MULTI-MODAL MULTI-SEMANTIC IMAGE RETRIEVAL

Kraisak Kesorn

**A thesis submitted in partial fulfilment of
the requirement of the degree of**

Doctor of Philosophy

**School of Electronic Engineering and Computer Science
Queen Mary, University of London**

2010

Declaration

The work presented in the thesis is the author's own

DATE: 1 October 2010

SIGNATURE: ______________________________

Table of Contents

List of Figures

List of Tables

List of Algorithms

ABSTRACT

The rapid growth in the volume of visual information, e.g. image, and video can overwhelm users' ability to find and access the specific visual information of interest to them. In recent years, ontology knowledge-based (KB) image information retrieval techniques have been adopted into in order to attempt to extract knowledge from these images, enhancing the retrieval performance. A KB framework is presented to promote semi-automatic annotation and semantic image retrieval using multimodal cues (visual features and text captions). In addition, a hierarchical structure for the KB allows metadata to be shared that supports multi-semantics (polysemy) for concepts. The framework builds up an effective knowledge base pertaining to a domain specific image collection, e.g. sports, and is able to disambiguate and assign high level semantics to '*unannotated*' images.

Local feature analysis of visual content, namely using Scale Invariant Feature Transform (SIFT) descriptors, have been deployed in the 'Bag of Visual Words' model (BVW) as an effective method to represent visual content information and to enhance its classification and retrieval. Local features are more useful than global features, e.g. colour, shape or texture, as they are invariant to image scale, orientation and camera angle. An innovative approach is proposed for the representation, annotation and retrieval of visual content using a hybrid technique based upon the use of an unstructured visual word and upon a (structured) hierarchical ontology KB model. The structural model facilitates the disambiguation of unstructured visual words and a more effective classification of visual content, compared to a vector space model, through exploiting local conceptual structures and their relationships. The key contributions of this framework in using local features for image representation include: first, a method to generate visual words using the semantic local adaptive clustering (SLAC) algorithm which takes term weight and spatial locations of keypoints into account. Consequently, the semantic information is preserved. Second a technique is used to detect the domain specific '*non-informative visual words*' which are ineffective at representing the content of visual data and degrade its categorisation ability. Third, a method to combine an ontology model with a visual word model to resolve synonym (visual heterogeneity) and polysemy problems, is proposed. The experimental results show that this approach can discover semantically meaningful visual content descriptions and recognise specific events, e.g., sports events, depicted in images efficiently.

Since discovering the semantics of an image is an extremely challenging problem, one promising approach to enhance visual content interpretation is to use any associated textual information that accompanies an image, as a cue to predict the meaning of an image, by transforming this textual information into a structured annotation for an image e.g. using XML, RDF, OWL or MPEG-7. Although, text and image are distinct types of information representation and modality, there are some strong, invariant, implicit, connections between images and any accompanying text information. Semantic analysis of image captions can be used by image retrieval systems to retrieve selected images more precisely. To do this, a Natural Language Processing (NLP) is exploited firstly in order to extract concepts from image captions. Next, an ontology-based knowledge model is deployed in order to resolve natural language ambiguities. To deal with the accompanying text information, two methods to extract knowledge from textual information have been proposed. First, metadata can be extracted automatically from text captions and restructured with respect to a semantic model. Second, the use of LSI in relation to a domain-specific ontology-based knowledge model enables the combined framework to tolerate ambiguities and variations (incompleteness) of metadata. The use of the ontology-based knowledge model allows the system to find indirectly relevant concepts in image captions and thus leverage these to represent the semantics of images at a higher level. Experimental results show that the proposed framework significantly enhances image retrieval and leads to narrowing of the semantic gap between lower level machine-derived and higher level human-understandable conceptualisation.

ACKNOWLEDGEMENTS

This thesis could not have been accomplished without the assistance of many people whose contributions I gratefully acknowledge. Firstly, I would like to express the deepest appreciation to my supervisor, Stefan Poslad, whose encouragement, guidance and support from the initial stage through to the conclusion enabled me to develop a deeper understanding of the subject. Without his support this thesis would not exist. In addition, I would like to express gratitude for all the support and help received from the School of Electronic Engineering and Computer Science, Queen Mary University of London.

It is also a pleasure to thank my research colleagues and friends for their support, in particular, Zekeng Liang, Zhenchen Wang, and Ioannis Barakos. Special thanks are also extended to Sutasinee Chimlek for her timely advice on visual content analysis techniques which accelerated my comprehension of the background knowledge in this area. I am also heartily thankful to Andrew Kell for his help in reading and correcting my written work throughout my studies.

Finally, I owe my deepest gratitude to the Royal Thai Government in supporting my PhD study at the School of Electronic Engineering and Computer Science, Queen Mary University of London, enabling me to live and experience another culture in the City of London, for the past four years. This tremendous opportunity will hopefully provide me with a sounder educational basis for my future career, broaden my vision and expand my knowledge, all of which I will endeavour to apply in my native country.

Lastly, I would like to thank my family for their endless love that formed the most important part of my growing-up and for always being there when I needed them, helping me to face difficulties. Their endless support and encouragement has made the journey easier and one that I will treasure for many years to come.

ABBREVIATIONS

Abbreviations	Meaning
API	Application Programming Interface
BVW	Bag of Visual Words
CBIR	Content-based Image Retrieval
CS	Classification Schemes
DDL	Data Definition Language
DoG	Difference of Gaussian
FIM	Frequent Itemsets Mining
IDF	Inverse Document Frequency
IMR	Image Retrieval System
IR	Information Retrieval
KB	Knowledge Base
KM	Knowledge Management
LIRE	Lucene Image Retrieval Framework
LM	Language Modelling
LSI/LSA	Latent Semantic Indexing/Analysis
NL	Natural Language
OVSS	Ontology-based for Visual Semantic Search
PCA	Principal Component Analysis
PIMR	Personalised Image Retrieval
pLSA	Probabilistic Latent Semantic Analysis
QBE	Query-By-Example
RDBMS	Relational Database Management
RDF	Resource Description Framework
RDFS	RDF Schema
SBIR	Semantic-Based Image Retrieval
SIFT	Scale Invariant Feature Transform
SLAC	Semantic Local Adaptive Clustering
SPARQL	SPARQL Protocol and RDF Query Language
SPI	System Programming Interface
SQRT	Square Root scheme
SVD	Singular Value Decomposition
SVM	Singular Vector Machine

Chapter 1

Introduction

There are billions of images on the World Wide Web (WWW), which are accessed by many millions of users globally. The continued rapid growth in digital visualisation makes it increasingly difficult to find, organise, access, and maintain users' visual information. When users require selective pieces of information from the Internet, they submit a query to, a possibly distributed, computer system that determines what is *relevant* to that query and what is not, based upon what is in its data store. This task is known as *Information Retrieval* (IR). The traditional method for retrieving data uses text information representation and is known as *text-based* or *keyword-based information retrieval*. Research in this technique began in the 1940's and several techniques have been developed and reached a high retrieval performance in term of the quality of results and computational efficiency (Tansley 2000).

CBIR research has evolved over many decades. Its main goal is to represent visual information to enable computer systems to understand the content of images or videos. Initially, it started through using low-level features, such as colour, texture, shape, structure, and spatial relationships, to represent visual content. However, most users are interested in the content of images at the semantic level, e.g. event, people, and location in images. Hence, there is often a considerable difference between users' frameworks for interpreting the semantics of the casual information and the aforementioned low-level features leading to the so called *semantic gap* - the gap between the semantic user view and the low-level feature abstraction from the visual content. As such, CBIR or the current keyword-based image retrieval approach is still far from enabling semantic-based access. Consequently, Semantic-Based Image Retrieval (SBIR) approaches became established a few years ago and soon became a notable theme for image and other multimedia information retrieval.

Research on SBIR has been conducted around various visual domain specific descriptors and topics such as detecting faces and people, landscapes, and cities (Szummer & Picard 1998; Vailaya et al. 1998; Paek & Chang 2000; Yang 2003). Today, a huge number of images are produced on a daily basis for news, sport, entertainment, and education. Of these, sport seems an incessant and popular domain, serving the interests of publishers, broadcasters, producers, sponsors, and an increasing global audience of Internet users in order to search, select and access visual information about sports events, athletes, genres, records and venues. Semantic descriptions of sport images are vital in order to cater for such diverse users' information needs.

Among the current semantic technologies, a knowledge based (ontological) approach is used to provide an explicit domain-oriented semantics in terms of defined concepts and their relationships that are not only machine-readable but also machine-processable. This supports semantic annotation, browsing, and retrieval of visual content. Ontologies describe visual content using well-structured concepts and relationships that are also human readable and meaningful.

1.1 Motivation

The Internet aids users in order to find information more easily, speedily, and at a lower cost of retrieval. However, human users usefully retrieve images at higher levels of semantics but this is still far from being achieved in practice. A knowledge-based model (KB) is a potential solution to resolve this issue. A KB is in general defined as capturing and organising the expertise and experience of a collective (Hampapur 2003). Since humans often understand things in the world more easily if they are represented in the form of classes and relationships, the data in a KB should be represented in a relationship-based model in order to support human decision-making, learning, reasoning and explanation. Therefore, an ontology model is deployed to represent KB in this research. This typically organises data into hierarchical structures which comprises several classes, instances and relationships. Typically, an image processing algorithm extracts visual data at a non-semantic level whereas users create queries at a higher semantic level. Therefore, a knowledge-based image retrieval system (IMR), based upon an ontology model, is very important to bridge these two levels together. This is able to support the mapping of low-level visual features to high-level semantic concepts and to support reasoning about data in order to promote semantic retrieval. An IMR needs to deal with two types of data, visual and textual information. First, the system needs to extract the metadata from images and to process the extracted metadata to represent the content of images effectively, even when text captions are not supplied. Second, an IMR needs to use text captions to enhance the image classification and interpretation process, even when these captions may be ambiguous. In addition, metadata that are extracted from visual content and text captions should be integrated to facilitate the Semantic-Based Image Retrieval system (SBIR). To achieve these, there are several challenges that such a system needs to overcome (Figure 1-1).

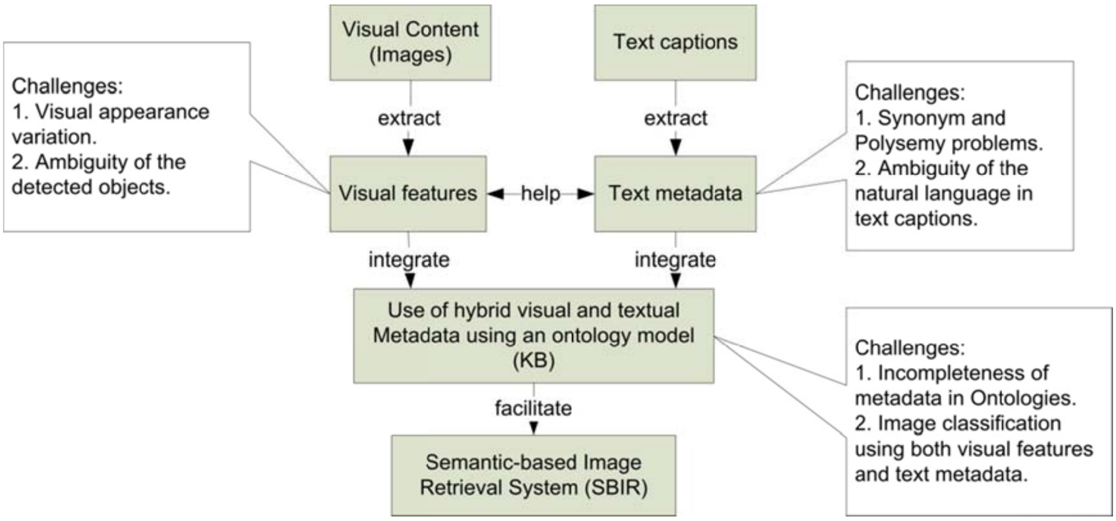

Visual Content (Images)
Text captions
Challenges:
1. Visual appearance variation.
2. Ambiguity of the detected objects.
extract
extract
Challenges:
1. Synonym and Polysemy problems.
2. Ambiguity of the natural language in text captions.
Visual features
help
Text metadata
integrate
integrate
Use of hybrid visual and textual Metadata using an ontology model (KB)
Challenges:
1. Incompleteness of metadata in Ontologies.
2. Image classification using both visual features and text metadata.
facilitate
Semantic-based Image Retrieval System (SBIR)

Figure 1-1 Motivations and challenges for IMR

1.1.1 Representing Visual Content and Classifying Images

In recent years, KBs have been adopted by multimedia retrieval techniques in order to extract knowledge from multimedia documents, enhancing the retrieval performance such as in (Sheikholeslami et al. 1998; Benitez et al. 2000; Tansley 2000). Knowledge extraction from images is of interest to many researchers in order to build more effective image KB systems, to extract and to manage large amounts of heterogeneous image data. The application of KB systems to image content retrieval presents several significant challenges including visual content representation, ambiguity of key objects and image classification.

- *Visual content representation*: visual data needs to be analysed and transformed into a format that represent the visual content effectively and can be used by KB systems. Typically, visual features are extracted by image processing algorithms and transformed into useful metadata. This metadata describes the content of visual document, so-called descriptors, which is manipulated and processed by standard information retrieval methods. Methods to represent and store this metadata in an efficient manner and for further use are needed. Image data contains a large number of unstructured visual features. The need to establish a good knowledge representation model to represent the features of interest is important for IMR systems.
- *Ambiguity of key objects and image classification*: low-level features of visual information can often be ambiguous. In comparison to words in text documents, multiple objects may share similar features, use synonyms, or an object may belong to several concepts (polysemy), for example, a 'horizontal bar' object can belong to high jump or pole vault event. Therefore, IMR systems should be able to handle these ambiguities properly in order to achieve a high image classification accuracy.

1.1.2 Ambiguity of Natural Language in Text Captions

One promising approach to enhance visual content understanding is to use any associated textual information with the image as a cue to interpret the meaning of an image by transforming this textual information into a structured annotation for the image. Text and image are two distinct types of information from different modalities, as they represent 'things' in different ways. However, there are some invariant and implicit connections between textual and visual information (Smeulders et al. 2000). In fact, the textual information surrounding images includes some form of human generated descriptions of images; thus, these should not be disregarded as they can be used to enhance image interpretation by supplementing image content with textual information associated with the image. However, the drawback of text captions is the ambiguity of the natural language used to describe visual content. In addition, some images do not have associated text captions. These challenges are highlighted:

- *Synonym and Polysemy*: these are the classical problems in the information retrieval system. Some researchers (Swain et al. 1997; Smith & Chang 1997) have solved the synonym problem by applying various clustering algorithms to their works but the polysemy problem still remains. Polysemy is one factor causing a poor precision performance for image retrieval.
- *Absence of text captions*: when text descriptions of images are not supplied, the system should be able to describe the high-level semantics of images based upon any distinctive low-level visual features. These visual features should be invariant to image scale, orientation, camera angle, and change in illumination. Consequently, the system can predict the semantics of the image more accurately.

1.1.3 Use of Hybrid Visual and Textual Metadata Models

The combination of textual information with image features information has been proposed to improve image search results, for example, (Swain et al. 1997; Zhao & Grosky 2002; Hu & Bagga 2004; Song et al. 2004; Smith & Chang 1997). These approaches focus on improving the retrieval performance to get more accurate results. There are some challenges in combining visual and textual metadata for IMR.

