the concepts of the vocabulary in a hierarchical or graph-based form and defines relationships among them. If the ontology plug-in in the evaluation framework is activated, the labels in the predicted set $\mathcal{P}$ are first checked against violations of real-world knowledge. The relations in the ontology are used to verify the co-occurrence of labels for one image. For example, an image cannot be considered at

the same time to be *indoor* and *outdoor*. These concepts are defined as *disjoint* in the ontology and the maximum costs of one are assigned as penalty instead of calculating the minimal costs to a label of $\mathcal{G}$.

6.3.3 Plug-in: Annotator Agreements

The annotator agreement describes the consistency in annotation among several human judges on a small set of photos (Nowak and Dunker 2009a). In case the plug-in for annotator agreements is activated, the matching procedure takes into account the subjectivity in determining a ground-truth. The empirically determined inter-annotator agreement is a value in the range of $[0, 1]$ that serves as scaling factor in Eq. 6.1. It lowers the costs for misclassified concepts depending on the subjectivity of the concept. The greater the disagreement on a concept computed over a validation set with several annotators, the lower the factor.

6.4 Experimental Work

We conduct two experiments to assess the quality of the semantic relatedness multi-label evaluation measures. The ranking experiment investigates the correlation among result lists that were calculated with the different semantic relatedness measures for a number of annotation systems. The stability experiment analyses the influence of noise in the ground-truth on the ranked result lists. For this experiment the binary ground-truth annotations were randomly flipped from zero to one and vice versa for 1%, 2%, 5% and 10% of the set. The evaluation score is calculated by using the altered ground-truths and the correlation of the rankings is analysed. The overall goal is to determine which semantic relatedness measure displays the best characteristics for multi-label evaluation. Next, the data on which the experiments are based is introduced, followed

by the configuration of the evaluation measure and a brief introduction on methods for the analysis of rank correlations.

6.4.1 Data

The experiments are carried out on the results of the runs of the ImageCLEF 2009 Photo Annotation Task (Nowak and Dunker 2009b). In this task (see Section 3.7.5), 13,000 Flickr photos were annotated with 53 visual concepts by 19 research teams in 73 run configurations and one random run. The visual concepts were part of the Consumer Photo Tagging Ontology defined by Nowak and Dunker (2009a). Each run is an unordered list containing the IDs of 13,000 test images followed by the confidence score, which gives an indication of the probability of each concept being present in an image. Initially, the confidence score is a floating point number between zero and one, where higher numbers denote higher confidence. In agreement with the participants, the confidence values were mapped to binary values using a threshold of 0.5 for the evaluation measures that need a binary decision about the presence of concepts. The utilisation of the ImageCLEF runs allows for a comparison of the semantic relatedness measures in a realistic annotation scenario and offers diverse and numerous configurations and systems.

In the experiments, the 15 introduced costmaps including the OS are plugged into the evaluation framework as described in Section 6.3. The scores for the 13,000 test images per run are averaged and ordered in a ranked list for each costmap.

6.4.2 Configurations

In the experiments, two configurations of the evaluation measure are investigated. In the first configuration, each costmap is included in the evaluation procedure together with the ontology plug-in and the agreement plug-in. This configuration is further

denoted as *complete measure*, as all parts of the evaluation framework are used. The second configuration explores the characteristics of each costmap without the other plug-ins. This means that the matching procedure is utilised to find matching labels and the costmap determines the costs between these labels. In the following, this configuration is referred to as *costmap measure*. As baseline for comparison, the *example-based* F-measure (F) is used to rank the results. It showed convincing characteristics in example-based multi-label evaluation as its score is not major influenced by random annotations or the number of labels annotated per image (Nowak et al. 2010b).

6.4.3 Correlation Analysis

The correlation between two different measures can be estimated by computing the Kendall τ coefficient between the respective rankings. The Kendall τ rank correlation coefficient (Kendall 1938) is a non-parametric statistic used to measure the degree of correspondence between two rankings.

Two identical rankings produce a correlation of +1, the correlation between a ranking and its perfect inverse is -1 and the expected correlation of two rankings chosen randomly is 0. The Kendall τ statistic assumes as null hypothesis that the rankings are discordant and rejects the null hypothesis when τ is greater than the $1 - \alpha$ quantile, with α as significance level. Melucci (2007) illustrated that it is likely that the Kendall τ statistic rejects the null hypothesis and decides for concordance, for example if the sample size is large. In his work, he compared the τ statistic with the Kolmogorov-Smirnov D statistic. He recommended to use several test statistics to support or revise a decision.

The Kolmogorov-Smirnov's D (Kolmogorov 1986) states as null hypothesis that the two rankings are concordant. It is less affected by the sample size, is sensitive to the

extent of disorder (in contrast to τ, which takes the number of exchanges into account) and tends to decide for discordance for instance in the case of two radically different retrieval algorithms (Melucci 2007).

For theses reasons, both statistic tests are applied in the experiments.

6.5 Results and Discussion

In the ranking experiment, the correlation between pairs of rankings of the ImageCLEF runs is analysed. For each relatedness measure, the ImageCLEF runs are evaluated, ordered into a ranked list and then the correlation is calculated between each pair of lists by exploiting the introduced rank correlation statistics.

The second experiment analyses the stability of the different relatedness measures concerning noise in the ground-truth. After evaluating the ImageCLEF runs, the rank correlation is investigated for each result list in comparison to the ranking with correct ground-truth and in comparison to the ranking at the previous stage of noise.

6.5.1 Ranking Results

Table 6.1 shows the results for the ranking experiment. In the upper triangle, the correlations for the complete measures are depicted and the lower triangle presents the Kendall τ coefficient for pairs of rankings of the costmap measures only. The last row and the last column shows the correlations to F. The cells that are coloured in grey, demonstrate the pairs of measures for which the Kolmogorov-Smirnov test supported the Kendall τ decision for concordance.

For the rankings of the *complete measures*, the coefficient is very high with an average correlation for all pairs of 0.92. For all costmap measures the correlation to other costmap measures is lower with an average of 0.86. In contrast, the Kolmogorov-

Smirnov test supports the decision in just 39% for the complete measures and in 21% for the costmap measures. The Kolmogorov-Smirnov test decides for concordance with the ranking of the F-measure in 6 of 15 cases for the complete measures, although the correlation coefficient of Kendall test is low. In case of the costmap measures, a correlation is supported in 3 of 15 cases.

The ranked result lists change more seriously in case of applying different costmap measures.

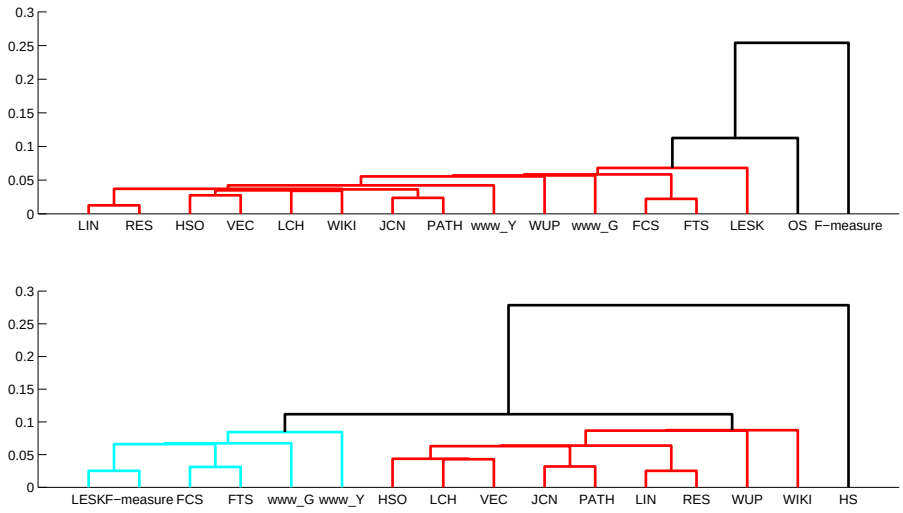

Figure 6.2: The upper dendrogram shows the results after hierarchical classification for the complete measures, the lower one for the costmap measures

Figure 6.2 visualizes the Kendall τ correlation coefficients in a dendrogram after applying a binary hierarchical clustering. A *dendrogram* is a tree visualisation in which each step of the hierarchical clustering is represented as a fusion of two branches into a single branch. The dendrogram shows a similar clustering for both configurations. In both cases, the highest correlation between any two measures can be found between RES and LIN. This leads to the conclusion that the differences between these measures are rather small and that the different way of scaling the information content between terms leads to an almost same ranking. In case of the costmap measures, HS has the lowest correlation to all other measures and falls into the outer cluster of the tree. For the

complete measures, the F-measure shows the lowest correlation. Considering that the F-measure does not take into account ontology, agreements and the matching procedure, which assigns fine-grained costs, this is not surprising. Interestingly the F-measure behaves very similar to LESK in case of the costmap measures only. The thesaurus-based measures behave quite similar, only LESK has a greater distance to the other ones and is clustered near the distributional methods. Also WIKI as the thesaurus-based measure with a different corpus stays close. For the distributional methods, FCS and FTS behave almost the same. In this experiment, the point of a better word coverage of FCS does not influence the results to a great extent. Summarizing, the dendrogram shows that the plug-ins of the framework while effecting the ranking more in case of the costmap measures, maintain the characteristics of the measures to each other with the exception of the F-measure.

