KNOWLEDGE MEDIA INSTITUTE

The Use of Ontologies for Improving Image Retrieval and Annotation

Technical Report kmi-08-08
September 2008

Ainhoa Llorente Coto

Llorente, A., Overell, S., Liu, H., Hu, R., Rae, A., Zhu, J., Song, D., and Rüger, S. (2008) Exploiting Term Co-occurrence for Enhancing Automated Image Annotation, Evaluating Systems for Multilingual and Multimodal Information Access. 9th Workshop of the Cross-Language Evaluation Forum

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

The Open University

The Use of Ontologies for Improving Image Retrieval and Annotation

Ainhoa Llorente Coto

Contents

Chapter 1

Introduction

1.1 Thesis Overview

Nowadays, digital photography is a common technology for capturing and archiving images due to the falling price of storage devices and the wide availability of digital cameras. Without efficient retrieval methods the search of images in large collections is becoming a painstaking work. Most of the traditional image search engines rely on keyword-based annotations because they lack the ability to examine image content. However, *"a picture is worth a thousand words"*, this means that up to a thousand words can be needed to describe the content depicted in a picture. This research proposes the use of highly structured annotations called ontologies to improve efficiency in image retrieval as well as to overcome the semantic gap that remains between user expectations and system retrieval capabilities.

This work focuses on automated image annotation which is the process of creating a model that assigns visual terms to images because manual annotation is a time consuming and inefficient task. Up to now, most of the automated image annotation systems are based on a combination of image analysis and statistical machine learning techniques. The main objective of this research is to evaluate whether the underlying information contained in an ontology created from the vocabulary of terms used for the annotation could be effectively used together with the extracted visual information in order to produce more accurate annotations.

1.2 Introduction

There are many market research firms involved in the estimation of the size of the digital information in the world. For instance, IDC forecasts [1] that

in 2006 the amount of digital information was of 161 hexabytes growing to 988 in 2010. One quarter of this digital universe corresponds to images (both moving and still) captured by more than 1 billion devices in the world that range from digital cameras and camera phones to medical scanners and surveillance security cameras. Managing this immense amount of images is not only a matter of having enough storage capacity but also a problem of having efficient retrieval methods. Some surveys have estimated that the cost of not finding information is of 5.3 millions of dollars a year for U.S. organizations.

Traditionally, there are two main trends in the process of retrieving images. The first one is called content-based image retrieval (CBIR), also known as query by image content (QBIC) or content-based visual information retrieval (CBVIR). "Content-based" means that the search will analyze the actual contents of the image by using image analysis techniques. The term "content" in this context refers to properties of the image called low-level features such as colours, shapes and textures. One of the limitations of CBIR is the so called semantic-gap [2]- the discrepancy between the information extracted from the visual features of images and the interpretation made by users. Another limitation is the incapability of dealing with abstract terms, how can be deduced from the colour, texture or shape something like happiness, sadness, hope or even despair?

Without the ability to examine image content, image retrieval systems must rely on text-based metadata. One example of that is Yahoo or Google image search engines which base their retrieval capabilities on searching the **context** of the images, that is to say, the image filename, text surrounding the web page or in the captions.

This approach depends heavily on the quality of the annotation. Thus, annotation is considered as a pre-stage of the retrieval process.

Some authors ([3], [4]) have stated that an efficient annotation of images with highly structured semantics could significantly lead to the improvement of the recall and precision of image retrieval.

Consequently, ontology-based retrieval systems should be taken into account.

Another limitation to overcome is the fact that images rarely include annotations because it is a time consuming process. As a result, very large collections of digital images without annotations continue to grow. One way of solving this is by automating this process, in the past there have been several attempts [2] of achieving this goal.

Automated image annotation can be defined as the process of modeling the work of a human annotator when assigning keywords to images based on their visual properties. These keywords are called visual concepts.

Different machine learning methods model the association between words and images (global features) or image regions (local features) for the automated annotation process. The problem is that while most models use the co-occurrence of image and words few analyze the dependence of annotation words on image.

The final objective of this research is to exploit ontological relationships among the keywords used for the annotation and demonstrate their effect on automated image annotation and retrieval.

Chapter 2

Overview of Image Retrieval

An image retrieval system is a computer system for browsing, searching and retrieving images from a large database of digital images. Several criteria can be considered in order to classify image retrieval systems:

- User interaction, browsing, typing text or inserting a image that is visually similar to the target image.
- Search performance, how the search engine actually searches. For instance, whether the search is accomplished through the analysis of visual features or through semantic annotations.
- Domain of the search, standalone search engine, only executes a search in a local computer versus Internet based search engine.

This analysis is focused on two major paradigms which are content-based image retrieval and semantic-based image retrieval. The latter is divided into the two groups depending on the nature of the metadata:

- Text-based metadata image retrieval
- Ontology-based image retrieval

2.1 Content-based image retrieval

CBIR or content-based image retrieval is the application of computer vision to the image retrieval problem. By content we understand colour, shape, texture, or any other information that can be derived from the image itself. These systems employ image processing technologies to extract visual features and then apply similarity measurements to them. Feature extraction algorithms extract features and store them in the form of multidimensional

vectors. Afterwards, similarity/dissimilarity measurement between two feature vectors is defined for each feature. In general, the distance between two vectors is equivalent to the dissimilarity between the corresponding images. Different distance metrics can be used depending on the considered feature. A good review about different dissimilarity measures use in CBIR can be found in [5].

Many multimedia retrieval systems [6] were developed in the 90s both for commercial and research purposes like QBIC [7], Virage [8], and many others like MARS, AMORE, Photobook or Excalibur. Some years later the basic concept of similarity search was transferred to several Internet image search engines including Webseek [9] and Webseer [10]. It is important to mention the efforts made to integrate CBIR with enterprize databases such as Informix datablades, IBM DB2 Extenders, or Oracle Cartridges with the objective of making CBIR more accessible to the industry. Smeulders et al. [2] give an exhaustive overview of the state of the art in CBIR before the year 2000. They identify three main categories based on user interaction: category search, target search and search by association.

- Category search or object detection which means to identify an object within an image. One example of this is face detection. Clearly, some problems that need to be overcome in order to achieve its fully automatization are viewpoint variation, illumination and occlusion.
- Target search or query by example, the user query is an image.
- Association search or browsing. It can be done by keyword or using visual features of the image.

Armitage and Enser [11] broaden notably the scope of the needs required by a human user when searching for an image.

More recently, the work by Datta et al. [12] analyzes some content-based image retrieval systems from the 2000s onwards. Finally, major research challenges [6] have been detected for the CBIR research community such as:

- Semantic search with emphasis on the detection of concepts in media with complex backgrounds.
- Multi-modal analysis and retrieval algorithms especially towards exploiting the synergy between the various media including text and context information.
- Experiential multimedia exploration systems toward allowing users to gain insight and explore media collections.

- Interactive search, emergent semantics, or relevance feedback systems.
- Evaluation with emphasis on representative test sets and usage patterns.

At the beginning of the 21st century researches started being aware that feature based similarity search algorithms were not as intuitive nor user-friendly as they had expected. One clear example of this is the QBIC system used by "The Hermitage Museum" search engine, it allows the user to adjust the relative weights among different features but it is quite cumbersome for the inexperienced user to set those weights since the interpretation of the meaning of the features could differ from the way the system encodes them. In addition to the difficulty to formulate an exact feature query, the semantic gap still remains. Another limitations are the low retrieval precision together with the requirement of advanced image processing and pattern recognition techniques. In order to overcome these drawbacks semantic-based techniques were introduced.

2.2 Metadata-based image retrieval

These image retrieval systems are based on metadata. There are different ways of classifying metadata. According to the *origin of the properties described* we find metadata that describe properties of the image itself in contrast with those describing the subject matter of the image. In the first group, properties like title, creator, resolution, image format, date and location can be considered. The new generation of digital cameras [13] are able to provide some of these data like category ("indoor", "outdoor"), time (date and timestamp), and location (GPS). In the second group the properties refer to the objects, persons or concepts depicted in the image. According to the *nature of the annotator* two groups can be considered:

- Annotations made by a human expert like the curator of an art museum. One drawback of this approach is the impossibility of annotating a huge collection of images.
- Annotations obtained by a computing system (automated image annotation). They are not as accurate as the one made by an expert.
- Folksonomies, social tagging systems that relies on the idea of the wisdom of the people. One representative example of this is `Flickr.com`. This approach overcomes the so much time consuming of a

manual annotation but the inconsistency in tag use can difficult the search through the entire collection of data.

Another classification scheme is related to the *structure of the metadata*. The simplest one is called **unstructured annotation** where simple tags or labels called keywords are used to describe the content of images. Another way of adding structure to the annotation is by using a controlled vocabulary that can range from a simple in-house list of words to lexical databases or thesauri; in this case we have **structured annotations**. Finally, one can find **highly structured annotations** by means of metadata schemas or ontologies which indicate how the terms in the vocabulary are linked to the image. According to the structure of the metadata, image retrieval systems can be classified in text-based image retrieval when the metadata used is unstructured or structure in opposition to the ontology-based image retrieval when the annotations are highly structured. In the first case the search will be accomplished based on a syntactic match while the latter will rely on a semantic match.

2.3 Text-based image retrieval

Image search is supported by augmenting images with keyword-based annotations and the search process always relies on keyword matching techniques.

The techniques most widely spread for creating the annotations that support this search are building keyword indices based on image content, embedding keyword-based labels into the image or extracting the annotations from the text surrounding images on the Internet, from the filename or even from the "alt" tag in HTML.

Some examples of keyword web-based search engines are Webshots (`www.webshots.com`), Ask images (`www.ask-images.com`), Google image, Altavista and Picsearch (`www.picsearch.com`). However, several limitations to this kind of search engines appeared rather soon. Firstly, users should have a complete domain knowledge in order formulate appropriate keywords for a valid query. Additionally, the difficulty in dealing with non-visual objects like expressing feelings or emotions still remains. Another limitation can be found in the subjectivity of the human annotator; different annotators will lead to different annotations. One example of that is the annotation game [14] developed by Von Ahn and Dabbish in which two different persons are prompt to label at the same time the same image and they do not get an score until they agree on the same annotation for a given image.

Finally, the infeasibility of having to describe visual content using simple words like shapes and textures of natural objects.

An ontology-based approach is proposed in order to overcome these problems. Thus, images are annotated with semantic tags that are defined and derived from a set of domain concepts or schemes called domain ontology. Consequently, the retrieval process is conducted at the abstract semantic level instead of the purely syntactic keyword matching level.

2.4 Ontology-based image retrieval

The Semantic Web provides new insights into the image retrieval problem, developing techniques to annotate the content of images by using ontologies.

An ontology is similar to a dictionary or glossary, but with greater detail and structure that enables computers to process its content. An ontology consists of a set of concepts, relations, and axioms that formalize a field of interest.

Halaschek-Wiener et al. [18] mention several reasons why ontologies can help image retrieval. The first reason can be found in the fact that ontologies provide the ability to model the semantics of what occurs in images such as object, events, etc. The expressivity of the current Web ontology standard, OWL, allows for affiliated searches based on logic and structural inference. Ontologies also provide an elegant mechanism to formally organize image content in small, logically contained groups (ontological concepts), while enabling them to be linked, merged, and distinguished with other concepts in logically contained groups. Additionally, they enable the ability to assert that many images refer to the same concepts through the use of URIs. This, in turn allows these disparate information pieces to be linked together through image depictions. Consequently, the use of ontologies provides an accepted standard that allows other individuals to process image content which has been previously annotated.

Systems whose main approach is to map low-level features of two ontological concepts [19], [20] and [21] have recently emerged. As a consequence, new tools which are closely tied to domain specific ontologies have been developed for annotation purposes [22], [23] and [24]. One example is Sculpteur [20], an ontology-based image retrieval that allows users to search and navigate semantically enriched multimedia. The ontology used is CIDOC that was created for accessing cultural heritage data. Current technologies only allow annotation with respect to preset ontologies or rely on application specific ontologies to be used as configuration mechanisms. The ability to annotate images with respect to any available ontology is

	iFIND [25]	**[26]**	**Behold [27]**
Organisation	Ms Research	Sony CSL	Imperial College
Location	China	Paris	UK
Process	Semiautomatic	Manual	Automatic
Image Analysis	Yes	Yes	Yes
Annotation	Keywords	Social Tags	Visual Concepts
Query	Keyword;QVE;Browsing	Keyword;QVE	Keyword
Search	Relevance Feedback	Textual;Image Similarity	Textual;Image Similarity
Weak Points	Poor Semantics	Misspelling Errors	Just based on visual concepts
Strong Points	Relevance Feedback	User-friendly Interface	Robust Image Processing

Table 2.1: CBIR and text metadata-based image retrieval

extremely important, as the notion of the Semantic Web heavily hinges on the development of multiple ontologies by various individuals, spanning many domains. While substantial progress has been made, further work in defining a more generic approach for annotating and managing digital images on the Web is needed.