- *A KB model to bridge both metadata*: the KB model should be designed to interlink both metadata together in order to facilitate the image classification and retrieval performance.
- *Incompleteness of metadata in the KB:* automatic ontology construction and metadata extraction from text descriptions or text captions are very challenging tasks to build a complete ontology due to several factors. First, these documents are written in an imprecise and inherently *vague style* using natural language. Second, they are often missing important information required by ontologies. Third, ontologies may require actual deployment and

to have some degree of openness rather than to be fixed at development in order to acquire the different ontology commitments of a diverse set of users whose conceptualisations may diverge from developers. Fourth, some important aspects cannot be modelled in present-day standard ontology languages e.g. uncertainty and gradual truth values. These cannot directly be represented in a strong ontology language representation such as OWL that hard-wire a specific logic, a description logic, into the ontology representation. These are significant reasons why a complete ontology cannot be built even when the system processes a large training set in order to acquire the metadata to populate the KB model. Therefore, in this thesis, ontology incompleteness refers to an absence of some semantic metadata and also relationships between concepts that cannot be represented in an ontology. The KB may be incomplete resulting in the failure of finding relevant information of the retrieval mechanism. Image retrieval systems operating solely on information in the KB, sometimes, are less effective than the systems using information directly from text captions. This is because of the inadequate coverage of annotations by a KB (Nagypál 2007). Therefore, image retrieval systems should be able to deal with information incompleteness in the KB.

These limitations drive the research objectives described in the following sections. Solutions to these problems are vital to achieve a good quality image retrieval system and are, as such, the main focus in this thesis.

1.2 Research Objectives

The central theme of this research is to improve IMR using semantics. A more specific problem is to investigate if the semantic annotation for ‘new’ unannotated image content can be semi-automatically derived from related images or text descriptions of the image. It is a complex problem and consists of several research sub-problems, as follows:

1. To represent the semantics of visual content using a hybrid technique of unstructured visual features and a hierarchical ontology-based (KB) that annotates and facilitates the retrieval mechanism. A major weakness of

existing frameworks is that visual features are disambiguated using a vector space model. This is more of a statistical calculation that does not represent the actual semantic relations properly. In other words, using a vector space model causes a loss in semantic information between visual features and, therefore, it cannot represent the semantics of visual content properly. The main objective is to use a structural model to disambiguate unstructured visual features in order to classify visual content more effectively, in comparison to a vector space model, by exploiting conceptual structures and relationships.

2. To investigate how to cross-link multi-media types (text and image) using multiple semantics. A technique for interlinking semantic metadata to data including different types of data, e.g. text captions and visual features, will be proposed. Visual features and text captions will be analysed and restructured into a semantic model in order to represent the context and content of images efficiently leading to an enhanced retrieval performance.
3. To handle the uncertainty for visual content categorisation. This is because visual content is often ambiguous. An example of the uncertainty of visual content in the sports domain is that a pole vault event requires a pole but an image of a pole vault may not show a pole. Therefore, a visual content analysis system should be able to handle this situation. A probabilistic model can be applied to handle this difficulty and to guide the classification module with a higher degree of confidence.
4. To handle the uncertainty of the knowledge model when it does not contain all the relevant information in a user query. The limitation of the existing systems is that they rely only on information in the ontology and will return no answers to users when there is no relevant information stored in the KB. In such a case, IMRs should try an alternative method to find the relevant information to user queries. In addition, an alternative method should still provide the capability of a semantic search also based on textual information derived image captions instead of only searching for relevant information in the KB since it may not provide the information needed.

1.3 Structure of this Thesis

This report is organised into eight chapters as follows.

Chapter 1, *Introduction*, gives an introduction, sets out the summary of the problem, the motivation, objectives, and presents the structure of this thesis.

Chapter 2, *Fundamentals*, indicates the trend of image retrieval research area in the past decade. This chapter introduces the fundamentals of image retrieval and ontology models. Two major techniques for image retrieval system, Content based Image Retrieval and Semantic-based Image Retrieval, are described.

Chapter 3, *Survey and Analysis of the State-of-the-Art Frameworks*, focuses on semantic frameworks for sport information retrieval systems including image, video, and audio. The requirements of IMRs are identified and summarised. A critical analysis of their problems and limitations is provided at the end of this chapter.

Chapter 4, *Overview of the Proposed Framework*, provides an overview of a proposed framework in this thesis. Missing requirements in existing IMR are identified and, then, various techniques used in the framework to achieve the goal and fulfil the identified requirements for image retrieval system are given. Finally, the technologies in order to build up a new solution are described.

Chapter 5, *Semantic-Based Image Retrieval System*, describes the details of the proposed framework. Two core models, an ontology model and MPEG-7, are compared, and the advantages and limitations of each model are shown. The ontology model used in this research will be presented as well as its structure, classes, and slots. This chapter introduces the two main components of the presented framework, visual analysis and linguistic analysis. In the visual analysis component, an ontology layer is built at the upper layer of the bag-of-visual word model to enhance a semantic-based classification and retrieval solution for visual content. For the linguistic analysis component, the techniques to extract metadata from text captions and transform these metadata into the ontology-based KB are described. Finally, the method for semantic retrieval is presented.

Chapter 6, *Handling Uncertainties in Visual Classification and in the KB*, demonstrates the use of the framework to cope with uncertainties in visual and textual information. Several types of uncertainties are identified. A Bayesian network is proposed to deal with the uncertainty of visual information whereas LSI (Latent Semantic Indexing) technique is applied to cope with the uncertainty of metadata in the KB.

Chapter 7, *Experimental Results Evaluation*, analyses and evaluates the empirical results of experiments to validate the presented framework proposed in Chapter 5 and 6. Before evaluating the proposed system, some hypotheses are established and the evaluations are conducted against these hypotheses. Later, the method to prepare the test collection for this experiment validation is described. Thereafter, the method to select the user queries is revealed which is set against certain conditions. The experimental results are compared with Lucene (full text-based search engine) and LIRE (content-based image search engine).

Chapter 8, *Conclusions and Future Work*, summarises the novelties, achievements, and limitations of the framework, and proposes some future directions of this research. The main direction of this future work is to modify the framework to support Personalised Image Retrieval System (PIMR).

Chapter 2

Fundamentals

Many documents comprise not only textual information but also visual data. An information retrieval (IR) system needs to support retrieving visual information, a so called '*Image Retrieval system*' (IMR). Hence, the two main areas of this thesis are IMR combined with a (ontological) knowledge based system. This chapter provides the background knowledge for these two areas.

Research concerning semantic visual information retrieval is still in an early stage but the interest in this research area is significantly increased in recent years and is therefore attracting greater attention. Surveys (Datta et al. 2008; Zhang 2007) indicate that researchers have put a huge effort into, and progressed significantly with respect to, making computers learn to understand, index, and annotate images. These aspects have inspired this thesis to investigate methods to improve the performance of IMR. In the next section, the basics of two major techniques for an image retrieval system namely content-based image retrieval and semantic-based image retrieval are described.

2.1 Main Processes for IMR

To improve the image retrieval system, the basis of the IMR process needs to be understood. The main processes of IMR are shown in Figure 2-1. This begins when a user formulates a query representing the requested information, or part of it, that is executed by the system.

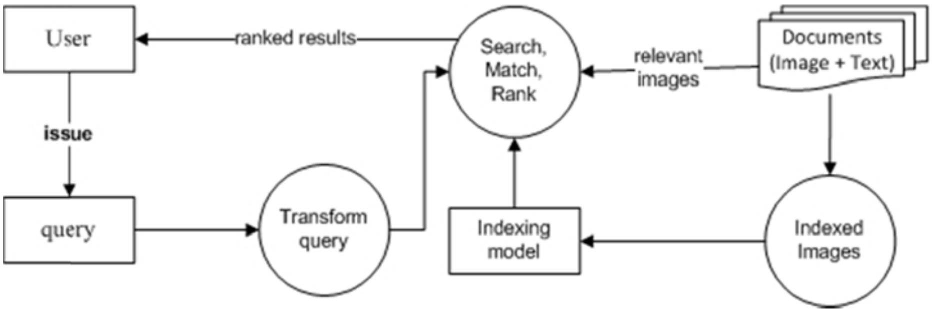


Figure 2-1 Basic image retrieval components

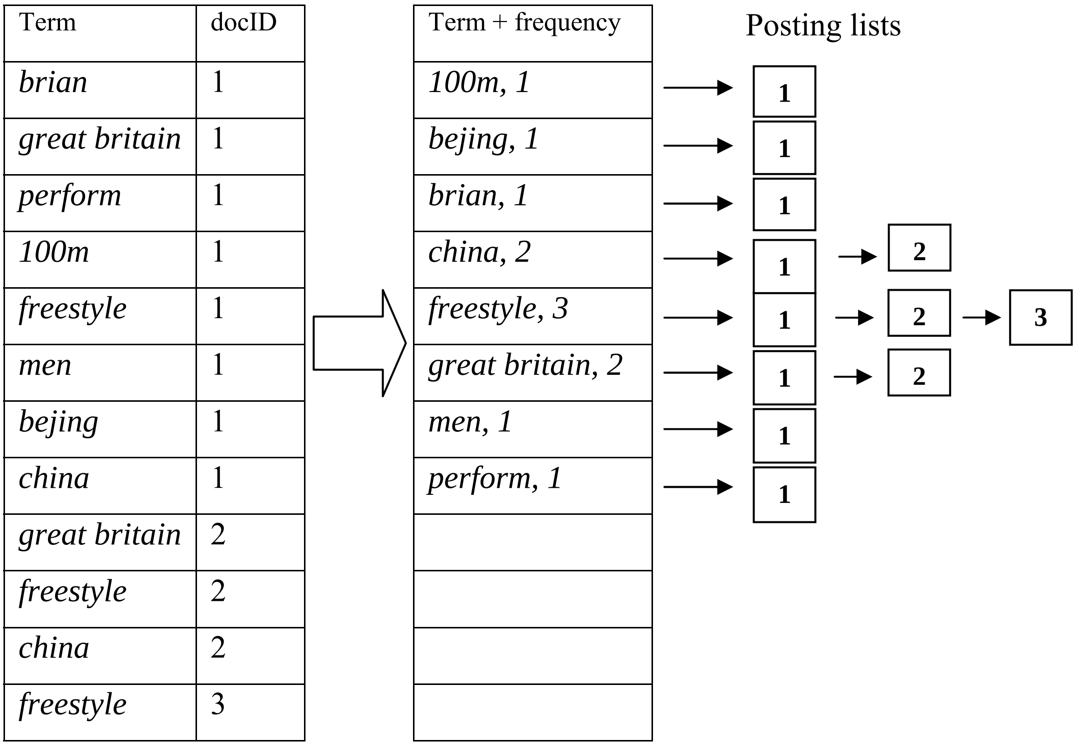
This query is submitted to the system and the system itself will automatically transform the user query into a system query. Usually, the IMR does not process the actual images themselves but rather a representation and abstraction of the images with respect to a particular information model which is created during the image indexing process. Finally, the image representations are matched against the query, and the matching representations are ranked according to an algorithm, and any related images returned to the user.

There are different types of queries such as keyword-based queries, Natural language queries, and QBE (Query-By-Example). A keyword-based query is a query expressed as a word or phrase that describes the information users are seeking. These keywords are used to match the keywords in documents. A Natural language query is a query which is expressed in the form of phrases or sentences in a human-understandable language such as English. In order to support spoken queries, both voice recognition and natural language query software are required. QBE is a method for query creation that allows the user to search for images based on an example in the form of sketching an image or a picture from a camera.

The indexing process, called the *indexer*, is performed in advance and avoids linearly scanning texts for key terms at run-time. The major steps to index documents (Manning et al. 2008) are:

- Collect the documents to be indexed.
- Tokenise the text, turning each document into a list of tokens.
- Perform linguistic pre-processing, e.g. stop word removal and normalisation, producing a list of normalised tokens that are the indexing terms.

- Index the documents in which specific terms occur by creating an inverted index, consisting of a *dictionary* and *postings*.



Term	docID
brian	1
great britain	1
perform	1
100m	1
freestyle	1
men	1
bejing	1
china	1
great britain	2
freestyle	2
china	2
freestyle	3

Term + frequency
100m, 1
bejing, 1
brian, 1
china, 2
freestyle, 3
great britain, 2
men, 1
perform, 1

Figure 2-2 Structure of the inverted indexing

The core indexing step is to organise the index list, e.g., to *sort* the terms into alphabetical order. Multiple occurrences of the same term from the same document are then merged. The *dictionary* also records some statistics such as the number of documents which contain each term. Each term in the list records that a term appeared in a particular document and is called *posting.* This list is then called a posting list (or inverted list). The dictionary in Figure 2-2 has been sorted alphabetically. Each postings list is sorted by document ID. This provides the basis for efficient query processing. This inverted index structure (Figure 2-2) is regarded as the most efficient structure for supporting ad hoc text searches (Manning et al. 2008). Keywords in a query will be matched with keywords in the dictionary and then the system retrieves files in the posting list.

In the case of large document collections, the resulting number of matching documents can far exceed the number a human user could possibly shift through.

Accordingly, it is essential for a search engine to rank the documents matching a query. To do this, the search engine computes, for each matching document, a score with respect to the query. Each term in a document is assigned a *weight* for that term using some popular method e.g. *tf-idf* weighting scheme. At this point, each document can be viewed as a *vector* with one component vector corresponding to each term in the dictionary, together with a weight for each component. To compute the similarity between a document and a query, a similarity measurement is applied such as *cosine similarity* (Manning et al. 2008). The resulting scores can then be used to select the top-scoring documents for a query. This allows a user to focus on more relevant data that is displayed at the top of the results list. Let $\{P_i\}_{i=1}^{N}$ be the set of all visual content in the collection. The similarity between the query (*q*) and the weighted φ associated with the visual content (*p*) in the collection is measured using the following inner product:

$$sim(p,q) = \frac{p \cdot q}{\|p\|\|q\|} \tag{1}$$

For IMR, the main processes to search images are similar to the processes in Figure 2-1. Rather than processing text, IMR processes mainly concern the visual data. The major techniques can be divided into two main approaches which are active among the researchers in this research area; Content-based Image Retrieval (CBIR), and Semantic-based Image Retrieval (SBIR).

2.2 Content-Based Image Retrieval (CBIR)

Visual feature extraction is the basis of most content-based image retrieval techniques. Because of perception subjectivity and the complex composition of visual data, there is no single best representation for any given visual feature. Multiple approaches have been introduced for each of these visual features. Each of them characterises the feature from a different perspective. Typically, the research on CBIR is based on two types of visual features (Alhwarin et al. 2008): global and local features. Global feature based algorithms aim at recognising concepts in visual content as a whole. First, global features (i.e. colour, texture, shape) are extracted and then statistical feature classification techniques (i.e. Naïve Bayes, SVM) are applied.

2.2.1 Global Features

A colour feature is one of the most widely used visual features in CBIR. It is simple to represent. Common colour features or descriptors in CBIRs include a colour-covariance matrix, colour histogram, colour moments, and colour coherence vector (Jing et al. 2003; Tong & Chang 2001; Zheng et al. 2004; Wang et al. 1999). The colour histogram is the most commonly used representation technique, statistically describing the combined probabilistic property of the three colour channels (RGB). Most of those colour features though efficient at describing colours, are often not directly related to any high-level semantics.

A texture feature refers to the patterns in an image that present the properties of homogeneity that do not result from the presence of a single colour or intensity value. Texture provides important information in image classification as it describes the content of many real-world images such as fruits, skins, clouds, trees, bricks, and fabrics. However, it is almost impossible to describe texture in words, because it is more a statistical and structural property. Texture features commonly used in image retrieval systems include spectral features, such as those obtained using Gabor filtering (Ma & Manjunath 1997) or wavelet transform (Wang et al. 2001), statistical features characterising texture in terms of local statistical measures, such as the Tamura texture features (Tamura et al. 1978). Among the various texture features, Gabor features and wavelet features are widely used for image retrieval and have been reported to match the results of human vision studies well (Ma & Manjunath 1997; Wang et al. 2001).

Shape features are important features of images though they have not been as widely used in CBIR as colour and texture features. Shape features, however, have shown to be useful in many domain specific images such as involving man-made objects. For colour images, however, it is difficult to apply shape features in contrast to colour and texture due to the inaccuracy of segmentation. Despite the difficulty, shape features are used in some systems and have shown a potential benefit for CBIR. For example, in (Mezaris et al. 2003), simple shape features such as eccentricity and orientation are used.