6.5.2 Results of Stability Experiment

Table 6.2 illustrates the results for the stability experiment for the complete measures and the costmap measures. The table shows the ranking correlation for the complete measures to the original ranking, the costmap measures to the original ranking, the complete measures to the previous stage of introduced noise and the costmap measures compared to the previous stage of noise from left to right. Again, the numbers in each cell denote the Kendall τ correlation coefficient and the gray cells highlight the combinations in which the Kolmogorov-Smirnov test supported the Kendall τ decision for concordance. In the last row, the results for the F-measure are presented. Please note that the F-measure is not computed using the evaluation framework and matching procedures. Therefore, there are no different results for costmap or complete measures.

For all measures, the correlation coefficient decreases with increasing amount of

	complete to original				costmap to original				complete to previous				costmap to previous			
	1%	2%	5%	10%	1%	2%	5%	10%	1%	2%	5%	10%	1%	2%	5%	10%
www_G	.947	.954	.958	.931	.755	.746	.698	.604	.947	.989	.974	.954	.755	.973	.914	.852
HSO	.990	.984	.953	.896	.722	.713	.687	.590	.990	.993	.968	.943	.722	.976	.927	.840
JCN	.996	.986	.956	.927	.689	.682	.671	.624	.996	.990	.970	.971	.689	.980	.942	.907
LCH	.989	.980	.938	.882	.730	.730	.673	.528	.989	.991	.958	.944	.730	.975	.891	.790
LESK	.991	.984	.955	.893	.730	.721	.667	.567	.991	.992	.970	.938	.730	.979	.900	.839
LIN	.983	.965	.919	.828	.739	.718	.641	.452	.983	.982	.954	.909	.739	.959	.859	.738
OS/HS	.986	.968	.910	.829	.712	.682	.582	.379	.986	.982	.941	.919	.712	.942	.826	.753
PATH	.993	.978	.956	.906	.698	.698	.675	.571	.993	.985	.978	.950	.698	.975	.939	.834
RES	.981	.961	.910	.831	.749	.726	.638	.480	.981	.981	.949	.921	.749	.958	.856	.763
VEC	.987	.975	.930	.853	.712	.704	.629	.478	.987	.988	.956	.923	.712	.968	.869	.761
WIKI	.987	.973	.927	.836	.728	.709	.653	.438	.987	.986	.953	.910	.728	.962	.876	.705
WUP	.985	.964	.910	.834	.756	.721	.624	.412	.985	.979	.945	.924	.756	.946	.841	.743
www_Y	.992	.981	.958	.912	.753	.747	.708	.610	.992	.989	.977	.954	.753	.976	.925	.855
FCS	.989	.974	.944	.893	.959	.933	.826	.644	.989	.985	.970	.949	.959	.974	.893	.819
FTS	.992	.974	.942	.876	.959	.921	.811	.621	.992	.982	.967	.935	.959	.962	.890	.810
F	.978	.954	.893	.773	-	-	-	-	.978	.976	.939	.879	-	-	-	-

Table 6.2: Kendall τ correlations for the complete and the costmap measures between the original ranking and the ranking with altered ground-truths are shown on the left. On the right, the correlations are shown when compared between the rankings of two noise stages. Cells in gray illustrate the combinations with concordance in the rankings according to the Kolmogorov-Smirnov test

noise. As it can be seen from the results of the ranking experiment, the Kendall τ coefficient is not very sensitive. It supports again a decision on correlation for every pair of rankings, although the correlation coefficient decreases to 0.38 at minimum. In the results of the Kolmogorov-Smirnov test, it is obvious that www_G changes the order of systems significantly by just introducing 1% noise for the complete measures compared to the original ranking. The OS measure shows a good stability as it keeps a concordant ranking until 10% of noise are introduced. All other complete measures remain stable in ranking until more than 2% of noise are included in the ground-truth. When the ranking of the complete measures is compared to the previous stage of noise, the OS and WUP remain stable over the four stages. www_G drops to discordance at the first stage, but is then concordant between the first and the second stage. The Kendall τ correlation coefficient is very high in this scenario, with over 0.9 correlation between the different stages. The costmap measures all behave the same when compared to the original ranking. The Kolmogorov-Smirnov test assigns concordance as long as not more than 2% noise are incorporated in the ground-truth. But the Kendall test shows at the same time, that the correlation coefficient varies significantly between the different measures at the stage of 10% noise. In case the costmap measures are compared to the previous stage of noise, the HS is again concordant. All other measures are stable in their ranking until more than 2% noise are incorporated. It is obvious that the correlation coefficient drops for all measure in the stage of 1% compared to the original except for FCS and FTS, but then rises again in the comparison between the other stages. The F-measure acts similar to most of the other measures by tolerating 2% noise without a major influence on the ranking, but with a drop in correlation with greater amount of noise.

In the following, we analyse an example for the ranking of www_G and www_Y in

the configuration of the complete measure after 1% noise was introduced. The Kendall test assigns a correlation of 0.947 and 0.992, respectively, but the Kolmogorov-Smirnov test just decides for concordance in case of www_Y. Having a look at the first 20 ranks of the system ranking after introducing 1% noise, for www_G the order changes to (1, 2, 3, 7, 9, 4, 5, 12, 6, 10, 14, 8, 11, 15, 13, 17, 18, 16, 19, 32). In contrast, the order of the first 20 ranks of www_Y is permuted to (1, 2, 3, 4, 5, 8, 6, 7, 9, 10, 11, 12, 15, 13, 16, 14, 17, 19, 18, 20). One can see from these sequences of numbers that in case of www_G the numbers are exchanged to a greater extent and swapped with more distant ranks as in case of www_Y. Summarizing the stability experiment, the measures are stable in their ranking for nearly all configurations until more than 2% of noise are introduced. www_G acts unstable from the beginning. The OS shows a longer stability in the ranking than the other measures. It has to be investigated whether it is sensitive enough in its ranking to cope with noise.

6.6 Conclusions

In this chapter, we studied the behaviour of semantic relatedness measures for the evaluation of multi-label image classification when the vocabulary of the collection adopts the hierarchal structure of an ontology. The 15 semantic relatedness measures are based on WordNet, Wikipedia, Flickr, or on the WWW and were compared to the *example-based* F-measure in two experiments, the ranking and the stability experiment.

The ranking experiment showed a correlation for the thesaurus-based measures HSO, JCN, PATH, and VEC and the image-based distributional measures FCS and FTS, in comparison to the baseline measure for the *complete measures*. In case of the *costmap measures*, the correlation only could be assigned to HSO, JCN, and PATH. The evaluation framework with all plug-ins, therefore, seemed to push the relatedness

measures closer to the baseline as the ontology plug-in incorporates penalties in case of violations. These penalties assigned the maximum costs; the same values that the F-measure assigned to incorrect classified labels.

Regarding the stability experiment, the above mentioned measures performed rather well and at least 2% noise could be incorporated without changing the order of systems significantly. The distributional *document-based* method www_G could not convince in its results, as it reacts unstable to a small amount of noise and has no confirmed correlation of both tests in the ranking experiment to the baseline or related measures.The OS showed the longest stability and therefore tends to be not sensitive enough for changes in annotations.

As final recommendation, we propose the utilisation of the FCS relatedness measure in the configuration of a *complete measure*. It behaves very similar to FTS, but although in our experiments no problem occurred with the word coverage, this can change for other concepts (Jiang et al. 2009). Even though the mentioned WordNet based measures showed promising results, the use of these measures presents some limitations. First, the vocabulary had to be adapted as not all words of the vocabulary were present in WordNet. Second, a prior disambiguation task is needed to find the sense for a given term.

A limitation of this work is the comparison of the relatedness measure ranking characteristics, which consider fine-grained costs between predicted and ground-truth annotations, using the F-measure as baseline that utilises binary scores.

As outcome of the experimental work presented in Section 6.4, the FCS relatedness measure was used as part of the OS [3] to evaluate the performance of the annotation algorithms submitted by the research teams participating in the 2010 edition of

[3] In what follows, it will be denoted as FCS-OS.

ImageCLEF (Nowak and Huiskes 2010)[4]. In total, 17 groups from 11 countries participated with 63 runs. The goal was to annotate 10,000 Flickr images, a subset of the MIR Flickr collection (Huiskes and Lew 2008), using 93 annotation words that were part of an ontology. Participants were offered three different configurations: *textual information* that consisted on EXIF tags and Flickr user tags; *visual information* that comprised the training set and their annotations; and the *multi-modal information* that was a combination of the previous two. Participants were to evaluate their results using three evaluation measures: MAP, F-measure, and the FCS-OS. With respect to the best configuration, the conclusion obtained was that the *multi-modal* always outperformed *visual* or *textual* configurations for teams that submitted runs in several configurations. In addition to that, the FCS-OS evaluation measure was more consistent with the values obtained by MAP and F-measure in the case of the *multi-modal* or *textual* configuration. As a final observation, significant differences were found among participants when using the OS-FCS measure.

[4]For a whole description of the evaluation conference, see Section 3.5.4.

Chapter 7

Conclusions and Discussion

This thesis has explored efficient ways to bridge the semantic gap in automated image annotation. Specifically, I have exploited the semantic relationships between words combined with statistical models based on the correlation between words and visual features to increase the effectiveness of probabilistic automated image annotation systems. To achieve this goal the following research questions have been investigated:

- *(i) How to successfully undertake the initial annotation of a scene?*
- *(ii) How to model semantic knowledge in an image collection?*
- *(iii) How to integrate semantic knowledge into the annotation process?*

This thesis can be divided into two main parts. The first part (Chapter 1–3) introduced the problem, highlighted related work, and the methodology. The second part (Chapter 4–6) presented my experimental work.