Benefits of ontology over keyword-based methods

The most important limitation of keyword-based methods is that they are unable to put the image information in context. The context is very difficult to model in keyword based queries. This situation becomes even worse when part of the context is spread across different media for instance in images and in text. In image collections indexed with keywords, a small subset of the controlled keyword set is associated with an image. The keywords themselves are unrelated atoms. If we consider the terms of the ontology to be our controlled keyword list, using an ontology and a structured description based on this ontology changes the annotation and querying process because it guides the annotation process using restrictions and default information. It also makes the relation between property values and agents explicit, telling which property value is connected using which property to which element of the subject matter or the image itself. For instance, let us consider the example “elephant under large tree”, if reduced to keywords, “large” can refer to the “elephant”, the “tree”, or even the image itself.

Finally, ontologies provide relations between the terms; in our example, default information like "elephants live in Africa" and inheritance can help. Inheritance provides a controlled means to widen or constrain a query.

2.5 Hybrid solutions for image retrieval

In practice, most image retrieval systems are a combination of the types mentioned in the previous sections as a way to overcome the current limitations of using "stand-alone" technologies. Some hybrid solutions are shown in Table 2.4 and Table 2.5. A poor performance of keyword-based systems in defining the content of an image can be improved by an optimal use of image processing techniques. The use of ontologies is always desirable because it allows the interchange of the annotations between different users or even systems. Additionally, if the data collection is well defined by the ontology the fact of adding image processing techniques will ensure an augmentation of the information obtained from visual features. Finally, one important requirement is to achieve a proper balance between automatic annotation and the quality of the metadata.

	AKTive Media [28]	**PhotoStuff [18]**	**Photocopain[19]**	**[29]**
University	Sheffield	Maryland	Southampton	N. Tsing Hua
Location	UK	USA	UK	China
Process	Manual	Manual	Semiautomatic	Automatic
Image Analysis	No	No	4 feature concepts	No
Annotation	Text and Ontology	Ontology and Visual Concepts	GPS; Social Tags;EXIF	Description of images in NL
Query	Keyword	Keyword	Keyword	NL
Search	Search for concepts	Search for concepts	Metadata search	Sentences in NL
Weak points	No visual features	No visual features	Simple image analysis	No visual features
Strong points	Regions annotation	Use of general ontologies	Richness of information	Close to human behaviour

Table 2.2: Ontology-based and text metadata image retrieval

Chapter 3

Automated Image Annotation

Automated image annotation, also known as image auto-annotation, consists of a number of techniques that aim to find the correlation between low-level visual features and high-level semantics. It emerged as a solution to the time-consuming work of annotating large datasets.

Most of the 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.

Automated image annotation can be divided with respect to the deployed machine learning method into co-occurrence models, machine translation models, classification approaches, graphic models, latent space approaches, maximum entropy models, hierarchical models and relevance language models. Another classification scheme makes reference to the way the feature extraction techniques treat the image either as a whole in which case it is called scene-orientated approach or as a set of regions, blobs or tiles which is called region-based or segmentation approach.

A very early attempt in using **co-occurrence** information was made by Mori et al. [30]. The process used by them starts by dividing each training image into equally rectangular parts ranging from 3x3 to 7x7. 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, conditional probability for each word and each cluster is estimated dividing 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 the carried out on the learning data. A new image is divided into parts, features are extracted, the nearest clusters are found for each divided part and an average of the conditional probability of the nearest clusters is calculated. Finally, words

are selected based on the largest average value of conditional probability.

Duygulu et al. [31] improved the co-occurrence method using a **machine translation model** that is applied in order to translate words into image regions called blobs in the same way as words from French might be translated into English. The dataset used by them, 5,000 images Corel dataset, has become a popular benchmark of annotation systems in the literature.

Monay and Gatica-Perez [32] introduced **latent variables** to link image features with words as a way to capture co-occurrence information. This is based on latent semantic analysis (LSA) which comes from natural language 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) [33].

Blei and Jordan [34] viewed the problem of modelling annotated data as the problem of modelling data of different types where one type describes the other. For instance, image and their captions, papers and their bibliographies, genes and their functions. In order to overcome the limitations of the generative probabilistic models and discriminative classification methods Blei and Jordan proposed a framework that is a combination of both of them. They culminated in **Latent Dirichlet Allocation**, [35] a model that follows the image segmentation approach and finds conditional distribution of the annotation given the primary type.

Jeon at al. [37] improved on the results of Duygulu et al. by recasting the problem as cross-lingual information retrieval and applying the **Cross-Media Relevance Model** (CMRM) to the annotation task. In addition to that, they showed that better ranked retrieval results could be obtained by using probabilistic annotation rather than hard annotation.

Lavrenko et al. [38] used the **Continuous-space Relevance Model** (CRM) to build continuous probability density functions to describe the process of generating blob features. The CRM model outperforms the CMRM model significantly.

Metzler and Manmatha [39] 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. [40] used a **Multiple Bernoulli Distribution** (MBRM), which outperforms CRM. MBRM differs from Continuous-space Relevance Model in the image segmentation and in the distribution of annotation words. CRM segments images into semantically-coherent regions while MBRM imposes a fixed-size rectangular grid (tiles) on each image. The advantage of this tile approach is that it reduces significantly the computa-

tional time. CRM models annotation words using a multinomial distribution opposed to MBRM which uses a multiple-Bernoulli distribution. This model focuses on the presence or absence of words in the annotation rather than in their prominence as it does the multinomial distribution. Image feature probabilities are estimated using a non-parametric kernel density estimation.

Other authors like Torralba and Oliva [41] focused on modelling **a global scene** rather than image regions. This 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. Torralba and Oliva supported the hypothesis that objects and their containing scenes are not independent. They learned global statistics of scenes in which objects appear and used them to predict presence or absence of objects in unseen images. Consequently, images can be described with basic keywords such as "street", "buildings" or "highways", using a selection of relevant low-level global filters.

Yavlinsky et al. [42] followed this approach using simple global features together with robust **non-parametric density estimation** and the technique of kernel smoothing. The results shown by Yavlinsky et al. are comparable with the inference network [39] and CRM [38]. Notably, Yavlinsky et al. showed that the Corel dataset proposed by Duygulu et al. [31] could be annotated remarkably well by just using global colour information.

One of the earliest work [47] in non-parametric approximation of density functions dates from 1974 when a collaboration between Stanford University and Jet Propulsion Lab gave birth to the segmentation of an image into meaningful regions following a statistical approach.

Chapter 4

Ontologies for Image Annotation

Gruber [48] defines an ontology as *"a formal specification of a shared conceptualization of a domain of interest"*.

There are several types of ontologies: general-purpose, domain-oriented, multimedia. Some examples of general purpose can be mentioned:

- DOLCE: (`http://www.loa-cnr.it/DOLCE.html`)
- The Upper Cyc Ontology: (`http://www.cyc.com/cyc-2-1/index.html`)
- IEEE Standard Upper Ontology: (`http://suo.ieee.org`)
- Suggested Upper Merged Ontology (SUMO): (`http://protege.stanford.edu/ontologies/sumoOntology/sumo_ontology.html`)

4.1 Multimedia Ontologies

With the advent of the Semantic Web a shared vocabulary is needed to annotate the vast collection of images and other multimedia resources. An ontology is necessary in order to provide the vocabulary with a set of relationships that enable sharing the knowledge between people and machines. The final goal is to obtain metadata, produced by annotation with respect to a shared ontological vocabulary, that will allow searching and navigating by concept.

The W3C Semantic Web Best Practices group [52] has been compiling a collection of vocabularies in RDF or OWL format that can be used for image annotation. One of the first problems that they had to face is the fact

that many vocabularies were prior to the Semantic Web so they need to be translated to RDF or OWL. Sometimes this translation is not immediate. That is the case of the standard MPEG-7.

The **Multimedia Content Description** standard, widely known as MPEG-7, standardizes ways (tools) to define multimedia Descriptors (Ds), Description Schemes (DSs) and the relationships between them. The descriptors are low-level features (visual or audio) while the description schemes are abstract description entities. It is represented in Description Definition Language (DDL) language. The main problem with MPEG-7 is that annotations are not interoperable. There are ambiguities due to complementary description tools.

In their paper [53] Bailer et al. claim that MPEG-7 profiles can only partly solve interoperability problems. There are several solutions that try to overcome the drawbacks of MPEG-7 by replacing it with a high quality multimedia ontology that should fulfils the following requirements:

- Reusability; design a core ontology for any multimedia related application.
- MPEG-7-Compliance; it should support most important description tools (decomposition, visual and audio descriptors).
- Extensibility; it should enable the inclusion of further media types and description tools.
- Modularity; it should enable the customization of multimedia ontology.
- High degree of axiomatization; it should ensure interoperability through machine accessible semantics.

Thus, several attempts have been made to translate MGEP-7 into a Semantic Web language such as DAML+OIL, RDFS or OWL. Finally, the MPEG-7 ontology by DMAG covers the whole standard, is an OWL Full ontology and contains 2372 classes and 975 properties.

4.2 Visual Ontologies

A Visual Ontology is an ontology which is based on the visual part of the standard MGEP-7. Some examples are the following:

- Visual Ontology (VO), [55] is an ontology for video retrieval that was built using two existing corpora WordNet and the visual part of

the standard MPEG-7 by creating links between visual and general concepts.

- Visual Descriptor Ontology (VDO) developed within the aceMedia project [51], [56] for semantic multimedia content analysis and reasoning, contains representations of MPEG-7 visual descriptors and models Concepts and Properties that describe visual characteristics of objects. The term descriptor refers to a specific representation of a visual feature (colour, shape, texture, etc.) that defines the syntax and the semantics of a specific aspect of the feature. For example, the dominant colour descriptor specifies among others, the number and value of dominant colours that are present in a region of interest and the percentage of pixels that each associated colour value has. Although the construction of the VDO is tightly coupled with the specification of the visual part of the standard MPEG-7, several modifications were carried out in order to adapt to the XML Schema provided by MPEG-7 to an ontology and the data type representations available in RDF Schema.

- The ontology designed by the group Mindswap of Maryland University [18] in order to describe the semantic of images, image regions (SVG), videos, frames, segments, and what they depict. This ontology is the default ontology of an image annotation tool called Photostuff. It is based on the visual part of the standard MPEG-7.

Other examples of ontologies that have been adapted to deal with visual resources are the following:

- Large-Scale Concept Ontology for Multimedia (LSCOM) [57] is a large standardized taxonomy for describing broadcast news video developed in a collaborative way by multimedia researchers, library scientists, and end users. Its final goal is to simultaneously optimize utility to facilitate end-user access, cover a large semantic space, make automated extraction feasible, and increase observability in diverse broadcast news video data sets. It has been widely used by the TRECVID community [58].

- CIDOC, Conceptual Reference Model (CRM) developed by CIDOC Documentation Standards Working Group is concerned with cultural heritage information describing concepts and relations relevant to all types of material collected and displayed by museums. It aims to support the exchange of relevant information across museums through coherent semantics and common vocabularies.

- OntoMedia ontology [59] is tailored to annotate cultural data, textual fiction and film. It has been built upon the ontologies ABC and CINDOC. It is multimodal and extensible. Thus, ABC ontology, [60] was developed for the cataloguing community, digital libraries.

4.3 Ontology Engineering

Gruber defines in his paper [61] some principles that should be considered in order to design ontologies for sharing knowledge. The main goal is to induce a hierarchy from the tags used for annotating the images with or without additional information from the context. Thus, the following options are considered:

- Background knowledge coming from:
  - Wordnet
  - Existing ontologies from the semantic web (on-line/off-line)
- No background knowledge.

During this process, questions such as how to represent, how to store and how to query the ontology should been answered.

Background Knowledge

Brewster et al. [62] state that the term background knowledge has been used loosely across a range of academic disciplines without receiving a precise definition. The most accurate definition that can be obtained from a dictionary is "information that is essential to understanding a situation or problem".

- WordNet [63] is a thesaurus created at Princeton University that organize its 90,000 English terms into synsets (groups of words that are synonymous with each other). Synsets are linked among them using hypernymy and hyponomy relationships forming a hierarchical semantic network.
- Re-using ontologies from the Semantic Web. In his position paper, Alani [64] outlines a method for creating automatically an ontology reusing existing on-lines ontologies. The fist step in the process is to write down a list of terms that represent the domain that is going to

be represented. The idea is to post these terms to Watson [65], a Semantic Web search engine, in order to retrieve some ontologies. After evaluating the results, some parts of the ontology will be extracted using segmentation techniques. Finally, these fragments will be merged together in order to create the final ontology. Specia and Motta [66] propose an approach for making explicit the semantics behind the tags used in social tagging systems such as delicious and Flickr using ontologies provided by the Semantic Web. Their approach is based on the work [67] done by Schmitz who obtained some promising initial results in inducing ontology from the Flickr tag vocabulary using a subsumption-based model.