The main advantages of global feature-based algorithms are that they are simple and fast. However, there are limitations in their reliability for object recognition under changes in image scaling viewpoints, illuminations, and rotation. Thus, local feature based algorithms are also being investigated and used.

2.2.2 Local Features

Visual features are often inconsistent due to variations in camera angle, orientation, camera viewpoint or change in illumination. In recent years, Lowe (1999) proposed a new approach for extracting distinctive invariant features from images that can be used to perform reliable matching between different views of an object or scene. This method has been called the *Scale Invariant Feature Transform (SIFT).* There are three major steps of visual feature extraction (Ke & Sukthankar 2004; Alhwarin et al. 2008):

- *Scale-space extrema detection*: searches over all scales and image locations using a Difference-of-Gaussian algorithm to identify potential points of interest (keypoints) that are invariant to scale and rotation.
- *Keypoint localisation*: candidate keypoints are localised by fitting a 3D quadratic function to the scale-space local sample point. The local extrema with low contrast that correspond to edges are eliminated because they are sensitive to noise.
- *Orientation assignment*: one or more orientations are assigned to each keypoint location with respect to a local image gradient direction. Future operations are performed on image data that has been transformed relative to an assigned orientation, scale, and location for each feature, thereby providing invariance to this transformation.

A local image descriptor is built for each keypoint based on the image gradients in its local neighbourhood. The region around a keypoint is divided into boxes i.e. 4x4 boxes. The gradient magnitudes and orientations within each box are computed and weighted by an appropriate Gaussian window, and the coordinate of each pixel and its gradient orientation are rotated relative to the keypoints orientation. Finally, a 128 dimensional vector (SIFT descriptors) is built. This descriptor is orientation invariant because it is calculated relative to the main orientation. To achieve the invariance against change in illumination, the descriptor is normalised to a unit length.

SIFT descriptors represent the detected keypoints of images in the form of a multi-dimensional, e.g., 128-dimensional, real-valued feature vector that can be used for further processing e.g. image classification. Recently, Mikolajczyk (2003) has compared several descriptors for image classification and found that SIFT descriptors performed the best. The main advantages of SIFT are:

- It has simple linear Gaussian derivatives. Hence, it is more stable to typical image perturbations, such as noise, than a higher Gaussian derivative or differential invariants (Csurka et al. 2004).
- There are far more components to these feature vectors e.g. 128 rather than 12 to 16. Hence this has a richer and potentially more discriminative representation.

In order to retrieve images based on visual features, users need to give an example image as a query or QBE. Then the system will transform the query to match with low-level features in the image repository. Images which have visual content similar to the query will be retrieved and provided to the user. However, this approach is quite difficult for non-expert users who may have problems in selecting an example image to represent the desired images.

2.3 Semantic-Based Image Retrieval (SBIR)

As stated in the previous sections, CBIR relies on multiple low-level features e.g. colour, shape and texture describing the image content. For image retrieval using QBE, the retrieval process consists of a query example image, input by a user. The image features are used to find images in the database which are the most similar to the query image. A drawback, however, is that these features cannot adequately represent the image semantics. Extensive experiments on CBIR systems show that low-level content descriptors often fail to describe the high level semantic concepts familiar to users (Zhou & Huang 2003).

However, users often want to search for images at a conceptual level e.g. images containing particular objects (target search). This is called "*Semantic-based Image Retrieval (SBIR)*". Image descriptions, in turn, are derived using low-level data-driven methods. A semantic search by a user and the low-level syntactic image descriptors may be disconnected. Since this problem is unresolved, this thesis is focused on different methods to associate higher level semantics with data-driven observables.

Numerous techniques were introduced to bridge the semantic gap between numerical image features and the richness of human semantics. Early IMR approaches are based on low-level features which fail to capture the underlying conceptual associations in images. Therefore, this thesis proposes a technique for reducing the "semantic gap" that comprises three main characteristics:

- Making use of both the visual content of images and the textual captions. Both visual and textual metadata are used together to allow the system to annotate an image properly.
- Using an ontology KB to define high-level semantics. The use of an ontology KB appears to be a promising way by which higher-level semantics can be incorporated into techniques that capture the semantics through automatic analysis.
- Generating a semantic template (ST) to support high-level image retrieval. Ontologies support for semantic image retrieval processes is added through defining their classes and relationships.

Since visual data cannot be used in its original form, it needs to be analysed and transformed into a format which can be used by Knowledge Management (KM) systems. In this thesis, the knowledge refers to the content of image (e.g. athlete name, sport type), context of image (e.g. where and when the image takes place) and image features (e.g. file size, file type, SIFT descriptors). Typically, knowledge is extracted by image processing algorithms and transformed into metadata. This metadata describes the content, context and visual features of an image document which is manipulated and processed by standard information retrieval methods. Image data contains large numbers of unstructured and dynamic visual features. How to establish a good knowledge representation model to represent visual content is very important for IMR. From much research (Swain et al. 1997; Zhao & Grosky 2002; Hu & Bagga 2004; Song et al. 2004; Smith & Chang 1997), an ontology KB seems to be an effective model to represent visual content enabling an IMR system to perform semantic search. In part through the emergence of the Semantic Web, ontologies have evolved as a key enabling technology to provide machine understandable semantics (Dasiopoulou et al. 2007). After Tim Berneres-Lee introduced the idea of the Semantic Web (Berners-Lee et al. 2001), ontologies have become more popular among researchers in IR research area. Informally, the ontology of a certain domain is about its terminology (domain vocabulary), all essential concepts in the domain, their classification, their taxonomy, their relations (including all important hierarchies and constraints) and domain axioms (Gasevic et al. 2009).

2.3.1 Knowledge (Ontology-based) Representation Techniques

There is no single "best" theory that explains large complex human knowledge organisations. Likewise, there is no single ideal knowledge representation technique suitable for all applications. When building a practical intelligent system, developers should select the best knowledge representation technique to suit the application. The knowledge representation models have been firstly developed in the field of artificial intelligence and then applied to other research areas. The most frequently used models to represent knowledge include object-attribute-value triplets, logic, and semantic networks. They are briefly described as follows (Gasevic et al. 2009).

- Object–attribute–value (O–A–V) triplet is a technique used to represent the facts about objects and their attributes. More precisely, an O–A–V triplet asserts an attribute value of an object. For example, the English phrase "The colour of the ball is yellow" can be written in O–A–V form as "Ball–colour–yellow".
- Logics aim at emulating the laws of thought by providing a mechanism to represent statements about the world. The representation language is defined by its syntax and semantics, which specify the structure and the meaning of statements, respectively. Different logics make different assumptions about what exist in the world (e.g. facts) and on the beliefs about the statement (e.g.

true/false/unknown). The most widely used and understood logic is First-Order Logic (FOL), which assumes (1) the existence of facts, objects (individual entities), and relationships among objects; and (2) the beliefs of true, false, and unknown for statements. For example, "$\forall x\, Athlete(x) \Rightarrow Person(x)$" means that "All Athletes are people".

- Semantic networks are graphs made up of objects, concepts, and situations in some specific domain of knowledge (the nodes in the graph), connected by some type of relationship (the link/arcs). Figure 2-3 shows an example. The network represented expresses several inter-related facts: Devid Beckham is an Athlete, which is a kind of human. A human has hands, and the number of hands is 2. The graphical notation used to represent semantic network is not standardised, but the things that can be typically represented in such network are concepts (classes), their instances (objects), attributes of concepts, relationships between classes, and objects, and values of attributes. As in object-oriented programming, objects/instances should normally have values for all attributes defined in the concept. There may be default values for some attributes (e.g., a human normally has 2 hands). Kind-of relationships denote inheritance. Also, parts of semantic networks are easily recognised as O–A–V triplets (e.g., "hands–number of–2").

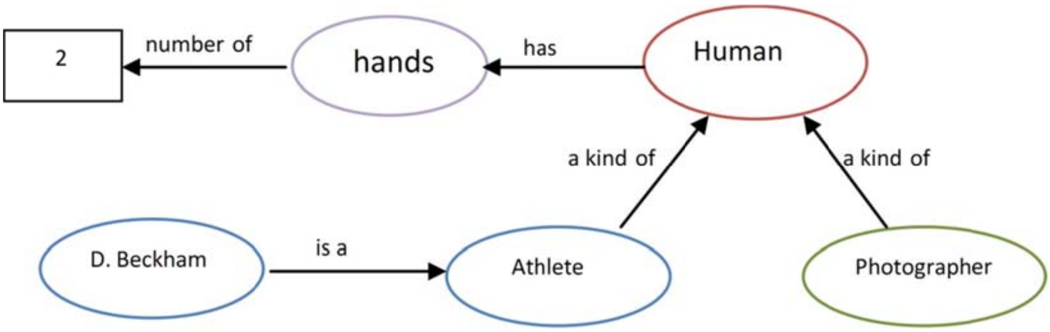


Figure 2-3 A simple semantic network

It is easy to insert new nodes, relationships, and values into a semantic network. The technique has proven to be intuitively clear and easy to learn, and is widely used by AI specialists and psychologists alike. However, semantic networks are known to be weak in representing and handling exceptions to the knowledge they represent. In large networks, all nodes must be carefully examined for exceptions. Simpler types of semantic network may also not explicitly model the richness in some relationships such as able-bodied athlete should have two and only two hands.

An ontology provides a useful way for formalising the semantics of the represented information. In principle, an ontology can actually be the semantic domain for an information system in a concrete and useful manner (Meersman 1999). For IMR, ontologies are used for reducing the semantic gap by building the knowledge for summarising, discovering, classifying, browsing and retrieving, and annotating images. Ontology-based frameworks are proposed for IMR in numerous collections. Ontologies for manual image annotation and semantic retrieval for animal collections have been presented in (Schreiber et al. 2001). In (Hollink et al. 2003), an ontology for considering art images has been presented. In (Sinclair et al. 2005), ontologies also have been applied successfully for handling museum collections. These frameworks have validated the assumption that ontologies could help improve information retrieval effectiveness e.g. it is possible to find relevant documents that are syntactically not similar to the query terms.

The usual way to use an ontology in an image retrieval system is to annotate images with the elements of an ontology. This annotation is usually called *semantic annotation* or *semantic metadata*. The process attempts to predict annotations of entire images using all the information present. This task is referred as an *annotation*. However, one might attempt to associate particular words with particular image substructures which is called *correspondence* (Barnard et al. 2003)**.** The annotation process is the process which assigns metadata in the form of captioning or keywords to an image. From the survey, there are three types of image annotation; manual, automatic, and semi-automatic annotation. Manual annotation refers to the annotation process carried out by human beings. The automatic annotation is done by a computer system for the entire process. Sometimes, this is called auto-annotation. The last type of annotation process is semi-automatic annotation. It depends on users' interaction to provide an initial query and feedback and a system's ability to use these annotations (Wenyin et al. 2001).

2.3.2 Advantages of Using Ontologies for IR

Gaservic (2009) has summarised the benefits of using ontologies in IR and IMR systems as follows.

2.3.2.1 Semantic similarity

Compared to syntax-based similarity measures in traditional textual information retrieval, ontology-based measures can exploit the ontology structure and thus measure semantic similarity. For example, the term list (“Sun LEE”, “Taekwondo”, “Gold medal”) has no syntactic similarity to the term list (“Sydney”, “2004”) but the two list are semantically similar. This is because Sun LEE was the Taekwondo winner (Gold medal) in the Sydney 2004 Olympic Games. The similarity can be obtained using the relationship between concepts e.g. Sun Lee-<winner>-Taekwondo-<participate>-Sydney 2004 Olympic Games.

2.3.2.2 Semantic annotation

Using ontology elements as semantic annotations, the knowledge stored in an ontology can be used to identify the potentially interesting elements to be included in a semantic annotation that may not be explicitly mentioned in the text. For example, if many entities, locations and athletes related to the Sydney 2004 Olympic Games appear in a text caption and also the time context is that of Sydney 2004, the annotation system can infer that the Sydney 2004 Olympic Games itself is relevant to an image. Then, it should be added to the semantic annotation, even if the phrase “Sydney 2004” does not appear in the text caption. These are some of the potential uses of ontologies in IMR. There are other potential uses of an ontology e.g. query expansion (Haubold et al. 2006; Natsev et al. 2007) but they will be not described in detail because they are considered to be out of the scope for this thesis.

2.4 MPEG-7 and Ontology-based KB

MPEG-7 has been established as a standard tool for describing multimedia since 2001. The goal of MPEG-7 is to enable advanced searching, indexing, filtering, and access of multimedia by enabling interoperability among devices and applications that deal with multimedia descriptors (MPEG Requirements Group 1999). The scope of the MPEG-7 standard is to define the syntax and the semantics used to create multimedia descriptions. MPEG-7 specifies four types of normative elements: Descriptors, Description Schemes (DSs), Description Definition Languages (DDLs), and coding schemes. These elements are used to deal with multimedia information e.g. low-level features such as colour, shape, motion, and audio as well as high-level features such as the title or the author's name. MPEG-7 defines the syntax of descriptors and description schemes using a DDL as an extension of the XML Schema language. An MPEG-7 specification can be divided into two main parts; *audio* and *video*. The audio and video parts define descriptors and description schemes for audio data, e.g. timbre, and video data, e.g. colour layout, respectively.

Although, MPEG-7 has become a key standard to multimedia research in searching, filtering and retrieval, MPEG-7 structure is not deployed directly for the KB in this thesis because of several reasons:

First, syntactic interoperability: although Semantic Web is a new trend for developing internet-based information system introduced by W3C, the combination of the use of Semantic Web and MPEG-7 can cause a lack of syntactic interoperability (Nack et al. 2005) because of the different languages used e.g. XML, the MPEG-7 (DDL), the resource description framework (RDF), RDF Schema (RDFS), and Web Ontology Language (OWL). The XML syntax underlying the MPEG-7 DDL facilitates platform and application independence and human and machine readability. However, because the MPEG-7 DDL merely adopts an XML Schema, i.e., it defines syntactic elements to represent structures in schemata form, the MPEG-7 DDL lacks particular media-based data types. Unlike RDF Schema or ontology-based modelling, the MPEG-7 DDL does not support the definition of semantic relations e.g. Beckham-plays-football. Thus, the MPEG-7 approach of fusing language syntax and schemata semantics is problematic and only a first step toward a language with which semantic descriptions of multimedia can be syntactically established. Identifying semantically relevant syntax elements in the semantic-related schema and including them in the MPEG-7 DDL is an open issue and would allow the Semantic Web to use the implicit semantics of low-level MPEG-7 binary descriptors.

Second, semantic interoperability: the MPEG-7 DDL is designed based on XML Schema rather than on RDF Schema. This choice was mainly political, because RDF Schema was not at the time MPEG-7 was being specified a W3C recommendation and thus was not chosen. Choosing XML Schema as the serialisation syntax has far-reaching consequences. As a syntax-oriented language, MPEG-7 DDL provides weak or light-weight semantic support, supporting only named attributes and unnamed hierarchical relationships (Nack et al. 2005). This DDL also provides inadequate reasoning services, particularly in subsumption-based reasoning on class and property hierarchies. Note that extensions to XML such as RDF, RDF-S and OWL can offer richer support for semantics and reasoning whilst also taking advantage of the use of the underlying XML serialisation as a standard data exchange format.

Third, no formal semantics is provided: MPEG-7 is based on XML Schema that defines syntax level aspects. Since no formal semantics are provided, the applications cannot access the meaning of the descriptions (Troncy & Carrive 2004). Finally, the goal of MPEG-7 in this semantics and interoperability context is unclear and one can wonder if it aims to be an exchange format or a real machine understandable and processable representation for multimedia descriptions.