In particular, Chapter 1 introduced the field of automated image annotation by revising several *classic probabilistic approaches*. Additionally, it discussed some limitations of these approaches that helped to establish a set of requirements that any efficient automated annotation algorithm should fulfil. These limitations were mainly

due to the *semantic gap* (Santini and Jain 1998) existing between the low and the high level features of an image. Chapter 2 presented an extensive overview of a new generation of *semantic-enhanced models* that attempt to address this issue. However, the overall performance of these approaches is not always stable owing largely to error propagation between the different parts of the model and to over-fitting when there is no sufficient data. Chapter 3 presented the methodology, i.e. the evaluation metrics and benchmark datasets adopted in the experimental phase of this research

With respect to the experimental work proposed in this thesis, each chapter focused on different aspects of the problem. Chapter 4 proposed a novel automated image annotation application that exploited the semantics of the collection through the utilisation of a combination of several semantic relatedness measures. Chapter 5 presented a state-of-the-art annotation algorithm, based on Markov Random Fields, which constitutes a good example of the *fully semantic integrated models* introduced in this thesis. Finally, Chapter 6, whose emphasis was placed on formulating new evaluation measures for automated image annotation, proposed a novel measure that has been successfully used to evaluate the annotation algorithms submitted to the ImageCLEF 2010 competition (Nowak and Huiskes 2010).

7.1 Achievements and Conclusions

This section details my major achievements and describes my conclusions. The discussion is organised around the three research questions formulated in Section 1.5.

- *(i) How to successfully undertake the initial annotation of the scene?*

  I provided two different algorithms that effectively undertook the initial identification of the objects contained in the scene (see Chapters 4 and 5). Both of

them considered the analysis of visual features as a fundamental part of their approach. In both cases, low-level features were combined with semantics in order to perform the identification of objects.

- *(ii) How to model semantic knowledge in an image collection?*

  I modelled the semantic knowledge according to two different ways. The first one was based on the utilisation of *semantic relatedness measures*, which provides an indication of the closeness between two words with respect to their meaning. Chapter 2 provided a comprehensive analysis of the sheer number of semantic measures proposed in the literature, with a special focus on those applied or applicable to enhance automated image annotation algorithms. These measures make use of various sources of knowledge, some internal and others external to the collection, which were explored more in detail in the aforementioned chapter. In particular, the best performance in automated image annotation applications was achieved with an adequate combination of internal and external measures as shown in Chapter 4.

  The second way of modelling semantic knowledge is related to the construction of a graph, sometimes as a part of a language model, other times as a combination of low and high level image features. For the first case, there exist several examples, which were explored in Chapter 2, such as Srikanth et al. (2005), Li and Sun (2006), Shi et al. (2006), Shi et al. (2007), and Fan et al. (2007). All of them use WordNet as a way to structure the hierarchy of concepts embedded in the language model. Alternatively, Markov Random Fields provide a very effective way of modelling semantic knowledge representing in a graph the image features and, at the same time, the annotation words.

- *(iii) How to integrate semantic knowledge into the annotation framework?*

  I have considered two different ways of integrating semantic knowledge into the annotation framework, one focuses on the *annotation process*, while the second focuses on the *evaluation stage*.

  With respect to the *annotation process*, this thesis has shown two different ways of accomplishing the semantic integration. Without any doubt, the final performance of the annotation method depends heavily on the selected method. Methods based on the *semantic-enhanced models* analysed in Chapter 2 perform two steps, one that deals with the initial identification of objects and a second that refines the results by pruning the non-related words or that incorporates a language model. In both cases, this is achieved by using a *conceptual fusion strategy* that takes into account the semantic dependencies between annotation words. However, errors may propagate from one stage to another reducing the overall performance of the combined approach. Another way of incorporating semantics into the annotation process corresponds to methods, such as the *fully semantic integrated models* that rely on simultaneously detecting the objects and modelling the correlations between them in one single step. This approach has several advantages compared to approaches based on the previous group, especially as their performance does not suffer from the propagation of errors between stages, as it is the case in *semantic-enhanced models*. An example of these approaches are Markov Random Fields as presented in Chapter 5. Results confirmed that these methods achieve higher performance than those based on *semantic-enhanced models*.

  Regarding the integration of semantic knowledge in the *evaluation stage*, this thesis investigated the effect of incorporating some semantic measures analysed

in Chapter 2 into the evaluation measure, the *ontology-based score* (OS) proposed by Nowak et al. (2010b). Some experiments were conducted to evaluate the stability of these measures, when noise was introduced, and their behaviour, when compared to baseline measures. The conclusion was that the best measures were the distributional ones based on Flickr.

7.1.1 Achievements

The major achievements of this thesis are: First, the Markov Random Field model (Chapter 5) that combined the detection of word correlation and image visual features in one single step. This model achieved, for the Corel 5k dataset, a *mean average precision* of 0.32, which compares favourably with the highest value ever obtained, 0.35, obtained by Makadia et al. (2008) albeit with different features. For the more realistic dataset used by the 2009 ImageCLEF competition, we were located in the position 21 out of 74 algorithms, with a *mean average precision* of 0.32. Furthermore, the strongest point of the model, to handle the detection in one single step, has shown several advantages compared to other approaches. First, it follows the principle of the *least commitment* as the learning and the optimisation is done in one single step for all the concepts. As a result, propagation of errors does not occur. Finally, the risk of *over-fitting* is significantly reduced as the entire samples are used efficiently in modelling the concepts and, at the same time, their occurrences.

The second major achievement corresponds to the *semantic-enhanced model* of Chapter 4. The achieved performance was comparable to state-of-the-art automated image annotation applications.

Third, the contribution to the research on evaluation measures. In particular, we defined a new evaluation measure, which can be used in annotation applications where

the vocabulary adopts the form of an ontology.

Thus, some experiments were conducted in order to understand what aspect of the annotation behaviour was more effectively captured by each measure. Finally, it concluded with the proposition of *distributional measures based on image information sources* such as Flickr, as they showed promising behaviour in terms of ranking and stability.

Other collateral achievements of this thesis are:

- The classification schema adopted for automated image annotation algorithms: *classic probabilistic models*, *semantic-enhanced models*, and *fully semantic integrated models*.

- The analysis of the limitations of *classic probabilistic models*. The study was conducted in Chapter 1 and it helped with the identification of gaps. Specifically, it identified that these models are likely to have limited success as a result of the semantic gap that exists between the low-level and high-level features of the image.

- The comprehensive review undertaken in Chapter 3 that discussed the evolution of the evaluation metrics and the benchmark datasets that are now adopted in the field.

Finally, this thesis successfully proves that the exploitation of the semantics between words combined with statistical models based on the correlation between words and visual features increases significantly the effectiveness of probabilistic automated image annotation systems.

7.2 Future Lines of Work

The work generated by this thesis opens up a series of interesting research paths. These paths are summarised as follows as they can define successful lines of research for the future:

7.2.1 Feature Selection using Global Features

As seen in Table 3.2, the best performing annotation algorithms correspond to models that make use of global features. In particular, the highest performing application was developed by Makadia et al. (2008). Such good performance was achieved by using a simple k-nearest neighbour algorithm combined with an effective and adequate selection of global features. Being aware of the bias of the Corel 5k set, they confirmed their good results with two additional and more realistic datasets.

Clearly, there is still a lot of work to be done with respect to feature selection by using global features and without any doubt, any advance done in this field will benefit the rest of automated image annotation applications.

7.2.2 Semantic Web applied to Automated Image Annotation

As mentioned in previous section, this thesis has only considered two types of semantic modelling. However, a comprehensive application of Semantic Web technologies might further improve the performance of these approaches. Until now, there have been some attempts to use concept hierarchies in automated image annotation such as Marszałek and Schmid (2007), Fan et al. (2008), and Srikanth et al. (2005).

To model the relationships between objects depicted in an image, the work of Biederman (1981) should be considered. Thus, he described the rules behind the human understanding of a scene. In this work, Biederman showed that perception and com-

prehension of a scene requires not only the identification of all the objects comprising it but also the specification of the relations among these entities. These relations are what mark the difference between a well-formed scene and an array of unrelated objects. Biederman introduced the notion of a schema, which is an overall representation of a picture that integrates the objects and relations and allows access to world knowledge about such settings. Thus, to comprehend a scene with a "boat", "water" and "waves" requires not only that these entities are identified but also to know that the boat is in the water and the water has got waves.

Consequently, a possible source of relationships between objects can be provided by the Semantic Web.

7.2.3 Combination of Low and High Level Features

New methodologies that combine successfully semantics and statistics are needed. Until now, the most effective technology is Markov Random Fields. Several configurations have been explored in this research (Chapter 5) but there is a need for additional work in the area. Additionally, the consideration of other kernels in the non-parametric density estimation may result in better results. Finally, a better choice of features as in Makadia et al. (2008) might undoubtedly improve the final performance of the model.

The impact of my work on the field is focused on stressing the benefits that the use of semantics between annotation words can bring to the image annotation problem as seen in Section 7.1.1. Additionally, the description of future lines of work, such as the identification that an adequate choice of global features can push the performance of annotation algorithms, may be useful to the research community.