No Background Knowledge

In this case, the process will consist in inducing a hierarchy among the terms using clustering techniques following a categorization approach. Each term is considered as a class and the final goal will be to cluster the terms into semantic categories.

4.4 Ontology Knowledge Extraction

Apart from annotation purposes another advantage of using an ontology is extracting the knowledge of the domain that is being categorized. The simplest approach is to measure the relationship among each pair of concepts (classes) that conforms to our ontology. This relationship can be estimated by relatedness or similarity measures. A similarity measure can be viewed as a kind of relatedness. In our case the approach to follow is the semantic relatedness between each pair of terms. Pedersen at al. [68] propose several general metrics such as:

- Baselines like path length which deals with finding the shortest path between concepts in a is-hierarchy.
- Path based measures such as:
  - distance to root in is-a hierarchy
  - shortest is-a path between concepts, scales by depth of taxonomy
  - upward, downward and horizontal paths using many relations
- Information content measures such as Resnik, Information Content (IC) of shared concept, IC of shared concept scaled by individual concept ICs, sum of individual ICs minus shared IC

- Gloss based measures such as:
  - Original Lesk
  - Extended gloss overlaps
  - Gloss vector

Gracia et al. [69] accomplish a similar work but generalizing to the case of several ontologies. Thus, this approach is close to Word Sense Disambiguation (WSD) which is the process of assigning a meaning to a particular word based on the context in which it occurs.

Stokoe at al. [70] describe a system that performs sense based IR which improve the precision over the standard term based vector space model.

In order to evaluate our results we can check with the Normalised Google Distance (NGD), a method created by Cilibrasi and Vitanyi [71] that is able to automatically extract the meaning of words from the WWW using Google page counts. In addition to that they conducted a massive experiment able to understand WordNet categories.

Chapter 5

Ontologies & Automated Image Annotation

Some examples on how to join these two fields, one belonging to the Computer Vision and the other to the Semantic Web can be found in the literature. Soo et al. [29] propose a framework that can facilitate image retrieval based on a sharable domain ontology and thesaurus. Apart from that, they use case-based learning using a natural language query parser to translate a natural query into query in RDF format. The parser is able to perform semantic annotation on the descriptive metadata of images and convert metadata automatically into RDF representation. Images are retrieved by matching the semantic and structural descriptions of the user query with the annotations. The collection used is a set of historical and cultural images that have been taken from Dr. Ching-chih Chens "First Emperor of China" CD-ROM defined and derived from a set of domain concepts. The ontology used is a Mandarin Chinese thesaurus. Hare at al. [73] name the mechanism of generating automatically semantics for multimedia entities as bottom-up approach in opposition to the top-down approach that consists of annotating images using ontologies. He considers that a combination of both approaches can lead to bridge the semantic-gap. Srikanth et al. exploit ontologies for achieving an automated annotation of images. [74]. They propose some methods that use a hierarchy defined on the annotation words in order to improve the performance of the annotation of translation models. The technique that uses for improving the automatic annotation of images is based on translation models. The effect of using the hierarchy in generating the visual vocabulary is demonstrated by improvements in the annotation performance of translation models. They use WordNet and the Corel collection of data. Saathoff et al. [16] propose an architecture for automated annotation of multimedia content that is independent

of specific algorithms but uses ontologies enriched with low-level features to label regions in images with semantic concepts. The ontology used is VDO (Visual Descriptor Ontology) modeled in RDFS. Some challenges should be considered:

- Different annotator might use different ontologies so they will have different annotations in the end.
- It is not a trivial task to translate the user query into semantic schema and this requires to have a significant amount of domain knowledge.
- Matching a query instance with each annotated image description can be extremely inefficient and tedious, above all, if the collection of images is rather large.

Chapter 6

Identification of Gaps

An interdisciplinary approach able to combine automated image annotation techniques with the use of background knowledge or highly structured annotations is needed.

However, the approach proposed in this thesis is slightly different because the ontologies are not used to annotate the content of the images as some authors propose but to aid with the process of auto-annotation itself.

The objective of this research is to enhance the automated annotation of images using ontologies.

The following gaps have been detected from the literature analysed in previous sections:

- Need of efficient image retrieval systems.
- Traditional search engines based on image processing techniques whose main drawbacks are low retrieval precision and difficulty to formulate an exact feature query.
- Need to combine the visual features of an image with the information extracted from its context that constitute the metadata.
- Semantic gap that is the lack of coincidence between the information extracted from the visual data and the interpretation that the same data have for the user in a given situation.
- Annotation is considered as an intermediate step to image retrieval.
- Solutions using keywords as metadata are rather poor.

- Difficulty in dealing with the semantics of an image, need to express the textual annotations as highly structure metadata.
- Lack of multimedia ontologies.
- Problem of annotation relying on the knowledge domain, if someone changes the ontology the results might be different.
- Manual annotation is a time consuming task that sometimes can not be accomplished because of the large size of image collections.
- Researchers have focused their efforts mostly on solving the problem of automated annotation of images using statistics without considering an interdisciplinary approach.
- Difficulty in dealing with an interdisciplinary solution that involves Information Retrieval, Computer Vision and Semantic Web.

Chapter 7

Hypothesis and Research Questions

The hypothesis formulated in this thesis arises from the limitations identified in the actual automated image annotation systems that constitute an intermediate stage for an image retrieval framework.

The proposed solution intends to decrease the semantic gap by using a combination of state of the art statistical techniques together with global visual features and ontologies to categorize the image content.

Based on the above discussion, the hypothesis to be tested is whether the use of ontologies increases the precision in the automated image annotation process and consequently improves image retrieval performance. The research in this thesis aims to contribute to the general research question:

Can the precision of statistical machine learning system that uses image analysis techniques be enhanced by using ontologies as background knowledge?

This question is divided into several research questions:

Q1. Does an improvement in the annotation necessarily yield an improvement on the retrieval?

Q2. What kind of background knowledge is necessary to use in order to achieve this goal?Which requirements should fulfill the ontology or the set of ontologies?

The immediate questions that come up is how to select the ontologies that better describe the content of a collection of images and whether these ontologies can be generalised to any domain.

Q3. How to extract the knowledge contained in the ontology and how to

use it?

Relatedness or semantic similarity measures should be considered among different concepts in the ontology.

Q4. How beneficial is this enrichment?

An ontology can be used to increase the number of words used during the annotation stage or as a framework for the automated annotation process.

Q5. Which is the best way to measure the effect of the use of ontologies in automated image annotation and retrieval?

Chapter 8

Corel Ontology

The starting point for the ontology is the collection of terms used for labelling the Corel Stock Photo Library. This vocabulary is made up of 374 visual terms such as:

```
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 ...
```

The goal of this ontology is to promote data interoperability, information search and retrieval, automated inferencing, and natural language processing. The ontology is defined in OWL language, a W3C Recommendation since 2004 and is designed following a top-down approach using the vocabulary of terms.

8.1 Suggested Upper Merged Ontology: SUMO

The main purpose of an Upper Ontology is to describe concepts that are meta, generic, abstract or philosophical, and hence are general enough to address at a high level a broad range of domain areas. However, concepts specific to particular domains are not included in an Upper Ontology but such an ontology provides a structure upon which ontologies for specific domains can be constructed. We select the Suggested Upper Merged Ontology (SUMO) which was created by the IEEE Standard Upper Ontology Working Group [75].

Despite the fact that SUMO was initially developed as a variant of KIF (a version of the first-order predicate calculus) it has been translated into various representation formats including OWL.

8.2 Defining the hierarchy: classes and subclasses

The hierarchy is intended to represent the various levels of generality of the vocabulary of terms. Based on their semantics the terms can be placed into an hierarchy of concepts although taking into account as well the visual similarities among members of the same group. As a starting point SUMO hierarchy has been used for the upper branches of the ontology but with the addition of more categories in the lower parts in order to be more specific. Thus, the achievement of a good domain knowledge of some categories has been a requirement in order to develop the hierarchy. For instance, for the category "animal" a clear understanding of the different families such as Anthozoa, Arthropod, Mammal, Reptile, etc. is necessary in order to place each term into the right position. The taxonomy created by Parr et al. [76] has been considered as an aid for defining the inner branches of the class "animal".

The main branches of the ontology come from the class "entity", which are:

```
Abstract Physical Process
```

The following step is to add subcategories until the final leaves of the ontology, which correspond to the vocabulary of terms, are reached.

Protege, an ontology editor tool, has been used to create the hierarchy by adding subclasses or subcategories to the main branches of the ontology.

8.3 First Version

At the end of my first year of PhD, a first version of the Corel Ontology was released. All the words of the vocabulary were contained in the ontology as classes or as the field *label* of the classes. The classes were denoted with words starting with an upper-case letter while in the vocabulary everything appears as lower case. Some words from the vocabulary were in plural so in that case, the corresponding class was named after the word in singular while the associated *label* field contained the real word of the vocabulary.

The final goal of this ontology was to prune irrelevant words in the annotation framework. In other to achieve this, relationships between each pair of terms were going to be added. Some examples of the relationships are the following:

- Polar "lives in" Arctic
- Camel "lives in" Desert
- Desert "has got" Dune
- Waves "is part of" Sea

For example, assume that the annotation framework has annotated an image with the keywords "polar desert ice", each pair of words (polar, desert)(polar, ice) and (desert, ice) are checked against the ontology in order to infer that *"a polar bear can_not_live in the desert"* and *"ice is_not_part_of the desert"*. As a consequence of this, "ice" will be removed from the set of annotations of the current image.

Unfortunately, the properties between classes were never fully implemented as it was impractical to perform it manually. Finally, this version contained some error such as mispellings or missing words that were corrected in the future version.

8.4 Current Version

The starting point for this ontology is the initial one where the relationships between were removed and the typos errors corrected. After writing a Java application called "ontologyPopulator", the current version is generated. In this version, the classes of ontology has got instances associated to them that match completely the words of the vocabulary. Relationships were added (See Figure) between each pair of instances with a value, obtained from the co-occurrence matrix explained in Section 10.1, that gives an estimation of the semantic similarity o dissimilarity between two terms. This ontology is attached in the annex of the present document.

Chapter 9

Baseline Experiment

This experiment consists in the replication of the work carried out by Yavlinsky et al.[42] as their work will be adopted as our annotation framework. Our target is to find ways to improve the accuracy of the previous method.

9.1 Feature Extraction

The features used in their experiment were a combination of colour feature, CIELAB, and texture feature, Tamura. CIE $L*a*b*$ (CIELAB) [77] is the most complete 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 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. [78] as the most fundamental texture feature and finally, directionality is a global property over a region. The process for extracting each feature is as follows, each image is divided into nine equal rectangular tiles, the mean and second central moment feature per channel are calculated in each tile. The resulting feature vector is obtained after concatenating all the vectors extracted in each tile. The feature information of this work has been obtained using a tool called "annotate" which generates low level (colour and texture) feature vectors.

9.2 Corel 5k Dataset

Over the past years Corel has been collecting images by photographers from around the world to create the Corel Stock Photo Library. This collection has grown to over 60,000 images covering an amazing array of topics such as wildlife, rural Africa, sunrises and sunsets, etc.

In this work, we use a subset of 5,000 images extracted from this collection in order to compare our results with previous experiments carried out by Duygulu et al. [31] which has turned out to be a reference benchmark in the literature for automated annotation of images. The collection of images is partitioned into two groups:

- A training set of 4,500 images
- A test set of 500 images

Each image has been annotated with a set of keywords ranging from three to five. These keywords constitute a vocabulary of 374 terms such as:

```
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 ...
```

9.3 Estimation of the probabilities

The main goal of this experiment is to build up a model by learning from the annotations of the training set that will enable us to guess the annotation keywords for each image of the test set. For each one of the keywords that belong to the vocabulary we build up a model which is called non-parametric model of distribution of image features.

Finally, images of the test set are analyzed and after applying the previous model the probabilities of each keyword being present in an image are estimated.