In contrast to MPEG-7, Knowledge-based models built using ontologies have become a very popular topic, not only in AI but also in other disciplines of computing. There are many definitions of the concept of ontology in AI and in computing (Guarino et al. 1995; Hendler 2001; Kalfoglou 2001). For example,

'*Ontology is a set of knowledge terms, including the vocabulary, the semantic interconnections, and some simple rules of inference and logic for some particular topic* (Hendler 2001)'

The important parts in Hendler's definition are the *semantic interconnections* and *inference and logic*. The former says that an ontology specifies the meaning of relations between the concepts used. The later part means that ontologies enable some forms of reasoning. In addition, an ontology facilitates accurate and effective communication of meaning. This opens up the possibility for knowledge sharing and reuse, which enables semantic interoperability between intelligent processes and applications (Gasevic et al. 2009). An ontology-based KB provides a number of useful features for knowledge representation in general. This thesis summarises the most important of these features based on the surveys from (Gruber 1993; Schreiber et al. 1994; Guarino et al. 1995; Chandrasekaran et al. 1999; McGuinness 2003).

First, vocabulary: an ontology provides a *vocabulary* (or the names) for referring to the terms in a subject area. Ontologies are different from such human-oriented vocabularies in that they provide *logical statements* that describe what the terms are, how they are related to each other, and how they can or cannot be related to each other. It is not the vocabulary as such that qualifies as an ontology, but the conceptualisations that the terms in the vocabulary are intended to capture (Chandrasekaran et al. 1999). In addition, ontologies are usually designed to specify terms with *unambiguous meanings,* with semantics independent of reader and context.

Second, taxonomy: a *taxonomy* (or *concept hierarchy*) is a hierarchical categorisation or classification of entities within a domain (Gasevic et al. 2009). It is also a clustering of entities based on common ontological characteristics. The vocabulary and the taxonomy of an ontology together provides a *conceptual framework* for discussion, analysis and information retrieval in a domain.

Third, knowledge sharing and reuse: the major purpose of ontologies is *knowledge sharing* and *knowledge reuse* by applications (Gasevic et al. 2009). This is because ontologies provide a description of the concepts and relationships that can exist in a domain and that can be shared and reused among intelligent agents and applications. There are many ways that knowledge-based system using ontologies can be shared and reused (Neches et al. 1991), for example:

- through the inclusion of source specifications—the content of one module is copied into another one at design time, is then possibly extended and revised, and is finally compiled into a new component;
- through the runtime invocation of external modules or services—one module invokes another, either as a method from a class library or through a Web service;

- through communication between intelligent processes such as agents—the messages that intelligent agents send to and receive from each other can have various kinds of knowledge as their content.

2.5 Summary

This chapter introduced the fundamentals of image retrieval and ontology KBs that are basis of the techniques described in later chapters. First, a classical IMR model has been described in detail and consists of four main processes: query issuing, query transforming, indexing (document representation), and matching and ranking. Next, two major techniques for IMR have been identified; CBIR (e.g. colour, texture, and shape), and SBIR (knowledge-based model). The major problem of CBIR is that these techniques cannot represent the meaning of an image effectively, the so called semantic *gap problem*. The initial idea to reduce the semantic gap using ontology-based models has been introduced. Next, knowledge representation techniques have been described and the advantages of using an ontology-based model for IMR are pointed out. Finally, a comparison between MPEG-7 and an ontology-based KB has been analysed to assess which method should be deployed for this framework. From the discussion in section 2.4, it is clear that the KB using ontology architecture is a better choice for designing and developing IMR for this research.

The scope of this thesis is to combine CBIR and SBIR techniques (intersection area in Figure 2-4) in order to facilitate image categorisation, knowledge based construction, and image annotations and retrieval. SIFT descriptors are exploited for image classification and an ontology model is deployed for the remaining tasks.

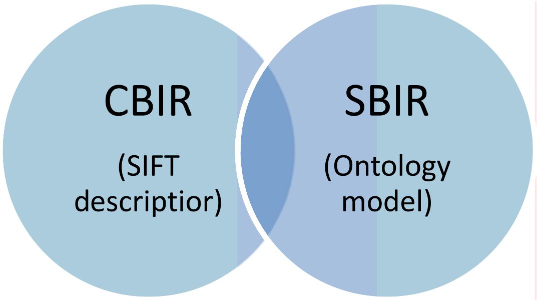


Figure 2-4 Main focus area of this thesis

Chapter 3

Survey and Analysis of the State-of-the-Art Frameworks

In this chapter, the state-of-the-art of image retrieval systems will be reviewed and analysed with respect to their best practices and limitations. The major challenges for IMR are first identified in section 3.1. This survey focuses on an analysis of ontology-based IMR with respect to the requirements identified in section 3.2. The survey is divided into two main parts. First, some recent works that use different knowledge representation techniques for knowledge-based IMR systems are described. Second, frameworks that use SIFT descriptors for visual content representation are reviewed. Finally, these state-of-the art frameworks are compared.

3.1 Problem Analysis of the Image Retrieval Systems

The goal of IMR system is to facilitate users to retrieve desired images easily. More recently, an enormous effort by researchers has been undertaken in this research area in order to invent more intelligent IMRs. There are, however, several factors inhibiting the use of image retrieval systems to retrieve images needed by users. Usually, text captions and low-level features are two main information sources to understand the meaning of an image. This kind of information has limitations in representing the meaning of an image to computer system. Typically, the major problems leading to a failure of IMR in finding the relevant images are as follows.

Natural language is often ambiguous. For example, many words can have the same meaning, the so called *synonymy problem*. Alternatively, one word can refer to more than one thing, the so called *polysemy problem*. For example, “Thailand” could be either a host country or the athlete's nationality. This can be called the *Natural language (NL) vagueness* problem. Synonymy, polysemy, and inflection of words can all cause algorithms, which use exact matches to search terms and to terms in image caption or file name, to fail.

There are many cases where specific concepts are not mentioned directly in the document text but are still relevant to documents semantically. For example, if users search for images about "Spain", traditional IMR is not able to retrieve images containing only "Barcelona" even though Barcelona is a city in "Spain". This kind of relationship is called an indirectly relevant relationship. Therefore, this is called the *indirectly relevant concepts* problem. For those systems that generate annotations based upon text captions, image captions however, they may not supply all information required by the concepts in an ontology KB. In such cases, the text caption of an image is said to be *absent* or *uncertain*. This problem lowers the precision with which the semantics of images can be defined. When text captions are not supplied, only visual features can be used to understand image content. The extracted visual features themselves cannot be used to represent the content of images directly. Thus, they need to be processed and transformed into another model to represent image content effectively. In several cases, images that have different visual appearances may be semantically similar e.g. eyes and nose. They are semantically similar as they are facial features. Hence, a visual features model should be invariant to *visual appearances*. In other words, it should tolerate *visual heterogeneity* (polysemy problem).

Further uncertainty occurs when the system tries to annotate an image based upon the detected objects in a scene. Some key objects might be not detected because of object recognition errors or because the input image is incomplete, e.g., a camera angle in an image may not capture all the important objects used to identify a particular scene. These uncertainties degrade the annotation power.

3.2 Formal Requirements of Image Retrieval Systems

From the major challenges described in the previous section, the requirements for a new IMR to meet those challenges are shown in TABLE 3-1. The requirements for the IMR can be divided into two main categories, *system* and *user* requirements.

TABLE 3-1: Requirements for a new IMR system based upon the general challenges for IMR identified in section 3.1

Requirements	Descriptions
System Requirements:	
NL Vagueness of text captions	IMR should handle the vagueness of natural language use in text captions and user queries
Indirect relevant concepts	IMR should exploit semantic relations to find any relevant concepts that are not mentioned explicitly in the captions or query.
Absence of text captions	IMR should be able to understand the meaning of an image even when the text captions are not supplied.
Visual heterogeneity (polysemy problem)	IMR should be invariant to the visual heterogeneity of an object and its low-level features, e.g. extracted visual features possibly found in different objects and scenes.
Uncertainty in image classification	IMR should be able to classify the content of an image even if some of the useful objects needed for interpretation are missing.
Ontology incompleteness	IMR should tolerate the incompleteness of semantic metadata, concepts and relationships in ontologies.
User requirements:	
Usability/ Learnability	IMR should be easy to use by non-expert users.
Result ranking	The results should be ranked by descending order according to the degree of similarity between a user's query and images in the collection.
Response time	IMR should not spend too long a time to return all relevant results to a user
Relevance Feedback	IMR should allow users to provide feedback to the system to say whether or not those results are relevant. Consequently, new search results should be more relevance to users' interests.

However, response time and relevance feedback are not the main focus in this thesis. In the next section, the existing systems will be analysed against these requirements.

3.3 Survey of State of the Art Frameworks

There have been hundreds of published papers concerning IMR in last decade. The current survey here is restricted to a discussion of ontology-based frameworks for IMR. Although this research focuses on image type visual content retrieval, this subsection also covers some related work with respect to text, video, and audio content because some of the ideas used by these types of multimedia retrieval can aid IMR.

Visual features are an important source of input to aid the system to understand image content. As mentioned in section 2.2 (p.13), local features e.g. SIFT descriptors are more robust than global features because they are invariant to camera angle, orientation, scaling or change in illumination. Therefore, the survey in this section also covers state of the art frameworks that use SIFT descriptors to represent visual content.

3.3.1 Ontology-Based KB Frameworks for IMR

Tansley *et al.* (2000) and colleagues proposed a data-driven approach for IMR that uses Web images and their surrounding textual annotations as the source of training data to bridge the semantic gap. The annotations come from many sources such as the surrounding text, file names, and alterative tags. This system allows users to query by both query-by-example and query-by-keyword which fulfils the user requirement in terms of useability because the system is easy to use. Using the WordNet thesaurus (Miller 1995), the system can solve the NL vagueness problem of text captions. The Image Thesaurus framework supports automatic metadata extraction from Web pages using a vision-based Web page analysis technique. From the experiments shown in the papers, however, the results have not been ranked according to the degree of relevance to the users' query. Such a ranking could be used as a tool to filter out the less important images to users. In addition, for the query-by-keyword part, only purely text queries are supported which means that no semantic search happens; only a traditional full-text search is executed. Since the framework relies on the surrounding text of an image, it cannot interpret the image content properly when the surrounding text is not supplied. As a consequence, the framework fails to associate a key term with a key region which leads to a failure to construct a complete knowledge base. In other words, the system cannot handle the uncertainties in the surrounding text of an image properly. The system uses only the generated knowledge base as a main source for finding relevant information which means it fails to find the relevant information for a user's query if the knowledge base is incomplete.

There are two key problems in using ontologies for information retrieval: semantic extraction from keywords and document indexing. Khan *et al.* (2004) proposed an automatic mechanism for the selection of concepts from a query which is expressed in the form of a NL query, addressing the first problem. For the second problem, the author uses an ontology-based technique to create a concept-based model corresponding to information selection requests. To improve the retrieval precision, query expansion is used to deal with natural type user queries. Although this framework is used to process the audio data, the innovative ideas in this work are valuable for the design of a framework for IMR. The results demonstrate the power of the proposed mechanism over keyword-based search techniques by providing many different levels of abstraction in a flexible manner leading to a greater accuracy in term of precision and recall (section 7.2.2 p.115). The NL vagueness of text captions and indirect relevant concept requirements are supported using a disambiguation algorithm. A result ranking function is used. Nonetheless, the query mechanism in this framework is very complicated. Therefore, the response time could be very slow to return the results to user. In other words, it might not be practical in a real situation because the system efficiency seems low. Ontology incompleteness is also neglected by the authors. This means the system will return null to a user when it cannot find the desired information in an ontology.

Schreiber *et al.* (2001) explored the use of knowledge contained in ontologies to index and search collections of images. The system comprises two main parts. First, a "Structure of a photo annotation" is an image annotation ontology that specifies an annotation's structure independent of the particular subject matter domain. This ontology provides a description template for annotation construction. Second, a "Subject matter vocabulary" provides the vocabulary and background knowledge describing features of the image's subject matter using a domain-specific ontology.

The system uses terminology from WordNet for annotations. To query images, the tool searches a database for annotations that have all properties specified in the target description filled with values that are equal to or are specialisations of the value in the target. A value is a specialisation if the domain of the related property is an RDFS class, and the value in the annotation is a subclass of the value in target description. However, the results from querying are not ranked. The annotation system uses an ontology model which offers some benefits over keywords. It guides the annotation process using restrictions and default information. Using an ontology, the system can find the indirectly relevant relationships between concepts in the knowledge base. The query tool uses a subsumption hierarchy as represented in the ontology. Users can create queries by selecting the concepts from that hierarchy. However, this seems difficult for some users who are not familiar with these hierarchies. In addition, users are not allowed to use keywords in queries that do not exactly match the concepts names in the subsumption hierarchy. Although the system exploits WordNet for describing the general terms of image features and colour of animals, the authors did not use WordNet to disambiguate a user's terminology. Another issue concerns comparing the results of the proposed system with WWW search engine e.g. Alta vista, could be considered unfair because Alta Vista uses machine generated indexes whereas the proposed system indexes data by hand. This is because indexing data manually usually provides better retrieval results than automatic indexing. However, indexing data manually is not scalable for use in a large IMR system. In addition, the proposed system relies solely on an ontology. It could fail to find relevant images when the ontology is incomplete. Further, the annotation process has been done manually. Accordingly, the system is not concerned with the uncertainty of image interpretation.

The proposed approach of Dasiopoulou *et al.* (2007) comprises two main modules, a semantic analysis module and a retrieval module. The main steps of the framework can be described briefly as the domain ontology provides the conceptualisation and vocabulary for structuring content annotations. Thus, semantic browsing is used, based on an ontology model, to facilitate query formulation. The analysis module is used to guide the analysis process and to support the detection of certain concepts defined in a domain ontology using low-level features. A domain analysis of content using ontologies uses common classes, resulting in a unified ontology-based framework that can handle visual content at a semantic level. With the presented framework, the analysis process starts by segmenting the input image and extracting low-level visual descriptors and spatial relations in accordance with the domain ontology definitions. Then, an initial set of hypotheses is generated by matching the extracted low-level descriptors against the ones in the objects prototypes including the ones in the domain knowledge definitions. To reach the final semantic annotation, spatial context domain knowledge is used, whereby the image semantics are extracted and the respective annotation metadata are generated. The user interface of this system, however, is a kind of image browsing rather than keyword querying, therefore the usability and learnability is not high. This is because some users might not really know how to use the system to browse for their desired pictures. The system exploits the low-level features of an image and matches the extracted low-level features to higher level conceptualisation. Thus, the system is able to interpret the image content without using text descriptions. However, for the case where some objects needed for interpretation are absent, this leads to a system failure to annotate the image. Therefore, the uncertainty in image classification requirement is partially fulfilled. Based on the information in the paper, the results are shown to the users without ranking. Users have to consider which images are relevant to their interests from the set of returned images by themselves. Its complex algorithm might affect the scalability of the system to query images from a huge image database. This issue needs to be evaluated. Furthermore, the system does not support the ontology incompleteness requirement. This means the system could fail to find the relevant images when the ontology is incomplete. Furthermore, it fails to cope with the uncertainty in image segmentation and with matching low-level features in an object at a higher semantic level, i.e., when the system cannot match any object concepts with the extracted low-level descriptors. In other words, it cannot cope with object recognition errors.

The Multimedia Thesaurus (MMT) (Tansley, 2000) is a network of concepts, semantic relationships between concepts, and media representation of concepts. A MMT is constructed for a collection of annotated images using a semi-automatic process. The concept network is a manually selected subset of the Dewey Decimal

Classification (DDC) schema. Annotated images are connected to concepts by matching the textual annotations of the images with the textual descriptions of the concepts using latent semantic indexing (LSI) of a large corpus of art documents. The semantic layer in MMT offers a facility for expressing explicit semantic relationships. A LSI algorithm links concepts leading to a system that partially supports NL vagueness in image descriptions. The query expansion algorithm can find the indirectly relevant concepts which are not explicitly mentioned in the query. Finally, the results from querying are ordered according to the distance of similarity rather than evaluated using classical measurements, precision and recall. The system expands the original query to a more sophisticated query. This might affect the response time to display the results to user. Users input a query using an example of an image to the system (QBE). It is questionable how the system can handle some media data that cannot be classified into any concepts or when one media feature can belong to multiple concepts. The uncertainty issue needs to be addressed. Finally, the network of concepts is only a main repository of knowledge for searching relevant data. Thus, the incompleteness of the knowledge is an issue of concern.