As stated in Section 1.2, the number of images is increasing dramatically on the

web. The actual state-of-the-art in automated image annotation allows collaborative photo sharing applications to benefit from them as they can provide any given image with a preliminary set of annotations that could be afterwards refined by the user. However, automated image annotation applications are not mature enough to be fully operative in the commercial domain. There is a lot of research still to be done in terms of computational expensiveness, improvement the effectiveness of applications, and feasibility of working with huge amounts of data. As a result, automated image annotation techniques are likely to be more important in future information systems.

Bibliography

Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., and Soroa, A. (2009). A Study on Similarity and Relatedness Using Distributional and WordNet-based Approaches. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Ames, M. and Naaman, M. (2007). Why we tag: motivations for annotation in mobile and online media. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 971–980.

Aslam, J. A. and Montague, M. (2001). Models for metasearch. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 276–284.

Athanasakos, K., Stathopoulos, V., and Jose, J. (2010). A Framework For Evaluating Automatic Image Annotation Algorithms. In *Proceedings of the 32nd European Conference on Information Retrieval*, vol. 5993, pp. 217–228, LNCS.

Baeza-Yates, R. and Ribeiro-Neto, B. (1999). *Modern Information Retrieval*. Addison Wesley.

Banerjee, S. and Pedersen, T. (2003). Extended Gloss Overlaps as a Measure of Semantic Relatedness. In *Proceedings of the Eighteenth International Confer-*

ence on Artificial Intelligence.

Barnard, K., Duygulu, P., and Forsyth, D. (2001). Clustering Art. In *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 434–441.

Barnard, K., Duygulu, P., and Forsyth, D. (2002). Modeling the Statistics of Image Features and Associated Text. In *Document Recognition and Retrieval IX, SPIE Electronic Imaging Conference.*

Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D., and Jordan, M. I. (2003). Matching Words and Pictures. *Journal of Machine Learning Research*, 3, 1107–1135.

Barnard, K. and Forsyth, D. (2001). Learning the Semantics of Words and Pictures. In *Proceedings of the International Conference on Computer Vision*, vol. 2, pp. 408–415.

Bartell, B. T., Cottrell, G. W., and Belew, R. K. (1994). Automatic combination of multiple ranked retrieval systems. In *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 173–181.

Biederman, I. (1981). On the semantics of a glance at a scene. In *Perceptual organization*, Erlbaum.

Blei, D. M. and Jordan, M. I. (2003). Modeling annotated data. In *Proceedings of International ACM Conference on Research and Development in Information Retrieval*, pp. 127–134.

Bollegala, D., Matsuo, Y., and Ishizuka, M. (2007). Measuring semantic similarity between words using web search engines. In *Proceedings of the 16th International Conference on World Wide Web*, pp. 757–766.

Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1), 13–47.

Carbonetto, P., de Freitas, N., and Barnard, K. (2004). A Statistical Model for General Contextual Object Recognition. In *Proceedings of the 8th European Conference on Computer Vision*, vol. 1, pp. 350–362.

Carneiro, G., Chan, A. B., Moreno, P. J., and Vasconcelos, N. (2007). Supervised Learning of Semantic Classes for Image Annotation and Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3), 394–410.

Carneiro, G. and Vasconcelos, N. (2005). Formulating Semantic Image Annotation as a Supervised Learning Problem. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 163–168.

Chen, H.-H., Lin, M.-S., and Wei, Y.-C. (2006). Novel association measures using web search with double checking. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, pp. 1009–1016.

Chen, S. F. and Goodman, J. (1996). An empirical study of smoothing techniques for language modeling. In *Proceedings of the 34th Annual Meeting on Association for Computational Linguistics*, pp. 310–318.

Cilibrasi, R. and Vitanyi, P. (2007). The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3), 370–383.

Clough, P. and Sanderson, M. (2004). The CLEF 2003 Cross-Language Image Retrieval Task. In *Workshop of the Cross-Language Evaluation Forum*.

Cusano, C., Ciocca, G., and Schettini, R. (2004). Image annotation using SVM. In *Proceedings of Internet Imaging IV*, vol. 5304, pp. 330–338, SPIE.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *IEEE Computer Vision and Pattern Recognition*, pp. 248–255.

Domínguez-Molina, J. A., González-Farías, G., and Rodríguez-Dagnino, R. M. (2003). *A practical procedure to estimate the shape parameter in the generalized Gaussian distribution*. Tech. rep., Universidad de Guanajuato.

Duygulu, P. (2003). *Translating images to words : A novel approach for object recognition*. Ph.D. thesis, Middle East Technical University.

Duygulu, P., Barnard, K., de Freitas, N., and Forsyth, D. A. (2002). Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In *Proceedings of the European Conference on Computer Vision*, pp. 97–112.

Escalante, H. J., Montes, M., and Sucar, L. E. (2007a). Improving Automatic Image Annotation Based on Word Co-occurrence. In *Proceedings of the 5th International Workshop on Adaptive Multimedia Retrieval*, pp. 57–70.

Escalante, H. J., Montes, M., and Sucar, L. E. (2007b). Word Co-occurrence and Markov Random Fields for Improving Automatic Image Annotation. In *Proceedings of the 18th British Machine Vision Conference*.

Everingham, M., Gool, L. V., Williams, C. K., Winn, J., and Zisserman, A. (2009). The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 88(2), 303–308.

Fan, J., Gao, Y., and Luo, H. (2007). Hierarchical classification for automatic image annotation. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp.

111–118.

Fan, J., Gao, Y., and Luo, H. (2008). Integrating Concept Ontology and Multitask Learning to Achieve More Effective Classifier Training for Multilevel Image Annotation. *IEEE Transactions on Image Processing*, 17(3), 407 –426.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874.

Feng, S. (2008). *Statistical models for text query-based image retrieval*. Ph.D. thesis, University of Massachusetts Amherst.

Feng, S. and Manmatha, R. (2008). A Discrete Direct Retrieval Model for Image and Video Retrieval. In *Proceedings of the International ACM Conference on Image and Video Retrieval*, pp. 427–436.

Feng, S., Manmatha, R., and Lavrenko, V. (2004). Multiple Bernoulli Relevance Models for Image and Video Annotation. In *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 1002–1009.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., and Ruppin, E. (2002). Placing search in context: the concept revisited. *ACM Transactions on Information Systems*, 20(1), 116–131.

Forsyth, D. A. and Ponce, J. (2003). *Computer Vision: A Modern Approach*. Prentice Hall.

Francis, N. W. and Kučera, H. (1982). Frequency Analysis of English Usage: Lexicon and Grammar. *Journal of English Linguistics*, 18(1), 64–70.

Freitas, A. and de Carvalho, A. (2007). A tutorial on hierarchical classification with applications in bioinformatics. *Intelligent Information Technologies: Concepts, Methodologies, Tools and Applications*.

Gabrilovich, E. and Markovitch, S. (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference for Artificial Intelligence*, pp. 1606–1611.

Gale, W. A., Church, K. W., and Yarowsky, D. (1992). One Sense Per Discourse. In *Proceedings of the workshop on Speech and Natural Language*, pp. 233–237.

Garg, N. (2009). *Co-occurrence Models for Image Annotation and Retrieval*. Master's thesis, Ecole Polytechnique Federale de Lausanne.

Godbole, S. and Sarawagi, S. (2004). Discriminative Methods for Multi-labeled Classification. *Advances in Knowledge Discovery and Data Mining*.

Gong, Y. and Xu, W. (2007). *Machine Learning for Multimedia Content Analysis*. Springer-Verlag New York, Inc.

Gracia, J. and Mena, E. (2008). Web-based Measure of Semantic Relatedness. In *Proceedings of 9th International Conference on Web Information Systems Engineering*, vol. 5175, pp. 136–150.

Grubinger, M., Clough, P., Müller, H., and Deselaers, T. (2006). The IAPR TC-12 Benchmark - A New Evaluation Resource for Visual Information Systems. In *Proceedings of the International Workshop OntoImage*, pp. 13–23.

Gurevych, I. (2005). Using the Structure of a Conceptual Network in Computing Semantic Relatedness. In *Proceedings of the 2nd International Joint Conference on Natural Language Processing*, vol. 3651, pp. 767–778.

Hanbury, A. and Serra, J. (2002). Mathematical morphology in the CIELAB space. *Image Analysis & Stereology*, 21, 201–206.

Hernández-Gracidas, C. and Sucar, L. E. (2007). Markov Random Fields and Spatial Information to Improve Automatic Image Annotation. In *IEEE Pacific-*

Rim Symposium on Image & Video Technology, vol. 4872, pp. 879–892.

Hirst, G. and St-Onge, D. (1998). Lexical chains as representations of context for the detection and correction of malapropisms. *WordNet: A Lexical Database for English*, pp. 305–332.

Hofmann, T. and Puzicha, J. (1998). *Statistical models for co-occurrence data.* Tech. rep., MIT.

Huiskes, M. J. and Lew, M. S. (2008). The MIR Flickr Retrieval Evaluation. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*, pp. 39–43.

Hull, D. (1993). Using statistical testing in the evaluation of retrieval experiments. In *Proceedings of International ACM Conference on Research and Development in Information Retrieval*, pp. 329–338.

Jelinek, F. and Mercer, R. L. (1980). Interpolated estimation of Markov source parameters from sparse data. In *Proceedings of the Workshop on Pattern Recognition in Practice.*

Jeon, J., Lavrenko, V., and Manmatha, R. (2003). Automatic Image Annotation and Retrieval using Cross-Media Relevance Models. In *Proceedings of International ACM Conference on Research and Development in Information Retrieval*, pp. 119–126.

Jiang, J. J. and Conrath, D. W. (1997). Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. In *Proceedings of International Conference Research on Computational Linguistics*, pp. 19–33.