The final result is a *nxm* matrix P, where n is the number of images in the test set and m is the number of terms in the vocabulary, and the value of each element *P(i,j)*, representing the intersection of an image of the test set with a word of the vocabulary. The value of each cell yields the probability of the word being present in the image:

$$\begin{pmatrix} city(0.018015) & bear(0.018356) & palm(0.044598) & ... & sun(0.053039) \\ city(0.009025) & bear(0.001158) & palm(0.012742) & ... & sun(0.007209) \\ city(0.000079) & bear(0.000276) & palm(0.000004) & ... & sun(0.163624) \\ ... & ... & ... & ... & ... \\ city(0.018015) & bear(0.000593) & palm(0.044598) & ... & sun(0.001158) \end{pmatrix}$$

The final annotations for each image are calculated by selecting the five words with the highest probability values. This implies sorting each row in the matrix for, finally, selecting the five firsts values.

9.4 Evaluation the results

The average value of the precision across all keywords is called *mean average precision*, and is used to evaluate the quality of the annotation algorithm. Queries made up of a combination of up to three keywords are to be made in order to try to retrieve the images annotated by them. For each keyword in the vocabulary, the images are ranked according to the recorded probability of that keyword, and the average precision of this ranking is calculated based on the manual annotations of the test images. Once, these values are obtained it will be easy to compare with other results contained in the literature of automated image annotation.

Chapter 10

Co-occurrence Data

Hofmann and Puzicha [80] determine the general setting described by the term *co-occurrence data* as follows. Suppose, two finite sets $X=\{x_1, x_2, ...,x_n\}$ and $Y=\{y_1, y_2, ...,y_m\}$ of abstract objects with arbitrary labelling, are given. As elementary observations pairs $(x_i, y_i) \in XxY$, that is, a joint occurrence of object x_i with object y_i. All data is numbered and collected in a sample set $S = \{(x_i(r), y_i(r), r) \text{ with } 1 \leq r \leq L\}$ with arbitrary ordering. The information in S is completely characterised by its sufficient statistics $n_{ij} = |\{(x_i, y_i, r) \in S\}|$ which measure the *frequency of co-occurrence* of x_i and y_i.

Depending on the displicine applied, there will be different interpretations. In *Computer Vision* X may correspond to image regions or the whole image and Y to features values. Likewise, in *Information Retrieval* X may correspond to a collection of documents and Y to a set of keywords. Hence n_{ij} denotes the number of occurrences of the word y_i in the document x_i.

The intrinsic problem of *Co-occurrence Data* is the sparseness of the data. 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. Typical state-of-the-art techniques in NLP apply *smoothing techniques* to deal with zero frequencies of unobserved events. Some techniques are, for example, *the back-off method*, *model interpolation* and the *similarity-based local smoothing*. An empirical comparison of smoothing techniques can be found in [81]. In *Information Retrieval* the proposals to deal with the sparseness of the data are, *cluster hypothesis*, Salton's *Vector Space Model* and *latent semantic indexing*.

According to *Fuzzy Set Theory* (when applied to *Information Retrieval*) [82], the degree of keyword co-occurrence in a textual dataset is a measure of the semantic relatedness and can be used to build a thesaurus. By analogy with our research, the thesaurus will be the whole collection where

	city	**mountain**	**sky**	**...**	**race**	**hawaii**
$\mathbf{I_1}$	0	0	0	-	1	0
$\mathbf{I_2}$	1	0	0	-	1	0
$\mathbf{I_3}$	1	0	1	-	0	0
...	-	-	-	-	-	-
$\mathbf{I_n}$	0	1	1	-	0	0

Table 10.1: Image-term Matrix

each entry will correspond to an image and the annotations will be the set of related keywords. A thesaurus can be constructed by defining a *co-occurrence matrix C*. In this matrix, c_{12}, the *normalised termed correlation index* between two keywords k_1 and k_2 can be defined by:

$$c_{12} = \frac{n_{12}}{n_1 + n_2 - n_{12}}$$

Where n_1 and n_2, are the number of images that contains the keyword k_1 and k_2 respectively, while n_{12} corresponds to the number of images containing both of them. Aditionally:

- $c_{12} = 0$ when $n_{12} = 0$; i.e., k_1 and k_2 do not co-occur (terms are mutually exclusive).
- $c_{12} > 0$ when $n_{12} > 0$; i.e., k_1 and k_2 co-occur (terms are non mutually exclusive).
- $c_{12} = 1$ when $n_{12} = n_1 = n_2$; i.e., k_1 and k_2 co-occur whenever either term occurs.

Thus, c_{12} oscillates between 0 and 1. Term correlation increases as c approaches 1.

10.1 Building a co-occurrence matrix

The starting point for calculating a co-occurrence matrix, given a vocabulary of terms, is building up an *image-term* matrix. For the Corel 5k dataset as described in Section 9.2, the *image-term* matrix A (Table 10.1) is a rectangular matrix of 4,500x374 dimension, where each row represents an image of the training set, each column a keyword from the vocabulary and each entry a_{ij} represents the number of times keyword j ocurrs in an

	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

Table 10.2: Co-occurrence Matrix

image I_i. Due to the fact that a keyword can only appears once in an image, this entry is binary. This matrix represents images as vectors in the *image-space*. Keywords are deemed similar to the extent that they occur in the same image. For instance, in the *image-space* the keywords "city" and "sky" and "mountain" and "sky" are similar as both share an image, the image I_3 and I_n respectively.

A *co-occurrence* matrix B (Table 10.2) is obtained after multiplying the *image-term* matrix A by its transpose ($B=A^TA$). The resulting *co-occurrence* matrix is a symmetric one (dimension 374x374) where each entry b_{ij} contains the number of times keyword i co-occurs with the keyword j. The elements in the diagonal represent the number of images annotated by each keyword. In this case, the matrix represents keywords as vectors in the *keyword-space*. Keywords are similar if they co-occurr with the same keywords. For instance, "city" is *semantically* similar to "sky" and to "race" although "sky" and "race" and not *semantically* similar as they never co-occurr together. However, different spaces yield different types of *semantic similarity*. A good definition of *semantic similarity* is provided in Section 11.6. This matrix B can be easily transformed into a matrix of *conditional probability* by dividing each element in a row by its L2 norm as suggested by Manning and Schtze in [83]. This process is equivalent to normalising the matrix using the Euclidean norm (L2 norm).

10.2 Word-to-word Co-occurrence and Automated Image Annotation

Escalante et al. propose in their paper [46] a Markov random field based on word co-occurrence information built on top of a k-nearest neighbour (k-NN) classifier (probabilistic annotation system) for improving the accu-

racy of automated image annotation. The k-NN is selected as a baseline for their experiments because it outperforms other state-of-the-art methods for automated image annotation. Despite the fact that they employ for their experiments a subset of the Corel dataset, they consider an external collection for extracting the co-occurrence information. The collection used is the IAPR-TC12 benchmark [84] provided by ImageCLEF evaluation conference. They use the captions of the 20,000 images of the collection in order to create a co-occurrence matrix. In order to calculate the matrix they count the number of documents in which two words from the vocabulary appear together. The captions are a few text lines provided by a human annotator indicating visual and semantic content. The only difference with this research is that they perform a smoothing technique *interpolation smoothing* over the matrix and they obtained conditional probabilities from the co-occurrence matrix in a different way. They consider that the conditional probability of two words are obtained by dividing each cell of the matrix by the number of documents in the external corpus. Additionally, they consider that if two words appear together in the caption of an image, they are visually related.

Another work similar to ours is that accomplished by Zhou et al. [44] who use as annotation framework the *Cross-Media Relevance Model* (CMRM) developed by Jeon at al. in [37]. They apply the *Automatic Local Analysis*, a method for performing query expansion in Information Retrieval as explained in [82]. They carry out some experiments with the Corel 5k dataset outperforming the state-of-the-art *Multiple Bernoulli Relevance Model* (MBRM) [40] improving the recall 21% and the precision 11% respectively.

Jin et al. [85] use as annotation framework, a *Translation Model*, [31] and *semantic similarity* as a way to prune irrelevant keywords within the Corel dataset. They measure the semantic similarity of the generated keywords by the annotation process, detect the noisy ones and discard them. They try different similarity measures using as knowledge base WordNet and manage to increase the accuracy of their resulting system.

Liu et al. [86] use a combination of correlation by WordNet and a correlation of statistical co-occurrence in order to expand the existing annotation image and to prune irrelevant keywords for each annotated image. Experiments conducted on the Corel dataset demonstrate the effectiveness and efficiency of their proposed solution.

Jin et al. [43] propose a new framework for automated image annotation that estimates the probability for a language model to be use for annotation an image. The use a word-to-word correlation which is taken into account through the *Expectation Maximisation* (EM) algorithm for finding optimal

language model for the given image.

10.3 Initial Experiments

Initially, we adopted a simple approach that intended to improve the accuracy of our annotation framework (See Chapter 9) by multiplying the matrices *P(i,j) x B'(j,k)*. Being *P(i,j)*, the probabilities matrix explained in Section 9.3 and *B'(j,k)*, the normalised co-occurrence matrix of Figure 10.2. As a result of this, a new matrix is obtained. Each row of the new matrix contains the modified probabilities for each one of the images of the test set. Again, the annotations for each image are obtained by sorting the probabilities and translating the five highest values into keywords. After realizing that the mean average precision was not improved, several corrections were tried such as changing the norm (using the Euclidean one) trying a naive smoothing technique that consisted in replacing each zero value by a small non-zero one. After several failed attempts, the method was discarded in favour of the proposed in Chapter 11.

Chapter 11

Using Second Order Statistics to Enhance Automated Image Annotation

11.1 Overview

The research challenge that we address in this work is to examine whether a traditional automated annotation system can be improved by using external knowledge. Traditional means any machine learning approach together with image analysis techniques. We use as a baseline for our experiments the work done by Yavlinsky et al. [42] who deployed non-parametric density estimation. We observe that probabilistic image analysis by itself is not enough to describe the rich semantics of an image. Our hypothesis is that more accurate annotations can be produced by introducing additional knowledge in the form of statistical co-occurrence of terms. This is provided by the context of images that otherwise independent keyword generation would miss. We test our algorithm with different datasets: Corel 5k, ImageCLEF 2008 and more recently TRECVID 2008. For the Corel dataset, we obtained statistically significant better results while our algorithm appeared in the top quartile of all methods submitted in ImageCLEF 2008. Regarding future work, we intend to apply Semantic Web technologies.

11.2 Motivation

The main challenge in automated image annotation is to create a model able to assign visual terms to an image in order to successfully describe it. The starting point for most of these algorithms is a training set of images that have already been annotated by humans. These annotations are unstructured textual metadata made up of simple keywords that describe the content of the images. Image analysis techniques are used to extract features from the images such as colour, texture and shape, in order to model the distribution of a term being present in the image. Features can be obtained from the whole image (global approach), from blobs, which are segmented parts of the image (segmentation approach) or from tiles, which are rectangular partitions of the image. The next step is to extract the same feature information from an unseen image in order to compare it with all the previously created models (one for each keyword). The result of this comparison yields a probability value of each keyword being present in the image which can be expressed as $p(keyword|image)$.

Several strategies can be adopted to produce the final output for these systems. One of them consists of an array of 1's or 0's, with the same length as the number of terms in the vocabulary, which indicates the presence or absence of the objects in the image. This is called *hard annotation* in contrast with *soft annotation*, which provides a probability score that gives some confidence for each concept being present or absent in the image. Other automated image annotation frameworks implement a strategy that assumes a fixed length for the annotations. For instance, if the length of the annotation is k, words with the top-k largest probability values are selected as annotations. Another way of achieving the same goal is to define a threshold that forces all the keywords with a higher probability than the threshold to be considered as annotations.

Independently of the method used to define the annotations, automated image annotation systems generate a set of keywords that helps to understand the scene represented in the image.

Finally, the performance of automated image annotation is measured by retrieving images that have been auto-annotated with respect to single-word queries. For each query, precision and recall are computed comparing the keywords generated by the system with the ground-truth or the annotations produced by human experts. The overall performance of the system [90] is estimated calculating the mean average precision and the number of words with recall greater than zero.

One limitation of these methods, as they generate keywords based on the correlation between words and image features, is the difficulty in dis-

tinguishing objects that are visually similar. How can the blue sea be distinguished from a blue sky? Without using additional information from the image context, a marble floor surface in a museum, as in Figure 11.1, can be confused with a layer of ice in the arctic because both of them have got similar colour and texture. Another limitation of traditional systems is that each word is produced independently from the other annotated words, without considering that these words represent objects that co-occur in the same scene.

11.3 Human understanding of a scene

Meaningful visual information comes in the form of scenes. Our 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. In addition to that, 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 recognize an object. In the previous example, if we had known that we were in a museum, we would have discarded the layer of ice in favour of the marble surface. On the other hand, the fact of knowing that, together with the unidentified object, there is a statue, it would have helped us to disambiguate the scene, and to think about a museum instead of the arctic.