Benitez (2000) proposed new methods for extracting semantic knowledge from annotated images. The MediaNet framework extends semantic knowledge frameworks such as thesauruses and Ontologies by including perceptual knowledge, and exemplifying concepts and relationships using multimedia. Perceptual Knowledge is discovered through grouping images into clusters based on their visual and text features. Semantic knowledge is extracted by disambiguating the senses of words in annotations using WordNet and image clusters. Finally, concepts networks of media, MediaNets, can be summarised by merging statistically similar concepts.

Extracted medianets (semantic networks) enable new ways of searching for images using multiple modalities. The medianets are used for expanding, refining, and translating user queries across different modalities in the image retrieval system. First, the proposed technique detects relevant concepts in incoming queries and adds other semantically and perceptually similar concepts. Then, images are retrieved and ordered based on how closely they match the incoming query and the relevant concepts. The results of experiments demonstrate an improved retrieval effectiveness using the techniques proposed to support semantic queries. The MediaNet retrieval system returns images that are both semantically related and visually similar to the input queries. In addition, the results from retrieving process have been ranked according to their relevance scores. The evaluation of the framework has been done using classical measures, precision and recall. Nevertheless, this framework does not provide a method to deal with the ambiguity of visual features and the uncertainty of image interpretation. Therefore, the system cannot annotate images effectively when the key visual features are missing.

3.3.2 Visual Features-Based Frameworks for Visual Content Representation

Typically, the object detection task, also known as object recognition, can be classified into two categories: global and local features based algorithms. However, there are limitations of the global features approach with respect to reliability of object recognition under changes in image scaling viewpoints, illuminations, and rotation. Thus, local features are also used. Several advantages of local features versus global features for object recognition and visual content categorisation have been given by Lee (2005). Local feature based algorithms focus mainly on *keypoints*. Keypoints are salient patches that contain rich local information about visual content. Moravec (1977) defined the concept of "point of interest" as distinct regions in images that can be used to match other regions in consecutive image frames. The use of the Harris corner detector (Harris & Stephens 1988) to identify interest points and to create a local image descriptor at each interest point from a rotationally invariant descriptor in order to handle arbitrary orientation variations has been proposed in (Schmid & Mohr 1997). Although this method is rotation invariant, the Harris corner detector is sensitive to changes in image scale (Alhwarin et al. 2008). Thus, it does not provide a good basis for matching images of different sizes. Lowe (1999) overcomes such problems by detecting the key locations over the image and its scales through the use of local extrema in a Difference-of-Gaussians (DoG). Lowe's descriptor is called the Scale Invariant Feature Transform (SIFT). SIFT algorithm is an algorithm for visual features extraction which are invariant to image scaling, translation, rotation, and partially invariant to illumination changes and affine projections. Further improvements for an object recognition technique based on SIFT descriptors are as follows. Ke *et al.* (2004) improved the SIFT technique by applying Principal Components Analysis (PCA) is used to reduce the dimensions of SIFT descriptors. Therefore, it makes local descriptors more distinctive, more robust to image deformations, and more compact, compared to the standard SIFT representation. Consequently, it increases image retrieval accuracy and matching speed. Recently, it has also been used in a special technique namely "Bag of Visual Words" (BVW). The BVW visual content representation has been drawn much attention by the computer vision communities, as it tends to code the local visual characteristics towards at the object level (Zheng et al. 2008). The main advantages of the BVW technique are its simplicity and its invariance to transformations, as well as, occlusion, and lighting (Csurka et al. 2004). There are hundreds of publications about visual content representation using the BVW model as it is a promising method for visual content classification (Tirilly et al. 2008), annotation (Wu et al. 2009), and retrieval (Zheng et al. 2008).

The BVW technique is motivated by an analogy to the "*bag of words"* representation for text categorisation (Joachims 1998; Tong & Koller 2002; Cristianini et al. 2002). A bag of keypoints corresponds to a histogram of the number of occurrences of particular image pattern in a given image. Keypoints are grouped into a large number of clusters so that those with similar descriptors are assigned into the same cluster. By treating each cluster as a "*visual word*" that represents the specific local pattern shared by the keypoints in that cluster, a visual-word vocabulary can describe all kinds of local image patterns. With its keypoints mapped into visual words, an image can be represented as a *"bag of visual words"*, or specifically, as a vector containing the (weighted) count of each visual word in that image, which can be used as a feature vector in image classification. By coding the statistics of local image regions independently, the BVW approach achieves the robustness in handling variable object appearances caused by changes in pose, image capturing conditions, scale, translation, clutter and occlusion etc. Hence, the effectiveness of the BVW approach motivates its deployment for object recognition to aid the visual content interpretation.

Using an analogy to words in the text document, the BVW model also has some similar limitations when representing visual content. Just as words in traditional text documents may be ambiguous, so may visual words in an image. The ambiguity lies in two areas: synonymy and polysemy. One possible way to disambiguate multiple word senses is to combine visual words into a larger unit. In other words, the collection (or co-occurrence) of several visual words is likely to be much less ambiguous. Yuan *et al.*(2007a) proposed a solution to tackle this issue using the "*visual phrase*" technique based upon a likelihood ratio test method and an improved frequent itemsets mining (FIM) algorithm. As a result, meaningful itemsets are discovered. Their experiments showed that a visual phrase lexicon represents images better than a visual words lexicon. The major weakness of visual phrase approach is that it merely considers the co-occurrence information among visual words but *neglects spatial information* amongst the visual words. To tackle this issue, Chen *et al.* (2009) proposed a novel method that combines SIFT descriptors with the real spatial constitution of image content called a "Gaussian Mixture Model" (GMM). GMM provides spatial weighting for visual words to facilitate content based image retrieval. The spatial constitution of an image is represented as a mixture of Gaussians in the feature space that decomposes an image into *n* regions. The spatial weighting scheme is achieved by weighting visual word according to the probability of each visual word belonging to each of the *n* regions in the image. The cosine similarity between spatial weighted visual word vector pairs is used as distance measurement between regions.

Tirilly *et al.* (2008) proposed a method that exploits spatial information of visual words into a *visual sentence* using Principal Component Analysis (PCA) and a method to eliminate *useless visual word* based on geometric properties of the keypoints and the use of probabilistic Latent Semantic Analysis (pLSA). In addition, a Language Modelling (LM) method is applied in order to classify keypoints of images. Experiments show that Tirilly's technique can significantly improve image classification, compared to SVM. However, the problem of this method is how to choose a consistent axis. Currently, there is no automatic method to choose this. The use of PCA over the coordinates of the visual words requires effective background elimination of visual words. Furthermore, applications of PCA tend to focus on images that contain only one object.

Besides the spatial information issue, another limitation with the BVW model is "*noise word removal*". In text-based information retrieval, noise words (also called *stop words)* represent frequently occurring, insignificant words that appear in a text document e.g. a, an, the, in, of, on, are, be, if, into, which etc. In contrast, in visual content processing, it is difficult to define what the noise words are, also called meaningless visual words or "*non-informative visual words*" because they are often domain specific, thus it is difficult to create a standard list for visual content. Yuan *et al.* (2007a) and Zheng *et al.* (2008) proposed a statistical significance measure of the visual phrase to detect redundant high-order word-sets and meaningless frequent word-sets.

Another issue is the simple clustering method is not sufficient to preserve the semantic information among visual words. Wu *et al.* (2009) proposed a novel method to preserve the semantic information during visual word generation process by learning a codebook from which semantically relates features that can be mapped to the same visual word. However, they do not perform noise reduction before constructing visual words which is the greatest obstacle that distracts the clustering algorithm.

The noisy visual words affect the performance of the clustering algorithm to represent the content of an image. Therefore, a major challenge with the BVW technique is handling noisy visual words. Tirilly *et al.* (2008) proposed a method to eliminate *useless visual words* based upon the geometric properties of the keypoints and the use of the probabilistic Latent Semantic Analysis (pLSA). However, the main disadvantage of this method is that it ignores the correlations between a specified word and other concepts in the collection. Some words might appear less in conjunction with one concept but appear more in conjunction with other concepts - these words could be the featured words. In such a case, deleting low-probability words will decrease the accuracy of categorisation. Rather than identifying unimportant visual words, Yuan *et al.* (2007b) tried to discover unimportant information at an upper layer (a larger unit) of visual words, *the meaningless phrases*, through measuring the likelihood ratio (the statistical significance measure) of those visual phrases. The *top-k* most meaningful word-sets with the largest likelihood ratio will be selected and the rest are considered as meaningless visual phrases and will be discarded. However, this method ignores the coherency (the ordering of visual words) of component visual words in a visual phrase.

Another important limitation of existing BVW models (Jiang & Ngo 2009; Yuan et al. 2007a; Zheng et al. 2008) is they do not take into account the spatial information between keypoints during visual words construction through applying a simple *k*-mean clustering algorithm. To tackle this issue, Wu *et al.* (2009) tried to preserve the semantic information in visual content during visual word generation through manually separating objects in a training phase. In this separation, any keypoints detected are considered to be relevant and put into the same visual word for each object category, so, that any linkage between visual words and a higher level semantic object category can be established.

Visual heterogeneity is perhaps the greatest obstacle for image and video categorisation and for retrieval systems which rely solely on visual appearance. In fact, different visual appearances might be semantically similar at a higher semantic conceptualisation. Recently, the use of probability distributions of visual word classes has been proposed (Zheng et al. 2008) based upon the hypothesis that semantically similar visual content will share a similar class probability distribution. However, this method is less useful when unrelated visual words exist that coincidently have a similar probability distribution. Yuan *et al.* (2007a, 2007b) overcomes the visual heterogeneity problem by proposing the use of a pattern summarisation technique that clusters correlated visual phrases (word-sets created from the Frequency Itemsets Mining algorithm, FIM) into phrase classes. Any phrases in the same class are considered as synonym phrases. Nevertheless, this method might not always be effective because the visual words constructed using the FIM algorithm ignore semantic information leading to less robustness and poorer quality visual words. These methods are mainly based on the use of a vector space model to represent visual content. Recently, a hierarchical model has been exploited by Jiang *et al.* (2009) to tackle the semantically similar visual content issue. A soft-weighting scheme is proposed to measure the relatedness between visual words. A hierarchical model is constructed using an Agglomerate clustering algorithm to capture the "*is-a*" relationship of visual words. However, the model of Jiang (2009) has some drawbacks. The hierarchical model is a binary model; there is no any multiple-parents relationship. Due to these limitations, the visual content analysis system inefficiently interprets the content of visual data and obtains a low visual content retrieval performance.

TABLE 3-2: How the problem requirements are supported in the state of the frameworks

State of the art frameworks	NL Vagueness of text captions	Indirectly relevant concepts	Absence text captions	Visual heterogeneity	Uncertainty in image interpretation.	Ontology incompleteness	Usability/Learnability	Result ranking
1) Image Thesaurus (2000)	✓	✓	✗	✗	✗	✗	✓	✗
2) Khan *et al.* (2004)	✓	✓	✗	✗	✗	✗	✓	✓
3) Schreiber *et al.* (2001)	✓	✓	✗	✗	✗	✗	P	✗
4) Dasiopoulou *et al.* (2007)	?	✓	✗	✗	✗	P	P	✗
5) Multimedia Thesaurus (2000)	P	✓	✗	✗	✗	✗	P	✓
6) MediaNet (2000)	✓	✓	✓	✗	✗	?	✗	✓

Legend: ✓= fulfils the requirement, ✗ = does not fulfil the requirement,
? = not know, P = partially fulfils the requirement.

3.4 Discussion

Based on the analysis in section 3.3.1, one can draw the conclusion that almost all systems address the issue of the vagueness of natural language and indirect relevant concepts. This is because it is one of the main motivations to use an ontology model in the first place, i.e., to avoid the problems that this kind of vagueness causes. Most of the systems combine various techniques such as natural language processing, full-text search and ontology reasoning. TABLE 3-2 classifies the related work against the various IMR requirements. From the analysis of the existing solutions, their major limitations can be defined as follows:

1) Almost all the surveyed frameworks exploit text captions and text descriptions to annotate images. In fact, text captions do not always accompany an image. Consequently, those frameworks will fail to generate annotations and metadata for that image. This leads to a system that has less robustness and stability. An IMR system should be able to analyse and interpret image content even if a text caption is not provided. This means the system should not solely rely on text captions but can also work on the image itself, i.e., low-level visual features can be used to retrieve those images where no text caption is supplied.
2) The surveyed systems cannot handle the uncertainty in object detection properly. For instance, when surveyed systems map the lower-level features to a higher level object conceptualisation, an extracted feature may possibly belong to multiple concepts of objects due to the visual heterogeneity (one visual appearance has multiple meanings). Therefore, an image representation model should support this requirement.
3) A related issue concerns handling image interpretation uncertainty. Basically, image content can be inferred from the detected objects in the scene shown in an image. However, no single object detection algorithm works perfectly. Some desirable objects might be not detected due to object detection errors or due to the lesser quality of an input image. Hence, an IMR system should be able to interpret the visual content properly even if some needed objects are missing.
4) The surveyed frameworks lack support for alternative searching mechanisms when one fails to find relevant data in the knowledge based model. These frameworks only rely on the information contained in the ontology model, regardless of the issue that an ontology model can rarely be built to cover all information in a domain of knowledge. Thus, IMR system should provide a backup method to compensate the ontology-based approach when it fails to find relevant information. The backup method enables the ontology-based image retrieval system to have more information tolerance. This means the

system can operate when there the required information does not exist in an ontology model or even when a domain ontology does not exist or is not accessible.

TABLE 3-3: Comparison of various surveyed systems against the type of limitations

Surveyed System	Uninformative visual word/phrase elimination	Semantic information preservation	Visual word senses disambiguation	
			Multiple visual words	Hierarchical structure
1) Yang et al. (2007)	Document frequency, X^2 statistics, Mutual information, and Pointwise mutual information	✘	✘	✘
2) Yuan *et al.* (2007a; 2007b)	Statistical significance measure	✘	Visual phrase	✘
3) Tirilly *et al.* (2008)	pLSA	✘	Visual sentence	✘
4) Zheng *et al.* (2008)	Statistical significance measure	✘	Visual phrase/ Visual synset	✘
5) Jiang *et al.* (2009)	✘	✘	✘	Agglomerate clustering
6) Wu *et al.* (2009)	✘	✓	✘	✘
7) Wang *et al.* (2009)	✘	✘	✘	Hierarchical Spatial Markov model

Another matrix can be drawn to classify the surveyed work with respect to the various limitations of existing BVW model as described in section 3.3.2. These limitations are highlighted as follows:

1) Semantic information can be lost during visual word construction when simple clustering algorithms are used, e.g. *k*-mean algorithm. Therefore, this process needs an alternative clustering algorithm which can group semantically relevant keypoints more effectively.
2) Uninformative visual word detection based solely on document frequency does not sensibly handle well visual words that appear in only a few concepts leading to them having a low document frequency and being considered as noisy visual words. In fact, sometimes, these could feature as visual words but they just

appear in a few concepts. To improve this, the noisy visual words detection process needs to be normalised.

3) Visual phrases and visual sentences disambiguate multiple word senses using a statistical computation. However this type of method does not represent the actual semantic relationship between visual words. A more effective model, to disambiguate visual word senses and to represent the semantics of visual content, is a hierarchical ontology model. Nevertheless, existing hierarchical clustering algorithms e.g. an Agglomerate clustering algorithm, is often impractical to capture semantic relationships between concepts of visual information as they do not represent the semantics of visual content efficiently.
4) A hierarchical model used in some frameworks is structured as a binary tree (Jiang & Ngo 2009) or a balanced tree that holds an equal number of child nodes in every parent node (Wang et al. 2009) but this is often not so practical. In fact, the number of child nodes in each parent is not necessarily equal and the relationship between nodes is not always a “is-a” relationship as proposed by Jiang (2009).