Jiang, Y.-G., Ngo, C.-W., and Chang, S.-F. (2009). Semantic context transfer across heterogeneous sources for domain adaptive video search. In *Proceedings*

of the 17th ACM international conference on Multimedia, pp. 155–164.

Jin, R., Chai, J. Y., and Si, L. (2004). Effective automatic image annotation via a coherent language model and active learning. In *Proceedings of the 12th International ACM Conferencia on Multimedia*, pp. 892–899.

Jin, Y., Khan, L., Wang, L., and Awad, M. (2005b). Image annotations by combining multiple evidence & WordNet. In *Proceedings of the 13th International ACM Conference on Multimedia*, pp. 706–715.

Jin, Y., Wang, L., and Khan, L. (2005a). Improving image annotations using WordNet. In *Proceedings of the International Workshop on Advances in Multimedia Information Systems*, vol. 3665, pp. 115–130.

Joachims, T. (1999). Making large-scale SVM learning practical. *Advances in Kernel Methods – Support Vector Learning*, pp. 169–184.

Kamoi, Y., Furukawa, Y., Sato, T., Kiwada, Y., and Takagi, T. (2007). Automatic Image Annotation Based on Visual Cognitive Theory. In *Proceedings of North American Fuzzy Information Processing Society*, pp. 239–244.

Kang, F., Jin, R., and Sukthankar, R. (2006). Correlated Label Propagation with Application to Multi-label Learning. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1719–1726.

Kendall, M. (1938). A New Measure of Rank Correlation. *Biometrika*, 30, 81–89.

Kolmogorov, A. N. (1986). On the empirical determination of a distribution law. *Selected Works of A.N. Kolmogorov*, 2, 139–146.

Kullback, S. and Leibler, R. (1951). On Information and Sufficiency. *Annals of Mathematical Statistics*, 22(1), 79–86.

Landauer, T. K., Foltz, P. W., and Laham, D. (1998). An Introduction to Latent Semantic Analysis. *Discourse Processes*, 25, 259–284.

Lavrenko, V., Feng, S., and Manmatha, R. (2004). Statistical Models For Automatic Video Annotation And Retrieval. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 17–21.

Lavrenko, V., Manmatha, R., and Jeon, J. (2003). A Model for Learning the Semantics of Pictures. In *Proceedings of the 16th Conference on Advances in Neural Information Processing Systems*.

Leacock, C. and Chodorow, M. (1998). Combining Local Context and WordNet Similarity for Word Sense Identification. *WordNet: A Lexical Reference System and its Application*, pp. 265–283.

Lee, J. H., Kim, M. H., and Lee, Y. J. (1993). Information retrieval based on conceptual distance in IS-A hierarchies. *Journal of Documentation*, 49(2), 188–207.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th Annual International Conference on Systems Documentation*, pp. 24–26.

Li, J. and Wang, J. Z. (2003). Automatic Linguistic Indexing of Pictures by a Statistical Modeling Approach. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(9), 1075–1088.

Li, W. and Sun, M. (2006). Automatic Image Annotation Based on WordNet and Hierarchical Ensembles. In *Proceedings of 7th International Conference on Intelligent Text Processing and Computational Linguistics*, vol. 3878, pp. 417–428.

Li, Y., Bandar, Z. A., and McLean, D. (2003). An Approach for Measuring Semantic Similarity between Words Using Multiple Information Sources. *IEEE Transactions on Knowledge and Data Engineering*, 15(4), 871–882.

Lin, D. (1998). An information-theoretic definition of similarity. In *Proceedings of the 15th International Conference on Machine Learning*, pp. 296–304.

Lindsey, R., Veksler, V. D., Grintsvayg, A., and Gray, W. D. (2007). Be Wary of What Your Computer Reads: The Effects of Corpus Selection on Measuring Semantic Relatedness. In *Proceedings of the 8th International Conference on Cognitive Modeling*, pp. 279–284.

Little, S., Llorente, A., and Rüger, S. (2010). An Overview of Evaluation Campaigns in Multimedia Retrieval. In *ImageCLEF: Experimental Evaluation in Visual Information Retrieval*, vol. 32, edited by H. Müller, P. Clough, T. Deselaers, and B. Caputo, pp. 507–525, Springer-Verlag.

Little, S. and Rüger, S. (2009). Conservation of effort in feature selection for image annotation. In *IEEE Workshop on Multimedia Signals Processing*.

Liu, J., Li, M., Ma, W.-Y., Liu, Q., and Lu, H. (2006). An adaptive graph model for automatic image annotation. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pp. 61–70.

Liu, J., Wang, B., Li, M., Li, Z., Ma, W., Lu, H., and Ma, S. (2007). Dual cross-media relevance model for image annotation. In *Proceedings of the 15th International Conference on Multimedia*, pp. 605–614.

Llorente, A., Little, S., and Rüger, S. (2009d). MMIS at ImageCLEF 2009: Non-parametric Density Estimation Algorithms. In *Working notes for the CLEF 2009 Workshop*.

Llorente, A., Manmatha, R., and Rüger, S. (2010a). Image Retrieval using Markov Random Fields and Global Image Features. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 243–250.

Llorente, A., Motta, E., and Rüger, S. (2009b). Image Annotation Refinement Using Web-Based Keyword Correlation. In *Proceedings of the 4th International Conference on Semantic and Digital Media Technologies*, vol. 5887, pp. 188–191, LNCS.

Llorente, A., Motta, E., and Rüger, S. (2010b). Exploring the Semantics Behind a Collection to Improve Automated Image Annotation. In *Proceedings of the 10th Workshop of the Cross–Language Evaluation Forum*, vol. 6242, pp. 307–314, LNCS.

Llorente, A., Overell, S., Liu, H., Hu, R., Rae, A., Zhu, J., Song, D., and Rüger, S. (2009c). Exploiting Term Co-occurrence for Enhancing Automated Image Annotation. In *Proceedings of the 9th Workshop of the Cross–Language Evaluation Forum*, vol. 5706, pp. 632–639, LNCS.

Llorente, A. and Rüger, S. (2008a). Can a probabilistic image annotation system be improved using a co-occurrence approach? In *Proceedings of the Workshop on Cross-Media Information Analysis, Extraction and Management*, vol. 437, pp. 33–42.

Llorente, A. and Rüger, S. (2009a). Using Second Order Statistics to Enhance Automated Image Annotation. In *Proceedings of the 31st European Conference on Information Retrieval*, vol. 5478, pp. 570–577, LNCS.

Llorente, A., Zagorac, S., Little, S., Hu, R., Kumar, A., Shaik, S., Ma, X., and Rüger, S. (2008b). Semantic Video Annotation using Background Knowledge

and Similarity-based Video Retrieval. In *TREC Video Retrieval Evaluation Notebook Papers*.

Lovász, L. (1993). Random walks on graphs: A survey. *Bolyai Society Mathematical Studies*, 2(2), 1–46.

Magalhães, J. (2008). *Statistical Models for Semantic-Multimedia Information Retrieval*. Ph.D. thesis, Imperial College.

Magalhães, J. and Rüger, S. (2006). Logistic regression of generic codebooks for semantic image retrieval. In *Image and Video retrieval*, pp. 41–50, Springer Berlin / Heidelberg.

Magalhães, J. and Rüger, S. (2007). Information-theoretic semantic multimedia indexing. In *Proceedings of the International ACM Conference on Image and Video Retrieval*, pp. 619–626.

Makadia, A., Pavlovic, V., and Kumar, S. (2008). A New Baseline for Image Annotation. In *Proceedings of the 10th European Conference on Computer Vision*, pp. 316–329.

Manjunath, B. S., Salembier, P., and Sikora, T. (2002). *Introduction to MPEG-7, Multimedia Content Description Interface*, vol. 1. John Wiley and Sons, Ltd.

Manning, C. D., Raghavan, P., and Schütze, H. (2009). *An Introduction to Information Retrieval*. Cambridge University Press.

Marszałek, M. and Schmid, C. (2007). Semantic hierarchies for visual object recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1 – 7.

Medelyan, O., Milne, D., Legg, C., and Witten, I. H. (2009). Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, 67(9),

716 – 754.

Melamed, I. D. (1998). *Empirical methods for exploiting parallel texts*. Ph.D. thesis, University of Pennsylvania.

Melucci, M. (2007). On Rank Correlation in Information Retrieval Evaluation. *ACM SIGIR Forum*, 41, 18–33.

Metzler, D. and Croft, B. W. (2005). A Markov Random Field model for Term Dependencies. In *Proceedings of the 28th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 472–479.

Metzler, D. and Manmatha, R. (2004). An Inference Network Approach to Image Retrieval. In *Proceedings of the 3rd International Conference on Image and Video Retrieval*, pp. 42–50.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of ACM*, 38(11), 39–41.

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Journal of Language and Cognitive Processes*, 6, 1–28.

Milne, D. and Witten, I. H. (2008). An effective, low-cost measure of semantic relatedness obtained from Wikipedia links. In *Proceedings of the first AAAI Workshop on Wikipedia and Artifical Intellegence*, pp. 25–30.

Mohammad, S. and Hirst, G. (2005). Distributional Measures as Proxies for Semantic Relatedness. Submitted for Publication.

Monay, F. and Gatica-Perez, D. (2003). On image auto-annotation with latent space models. In *Proceedings of the 11th International ACM Conference on Multimedia*, pp. 275–278.