An attempt to identify the rules behind the human understanding of a scene was made by Biederman in [91]. 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. Biederman introduces the notion of a *schema*, which is an overall representation of an image that integrates the objects and their relations. For example, the action of recognising a scene with "boat", "water" and "waves" (Figure 11.3) 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.

Biederman considers that a scene is characterised by five relations among objects:

- Support: Most objects rest on a surface.
- Interposition: An opaque object will occlude the contours of an object behind it.

- Probability: It refers to the likelihood of a given object being in a given scene.
- Position: Objects that are likely to occur in a given scene often occupy specific positions.
- Size: It applies to the relation between the dimensions of the objects in the image.

Our hypothesis is that the constraints derived from the schema of the image can be used to improve an automated image annotation algorithm. In this paper we mainly focus on the probability relation, while the use of other relationships is currently under consideration.

11.4 Limitations of probabilistic approaches

As a first step to understand what needs to be improved, we analysed different cases in which wrong keywords were assigned by a machine learning approach. The result of the study is the identification of two main categories of inaccuracies.

The first group corresponds to problems recognizing objects in a scene. This happens when a marble floor surface in a museum is confused with a layer of ice or when waves in the sea are taken for wave-like sand dunes in a desert. These problems are a direct consequence of the use of correlation between low-level features and keywords, as well as the **difficulty in distinguishing visually similar concepts**. One way to tackle these problems is to refine the image analysis parameters of the system, but this task is out of the scope of this work. Duygulu et al. also addressed these problems, suggesting that they are the result of working with vocabularies not suitable for research purposes. In their paper [31], 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", which depend on the features selected.

In the second group of inaccuracies we find different levels of incoherence among tags, that range from the improbability to the impossibility of two objects being together in the real world. This problem is the result of **each annotated word being generated independently** without considering their context.

Other inaccuracies come from the **improper use of compound names** in some data collections. Compound names are usually handled as two independent words. For instance, in the Corel 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 never appear apart sufficiently often in the learning set, the system is unable to disentangle them. Methods for handling compound names can be found in the work done by Melamed [92].

Finally, it is important to mention the **over-annotation** problem. This situation happens when the ground-truth is made up of less words than the annotations. One example is shown in Figure 11.4 where the ground-truth is "bear","black", "reflection" and "water", although the annotation system assigns additionally the word "cars". Over-annotation decreases the accuracy of the image retrieval as it introduces irrelevant words inside the annotations. This problem was also detected by Jin et al. [43] who proposed a system with flexible annotation length in order to avoid the over-annotation.

Our work attempts to overcome the limitations of words being generated independently by applying statistical analysis techniques. In order to go from low-level features to the high-level features (semantics) of an image, semantic constraints should be considered, such as relations among entities and likelihood of each entity being present in a given scene.

11.5 Examples of inaccurate annotations in the Corel dataset

In the following we provide some key examples produced by the annotation system developed by Yavlinsky et al. [42]

Inaccuracy or Imprecision: Inaccuracy or imprecision occurs in automated image annotation systems when wrong annotations have been generated for a given image. The examples below represent the worst case, where none of the keywords produced by the system represent any object in the image.

In Figure 11.1, the scene represents a museum with some pieces of art in the background. A marble floor surface is confused with a layer of ice and a piece of art such as a statue with a bear. The ground-truth is: "art", "museum" and "statue".

In Figure 11.2, the scene describes a sunset at a piece of land bordering the sea; some houses can be distinguished in the background of the image. The ground-truth is: "house", "shore", "sunset" and "water". However, waves in the water are taken for dunes (similar texture), houses for hills, sunset for sand (same colour). Due to the fact that the system is unable

Figure 11.1: System annotations: “snow”, “water”, “ice”, “bear”, “rocks”.

Figure 11.2: System annotations: “sky”, “hills”, “dunes”, “sand”, “people”.

to decipher the scene represented in the image, this kind of inaccuracy can not be resolved by modelling the context. Processing these nonsensical data with statistical co-occurrence information will produce a nonsensical output.

Incoherence and Improbability: Incoherence and improbability appears when there is a lack of cohesion among annotation words.

In Figure 11.3, the image constitutes an example of incoherence. The image shows a boat in the water making waves. The ground-truth is: “boats”, “water” and “waves” while the system has assigned the following words: “water”, “desert”, “valley”, “people” and “street”. Clearly, there are some keywords that are incoherent. A “street” can not be in the “desert”. “Water” is normally not found in a “desert”. “Valley” is *a low area between hills or mountains, typically with a river or stream flowing through it*, according to the Oxford Dictionary of English [93]. Thus, the following incompatibilities between words are found: desert-valley; desert-water; desert-street.

The final example of Figure 11.4 represents an improbability. This situation happens when there is a certain unlikelihood of some keywords appearing together. The image shows the reflection of a black bear on the water. The ground-truth is: “bear”, “black”, “reflection” and “water”. The annotations match the ground-truth except for the last word (“cars”).

Figure 11.3: System annotations: “water”, “desert”, “valley”, “people”, “street”.

Figure 11.4: System annotations: “water”, “bear”, “black”, “reflection”,“cars”.

In this case, there is an unlikelihood between “cars” and “bear” because the probability of both terms appearing together in the real life is really low.

11.6 Semantic similarity

Our work aims at overcoming the limitations of automated image annotation about words being independently generated with respect to each other. In order to guarantee that all the keywords that annotate an image are coherent between each other we consider that, as they share the same context (which is the scene depicted in the image), they share a certain degree of semantic similarity.

Among all the many uses of the concept semantic similarity we refer to the definition by Miller and Charles [94] who consider it *“as the degree of contextual interchangeability or the degree to which one word can be replaced by another in a certain context”*. Consequently, two words are similar if they refer to entities that are likely to co-occur together like “forest” and “tree”, “sea” and “waves”, “desert” and “dunes”, etc.

Semantic similarity can be represented in several ways using ontologies

(topological similarity) or using statistical analysis techniques such as vector space models to correlate words and contexts (images). Due to its simplicity, our algorithm uses the approach of vector space models. In the vector space model words are represented as vectors. Different spaces can be considered, although we focus on the word space.

We take advantage of the analogy between natural language and image retrieval based on textual searches. The starting point is an image-term matrix. This matrix is a rectangular matrix of 1500 x 374 dimension, where each row represents an image of the training set and each column a term of the vocabulary. Each cell indicates the presence or absence of a term in an image. A co-occurrence matrix is obtained after multiplying an image-term matrix by its transpose. The resulting co-occurrence matrix is a symmetric matrix A where each entry aij contains the number of times word i co-occurs with word j. The elements in the diagonal represent the number of times a word annotates an image. This matrix A is transformed into a conditional probability distribution after being normalised, dividing each element of a column by its Euclidean norm as suggested by Manning and Schtze in [83].

11.7 Experiments

We use as a baseline for our experiments the framework developed by Yavlinsky et al. who used global features together with a non-parametric density estimation. Their experimental set-up [42] is similar to that defined by Duygulu et al. [31], which is considered a benchmark for automated image annotation systems. The algorithm of Yavlinsky et al. was tested on a dataset of 5,000 images from 50 Corel Stock Photo CDs that comprises a training set of 4,500 images and a test set of 500 images. Images of the training set were annotated by human experts using a set of keywords ranging from three to five from a vocabulary of 374 terms. Low-level features, CIELAB [77] colour and Tamura [78] texture, were extracted and combined after segmenting the images into nine equal tiles. The output of their experiment is a set of five tags per image that correspond to the keywords with highest probability. The system was evaluated on a subset of 179 keywords that were selected based on their capacity for annotating more than one image from the test set. The evaluation measures achieved showed state-of-the-art performance for the Corel dataset as evidenced in a review by Magalhães and Rüger [95].

Description of the algorithm

The input for our algorithm are the top five keywords and their associated probabilities per image generated by the framework of Yavlinsky et al. [42].

The context of the images is computed using statistical co-occurrence of pair of words appearing together. This information is represented as a co-occurrence matrix described in Section 11.6.

The algorithm used to enhance image annotation is the following:

```
For each image I in Testset:
  if Probabilities(I) > threshold1:
    For all pairs of keywords A, B in best5KeywordsOf(I):
      if dissimilar(A, B):
        LowerProbability(B)
        For each keyword C in keywordSet:
          if keywordRelatedTo(B, C):
            LowerProbability(C)
```

Our system is tailored for the cases when there is incoherence or improbability among the annotation keywords. If the system is unable to appropriately annotate the image, like in the example of the statue, there is no space for improvement using our algorithm. Consequently, our algorithm takes into consideration only the images for which the underlining system is *"confident enough"* i.e. at least one of the keywords has greater probability than a threshold (*threshold1*) which is estimated empirically. Then, system checks all the annotations in order to detect incoherence between each pair of keywords with the help of the correlation matrix. This is achieved with the help of the function the function *dissimilar(A,B)*. Note that the first parameter A of this function has always a greater probability than the second argument B. We consider that two terms are *semantically dissimilar* if the correlation value is lower than a threshold (*threshold2*), which we estimated empirically. If the system finds that the keywords A and B are incoherent, it will lower the probability of the keyword associated to the lowest probability (B). Furthermore, the probability of each keyword C semantically similar to B is also lowered. On the contrary, two terms are *semantically similar* or related if the correlation value is greater than a threshold (*threshold3*), which again was estimated empirically. After modifying the probability values of some keywords, new and more precise annotations are produced.

11.8 Results: Corel dataset

In the first place, we adopted a standard annotation database, the Corel 5k dataset, because it is considered a benchmark for comparing algorithms in automated image annotation. We evaluated the performance of our algorithm (*Enhanced Method*) comparing it with the deployed by Yavlisnky et al. (*Trad. Method*) under two different metrics, the image annotation and the ranked retrieval. Under the image annotation metric, automated image annotation is defined as the top five annotation words assigned by the algorithm. Recall and precision of every word in the test set are computed. This metric is based on comparing the keywords automatically generated for the test set with the human-produced ground-truth ignoring rank order. Carneiro et al. explained in their paper [96] how to calculate this metric. For a given word, assuming that there are *Wh* human annotated images in the test set and the system annotates *Wauto*, of which *Wc* are correct, the per-word recall and precision are given by recall = Wc / Wh and precision = Wc / Wauto, respectively. Finally, the values of recall and precision are averaged over the set of words that appear in the test set to get the **average per-word precision** *P* and **average per-word recall** *R*. The number of **words with non-zero recall** *NZR*, words with *Wc* greater than zero, is also considered as it provides an indication of how many words the system has effectively learned.

The performance of rank retrieval is also evaluated by measuring precision and recall. Given a query term and the top n image matches retrieved from the database, recall is the percentage of all relevant images contained in the retrieved set, and precision is the percentage of n which are relevant. Relevant means that the ground-truth annotation of the image contains the query term. Under the ranked retrieval metric, 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. MAP is calculated on a subset of 179 keywords that were selected based on their capacity for annotating more than two images from the test set. A comparison of the results using both methods is presented in Table 11.1.

The mean average precision (MAP) of our algorithm is 0.2922 which gives statistically significant better results than the value obtained by Yavlinsky et al., which were comparable to state-of-the-art automated image annotation. In order to demonstrate the statistically significant better results we ran a sign-test [97] comparing the average precision per query (word) of both methods and we obtained 1% level of significance. Interestingly, our algorithm is able to increase the number of words with non-zero recalling

Metric 1	**Trad. Method**	**Enhanced Method**
Words with NZR	86	91
Precision	0.1036	0.1101
Recall	0.1260	0.1318
Metric 2	**Trad. Method**	**Enhanced Method**
MAP	0.2861	0.2922

Table 11.1: Comparative results for the Corel dataset

Figure 11.5: A booby bird.

from 86 to 91 as well as the precision and recall under Metric 1.

11.9 Enhanced annotations for Corel dataset

In Figures 11.5, 11.6, 11.7 and 11.8, we show some examples of the better performance achieved by our algorithm with the Corel 5k dataset.

Figure 11.5 represents a very successful example. The initial annotations obtained by a traditional method are: "birds", "snow", "nest", "rodent" and "rabbit". There is a semantic similarity between pair of words such as birds-nest and rodent-rabbit. The system checks all the probability values and selects the only word with a probability greater than the threshold, which is "birds"; the rest are rather low and consequently their value is decreased by the algorithm. As a result of this, new words come up like "water", "booby" and "flight". The only wrong word that is not pruned is "snow".

The initial annotations of Figure 11.6 are: "water", "sky", "beach", "mountain" and "valley". "Water" and "sky" are kept as annotations because they present high probabilities however "mountain" and "valley" are

Figure 11.6: A sand castle.

pruned by the system because they have low probabilities and are related. These words are replaced by the successful "sand" and "people".