3.5 Summary

This chapter addressed the main challenges of IMRs in section 3.1. These challenges can be considered as important requirements that IMRs should fulfil. In section 3.2, the formal requirements for IMRs are identified. Section 3.3 has reviewed the relevant literature for knowledge-based frameworks for IMR and visual features-based frameworks for image representation. Finally, these state of the art frameworks have been analysed, compared and discussed based on the requirements of IMR mentioned in TABLE 3-1.

TABLE 3-2 and TABLE 3-3 compare the surveyed frameworks with respect to IMR requirements and limitations. The surveyed frameworks are at best only a partial solution to fulfil the application domain requirements given in TABLE 3-1. In order to support all those requirements, the framework presented in the next chapter is proposed.

Chapter 4

Overview of the Proposed Framework

As discussed in Chapter 3, there is no state-of-the-art framework that can fulfil all of the identified requirements. Therefore, a new approach is proposed to support those requirements. This chapter provides a basic, intuitive understanding of this proposed framework and in the subsequent sections, more details will be revealed. The main goal of designing this framework is to exploit background knowledge from exiting images and associated text captions which is extracted and stored in ontologies in order to increase the IMR effectiveness. Although state-of-the-art systems solve several challenges, there are still some issues which remain unresolved as stated in Chapter 3 (TABLE 3-2 and TABLE 3-3). The solution devised in this research addresses those limitations, and those are the main contributions of this thesis.

4.1 High-Level Architecture View

The main components of the proposed framework are: (1) the knowledge acquisition and analysis component, (2) the knowledge storage component, and (3) The image retrieval and filtering component. The components of the architecture shown in Figure 4-1 and are described in more detail below.

The knowledge acquisition and analysis component: visual and textual information are used to establish a KB for the presented framework. HTML documents will be parsed to extract text information (*Linguistic analysis*) using shallow natural language processing (NLP) to generate useful information that can improve the quality of the text-to-ontology mapping.

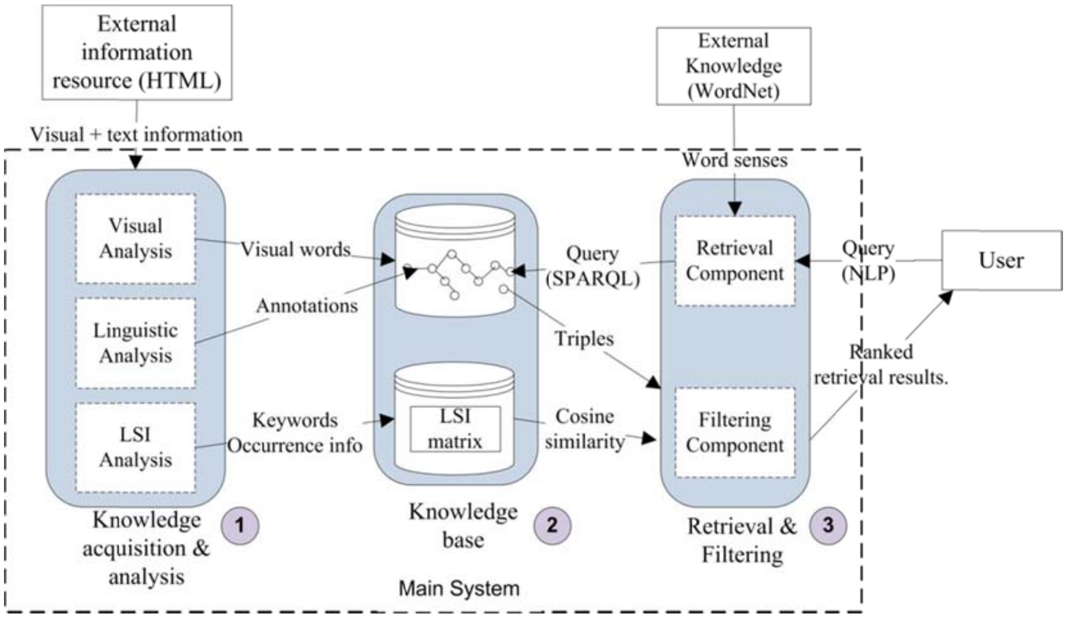


Figure 4-1 Overview of the ontology-based image retrieval system

This information includes various types of text snippets such as athlete name, host country, host city, Olympic Games event, sport name, date, time etc. The information created by NLP is stored in the form of NLP annotations (XML format). These annotations are used in the next step in order to create an initial metadata representation. Next, these text snippets are matched with ontology concepts and instances. During this matching process, text snippets are disambiguated using WordNet. Finally, the initial set of metadata is expanded using semantic rules and probability model in order to find the ontology concepts that are relevant to an image but which may not be explicitly mentioned in the text caption. This step is called *Knowledge discovery*.

Visual analysis: images are processed in order to extract visual information using a Difference of Gaussian (DoG) detector (Lowe 1999). DoG is used to automatically detect keypoints from images. Then, the detected keypoints are converted to a Scale-Invariant Feature Transform (SIFT) descriptor which is a 128-dimensional real-valued feature vector (Lowe 2004). These SIFT descriptors are further quantised into clusters called *visual words* using SLAC clustering algorithm (AlSumait & Domeniconi 2008). These visual words are then further processed to construct BVW that counts the numbers of visual words assigned to each cluster. To disambiguate visual words, an ontology model is used to group relevant visual words into the same cluster (also called *bag*) and each bag is used to determine the object category e.g. athlete, horizontal bar, javelin, or pole. Finally, all detected objects in an image are combined together in order to categorise an image using a Bayesian network model. The main advantage of the Bayesian network is that it can handle an uncertainty when some objects are missing to aid the classification. It can predict what the content (sport type) in an image should be, based upon a statistical calculation of previous data.

Besides the visual analysis component, a *Latent Semantic Indexing* (LSI) algorithm is deployed to find the interrelationships between terms and images. After any textual information is parsed from the HTML documents, stop words (unimportant words) are removed. Then the remaining keywords are stemmed to link variations in words to a common base or root form. LSI creates a term-image matrix which contains the numbers of terms (row) that have appeared with the image (column). This frequency is used to determine the degree of importance of those terms to the image. Each term will be assigned a weight to show the importance of that term to the image regardless the length of text captions. This information is useful when the presented IMR fails to find relevant images due to the incompleteness of semantic metadata in ontology.

The Knowledge base: the semantic metadata generated by knowledge acquisition and the analysis component is stored in a RDBMS in the first instance. A RDBMS, MySQL, is used in this framework. To be able to use this data in a semantic context, this metadata is mapped to the ontology to give the data a well-defined meaning. RDBMS offers a robust management system to enable semantic metadata to be shared, exchanged, and integrated from different sources and enables applications to use data in different contexts. The semantic metadata model itself is represented in RDF[1], the Resource Description Framework, which can be represented as a directed graph consisting of nodes and directed arcs linking pairs of nodes. RDF was chosen as it represents a compromise between supporting named relationships between concepts, being efficient to parse and supporting a standard query language.

[1] The Resource Description Framework (RDF) adds support for named associations between concepts to XML (http://www.w3.org/TR/REC-rdf-syntax/ RDF)

Although, other representations such as OWL[2], the Web Ontology Language, are more semantically expressive, OWL is far more complex to process and query. During the mapping process, keywords from RDBMS are used to disambiguate the word senses using external knowledge resource e.g. WordNet.

The image retrieval and filtering: the query and retrieval process deals with keywords from users. The main functions of this process are to eliminate stop words and to generate a SPARQL query automatically and to retrieve the relevant images in response to user queries. This can be done by performing a SPARQL query on RDF file. The SPARQL[3] query language is a W3C recommendation for querying data from RDF documents which forms part of the knowledge base. SPARQL query language is selected to retrieve information from RDF files because it has the important features of a query language for RDF compared to RDQL[4]. It supports INSERT, UPDATE, and DELETE commands which are not supported in RDQL. In addition, a nested SELECT, can be processed using SPARQL and the results can be ranked (ascending or descending) which is very useful for information retrieval system. A query returns a list of instance tuples that satisfies the query. To ensure that the results are relevant to the query, a cosine similarity measurement is also performed to measure the similarity between a query and semantic metadata in the KB. Combining the results with LSI could solve the incompleteness of semantic metadata in ontology. The decision whether or not the data is relevant to the user query is not binary i.e. relevant or not relevant, but rather, it is probabilistic i.e. information receives a relevance rank.

4.2 Implementation

The architecture is realised using several technologies in order to achieve a good performance and a good quality of results. In detail, the following technological decisions were made:

[2] The Web Ontology Language (OWL), http://www.w3.org/TR/owl-ref/. OWL adds support to RDFS for range and domain constraints, existence and cardinality constraints, transitive, inverse and symmetrical properties and for logic.

[3] SPARQL query, (http://www.w3.org/TR/rdf-sparql-query)

[4] RDQL query, http://www.w3.org/Submission/RDQL

- Windows XP is used as the computer operating system for the experiments in this thesis.
- For web server technology, an Apache Tomcat Server[5] version 5.5 is deployed.
- For storing extracted information from text captions temporarily, a relational database is used. Extracted information in this application includes the keyword list and term frequency for LSI. This is because they are just simple syntactical information, and there is no need to explore the rationale for this as it is, only retrieved for indexing purposes. In this thesis, the open source database, MySQL server version 5.0 is used to store this extracted information.
- For the main KB, Jena[6] API version 2.5 is used to construct an ontology KB. Jena provides off-the-shelf methods for creating ontology classes, properties, and assigning instances to these classes and properties. It is able to generate RDF/XML files using Jena for further processing such as querying information from RDF file using SPARQL.
- During the LSI algorithm implementation, Matlab is deployed in this research to perform complex computations e.g. Matrix operations and the Singular Value Decomposition (SVD) computation. Matlab is an extremely powerful tool to deal with statistic and matrix operation. It allows you to use its libraries in an easy and convenient way. Java code is able to access Matlab functions using a Java API or through executing Matlab's executable (.exe) files.
- For metadata generation, the ESpotter[7], Natural Language Processing (NLP) tool, is used. It is able to recognise the name of person, organisation, date, and location from various input types of documents e.g. Text or HTML documents. In addition, it allows a user to export the recognised entities into

[5] Apache Tomcat Server (http://tomcat.apache.org)

[6] Jena API (http://jena.sourceforge.net)

[7] ESpotter (http://kmi.open.ac.uk/people/jianhan/ESpotter)

an XML file for further processing. Another popular framework for NLP is GATE framework[8]. GATE has several more functions than ESpotter e.g. exporting metadata to Oracle, PostgreSQL, or ontology. Although GATE framework comes with several powerful functions more than ESpotter, ESpotter is preferred because it is simple to use and serves the purpose. In addition, based upon experience, GATE is a complicated framework to export metadata, as XML, for further processing.

- For the image processing module, Matlab is deployed to complete this task. A Difference of Gaussian (DoG) detector to automatically detect keypoints from images. The detected keypoints are represented using SIFT descriptor which is a 128-dimensional real-valued feature vector.
- For the image classification, the Weka[9] framework version 3.6 is employed in order to perform the image classification task. Weka provides several built-in classifiers e.g. Naive Bayes and Singular Vector Machine (SVM). The classifying results are stored in CSV file format for further processing.
- For Web GUI, the graphical user interface of the prototype is a Google like interface, implemented using HTML and Java Server Page (JSP) technology. During searching, the web GUI is responsible for parsing the textual user query into a semantic query representation. The semantic query is later processed by the searching algorithm implemented in Java.

4.3 Summary

In this chapter, an overview of a proposed framework has been provided. The high-level system architecture of the presented solution was introduced and comprises three main parts, knowledge acquisition and analysis, a knowledge base, and image retrieval and filtering. Two main sources of knowledge for the framework are text captions and low-level features of images. Text captions are processed using NLP that extract metadata which is transferred to the knowledge base. Visual data is analysed and low-level features are extracted using DoG and then transformed into SIFT

[8] GATE (http://www.gate.ac.uk)

[9] The Weka framework (http://www.cs.waikato.ac.nz/ml/weka)

descriptors. SIFT descriptors are used to generate visual words which when incorporated with an ontology model can be used to recognise objects in an image (section 5.3.1, p.58). Semantic metadata is represented using RDF and can be queried using SPARQL to retrieve relevant content. Finally, several technologies were identified that need to be used and integrated in order to realise the architecture and to achieve the goals of this research that were given in Chapter 1.

In the next chapter, the designing of the KB model will be explained in order to support the main three parts of the proposed framework. More detail of each part of the framework will be described as well as its construction and deployment process.

Chapter 5

Semantic-based Image Retrieval System

The goal of this chapter is to attempt to build an effective Knowledge based (KB) system that can store any extracted image metadata in a form to enhance the retrieval performance for IMR. One of the basic features of KB system is the knowledge representation that holds knowledge in a machine-processable format and to some extent a human-processable format – the latter may require some sort of data transformation. To achieve this, an ontological type of knowledge base has been adopted to this framework. It is not easy, however, to build, maintain and deploy an ontology due to its complexity. Hence, strategies and tools for ontology creation, deployment and evaluation are introduced. Since this thesis does not focus on the ontology maintenance phase, the strategy to maintain the ontology will not be addressed. The representation for ontologies used in this research is first described. Later, the method to deploy an ontology that involves extracting low-level features from images and matching those low-level features to higher-level semantics will be described. Finally, the strategy to perform semantic image retrieval is revealed.

5.1 Knowledge-based Model for IMR

A KB can be represented by semantic networks of nodes and arcs. Nodes represent instances or concepts (classes) e.g. “Agassi” or “Athlete” and the links represent relationships between nodes e.g., “Agassi” *is-a* “Athlete”. Creating the KB model, also called the *Conceptualisation*, is probably the most important part and the most complex task of the ontology building process.

5.1.1 Knowledge-based Model Designing Steps

Building of an ontology is a part of a knowledge representation process. As such, it relies on common understanding of how people represent, understand and acquire knowledge. In order to manipulate facts and ideas, people tend to impose a structure on their knowledge: Similar things are grouped together according to certain common attributes or characteristics which they process and then use to describe that whole group. That is called a *Concept*. An ontology is composed of concepts and their relationships. To produce a formal ontology, an ontology representation language is selected in order to formalise an ontology conceptualisation and produce a hierarchy of concepts (organisation of concepts into a "kind-of" relation). The process of conceptualisation of a domain in the following steps (Poslad 2009):

- Defining the concept taxonomy. This is a core to most knowledge representation languages. It defines the nature of categories in terms of generalisation and specialisation.
- Defining a set of relations used between concepts. This set of relations can itself be organised into a hierarchy. In addition, properties of concepts are also need to be defined.
- Defining constraints for the values in a relation, e.g., how many values there can be, constraints on the value, e.g., positive integer etc.
- Defining axioms on relations and concepts, e.g., a proposition that is always true such as "is-owner" is the inverse of "belongs-to".

Typically, a created knowledge model requires a process of refinement in order to improve and validate it. This process is repeated until the system has achieved the desired level of performance.

5.1.2 Open (rather than Closed) Knowledge Model

A closed knowledge-based model refers to a model that relies only on metadata defined in the model, e.g., a RDBMS KB model. A closed world KB model implies that data not present in the KB is *false*, while an open world KB model states that the data is not presented in the KB is *unknown* (Poslad 2009). A closed KB model is more useful in domains where its knowledge can be fixed before deployment. It is less useful in some domains because it limits the scope of information that can be searched. It returns an empty result for a user query when relevant metadata in the KB is not present. An example of a framework that tries to tackle the limitation of a closed KB is Llorente and Rüger (2009). They proposed a method for image annotation that overcomes the limitation of a closed KB (WordNet) using semantic relatedness measures based upon keyword correlation on the Web. The most important benefit of this approach is that it is not limited to the scope of topics provided by a training set but annotation keywords come from a web-based search engine. The KB in this thesis is designed as an open KB model. Hence, it does not rely only on the metadata presented in the KB. Unknown terms in a query will be forwarded to a LSI module in order to perform a second search on LSI vector space model. A LSI model provides term frequency information which can be used for an implicit semantic search when this information is not contained in the ontology KB.