Monay, F. and Gatica-Perez, D. (2004). PLSA-based image auto-annotation: con-

straining the latent space. In *Proceedings of the 12th International ACM Conference on Multimedia*, pp. 348–351.

Morgan, W., Greiff, W., and Henderson, J. (2004). Direct maximization of average precision by hill-climbing, with a comparison to a maximum entropy approach. In *Proceedings of North American Chapter of the Association for Computational Linguistics - Human Language Technologies*, pp. 93–96.

Mori, Y., Takahashi, H., and Oka, R. (1999). Image-to-word transformation based on dividing and vector quantizing images with words. In *International Workshop on Multimedia Intelligent Storage and Retrieval Management.*

Morris, J. and Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1), 21–48.

Müller, H., Marchand-Maillet, S., and Pun, T. (2002). The Truth about Corel - Evaluation in Image Retrieval. In *Proceedings of the International ACM Conference on Image and Video Retrieval*, pp. 38–49.

Naphade, M., Smith, J. R., Tesic, J., Chang, S.-F., Hsu, W., Kennedy, L., Hauptmann, A., and Curtis, J. (2006). Large-Scale Concept Ontology for Multimedia. *IEEE MultiMedia*, 13(3), 86–91.

Nowak, S. and Dunker, P. (2009a). A Consumer Photo Tagging Ontology: Concepts and Annotations. In *THESEUS/ImageCLEF Pre-Workshop.*

Nowak, S. and Dunker, P. (2009b). Overview of the CLEF 2009 Large Scale - Visual Concept Detection and Annotation Task. In *Proceedings of the 10th Workshop of the Cross–Language Evaluation Forum*, vol. 6242, pp. 94–109, LNCS.

Nowak, S. and Huiskes, M. (2010). New Strategies for Image Annotation: Overview of the Photo Annotation Task at ImageCLEF 2010. In *Working notes for the CLEF 2010 Workshop*.

Nowak, S., Llorente, A., Motta, E., and Rüger, S. (2010a). The Effect of Semantic Relatedness Measures on Multi-label Classification Evaluation. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, pp. 303–310.

Nowak, S. and Lukashevich, H. (2009). Multilabel Classification Evaluation using Ontology Information. In *Proceedings of ESWC Workshop on Inductive Reasoning and Machine Learning on the Semantic Web*.

Nowak, S., Lukashevich, H., Dunker, P., and Rüger, S. (2010b). Performance Measures for Multilabel Classification — A Case Study in the Area of Image Classification. In *Proceedings of the ACM Conference on Multimedia Information Retrieval*, pp. 35–44.

Overell, S., Llorente, A., Liu, H., Hu, R., Rae, A., Zhu, J., Song, D., and Rüger, S. (2008). MMIS at ImageCLEF 2008: Experiments combining Different Evidence Sources. In *Working notes for the CLEF 2008 Workshop*.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). *The PageRank Citation Ranking: Bringing Order to the Web*. Tech. rep., Stanford University.

Patwardhan, S. (2003). *Incorporating Dictionary and Corpus Information into a Context Vector Measure of Semantic Relatedness*. Master's thesis, University of Minnesota.

Patwardhan, S., Banerjee, S., and Pedersen, T. (2003). Using Measures of Semantic Relatedness for Word Sense Disambiguation. In *Proceedings of the Fourth*

International Conference on Intelligent Text Processing and Computational Linguistics, pp. 241–257.

Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). WordNet::Similarity - Measuring the Relatedness of Concepts. In *Proceedings of the 19th American Conference on Artificial Intelligence*.

Peters, C. and Braschler, M. (2001). Cross-Language System Evaluation: the CLEF Campaigns. *Journal of the American Society for Information Science and Technology*, 52(12), 1067–1072.

Ponzetto, S. and Strube, M. (2007a). An API for Measuring the Relatedness of Words in Wikipedia. In *Companion Volume to the Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, pp. 49–52.

Ponzetto, S. and Strube, M. (2007b). Knowledge Derived From Wikipedia For Computing Semantic Relatedness. *Journal of Artificial Intelligence Research*, 30, 181–212.

Qi, G.-J., Hua, X.-S., Rui, Y., Tang, J., Mei, T., and Zhang, H.-J. (2007). Correlative multi-label video annotation. In *Proceedings of the 15th International Conference on Multimedia*, pp. 17–26.

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity in a Taxonomy. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pp. 448–453.

Resnik, P. (1999). Semantic Similarity in a Taxonomy: An Information-based Measure and its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence*, pp. 95–130.

Rubenstein, H. and Goodenough, J. B. (1965). Contextual correlates of synonymy.

Communications of ACM, 8(10), 627–633.

Sahami, M. and Heilman, T. D. (2006). A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th International Conference on World Wide Web*, pp. 377–386.

Sanderson, M. and Zobel, J. (2005). Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 162–169.

Santini, S. and Jain, R. (1998). Beyond query by example. In *IEEE Second Workshop on Multimedia Signal Processing*, pp. 3–8.

Schreiber, G., Dubbeldam, B., Wielemaker, J., and Wielinga, B. (2001). Ontology-Based Photo Annotation. *IEEE Intelligent Systems*, 16(3), 66–74.

Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press.

Shaw, J. A. and Fox, E. A. (1994). Combination of Multiple Searches. In *Proceedings of the Second Text REtrieval Conference*, pp. 243–252.

Shen, X., Boutell, M., Luo, J., and Brown, C. (2004). Multi-label machine learning and its application to semantic scene classification. In *Proceedings of IS&T/SPIE's 16th Annual Symposium on Electronic Imaging: Science and Technology*, vol. 5307, pp. 188–199.

Shi, R., Chua, T.-S., Lee, C.-H., and Gao, S. (2006). Bayesian Learning of Hierarchical Multinomial Mixture Models of Concepts for Automatic Image Annotation. In *Proceedings of the ACM Conference on Image and Video Retrieval*, pp. 102–112.

Shi, R., Lee, C.-H., and Chua, T.-S. (2007). Enhancing image annotation by integrating concept ontology and text-based Bayesian learning model. In *Proceedings of the 15th International Conference on Multimedia*, pp. 341–344.

Simpson, J. and Weiner, E. (1989). *The Oxford English Dictionary*. Clarendon Press.

Smeaton, A. F., Over, P., and Kraaij, W. (2006). Evaluation Campaigns and TRECVID. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pp. 321–330.

Srikanth, M., Varner, J., Bowden, M., and Moldovan, D. (2005). Exploiting ontologies for automatic image annotation. In *Proceedings of the 28th International ACM Conference on Research and Development in Information Retrieval*, pp. 552–558.

Stathopoulos, V. and Jose, J. (2009). Bayesian Mixture Hierarchies for Automatic Image Annotation. In *Proceedings of the 31st European Conference on Information Retrieval*, vol. 5478, pp. 138–149, LNCS.

Stathopoulos, V., Urban, J., and Jose, J. (2008). Semantic relationships in multimodal graphs for automatic image annotation. In *Proceedings of the 30th European Conference on Information Retrieval*, vol. 4956, pp. 490–497, LNCS.

Tamura, H., Mori, S., and Yamawaki, T. (1978). Textural Features Corresponding to Visual Perception. *IEEE Transactions on Systems, Man and Cybernetics*, 8(6), 460–473.

Tang, J. and Lewis, P. (2007). A Study of Quality Issues for Image Auto-Annotation with the Corel Data-Set. *IEEE Transactions on Circuits and Systems for Video Technology*, 1(3), 384–389.

Tollari, S., Detyniecki, M., Fakeri-Tabrizi, A., Amini, M.-R., and Gallinari, P. (2008). UPMC/LIP6 at ImageCLEFphoto 2008: On the Exploitation of Visual Concepts (VCDT). In *Working notes for the CLEF 2008 Workshop*.

Torralba, A. and Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, 14(3), 391–412.

Tsoumakas, G. and Vlahavas, I. (2007). Random k-labelsets: An ensemble method for multilabel classification. In *Proceedings of the 18th European Conference on Machine Learning*, pp. 406–417.

Turney, P. D. (2001). Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL. In *Proceedings of the 12th European Conference on Machine Learning*, pp. 491–502.

van Rijsbergen, C. J. (1977). A theoretical basis for the use of co-occurrence data in information retrieval. *Journal of Documentation*, 33, 106–119.

van Rijsbergen, C. J. (1979). *Information Retrieval*. Butterworths.

Viitaniemi, V. and Laaksonen, J. (2007). Evaluating Performance of Automatic Image Annotation: Example Case by Fusing Global Image Features. In *International Workshop on Content-Based Multimedia Indexing*, pp. 251 –258.

Wang, C., Jing, F., Zhang, L., and Zhang, H.-J. (2006). Image Annotation Refinement using Random Walk with Restarts. In *Proceedings of the 14th Annual ACM International Conference on Multimedia*, pp. 647–650.

Wang, Y. and Gong, S. (2007). Refining image annotation using contextual relations between words. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, pp. 425–432.

Weinberger, K. Q., Slaney, M., and Zwol, R. V. (2008). Resolving Tag Ambiguity.

In *Proceedings of the 16th ACM International Conference on Multimedia*, pp. 111–120.

Westerveld, T. and de Vries, A. P. (2003). Experimental Evaluation of a Generative Probabilistic Image Retrieval Model on 'Easy' Data. In *Proceedings of the SIGIR Multimedia Information Retrieval Workshop*.

Wilkins, P., Ferguson, P., and Smeaton, A. F. (2006). Using Score Distributions for Querytime Fusion in Multimedia Retrieval. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pp. 51–60.