The image represented in Figure 11.7 is located in Kauai, one of the islands belonging to Hawaii. The initial words detected are "sky", "clouds", "ice", "tower" and "stick". From all of them, those which the highest probability are maintained while "ice", "tower" and "stick" are discarded. The final annotations contain two mistaken words: "ruins" and "stone". The last example shows a landscape characteristic of Scotland. All initial generated annotations are right: "scotland", "water", "mountain" and "sky" except "train" which is pruned due to its low probability. Finally, "hills" is successfully detected.

Figure 11.7: Kauai.

Figure 11.8: Scotland.

11.10 Other datasets: ImageCLEF 2008 and TRECVID 2008

Our algorithm was also tested with the collection of images provided by ImageCLEF 2008 for the Visual Concept Detection Task (VCDT) in [98]. This collection [84] was made up of 1,800 training images and 1,000 test

Metric 1	**Trad. Method**	**Enhanced Method**
EER	0.288186	0.284425
AUC	0.776546	0.779423
Metric 2	**Trad. Method**	**Enhanced Method**
MAP	0.588489	0.589168

Table 11.2: Comparative results of ImageCLEF 2008

images, taken from locations around the world and comprising an assorted cross-section of still natural images. The results are presented under the evaluation metric followed by the ImageCLEF organisation which is based on ROC curves and under the image annotation metric. ROC curves [99] 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. The area under the ROC curve, AUC, is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one.

The results obtained are represented in Table 11.2.

Such good results were obtained because the collection uses a vocabulary of 17 words which denotes concepts quite general, such as “indoor”, “outdoor”, “person”, “day”, “night”, “water”, “road or pathway”, “vegetation”,“tree”, “mountains”, “beach”, “buildings”, “sky”, “sunny”, “partly cloudy”, “overcast” and “animal”. In addition to that, our algorithm performed rather well appearing in the top quartile of all methods submitted in ImageCLEF 2008, however it failed to provide signicant improvement over the automated image annotation method. An explanation for this can be found in the small number of terms of the vocabulary that hinders the functioning of the algorithm and in the nature of the vocabulary itself, where instead of incoherence we have mutually exclusive terms and almost no semantically similar terms.

Regarding the TRECVID video retrieval benchmarking event [58], we participated in the High-level Feature Extraction task. During this task, participants were entitled to return for each high-level semantic feature a list of at most 2000 shots from the test collection of videos (Sound and Vision data), ranked according to the highest possibility of detecting the presence of the feature. The features were a list of 20 concepts drawn from the large LSCOM [57] feature set. Unfortunately, it is too early to draw any conclusions about our performance in TRECVID as we are still waiting for the evaluation results that are going to be provided by the organisation.

11.11 Conclusions and Future work

The main goal of this work is to improve the accuracy of a traditional automated image annotation system based on a machine learning method. We have demonstrated that building a system that models the image context on top of another that is able to accomplish the initial identification of the objects increases significantly the mean average precision of an automated annotation system. Experiments has been carried out with three datasets, Corel 5k and ImageCLEF 2008 and TRECVID 2008. Our algorithm shows that modelling a scene using co-occurrence values between pair of words and using this information appropriately, help to achieve better accuracy. However, it only obtained statistically better results than the baseline machine learning approach in the case of the Corel dataset where the vocabulary of terms were big enough. An explanation for this can be found in the small number of terms of the vocabulary that hinders the functioning of the algorithm. This has sense as a big vocabulary allows us to exploit properly all the knowledge contained in the image context. This is in tune with the opinion of most researches [100] as they believe that hundreds or thousands of concepts would be more appropriate for general image or video retrieval tasks.

Another important conclusion is the nature of the vocabulary, if it is quite general like in the case of the ImageCLEF 2008, the accuracy increases notably. On the other hand, the vocabulary used for annotating the Corel dataset is much more specific and consequently the algorithm decreases its accuracy as it needs to be precise enough to distinguish between animals belonging to the same family such as “polar bear”, “grizzly” and “black bear”.

Chapter 12

ImageCLEF 2008

ImageCLEF is the cross-language image retrieval track run as part of the Cross Language Evaluation Forum (CLEF) campaign. It evaluates retrieval of images described by text captions based on queries in a different language; both text and image matching techniques are potentially exploitable. In this paper, we describe the experiments of the MMIS group at ImageCLEF 2008 where we participate in the following tasks: Visual Concept Detection Task (VCDT), ImageCLEFphoto and ImageCLEFWiki. All experiments were performed in a single framework of independently testable and tuneable modules as seen in Figure 12.1.

12.1 Visual Concept Detection Task

The dataset used in this task is a subset of the IAPR TC-12 Benchmark [84] which consists of 20,000 still natural images taken from locations around the world and comprising an assorted cross-section of still natural images. This includes pictures of different sports and actions, photographs of people, animals, cities, landscapes and many other aspects of contemporary life. The objective of the VCDT is to detect the presence or absence of 17 visual concepts in the 1,000 images that constitute the test set, given a training set of 1,800 images. The training set is classified according to the concept hierarchy of Figure 12.2 along with their classification scores (0 when absent and 1 when present). It is the only data that can be used to train concept detection/annotation techniques. Once an object is detected in an image of the test set, some confidence scores are provided, the higher the value the greater the confidence of the presence of the object in the image. This task will help in solving the photographic retrieval task of ImageCLEF 2008.

As shown in Figure 12.2, our vocabulary is made up of 17 visual concepts

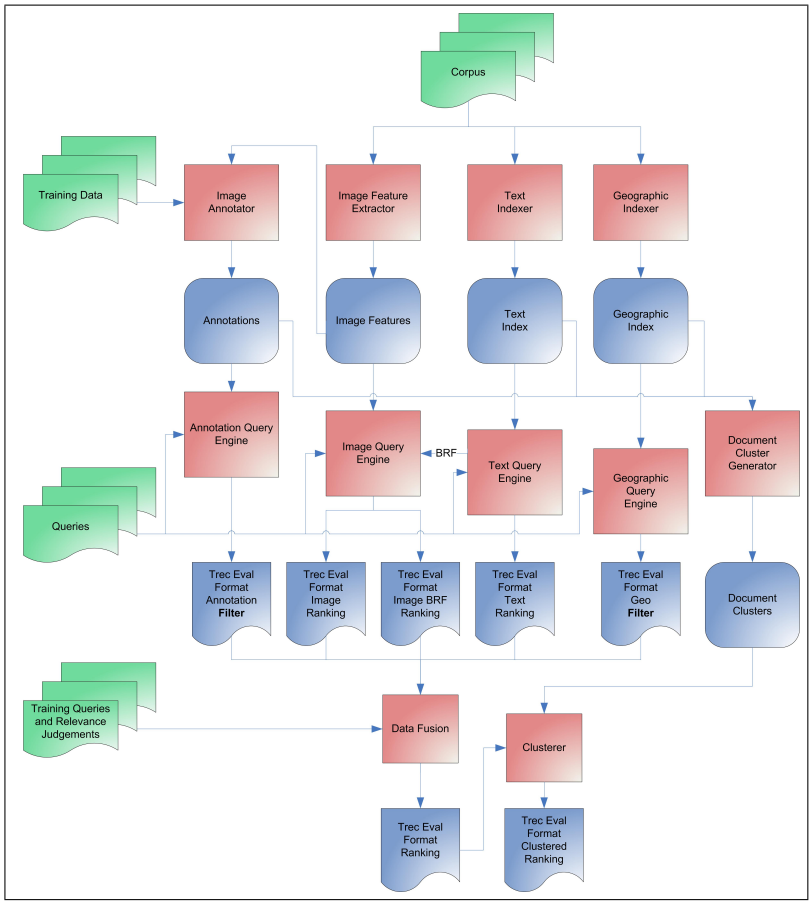

Figure 12.1: Framework developed by MMIS

and adopts a hierarchical structure. In the first level, we find two general concepts like "indoor" and "outdoor" which are mutually exclusive while in lower levels of the hierarchy we find more specific concepts that are subclasses of the previous ones. Some concepts can belong to more than one class, for instance, a "person" can be part of an "indoor" or "outdoor" scene but others are mutually exclusive, a scene can not represent "day" and "night" at the same time.

For the VCDT task, we submitted four different runs (an algorithm per run), all of them corresponds to automatic runs dealing with visual information. The algorithms that obtained better results are described below:

- Traditional Algorithm: this algorithm corresponds to the work car-

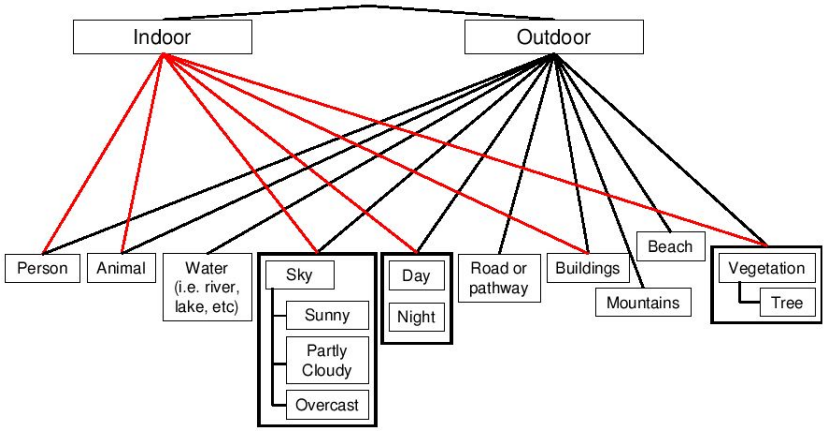

Figure 12.2: Hierarchy of the visual concepts

ried out by Yavlinsky et al. [42]. Their method is based on a supervised learning model that uses a Bayesian approach together with image analysis techniques. The algorithm exploits simple global features together with robust non-parametric density estimation using the technique of kernel smoothing in order to estimate the probability of the words belonging to a vocabulary being present in each one of the images of the test set. This algorithm was previously tested with the Corel dataset.

- Enhanced Algorithm: this second algorithm is described in detail in Section 11.7. The input is the annotations achieved by the algorithm developed by Yavlinsky et al. together with a matrix 17 x 17 that represents the probabilities of all the words of the vocabulary being present in the images.

The third algorithm uses image similarity measures while the fourth one is a combination of the results achieved by the other three.

Evaluations and Results

The evaluation metric followed by the ImageCLEF organisation is based on ROC curves. Initially, a receiver operating characteristic (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 [99] were applied to information retrieval in order to represent the fraction of true positives (TP) against the fraction of false positives

Algorithm	EER	AUC
Enhanced automated image annotation	0.284425	0.779423
Automated image annotation	0.288186	0.776546
Combined algorithm	0.318990	0.736880
Dissimilarity measures algorithm	0.410521	0.625017

Table 12.1: Comparative results of MMIS group

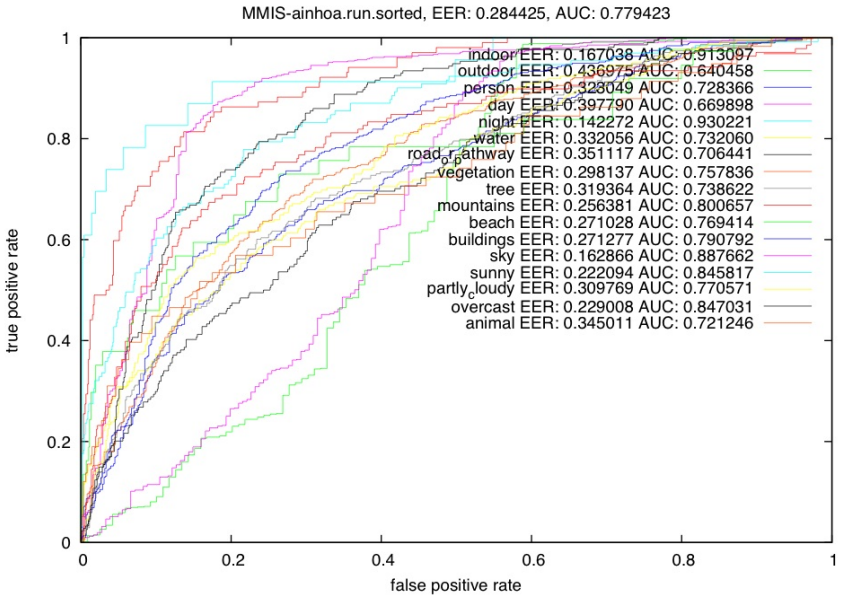

Figure 12.3: ROC curves for our best annotation algorithm

(FP) in a binary classifier. The Equal Error Rate (EER) is the error rate at the threshold where FP=FN. The area under the ROC curve, AUC, is equal to the probability that a classifier will rank a randomly chosen positive instance higher than a randomly chosen negative one. The results obtained by the four algorithms developed by our group are represented in Table 12.1. Our best result corresponds to the "Enhanced Automated Image Annotation" algorithm as seen in Figure 12.3.