5.1.3 Knowledge-based Development Tools

For editing the concepts and properties of an ontology, several graphical ontology editors are available for this task such as Protégé[10], SWOOP[11], and KAON[12]. Currently, ontology development and ontology population still remain predominantly a manual, human resource intensive task. These existing graphical ontology editors have various problems with ontologies containing many instances (Nagypál 2007). First, they do not scale well e.g. they are simply too slow on big ontologies. Second, their GUI does not support browsing and editing of a large number of entities. Protégé-frame was selected to design an ontology manually. During the deployment phase, the ontology created is parsed and processed using the Jena. Protégé is also used for visualising the structure of the KB in a graphical way because it is easier for human readers to understand the ontology structure using visual information rather than only text descriptions.

[10] Protégé ontology editor (http://protege.stanford.edu)

[11] Swoop ontology editor (http://code.google.com/p/swoop)

[12] Kaon ontology editor (http://kaon.semanticweb.org)

5.2 Knowledge-base Development for the Sports Domain

The structures and relationships for the sports domain ontology in this framework are specified based upon the structure of sports information used by the Olympic organisation website. Although there are several sports genres in the Olympic Games, this research focuses only on the genres for Athletics sports as this is sufficient to bring a number of challenges to the proposed system. First, the visual appearances of events in the Athletic sport are quite similar. This is very challenge to the system to categorise them properly based upon the extracted low-level features. Second, some objects appearing in images are shared between two or more events e.g. a horizontal bar can appear in the high jump and the pole vault event. As such, they are ambiguous. This brings another challenge to the system in order to annotate an image properly and to deal with a polysemy issue. After surveying data at the Olympic organisation website, three main classes of ontology have been defined, a Sport Domain, Visual Features and an Image Annotation ontology (Figure 5-1). The *Sport Domain ontology* provides the vocabulary and background knowledge describing image content. The *Visual features ontology* provides low-level information such as SIFT descriptors, resolution, image size. This ontology is designed to aid image interpretation. The *Image Annotation ontology* specifies an annotation's structure independent of the particular subject matter domain (sports domain in this case). This ontology provides the description template for annotation construction.

5.2.1 Sport Domain Ontology

The Domain ontology describes the vocabulary and background knowledge of the subject domain. In this ontology, there are several main classes e.g. *Athletes, Sports (Athletics* disciplines only*), Sport Equipment, Events (Olympic Games)* with numbers of properties (slots in the Protégé terminology) to correspond with two aspects from the Image Annotation ontology.

- *Athletes* provide information about athletes e.g. Name, Gender, Nationality, Participation and Sport and Medal. Participation refers to the Olympic Games in which an athlete has participated. Since an athlete can participate in more than one Olympic Games, this property can contain multiple values. It connects to the Olympic Games class in order to retrieve the title and year of the Olympic Games in which that athlete participated. For the Sport and Medal property, it stores sport genres that an athlete won in the Olympic Games and the medal e.g. Gold, Silver, and Bronze. This property also connects to the Sport class in order to retrieve Sport information, e.g. name and type of sport.

- *Sport (Athletics)* provides information about sport genres. This thesis focuses on the Athletics sport only. It contains sport name, type (i.e. Field, Track, Road, or Combined), distance of each Athletics sport (if applicable) and equipment. The equipment property is connected to the Sport equipment class in order to obtain equipment information.
- *Sport Equipment* contains sport equipment information. This class is shared by the Sport (Athletics) class using is-a relationship. For example, a Bar (horizontal bar) in the pole vault event *is-a* Bar in the Sport Equipment class.
- *Events (Olympic Games)* contains information about the Olympic Games e.g. opening date, closing date, host city and country, formal title i.e. XXX Olympic Summer Games, short title i.e. London 2012, and its official website.

5.2.2 Visual Features Ontology

Visual Features: this class represents the metadata about the representation of an image as a whole e.g. format (jpg, bmp), size, resolution of a picture, and low-level features i.e. SIFT descriptor (section 2.2.2, p.15).

5.2.3 Image Annotation Ontology

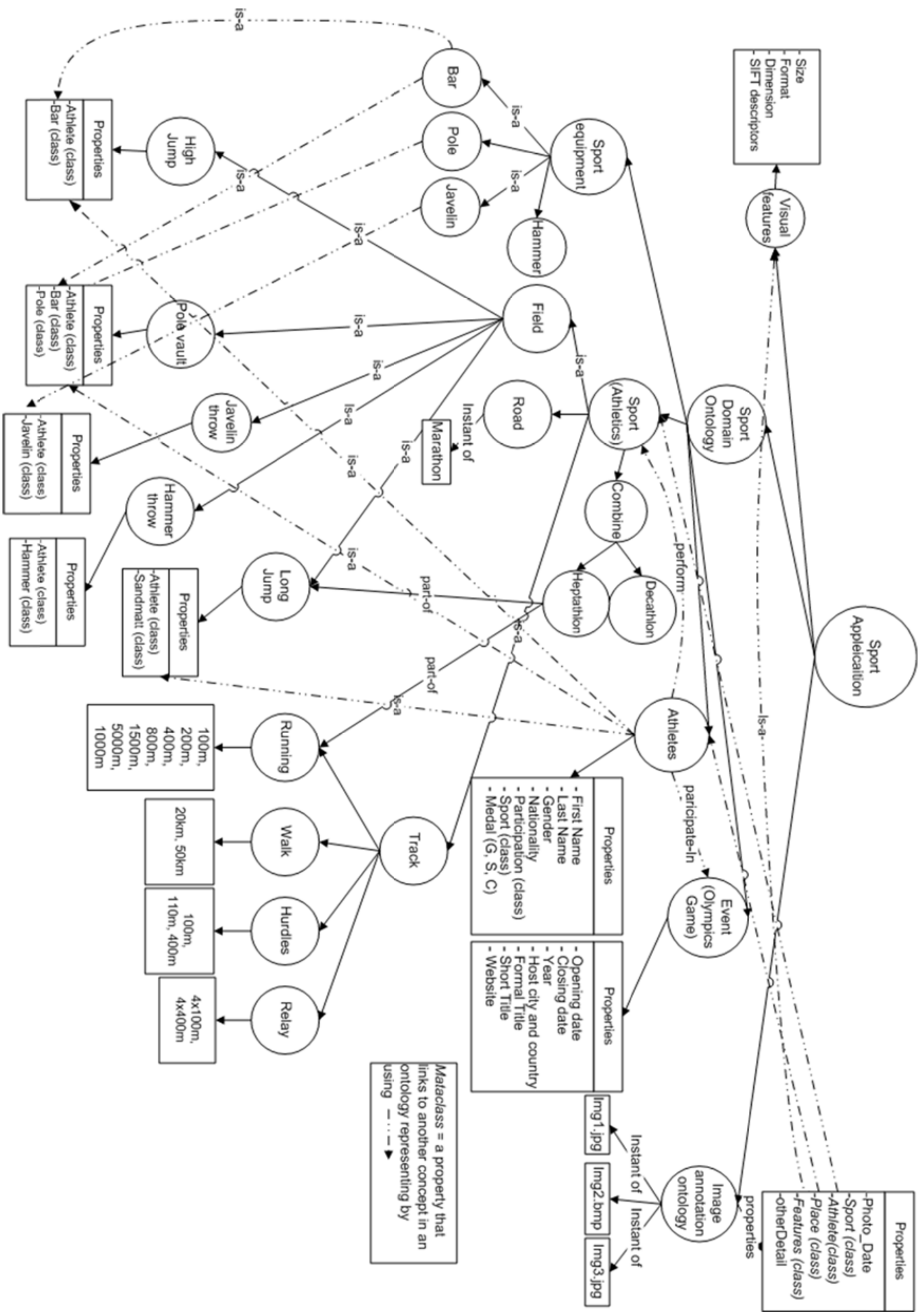
When looking at an image, the following aspects are distinguishable:

- *What does the photo depict?* For example, the image shows that an athlete performed a high jump at the Athens 2004 event. This kind of information is called a photo’s *subject matter feature*.
- *When and where was the photo taken?* This is called *Image information*. This metadata relates to the image such as place and the date a photo was recorded.

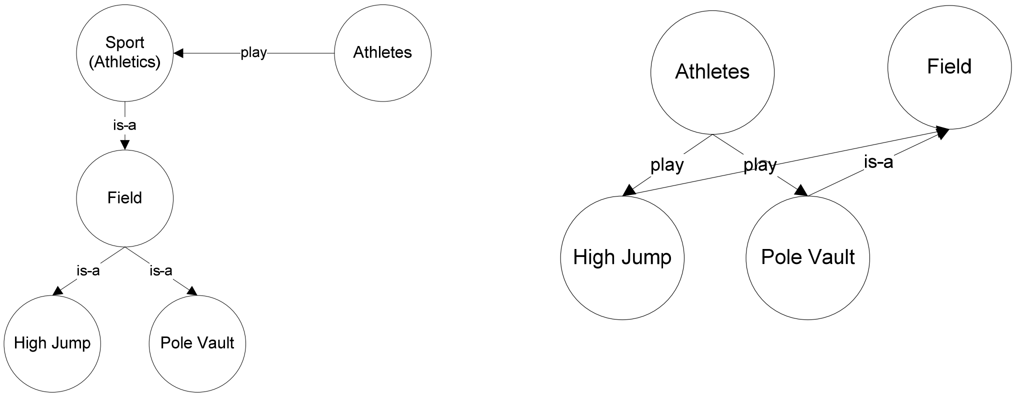
The main purpose of the Image Annotation Ontology design is to store the annotations of images for the sports domain corresponding to the aspects mentioned above. This ontology provides a template for sports image annotation. The Image Annotation Ontology and the Domain Ontology are linked together via properties defined as a *metaclass* in the Image Annotation Ontology. A metaclass refers to a property that is a concept in the Sport Domain ontology. In Figure 5-1, a metaclass is the property that have (*class*) after the properties' name e.g. Sport(class) and Athlete(class). To design an ontology for the KB, sports data was collected from the Olympic organisation website and the sport taxonomy is shown in TABLE 5-1. Later, this taxonomy will be transformed into the ontology model.

TABLE 5-1 Sport taxonomy

1. Olympic Games - Opening date, Closing date, Year - Host city and country - Formal Title - Short Title - Website **2. Athletes** - First Name, Last Name - Gender (Male, Female) - Nationality - Sport and Year (Multi-value) - Participation and Medal (Multi-value) **3. Sports (Athletics)** **3.1 Combined Events** - Decathlon men - heptathlon women	**3.2 Road Events** - Marathon (men, women) **3.3 Track Events** - Running (men, women) (100m, 200m, 400m, 800m, 1500m, 5000m, 1000m) - Hurdles (men, women)(110m, 400m) - Walk (men, women) (20km, 50km) - Relay (men, women) (4x100m, 4x400m) **3.4 Field Events** - Discus throw (men, women) - Hammer throw (men, women) - High jump (men, women) - Javelin throw (men, women) - Long jump (men, women) - Pole vault (men, women) - Shot put (men, women) - Triple jump (men, women)


Sport Application
Sport Domain Ontology
Visual features
-Size
-Format
-Dimension
-SIFT descriptors
Sport equipment
Bar
Pole
Javelin
Hammer
is-a
Field
Road
Sport (Athletics)
Instant of
Marathon
Combine
Decathlon
Hepathlon
Track
High Jump
Pole vault
Javelin throw
Hammer throw
Long Jump
Running
Walk
Hurdles
Relay
Properties
-Athlete (class)
-Bar (class)
-Pole (class)
-Javelin (class)
-Hammer (class)
-Sandmatt (class)
part-of
perform
100m, 200m, 400m, 800m, 1500m, 5000m, 1000m
20km, 50km
100m, 110m, 400m
4x100m, 4x400m
Athletes
paricipate-In
- First Name
- Last Name
- Gender
- Nationality
- Participation (class)
- Sport (class)
- Medal (G, S, C)
Event (Olympics Game)
- Opening date
- Closing date
- Year
- Host city and country
- Formal Title
- Short Title
- Website
Image annotation ontology
properties
-Photo_Date
-Sport (class)
-Athlete(class)
-Place (class)
-Features (class)
-otherDetail
Img1.jpg
Img2.bmp
Img3.jpg
Mataclass = a property that links to another concept in an ontology representing by using — · — ▸

B



(a) ☑

(b) ☒

Figure 5-2 Ontology design choices.

The hierarchy in the KB is designed as follows: Level *i* of the hierarchical model is a general concept of level *i*-1. Some object classes have been decomposed into sub-classes that are the sub-classes of the parent class. Edges represent several types of relationships between classes e.g.   is-a, participate-in, part-of, and instance-of. It is noticeable that some classes e.g. visual features and sport equipment are not shown in the sport taxonomy. This is because they are added to an ontology in order to aid visual content categorisation and retrieval. The ontology structure in Figure 5-1 is designed as follows. First, redundant concepts are minimised. For example, the Sport Equipment and Athlete concept are designed as main concepts rather than as subclasses of sport events e.g. high jump, pole vault, long jump or javelins throw. This is because subclassing the Sport equipment and Athletes concepts from sport events can lead to concept redundancy (every sport event has the Sport Equipment and Athletes concepts). Second, the ontology structure in Figure 5-1more efficiently facilitates sport image retrieval. Figure 5-2 illustrates an example of ontology design options for the KB. In Figure 5-2, (a) is selected as the structural model used for this framework because it is a more efficient model for SBIR than Figure 5-2 (b). For instance, "*Find all athletes who play Field sport events*", the scheme in Figure 5-2 (b) may make it more difficult to find an answer. For Figure 5-2 (a), the system easily find the answers for this query by following the route *Athletes-<play>-Sport-<is-a>-Field* and then all concepts under the Field concept will be retrieved. In contrast, the structure in Figure 5-2 (b) can make it more complex to answer this simple query.

This is because the system needs to follow *all outgoing links* of all sport events until it finds the *Field* concept. Thus, this structure is less effective and scalable when the ontology covers a large number of sports. The search performance will be degraded. In addition, the ontology structure in Figure 5-2 (b) does not model the concepts and relationships in the real world satisfactorily. Typically, the upper level concepts are a generalisation of the lower level concepts. In Figure 5-2 (b), the Athletes concept is not a generalisation of high jump and pole vault event and they do not share any common properties. Therefore, the structure and relationships between concepts in Figure 5-2 (a) is selected to represent sports domain information.

5.3 Knowledge-based Acquisition

The main focus of this process is to extract knowledge from *visual data* and *text captions* and to store this extracted knowledge in a semantic model. First, the low-level features are extracted and processed using a BVW technique in order to detect objects in images. Then, the extracted visual information is mapped to higher-level semantic conceptualisations based on the ontology model. Later, image captions are exploited to enhance the semantic interpretation of the extracted visual features.

5.3.1 Semantic Visual Analysis

In this section, techniques for visual content analysis and classification are demonstrated. This section explains a process to compute higher level representations from lower level ones. There are two main visual content analysis and interpretation processes:

- *Signal processing* transforms a raw image data into primitive objects (person, tree, ball, horizontal bar etc.) using low-level image processing. Low-level image processing comprises several steps and is often called “analysis” e.g. image analysis.
- *Higher level descriptions* are identified based on the primitive objects and specific prior knowledge relevant (the facts that are not explicit in the data e.g. knowledge about a sports event) for the interpretation. This information is combined together in order to aid image classification e.g. an athlete, a pole, and a horizontal bar characterises a “pole vault” event.