Wu, L., Hua, X.-S., Yu, N., Ma, W.-Y., and Li, S. (2008). Flickr Distance. In *Proceedings of the 16th ACM International Conference on Multimedia*, pp. 31–40.

Wu, L., Li, M., Li, Z., Ma, W.-Y., and Yu, N. (2007). Visual Language Modeling for Image Classification. In *Proceedings of the International Workshop on Multimedia Information Retrieval*, pp. 115–124.

Wu, Z. and Palmer, M. (1994). Verb Semantics And Lexical Selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pp. 133–138.

Xiang, Y., Zhou, X., Chua, T.-S., and Ngo, C.-W. (2009). A Revisit of Generative Model for Automatic Image Annotation using Markov Random Fields. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, pp. 1153–1160.

Yang, J. and Hauptmann, A. G. (2008). (Un)Reliability of video concept detection. In *Proceedings of the International ACM Conference on Image and Video Retrieval*, pp. 85–94.

Yavlinsky, A., Schofield, E., and Rüger, S. (2005). Automated Image Annotation Using Global Features and Robust Nonparametric Density Estimation. In *Proceedings of the International ACM Conference on Image and Video Retrieval*, pp. 507–517.

Yilmaz, E., Kanoulas, E., and Aslam, J. A. (2008). A simple and efficient sampling method for estimating AP and NDCG. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 603–610.

Zagorac, S., Llorente, A., Little, S., Liu, H., and Rüger, S. (2009). Automated Content Based Video Retrieval. In *TREC Video Retrieval Evaluation Notebook Papers*, edited by NIST.

Zhai, C. and Lafferty, J. (2001). A study of smoothing methods for language models applied to Ad Hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 334–342.

Zhou, X., Wang, M., Zhang, Q., Zhang, J., and Shi, B. (2007). Automatic image annotation by an iterative approach: incorporating keyword correlations and region matching. In *Proceedings of the International ACM Conference on Image and Video Retrieval*, pp. 25–32.

Appendix A

Datasets

A.1 Corel 5k Dataset

All the annotation words form a vocabulary of 374 words, which are listed as follows:

city mountain sky sun water clouds tree bay lake sea beach boats people branch leaf grass plain palm horizon shell hills waves birds land dog bridge ships buildings fence island storm peaks jet plane runway basket flight flag helicopter boeing prop f-16 tails smoke formation bear polar snow tundra ice head black reflection ground forest fall river field flowers stream meadow rocks hillside shrubs close-up grizzly cubs drum log hut sunset display plants pool coral fan anemone fish ocean diver sunrise face sand rainbow farms reefs vegetation house village carvings path wood dress coast sailboats cat tiger bengal fox kit run shadows winter autumn cliff bush rockface pair den coyote light arctic shore town road chapel moon harbor windmills restaurant wall skyline window clothes shops street cafe tables nets crafts roofs ruins stone cars castle courtyard statue stairs costume sponges sign palace paintings sheep valley balcony post gate plaza festival temple sculpture museum hotel art fountain market door mural garden star butterfly angelfish lion cave crab grouper

pagoda buddha decoration monastery landscape detail writing sails food room entrance fruit night perch cow figures facade chairs guard pond church park barn arch hats cathedral ceremony crowd glass shrine model pillar carpet monument floor vines cottage poppies lawn tower vegetables bench rose tulip canal cheese railing dock horses petals umbrella column waterfalls elephant monks pattern interior vendor silhouette architecture blossoms athlete parade ladder sidewalk store steps relief fog frost frozen rapids crystals spider needles stick mist doorway vineyard pottery pots military designs mushrooms terrace tent bulls giant tortoise wings albatross booby nest hawk iguana lizard marine penguin deer white-tailed horns slope mule fawn antlers elk caribou herd moose clearing mare foals orchid lily stems row chrysanthemums blooms cactus saguaro giraffe zebra tusks hands train desert dunes canyon lighthouse mast seals texture dust pepper swimmers pyramid mosque sphinx truck fly trunk baby eagle lynx rodent squirrel goat marsh wolf pack dall porcupine whales rabbit tracks crops animals moss trail locomotive railroad vehicle aerial range insect man woman rice prayer glacier harvest girl indian pole dance african shirt buddhist tomb outside shade formula turn straightaway prototype steel scotland ceiling furniture lichen pups antelope pebbles remains leopard jeep calf reptile snake cougar oahu kauai maui school canoe race hawaii

In what follows, the list of topics represented by the 5,000 images of the Corel 5k dataset are enumerated:

Air shows

Bears

Fiji (island of Pacific)

Tigers

Foxes and Coyotes

Greek Isles

Los Angeles

Underwater Reefs

Hong Kong

Denmark

Israel

English country gardens

Holland

Images of Thailand

New York City

Ireland

Ice and Frost

Images of France

Wildlife of Galapagos

North America Deer

Arabian Horses

Flowers

African special animals

Peru

Images of Death Valley

California Coasts

Tropical Plants

Swimming Canada

Land of the Pyramids

Nesting Birds

Alaskan Wildlife

Rural France

Steam Trains

Polar Bears

Nepal

Indigenous People

Spirit of Buddha

Auto racing

Bridges

Bonny Scotland

Canadian Rockies

Zimbabwe

Mayan and Aztec Ruins

Namibia

Aviation Photography

Beaches

North American Wildlife

Hawaii

Table A.1: Plural words in Corel 5k vocabulary

animals	bulls	crops	flowers	mushrooms	plants	roofs	shops	tusks
antlers	cars	crystals	foals	needles	poppies	ruins	shrubs	vegetables
birds	carvings	cubs	hats	nets	pots	sailboats	sponges	vines
blooms	chairs	designs	hills	paintings	pups	sails	stems	waterfalls
blossoms	chrysanthemums	dunes	horns	peaks	rapids	seals	swimmers	waves
boats	clouds	farms	horses	pebbles	reefs	shadows	tables	whales
buildings	crafts	figures	monks	petals	rocks	ships	tracks	windmills

Table A.2: WordNet wrong senses disambiguated for the Corel 5k dataset

Word	Disambiguated Sense	Actual Sense
aerial	a pass to a receiver downfield from the passer	an antenna
albatross	something that hinders or handicaps	a bird
balcony	an upper floor in an auditorium	a balustrade
bengal	a region in the northeast of the Indian subcontinent	a bengal tiger
black	the quality of the achromatic colour of least lightness	a black bear
blooms	the organic process of bearing flowers	flowers
booby	an ignorant or foolish person	a bird
branch	a division of a larger organisation	a division of a stem of a plant
canal	an indistinct surface feature of Mars	a strip of water
church	a group of Christians	a place for worship
clouds	a collection of particles	a suspended mass of water
column	a line of units following one after another	a pillar
coral	a colour averaging a deep pink	the skeleton of a coral
crystals	a solid formed by the solidification of a chemical	glass
cubs	an awkward youth	the young of a fox, bear, or lion
designs	the act of working out the form of sth.	a decorative work
detail	an isolated fact	a decorative feature of a work of art
display	sth. intended to communicate an impression	sth. shown to the public
dock	an enclosure in a court of law	a harbour
fawn	a colour varying around a light grey-brown colour	a young deer
figures	a diagram illustrating textual material	a model of a bodily form
fly	the insect	a bird
formula	a mathematical formula	a formula one car
horns	a noisemaker	outgrowths on the heads of ungulates

Continued on next page

Table A.2 – continued from previous page

Word	Disambiguated Sense	Sense Attributed
hut	temporary military shelter	small crude shelter used as a dwelling
kit	a case for containing a set of articles	a young cat or fox
land	the land on which real estate is located	ground or fields
lichen	an eruptive skin disease	the plant
light	electromagnetic radiation	the quality of emitting light
lynx	a text browser	short-tailed wildcats
marine	a member of the U.S. Marine Corps	sth. found in the sea
market	the world of commercial activity	the marketplace
model	a hypothetical description of a complex entity	the latest model of a car
nets	a computer network	a trap made of netting to catch fish
pack	a large indefinite number	a group of hunting animals
palm	the inner surface of the hand	a palm tree
path	a course of conduct	a trail or track
pattern	a perceptual structure	a decorative work
peak	the most extreme possible amount or value	the top of a mountain
pillar	a fundamental principle	a column
plant	buildings for carrying on industrial labour	vegetation
pool	an excavation filled with water	a small lake
post	the position where a guard stands	a pole or a pillar
prop	a support placed beneath sth. to keep it from shaking	a propeller
pyramid	a polyhedron	a monument
railroad	the organisation responsible for operating trains	the rail tracks
range	an area in which something acts	a series of mountains
reflection	a calm consideration	the image reflected by a mirror
relief	the feeling that comes when sth. burdensome is removed	sculpture
remains	any object that is left unused	the ruins
ruins	an irrecoverable state of devastation and destruction	a ruined building
run	a score in baseball	a race run on foot
runway	parallel bars making the railway	surface where planes take off and land
seals	fastener	a marine mammals
shadows	shade within clear boundaries	something existing in perception only

Continued on next page

Table A.2 – continued from previous page

Word	Disambiguated Sense	Sense Attributed
shell	ammunition	hard outer covering of many animals
sign	a perceptible indication of sth. not apparent	a public display of a message
sphinx	an inscrutable person	a monument
stems	a word after all affixes are removed	the stem of a plant
steps	any manoeuvre made as part of progress toward a goal	stairs
stick	an implement consisting of a length of wood	a branch of a tree
table	a set of data arranged in rows and columns	a piece of furniture
tails	the posterior part of the body of a vertebrate	the rear part of an aircraft
tiger	a fierce or audacious person	large feline of forests in most of Asia
trail	a mark left by something that has passed	a path or track
water	binary compound	the earth's surface covered with water
whales	a very large person	a cetacean mammal
wood	the substance under the bark of trees	a forest