Conclusions

The enhanced automated image annotation method performed well appearing in the top quartile of all methods submitted, however it failed to provide signicant improvement over the automated image annotation method. An explanation for this can be found in the small number of terms of the vocabulary that hinders the functioning of the algorithm and another in the nature of the vocabulary itself, where instead of incoherence we have mutually exclusive terms and almost no semantically similar terms.

12.2 Photo Retrieval Task

The goal of this task is, given some multimedia queries, a collection of 20,000 IAPR TC-12 images and the set of annotations generated by VCDT, to present the top image results of a ranked list that will ideally contain diverse items representing different sub-topics. In addition to the annotations each image is accompanied by an alphanumeric caption stored in a semi-structured format. This task intends to take a different approach to evaluation by studying image clustering as a good image search engine ensures that duplicate or near duplicate documents retrieved in response to a query are hidden from the user.

Image Clustering

We propose a simple method of re-ordering the top of our rank based on document annotations. We consider three sources of annotations: Automated annotations assigned to images, words matched to WordNet and locations extracted from text. WordNet is a freely available semantic lexicon of 155,287 words mapping to 117,659 semantic senses [101]. In our experiments we compare two sources of annotations: automated image annotations (Image clustering) and words matched to WordNet (WordNet clustering).

In Image clustering all the images have been annotated with concepts coming from the Corel ontology. This ontology was created using SUMO and enriching it with a taxonomy of animals created by the University of Michigan as described in Chapter 8. After that, the ontology was populated with the vocabulary of 374 terms used for annotating the Corel dataset. Among many categories, we can find animal, vehicle and weather.

For example, if the cluster topic is “animal” we will split the ranked list of results into sub ranked lists, one corresponding to every type of animal and an additional uncategorised rank. These ranks will then be merged to

form a new rank, where all the documents at rank 1 in a sub list appear first, followed by the documents at rank 2, followed by rank 3 etc. The documents of equal rank in the sublists are ranked amongst themselves based on their rank in the original list. This way the document at rank 1 in the original list remains at rank 1 in the re-ranked list. We only maximise the diversity of the top 20 documents, after the 20th document the other documents maintain their ordering in the original list.

12.3 Wikipedia Task

The goal of this task is given a multimedia query -called topic- describing a user's multimedia information need, find as many relevant images as possible from the Wikipedia image collection. The collection to be used in this task has been created and employed by the INEX Multimedia Track (2006-2007). The collection consists of approximately 150,000 Wikipedia images provided by Wikipedia users. Each image is associated with user-generated alphanumeric, unstructured metadata in English. These metadata usually contain a brief caption or description of the image, the Wikipedia user who uploaded the image, and the copyright information. These descriptions are highly heterogeneous and of varying length. The topics are multimedia queries that can consist of a textual, visual and a conceptual part, with the latter two parts being optional. Thus, the 75 topics are expressed in XML format following the INEX MM initiative and they include the following fields:

- Title: query by keywords.
- Concept (optional): query by one or more concepts coming from the MediaMill 101 concepts.
- Image (optional): query by one or more images.
- Narrative: description of the information need where the definitive definition of relevance and irrelevance are given.

The topics for this year are a combination of the topics previously used in INEX MM and ImageCLEF photo tasks and the topics created by this year's task participants.

Experiments

Our experiments consist in for a given query, extracting the information contained in their fields, and depending on which fields are present, combin-

ing the data coming from image indexing, textual indexing and conceptual indexing and returning ranked lists of (up to) the top 1000 images ranked in descending order of similarity following the standard TREC format.

Conceptual indexing

The objective of this subtask is to perform a conceptual indexing based on the "concept" field of the query. For a given concept a ranked list of 1000 images from the collection are returned, ordered according to their similarity value. We used the concept classifiers [102] provided by the University of Amsterdam, who evaluated generic video indexing performance on 85 hours of international broadcast news data, from the TRECVID benchmark, using a lexicon of 101 semantic concepts. These concepts are called the MediaMill concepts and have been created taking into consideration the LSCOM [57] ontology. Despite the fact that the UvA classifier only provides some confident scores for only 146,151 images of the whole collection and the performance of these classifiers on the broad collection of Wikipedia images varies greatly, we believe it may still be a useful source of information. Especially if we take into consideration that this conceptual indexing is to be combined with the rest of the data indexed.

Bibliography

[1] Gantz, J.F., Reinsel, D., Chute, C., Schlichting, W., McArthur, J., Minton, S., Xheneti, I., Toncheva, A., Manfrediz, A.: The expanding digital universe. Technical report, International Data Corporation (IDC) (March 2007)

[2] Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions On Pattern Analysis and Machine Intelligence **22**(12) (2000) 1349–1380

[3] Hollink, L.: Semantic annotation for retrieval of visual resources. PhD thesis, Vrije Universiteit Amsterdam (2006)

[4] Hare, J.S., Lewis, P.H., Enser, P.G.B., Sandom, C.J.: Mind the gap: another look at the problem of the semantic gap in image retrieval. In: Multimedia Content Analysis, Management and Retrieval. Volume 6073., SPIE (2006) 607309–1

[5] Liu, H., Song, D., Rüger, S., Hu, R., Uren, V.S.: Comparing dissimilarity measures for content-based image retrieval. In: Asia Information Retrieval Symposium (AIRS). (2008) 44–50

[6] Lew, M.S., Sebe, N., Djeraba, C., Jain, R.: Content-based multimedia information retrieval: State of the art and challenges. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP) **2**(1) (February 2006) 1–19

[7] Flickner, M., Sawhney, H., Niblack, W., Ashley, J., Huang, Q., Dom, B., Gorkani, M., Hafner, J., Lee, D., Petkovic, D., Steele, D., Yanker, P.: Query by image and video content: the QBIC system. MIT Press, Cambridge, MA, USA (1997)

[8] Bach, J.R., Fuller, C., Gupta, A., Hampapur, A., Horowitz, B., Humphrey, R., Jain, R., Shu, C.F.: Virage image search engine: An

open framework for image management. In: Storage and Retrieval for Image and Video Databases (SPIE). (1996) 76–87

[9] Smith, J., Chang, S.: Multi-stage classification of images from features and related text. In: Proceedings of the 4th Europe DELOS workshop. (1997)

[10] Frankel, C., Swain, M.J., Athitsos, V.: Webseer: An image search engine for the world wide web. Technical report, The University of Chicago, Chicago, IL, USA (1996)

[11] Armitage, L.H., Enser, P.G.: Analysis of user need in image archives. Journal of Information Science **23**(4) (August 1997) 287–299

[12] Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Transactions on Computing Surveys **40**(2) (2008) 1–60

[13] Benitez, A.B., Chang, S.F.: Perceptual knowledge construction from annotated image collections. In: IEEE International Conference On Multimedia & Expo (ICME-2002), Lausanne, Switzerland (August 2002)

[14] von Ahn, L.v., Dabbish, L.: Labeling images with a computer game. In: Proceedings of the 2004 conference on Human factors in computing systems (CHI), ACM Press (2004) 319–326

[15] Petridis, K., Anastasopoulos, D., Saathoff, C., Timmermann, N., Kompatsiaris, Y., Staab, S.: M-ontomat-annotizer: Image annotation linking ontologies and multimedia low-level features. In: Demos and Posters of the 3rd European Semantic Web Conference (ESWC). (2006) 633–640

[16] Saathoff, C., Timmermann, N., Staab, S., Petridis, K., Anastasopoulos, D., Kompatsiaris, Y.: M-ontomat-annotizer: Linking ontologies with multimedia low-level features for automatic image annotation. In: Demos and Posters of the 3rd European Semantic Web Conference (ESWC). (2006)

[17] Athanasiadis, T., Tzouvaras, V., Petridis, K., Precioso, F., Avrithis, Y., Kompatsiaris, Y.: Using a multimedia ontology infrastructure for semantic annotation of multimedia content. In: Proceedings of International Workshop on Knowledge Markup and Semantic Annotation (SemAnnot), Springer (November 2005)

[18] Halaschek-Wiener, C., Schain, A., Grove, M., Parsia, B., Hendler, J.: Management of digital images on the semantic web. In: Proceedings of the International Semantic Web Conference (ISWC). (2005)

[19] Tuffield, M.M., Harris, S., David, Chakravarthy, A., Brewster, C., Gibbins, N., Hara, K.O., Ciravegna, F., Sleeman, D., Shadbolt, N.R., Wilks, Y.: Image annotation with photocopain. In: Proceedings of the 15th World Wide Web Conference (WWW). (2006)

[20] Addis, M., Boniface, M., Goodall, S., Grimwood, P., Kim, S., Lewis, P., Martinez, K., Stevenson, A.: Sculpteur: Towards a new paradigm for multimedia museum information handling. In: Proceedings of International Semantic Web Conference (ISWC). (2003)

[21] Dupplaw, D., Dasmahapatra, S., Hu, B., Lewis, P., Shadbolt, N.: Multimedia distributed knowledge management in miakt. In: Knowledge Markup and Semantic Annotation, 3rd International Semantic Web Conference (ISWC). (2004)

[22] Schreiber, A.T.G., Dubbeldam, B., Wielemaker, J., Wielinga, B.: Ontology-based photo annotation. IEEE Intelligent Systems **16**(3) (2001) 66–74

[23] Hollink, L., Schreiber, G., Wielemaker, J., Wielinga, B.: Semantic annotation of image collections. In: Workshop on Knowledge Markup and Semantic Annotation (KCAP). (2003)

[24] Lafon, Y., Bos, B.: Describing and retrieving photos using rdf and http. w3c note (2002)

[25] Zhang, H., Wenyin, L., Hu, C.: ifind- a system for semantics and feature based image retrieval over internet. In: Proceedings of the 8th International ACM Conference on Multimedia, New York, NY, USA, ACM Press (2000) 477–478

[26] Aurnhammer, M., Hanappe, P., Steels, L.: Augmenting navigation for collaborative tagging with emergent semantics. In: Proceedings of the International Semantic Web Conference (ISWC), Springer, LNCS (November 2006)

[27] Yavlinsky, A.: Behold: a content based image search engine for the world wide web. Technical report, Imperial College, MMIS group (2006)

[28] Chakravarthy, V.L.A.: Cross-media document annotation and enrichment. In: Proceedings of the 1st Semantic Authoring and Annotation Workshop (SAAW). (2006)

[29] Soo, V.W., Lee, C.Y., Li, C.C., Chen, S.L., chih Chen, C.: Automated semantic annotation and retrieval based on sharable ontology and case-based learning techniques. In: Proceedings of the 3rd ACM/IEEE-CS Joint Conference on Digital libraries (JCDL), Washington, DC, USA, IEEE Computer Society (2003) 61–72

[30] Mori, Y., Takahashi, H., Oka, R.: Image-to-word transformation based on dividing and vector quantizing images with words. In: International Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM). (1999)

[31] Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.A.: Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In: European Conference on Computer Vision, London, UK, Springer-Verlag (2002) 97–112

[32] Monay, F., Gatica-Perez, D.: On image auto-annotation with latent space models. In: Proceedings of the 11th International ACM Conference on Multimedia (MM), New York, NY, USA, ACM (2003) 275–278

[33] Monay, F., Gatica-Perez, D.: Plsa-based image auto-annotation: constraining the latent space. In: Proceedings of the 12th International ACM Conference on Multimedia (MM), New York, NY, USA, ACM (2004) 348–351

[34] Blei, D.M., Jordan, M.I.: Modeling annotated data. In: Proceedings of International ACM Conference on Research and Development in Information Retrieval (SIGIR), New York, NY, USA, ACM (2003) 127–134

[35] Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3** (2003) 993–1022

[36] Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D., Jordan, M.: Matching words and pictures. Journal of Machine Learning Research **3** (2003) 1107–1135

[37] Jeon, J., Lavrenko, V., Manmatha, R.: Automatic image annotation and retrieval using cross-media relevance models. In: Proceedings

of International ACM Conference on Research and Development in Information Retrieval (SIGIR), New York, NY, USA, ACM Press (2003) 119–126

[38] Lavrenko, V., Manmatha, R., Jeon, J.: A model for learning the semantics of pictures. In: Advances in Neural Information Processing Systems (NIPS). (2003)

[39] Metzler, D., Manmatha, R.: An inference network approach to image retrieval. In: Proceedings of the 3rd International Conference on Image and Video Retrieval (CVIR). (2004) 42–50

[40] Feng, S.L., Manmatha, R., Lavrenko, V.: Multiple bernoulli relevance models for image and video annotation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) **02** (2004) 1002–1009