A.2 TRECVID 2008 Dataset

The following list details the vocabulary formed by the annotation words. In total, there are 20 words, each one is accompanied by a short description:

001 Classroom: a school- or university-style classroom scene. One or more students must be visible. A teacher and teaching aids (e.g. blackboard) may or may not be visible;

002 Bridge: a structure carrying a pathway or roadway over a depression or obstacle. Such structures over non-water bodies such as a highway overpass or a catwalk (e.g., as found over a factory or warehouse floor) are included;

003 Emergency_Vehicle: external view of, for example, a police car or van, fire truck or ambulance. There may be other sorts of emergency vehicles. Included may be

UN vehicles, but NOT military vehicles;

004 Dog: any kind of dog, but not wolves;

005 Kitchen: a room where food is prepared, dishes washed, etc.

006 Airplane_flying: external view of a heavier than air, fixed-wing aircraft in flight gliders included. NOT balloons, helicopters, missiles, and rockets;

007 Two people: a view of exactly two people (not as part of a larger visible group);

008 Bus: external view of a large motor vehicle on tires used to carry many passengers on streets, usually along a fixed route. NOT vans and SUVs;

009 Driver: a person operating a motor vehicle or at least in the driver's seat of such a vehicle;

010 Cityscape: a view of a large urban setting, showing skylines and building tops. NOT just street-level views of urban life;

011 Harbor: a body of water with docking facilities for boats and/or ships such as a harbor or marina, including shots of docks. NOT shots of offshore oil rigs, piers that do not look like they belong to a harbor or boat dock;

012 Telephone: any kinds of telephone, but more than just a headset must be visible;

013 Street: a regular paved street NOT a highway, dirt road, or special type of road or path;

014 Demonstration_Or_Protest: an outdoor, public exhibition of disapproval carried out by multiple people, who may or may not be walking, holding banners or signs;

015 Hand: a close-up view of one or more human hands, where the hand is the primary focus of the shot;

Table A.3: Concepts used as annotation words for ImageCLEF 2008

ID	Annotation Word
0	indoor
1	outdoor
2	person
3	day
4	night
5	water
6	road or pathway
7	vegetation
8	tree
9	mountains
10	beach
11	buildings
12	sky
13	sunny
14	partly cloudy
15	overcast
16	animal

016 Mountain: a landmass noticably higher than the surrounding land, higher than a hill, with the slopes visible;

017 Nighttime: a shot that takes place outdoors at night. NOT sporting events under lights;

018 Boat_Ship: exterior view of a boat or ship in the water, e.g. canoe, rowboat, kayak, hydrofoil, hovercraft, aircraft carrier, submarine, etc.

019 Flower: a plant with flowers in bloom; may just be the flower;

020 Singing: one or more people singing - singer(s) visible and audible, solo or accompanied, amateur or professional.

A.3 ImageCLEF 2008 Dataset

The vocabulary was made up of the following 17 words as seen in Table A.3

The words are presented graphically adopting the hierarchical representation of an ontology as seen in Figure A.1.

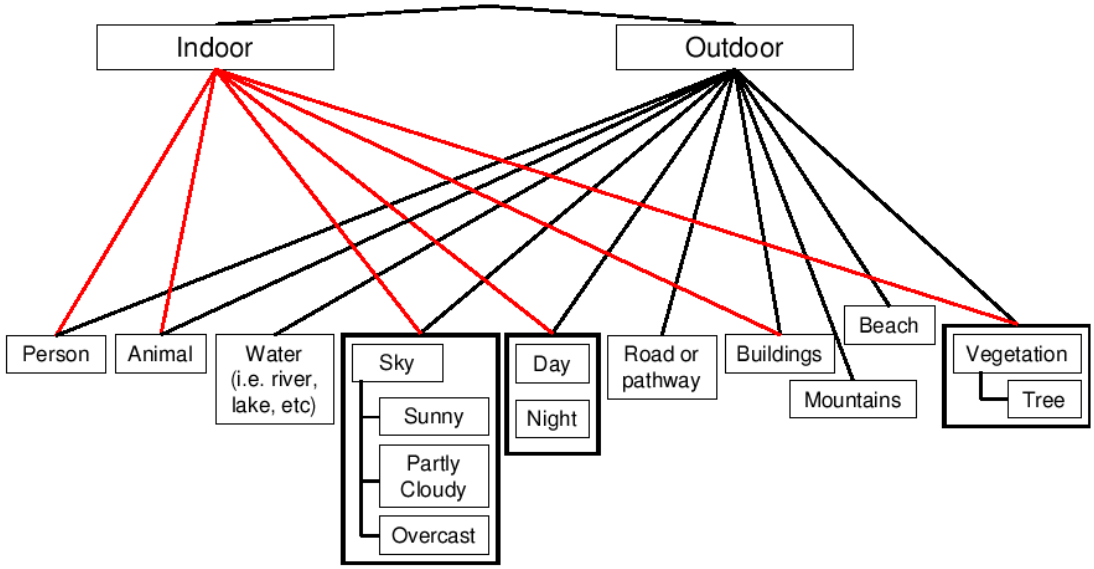

Figure A.1: Ontology containing the annotation words for ImageCLEF 2008

A.4 ImageCLEF 2009 Dataset

The vocabulary was made up of the following 53 words (Table A.4). All words refer to a holistic visual impression of the associated image. The annotation words refer to the impression of the whole image. Each image contains multiple annotations. Some of them are modelled as disjoint, meaning that if concept a is present in an image concept b cannot be present at the same time. Other concepts are modelled as optional.

The Consumer Photo Tagging Ontology developed by (Nowak and Dunker 2009a) was used for the competition. All annotation words are instances of the categories of the ontology as seen in Figure A.2. However, not all the concepts in the ontology are used as annotations as their main purpose is to help in structuring the knowledge.

Table A.4: Concepts used as annotation words

ID	Annotation Word	Category in Ontology
0	Partylife	SceneDescription.AbstractCategories.Partylife
1	Familiy_Friends	SceneDescription.AbstractCategories.FamiliyFriends
2	Beach_Holidays	SceneDescription.AbstractCategories.BeachHolidays
3	Building_Sights	SceneDescription.AbstractCategories.BuildingsSights
4	Snow	SceneDescription.AbstractCategories.SnowSkiing
5	Citylife	SceneDescription.AbstractCategories.Citylife
6	Landscape_Nature	SceneDescription.AbstractCategories.LandscapeNature
7	Sports	SceneDescription.Activity.Sports
8	Desert	SceneDescription.AbstractCategories.Desert
9	Spring	SceneDescription.Seasons.Spring
10	Summer	SceneDescription.Seasons.Summer
11	Autumn	SceneDescription.Seasons.Autumn
12	Winter	SceneDescription.Seasons.Winter
13	No_Visual_Season	SceneDescription.Seasons.NoVisualCue
14	Indoor	SceneDescription.Place.Indoor
15	Outdoor	SceneDescription.Place.Outdoor
16	No_Visual_Place	SceneDescription.Place.NoVisualCue
17	Plants	LandscapeElements.Plants
18	Flowers	LandscapeElements.Plants.Flowers
19	Trees	LandscapeElements.Plants.Trees
20	Sky	LandscapeElements.Sky
21	Clouds	LandscapeElements.Sky.Clouds
22	Water	LandscapeElements.Water
23	Lake	LandscapeElements.Water.Lake
24	River	LandscapeElements.Water.River
25	Sea	LandscapeElements.Water.Sea
26	Mountains	LandscapeElements.Mountains
27	Day	SceneDescription.TimeOfDay.Day
28	Night	SceneDescription.TimeOfDay.Night
29	No_Visual_Time	SceneDescription.TimeOfDay.NoVisualCue
30	Sunny	SceneDescription.TimeOfDay.Sunny
31	Sunset_Sunrise	SceneDescription.TimeOfDay.SunsetOrSunrise
32	Canvas	Representation.Canvas
33	Still_Life	Representation.StillLife
34	Macro	Representation.MacroImage
35	Portrait	Representation.Portrait
36	Overexposed	Representation.Illumination.Overexposed
37	Underexposed	Representation.Illumination.Underexposed
38	Neutral_Illumination	Representation.Illumination.Neutral
39	Motion_Blur	Quality.Blurring.MotionBlur
40	Out_of_focus	Quality.Blurring.OutOfFocus
41	Partly_Blurred	Quality.Blurring.PartlyBlurred
42	No_Blur	Quality.Blurring.NoBlurDetectable
43	Single_Person	PicturedObjects.Persons.Single
44	Small_Group	PicturedObjects.Persons.SmallGroup
45	Big_Group	PicturedObjects.Persons.BigGroup
46	No_Persons	PicturedObjects.Persons.NoPersons
47	Animals	PicturedObjects.Animals
48	Food	PicturedObjects.Food
49	Vehicle	PicturedObjects.Vehicles
50	Aesthetic_Impression	Quality.Aesthetics.AestheticImpression
51	Overall_Quality	Quality.Aesthetics.HighGradeOverallQuality
52	Fancy	Quality.Aesthetics.Fancy

Figure A.2: Consumer Photo Tagging Ontology