[41] Torralba, A., Oliva, A.: Statistics of natural image categories. Network: Computation in Neural Systems **14**(3) (2003) 391–412

[42] Yavlinsky, A., Schofield, E., Rüger, S.: Automated image annotation using global features and robust nonparametric density estimation. In: International ACM Conference on Image and Video Retrieval (CIVR). (2005) 507–517

[43] Jin, R., Chai, J.Y., Si, L.: Effective automatic image annotation via a coherent language model and active learning. In: Proceedings of the 12th International ACM Conferencia on Multimedia (MM), New York, NY, USA, ACM (2004) 892–899

[44] Zhou, X., Wang, M., Zhang, Q., Zhang, J., Shi, B.: Automatic image annotation by an iterative approach: incorporating keyword correlations and region matching. In: International ACM Conference on Image and Video Retrieval (CIVR), New York, NY, USA, ACM (2007) 25–32

[45] Naphade, M.R., Kozintsev, I.V., Huang, T.S.: A factor graph framework for semantic video indexing. IEEE Transactions On Circuits And Systems For Video Technology **12** (2002)

[46] Escalante, H., Montes y Gomez, M., Sucar, L.: Word co-occurrence and markov random fields for improving automatic image annotation. In: Proceedings of the 18th British Machine Vision Conference (BMVC). (2007)

[47] Feldman, J., Yakimovsky, Y.: Decision theory and artificial intelligence: A semantics based region analyzer. Artificial Intelligence **5**(4) (1974) 349–371

[48] Gruber, T.R.: A translation approach to portable ontology specifications. Knowledge Acquisition **5**(2) (June 1993) 199–220

[49] Khan, L.: Standards for image annotation using semantic web. Computer Standards & Interfaces **29**(2) (2007) 196–204

[50] van Ossenbruggen, J., Hardman, L., Geurts, J., Rutledge, L.: Towards a multimedia formatting vocabulary. In: Proceedings of the 12th international conference on World Wide Web (WWW), New York, NY, USA, ACM Press (2003) 384–393

[51] Petridis, K., Bloehdorn, S., Saathoff, C., Simou, N., Dasiopoulou, S., Tzouvaras, V., Handschuh, S., Avrithis, Y., Kompatsiaris, Y., Staab, S.: Knowledge representation and semantic annotation of multimedia content. In: IEEE Proceedings on Vision Image and Signal Processing. Volume 153. (June 2006) 255–262

[52] W3C: Image annotation on the semantic web: Vocabularies overview (2006)

[53] Bailer, W., Schallauer, P., Hausenblas, M., Thallinger, G.: Mpeg-7 based description infrastructure for an audiovisual content analysis and retrieval system. In: Proceedings of Conference on Storage and Retrieval Methods and Applications for Multimedia. Volume 5682. (2005) 284–295

[54] Naphade, M., Kennedy, L., Kender, J., Chang, S., Smith, J., Over, P., Hauptmann, A.: A light scale concept ontology for multimedia understanding for trecvid 2005 (lscom-lite). Research report, IBM (2005)

[55] Hollink, L., Worring, M.: Building a visual ontology for video retrieval. In: Proceedings of the 13th International ACM Conference on Multimedia (MM), New York, NY, USA, ACM Press (2005) 479–482

[56] Bloehdorn, S., Petridis, K., Saathoff, C., Simou, N., Tzouvaras, V., Avrithis, Y., Handschuh, S., Kompatsiaris, Y., Staab, S., Strintzis, M.G.: Semantic annotation of images and videos for multimedia analysis. In: Proceedings of the Second European Semantic Web

Conference (ESWC). Volume 3532 of Lecture Notes in Computer Science., Springer (2005) 592–607

[57] Naphade, M., Smith, J.R., Tesic, J., Chang, S.F., Hsu, W., Kennedy, L., Hauptmann, A., Curtis, J.: Large-scale concept ontology for multimedia. IEEE MultiMedia **13**(3) (2006) 86–91

[58] Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: Proceedings of the 8th ACM international workshop on Multimedia information retrieval (MIR), New York, NY, USA, ACM (2006) 321–330

[59] Jewell, M.O., Lawrence, K.F., Tuffield, M.M., Bennett, P.A., Millard, D.E., Nixon, M.S., Schraefel, Shadbolt, N.R.: Ontomedia: An ontology for the representation of heterogeneous media. In: Proceedings of Multimedia Information Retrieval Workshop, Brazil (2005)

[60] Lagoze, C., Hunter, J.: The abc ontology and model. In: Proceedings of the International Conference on Dublin Core and Metadata Applications (DC), National Institute of Informatics, Tokyo, Japan (2001) 160–176

[61] Gruber, T.R.: Toward principles for the design of ontologies used for knowledge sharing. In Guarino, N., Poli, R., eds.: Formal Ontology in Conceptual Analysis and Knowledge Representation, Deventer, The Netherlands, Kluwer Academic Publishers (1993)

[62] Brewster, C., Ciravegna, F., Wilks, Y.: Background and foreground knowledge in dynamic ontology construction. In: Proceedings of the Semantic Web Workshop, Toronto, August 2003, SIGIR (2003)

[63] Miller, G.A.: Wordnet: A lexical database for english. Communications of ACM **38**(11) (1995) 39–41

[64] Alani, H.: Position paper: ontology construction from online ontologies. In: Proceedings of the 15th International Conference on World Wide Web (WWW), New York, NY, USA, ACM Press (2006) 491–495

[65] d'Aquin, M., Sabou, M., Dzbor, M., Baldassarre, C., Gridinoc, L., Angeletou, S., Motta, E.: Watson: A gateway for the semantic web. In: Proceedings of the European Semantic Web Conference (ESWC). Volume 4519 of Lecture Notes in Computer Science., Springer-Verlag (July 2007)

[66] Specia, L., Motta, E.: Integrating folksonomies with the semantic web. In: Proceedings of 4th European Semantic Web Conference (ESWC). (July 2007)

[67] Schmitz, P.: Inducing ontology from flickr tags. In: Proceedings of International Conference on World Wide Web (WWW). (May 2006)

[68] Pedersen, Banerjee, Patwardhan: Maximizing semantic relatedness to perform word sense disambiguation. Technical report, University of Minnesota (2003)

[69] Gracia, J., Trillo, R., Espinoza, M., Mena, E.: Querying the web: a multiontology disambiguation method. In: Proceedings of the 6th International Conference on Web Engineering (ICWE), New York, NY, USA, ACM Press (2006) 241–248

[70] Stokoe, C., Oakes, M.P., Tait, J.: Word sense disambiguation in information retrieval revisited. In: Proceedings of International ACM Conference on Research and Development in Information Retrieval (SIGIR), New York, NY, USA, ACM Press (2003) 159–166

[71] Cilibrasi, R., Vitanyi, P.: Automatic meaning discovery using google. Technical report, Centrum Wiskunde & Informatica (CWI) (2004)

[72] Enser, P.G., Sandom, C.J., Lewis, P.H.: Automatic annotation of images from the practitioner perspective. In: International ACM Conference on Image and Video Retrieval (CIVR). Volume 3568. (2005) 497–506

[73] Hare, J.S., Sinclair, P.A.S., Lewis, P.H., Martinez, K., Enser, P.G.B., Sandom, C.J.: Bridging the semantic gap in multimedia information retrieval: Top-down and bottom-up approaches. In: Mastering the Gap: From Information Extraction to Semantic Representation (ESWC). (2006)

[74] Srikanth, M., Varner, J., Bowden, M., Moldovan, D.: Exploiting ontologies for automatic image annotation. In: Proceedings of the 28th International ACM Conference on Research and Development in Information Retrieval (SIGIR), New York, NY, USA, ACM Press (2005) 552–558

[75] Niles, I., Pease, A.: Towards a standard upper ontology. In: Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS), New York, NY, USA, ACM Press (2001) 2–9

[76] Parr, C., Sachs, J., Parafiynyk, A., Wang, T., Espinosa, R., Finin, T.: ETHAN: the Evolutionary Trees and Natural History Ontology. Technical report, University of Maryland, Baltimore County (November 2006)

[77] Hanbury, A., Serra, J.: Mathematical morphology in the CIELAB space. Image Analysis & Stereology **21** (2002) 201–206

[78] Tamura, H., Mori, T., Yamawaki, T.: Textural features corresponding to visual perception. IEEE Transactions on Systems, Man and Cybernetics **8**(6) (June 1978) 460–473

[79] Jelinek, F., Mercer, R.L.: Interpolated estimation of markov source parameters from sparse data. In: Proceedings of the Workshop on Pattern Recognition in Practice. (1980)

[80] Hofmann, T., Puzicha, J.: Statistical models for co-occurrence data. Technical report, MIT (1998)

[81] Chen, S.F., Goodman, J.: An empirical study of smoothing techniques for language modeling. In: Proceedings of the 34th annual meeting on Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (1996) 310–318

[82] Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley (May 1999)

[83] Manning, C.D., Schütze, H.: Foundations of statistical natural language processing. MIT Press (1999)

[84] Grubinger, M., Clough, P., Müller, H., Deselears, T.: The IAPR TC-12 Benchmark - a new evaluation resource for visual information systems. In: International Workshop OntoImage. (2006) 13–23

[85] Jin, Y., Khan, L., Wang, L., Awad, M.: Image annotations by combining multiple evidence & wordnet. In: Proceedings of the 13th International ACM Conference on Multimedia (MM), New York, NY, USA, ACM Press (2005) 706–715

[86] Liu, J., Li, M., Ma, W.Y., Liu, Q., Lu, H.: An adaptive graph model for automatic image annotation. In: Proceedings of the 8th ACM international workshop on Multimedia information retrieval (MIR), New York, NY, USA, ACM (2006) 61–70

[87] Setia, L., Teynor, A., Halawani, A., Burkhardt, H.: Image classification using cluster cooccurrence matrices of local relational features. In: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval (MIR), New York, NY, USA, ACM (2006) 173–182

[88] Galleguillos, C., Rabinovich, A., Belongie, S.: Object categorization using co-occurrence, location and appearance. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2008)

[89] Hardoon, D.R., Saunders, C., Szedmak, O.: A correlation approach for automatic image annotation. In: International Conference on Advanced Data Mining and Applications. Volume 4093. (2006) 681–692

[90] Kwasnicka, H., Paradowski, M.: On evaluation of image auto-annotation methods. In: International Conference on Intelligent Systems Design and Applications (ISDA), Washington, DC, USA, IEEE Computer Society (2006) 353–358

[91] Biederman, I.: On the semantics of a glance at a scene. In: Perceptual organization. Erlbaum (1981)

[92] Melamed, I.D.: Empirical methods for exploiting parallel texts. PhD thesis, University of Pennsylvania, Philadelphia, PA, USA (1998)

[93] Simpson, J., Weiner, E., eds.: The Oxford English Dictionary. Clarendon Press (1989)

[94] Miller, G.A., Charles, W.G.: Contextual correlates of semantic similarity. Journal of Language and Cognitive Processes **6** (1991) 1–28

[95] Magalhaes, J., Rüger, S.: Information-theoretic semantic multimedia indexing. In: International ACM Conference on Image and Video Retrieval (CIVR), New York, NY, USA, ACM (2007) 619–626

[96] Carneiro, G., Chan, A.B., Moreno, P.J., Vasconcelos, N.: Supervised learning of semantic classes for image annotation and retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence **29**(3) (2007) 394–410

[97] Hull, D.: Using statistical testing in the evaluation of retrieval experiments. In: Proceedings of International ACM Conference on Research and Development in Information Retrieval (SIGIR), New York, NY, USA, ACM (1993) 329–338

[98] Clough, P., Müller, H., Sanderson, M.: The clef 2004 cross-language image retrieval track. In: Fifth Workshop of the Cross-Language Evaluation Forum (CLEF), Heidelberg, Germany, Lecture Notes in Computer Science (LNCS), Springer (2005)

[99] Fawcett, T.: An introduction to ROC analysis. Pattern Recognition Letters **27**(8) (2006) 861–874

[100] Hauptmann, A., Yan, R., Lin, W.H.: How many high-level concepts will fill the semantic gap in news video retrieval? In: Proceedings of the 6th ACM International Conference on Image and Video Retrieval (CVIR), New York, NY, USA, ACM (2007) 627–634

[101] : WordNet: An Electronic Lexical Database. The MIT Press (1998)

[102] Snoek, C.G.M., Worring, M., van Gemert, J.C., Geusebroek, J.M., Smeulders, A.W.M.: The challenge problem for automated detection of 101 semantic concepts in multimedia. In: Proceedings of the 14th International ACM Conference on Multimedia (MM), New York, NY, USA, ACM (2006) 421–430