Therefore, this section starts by describing a technique for automatic object detection in an image. Then an image classification technique for higher level conceptualisation is described.

Since visual features alone may not be sufficient to allow computer system to analyse and interpret the meaning of visual data (Moller & Neumann 2008), other useful cues to guide these are needed e.g. any accompanying textual information or an external knowledge base. However, there are some situations when there is no textual information supplied. Therefore, the main objective of this section is to deal with this situation. Therefore, this section presents a framework to represent a higher level conceptualisation of visual data derived from lower level features.

This thesis exploits the BVW model to aid object recognition and image classification. The main advantage of the BVW model is its invariance to camera angle, image scale and orientation, as well as, occlusion, and lighting (Csurka et al. 2004). However, major limitations of existing BVW models (section 3.3.2, p.34) include: they can include many non-informative visual words; they do not preserve the semantics during visual word construction; and they are variant to visual appearance. Hence, this thesis proposes a method to generate a new representation model which resolves the above difficulties and enhances image retrieval efficiency. This technique represents the processes of the Visual Analysis module in Figure 4-1 (p.44). There are five main steps to perform the visual analysis which are described as follows:

1) *Feature detection*: to extract several local patches which are considered as candidates for the basic elements, the main visual words. Interest point detectors detect the "keypoints", the salient patches, in an image. In this thesis, the Difference of Gaussian (DoG) detector (Lowe 2004) is used for the automatic detection of keypoints from images. The DoG detector provides a close approximation to the scale-normalized Laplacian of Gaussian that produces the most stable image features compared to a range of other possible image functions, such as the gradient, Hessian, and Harris corner detector.

2) *Feature representation*: each image and object are abstracted from several local patches. Feature representation methods deal with how to represent the patches as numerical vectors. These methods are called feature descriptors. A SIFT descriptor is deployed in this framework. SIFT converts each patch into a multi-dimensional (128) vector. After this step, each visual content is a collection of vectors of the same dimension (128 of SIFT) where the order of different vectors is of no importance.
3) *Visual words construction*: converts vectors representing patches to "visual words" which produces a BVW model represented in the form of vector (histogram). A visual word can be considered as representation of several similar patches. In this thesis, the SLAC algorithm (AlSumait & Domeniconi 2008) is exploited to cluster the vectors. Each cluster is considered as a visual word that represents a specific local pattern shared by the keypoints in that cluster. The number of the clusters is the bag of visual words' size. This representation is analogous to the bag-of-words document representation in terms of form and semantics because a BVW representation can be converted into a visual-word vector similar to the term vector of a text document.
4) *Non-informative visual words identification*: some of the generated visual words may not be useful to represent visual content. Hence, this kind of visual word needs to be detected and removed in order to reduce the size of visual word feature space and to reduce the computation cost. This can be done using a Chi-square model. After non-informative visual words removal, the remaining visual words are called "*informative visual words*".
5) *Visual word mapping to an ontology model*: the informative visual words are then mapped to a hierarchical model which describes the visual content more explicitly and efficiently than the feature space model using conceptual structures and relationships. This structured model is able to disambiguate visual word senses effectively. Hence, it can more accurately classify images.

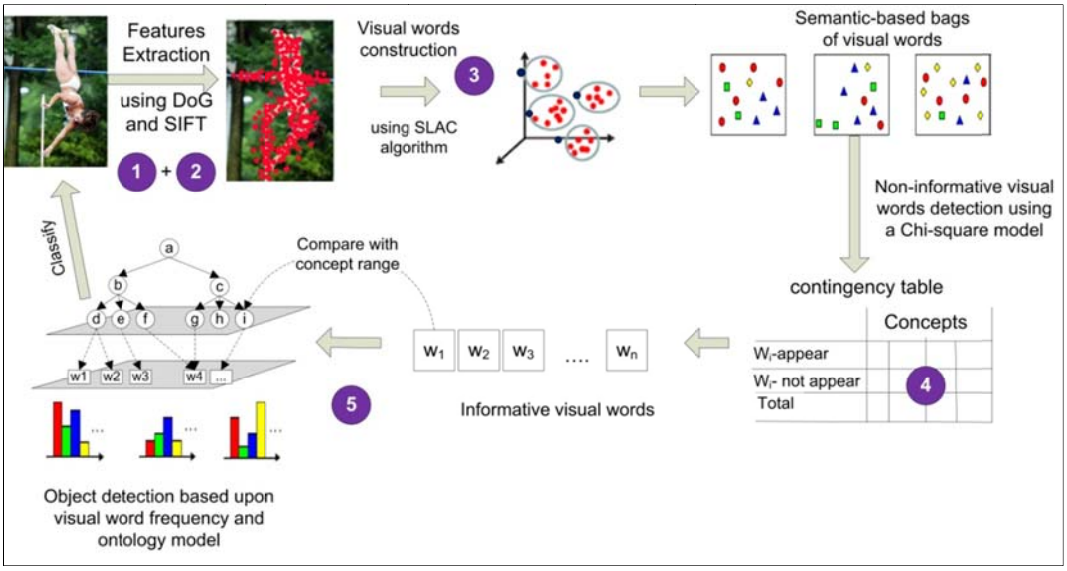

Features Extraction
using DoG and SIFT
1
+
2
Visual words construction
3
using SLAC algorithm
Semantic-based bags of visual words
Non-informative visual words detection using a Chi-square model
contingency table
Concepts
W_i-appear
W_i- not appear
Total
4
Classify
a
b
c
d
e
f
g
h
i
Compare with concept range
w1
w2
w3
w4
...
W1
W2
W3
....
Wn
5
Informative visual words
Object detection based upon visual word frequency and ontology model

Figure 5-3 Architecture of the visual content analysis and classification framework

In this thesis, the first two steps will not be discussed further here since this work makes no major contribution to those areas. Instead, the focus of this thesis is more towards preserving semantic information during the BVW construction process starting with the training phase. This is followed by semantic visual word construction, non-informative visual word elimination, and a semantic visual word mapping to an ontology model. Figure 5-3 illustrates the main steps of the visual analysis module.

5.3.1.1 Training Phase: Semi-Supervised Learning

Since visual content may contain noise from its background, the visual objects of interest are manually separated from the background. The objects in the visual content are the extracted keypoints with respect to the local appearance of those objects. These keypoints are considered relevant because they are from the same object. This method can eliminate noise from the background. Then, only the keypoints of objects will be further processed to generate visual words (Wu et al. 2009). The linkage between the visual words and high level semantics for an object category can be obtained, which serves to connect low level features to high level semantic objects. It is noted that when performed manually, such object separation is not an efficient method for a large-scale multimedia system. However, this method is applied for training only in order to allow the system to learn the proper sets of visual words. In each object category, all the related objects are clustered using the SLAC clustering algorithm in order to generate visual words. The main benefit of the SLAC clustering algorithm over the *k*-mean algorithm is that it does not need to specify the cluster numbers (*k* value) and it takes the term weighting and spatial information, i.e., the distance between keypoints, into account. As a consequence, a set of visual words (ϖ) and $\{\varpi_i \in C_i\}$ are obtained for each object category C_i. Different visual words represent different views of different parts of an object.

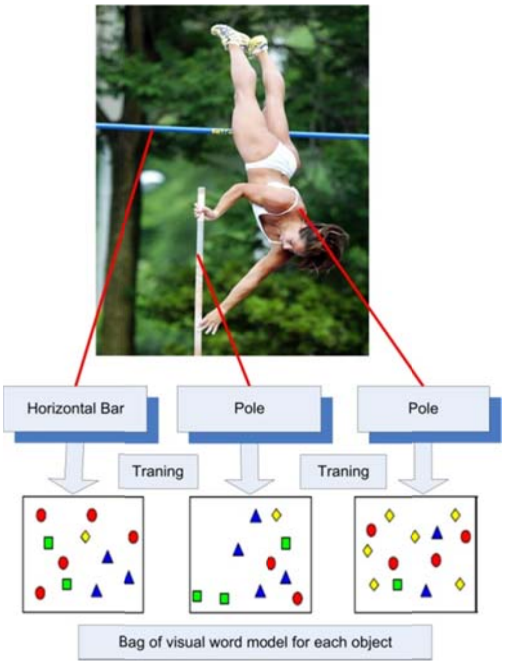


Figure 5-4 Objects in visual content are manually separated in order to train the system to generate visual words for each object and to obtain a distance metric for each object

Having obtained bag of visual words, the concept range of each object will be calculated.

DEFINITION 1. The range of concept of the key object

The range (r_i) of a concept i is the maximum distance of an visual word's centroid (v) to the concept's centroid (c_i) and can be calculated using the following formula:

$$r_i = \max|v - c_i| \ , \quad v \in \varpi \tag{2}$$

The concept range is useful for the visual word sense disambiguation and image classification. The main difference between the training phase and the testing phase is that in the testing phase objects are not separated from the background. It is not practical to manually separate objects from the background for all the test data. Hence some noise arises during the keypoints detection process. There are often many generated visual words that are considered uninformative in representing visual content. Hence, these non-informative visual words need to be detected and eliminated. Figure 5-5 shows the example of keypoints detected from images. Green circles in images represent various scales of the keypoints. These are analysed in order to make them invariant to scale, illumination, orientation, and camera angle

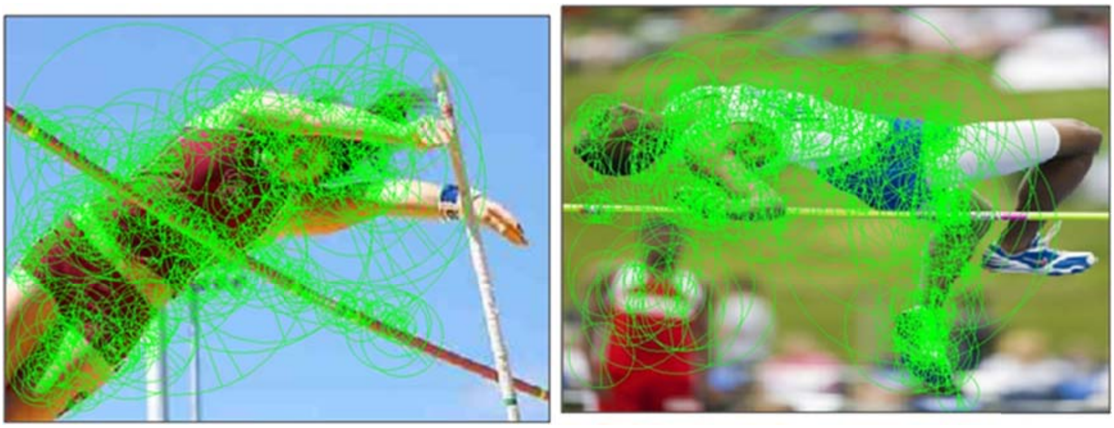


Figure 5-5 Examples of images with the detected keypoints

5.3.1.2 Visual Word Generation using the SLAC Clustering Algorithm

A simple method e.g., the *k*-mean clustering algorithm performs clustering over all the vectors. Each cluster is considered as a visual word that represents a specific local pattern shared by the keypoints in that cluster. However, the major drawback of the *k*-mean algorithm is that it appears to be unaware of the spatial location of keypoints. This loses semantic information between the low level features and the high level semantics of objects in the visual content. To overcome this issue, a method is proposed that can find the semantically similar keypoints and cluster them into the same group using a similarity matrix based on the SLAC algorithm.

SLAC supports subspace clustering, an extension of traditional clustering, that captures any local feature relevance within a cluster. To find semantically similar keypoints (φ) and to cluster these, learning kernel methods, local term weightings and semantic distance are deployed. A kernel represents the similarity between documents and terms. From data mining, φ should be mapped to nearby positions in the feature space. To represent the whole corpus of N documents, the document-term matrix, $\mathcal{D}$, is constructed. $\mathcal{D}$ as a $N \times D$ matrix whose rows are indexed by documents (images) and whose columns are indexed by keypoints. The numerical values in $\mathcal{D}$ are a frequency of term i in document d. The key idea of the technique in this section is to use the semantic distance between pairs of keypoints, through defining a local kernel for each cluster as follows:

$$K_j(d_1, d_2) = \phi(d_1) Sem_j Sem_j^T \phi(d_2)^T, \tag{3}$$

$$Sem_j = R_j P \tag{4}$$

where d is a document in the collection, and $\phi(d)$ is document vector, Sem_j is a semantic matrix which provides additional refinements to the semantics of the representation. P is the proximity matrix (Figure 5-6 (a)) defining the semantic similarities between the different terms and R_j is a local term-weighting diagonal matrix (Figure 5-6 (b)) corresponding to cluster j, where w_{ij} represents the weight of a keypoint i for cluster j, for $i = 1,..,D$. One simple way to compute weights to w_{ij} is to use the inverse document frequency (*idf*) scheme. However, the *idf* weighting scheme concerns only the document frequency without taking the distance between keypoints into account. In other words, the *idf* weighting scheme does not involve the inter-semantic relationships among terms. Therefore, a new weighting measure based on the local adaptive clustering (LAC) algorithm is utilised to construct matrix R.

$P=$		k_1	k_2	k_3	...	k_n
	k_1	-	x_{12}	x_{13}	...	x_{1n}
	k_2	x_{21}	-	x_{23}	...	x_{2n}
	k_3	x_{31}	x_{32}	-	...	x_{3n}
	...	...	...	...	...	...
	k_n	x_{n1}	x_{n2}	x_{n3}	...	x_{nn}

(a)

$$R_j = \begin{pmatrix} w_{j1} & 0 & \ldots & 0 \\ 0 & w_{j2} & \ldots & 0 \\ : & : & : & : \\ 0 & 0 & \ldots & w_{jD} \end{pmatrix}$$

(b)

Figure 5-6 (a) Example of a proximity matrix; (b) Local term-weighting diagonal matrix (R_j)

LAC gives less weight to data which are loosely correlated and this has the effect of elongating distances along that dimension (Domeniconi et al. 2007). In contrast, any features along which data are strongly correlated receive a larger weight which has the effect of constricting distances along that dimension. Equation (5) shows the LAC term weight calculation.

$$w_{ij} = \frac{\exp\left(-\frac{1}{|S_j|} \sum_{x \in S_j} (c_{ji} - x_i)^2 / h \right)}{\sum_{i=1}^{D} \exp(-\frac{1}{|S_j|} \sum_{x \in S_j} (c_{ji} - x_i)^2 / h)}; \qquad (5)$$

where a set S_j of N points x in the D-dimensional Euclidean space, c_{ij} is a center of i component of vector j, and the coefficient $h \geq 0$ is a parameter of the procedure which controls the relative differences between feature weights. In other words, h controls how much the distribution of weight values will deviate from the uniform distribution.

P has nonzero off-diagonal entries, $P_{ij} > 0$, when the term i is semantically related to the term j. To compute P, the Generalized Vector Space Model (GVSM) (Wong et al. 1985) is deployed to capture the correlations of terms by investigating their co-occurrences across the corpus based on the assumption that two terms are semantically related if they frequently co-occur in the same documents. Since P holds a similarity measure between terms in the form of co-occurrence information, it is necessary to transform this to a distance measure before utilising it. Equation (6) shows the transformation formula:

$$P_{ij}^{dist} = 1 - (P_{ij} / \max(P)) \tag{6}$$

where P_{ij}^{dist} is a distance information, *max(P)* is the maximum entry value in the proximity matrix. Consequently, a *semantic dissimilarity matrix* for cluster *j* is a $D \times D$ matrix given by Equation (7).

$$Sem_j^{dissim} = R_j P^{dist} \tag{7}$$

This represents semantic dissimilarities between terms with respect to the local term weightings. The SLAC algorithm starts with *k* initial centroids and equal weights. It partitions the data points, re-computes the weights and data partitions accordingly, and then re-computes the new centroids. The algorithm iterates until convergence or a maximum number of iterations are exceeded. The SLAC uses a semantic distance. A point *x* is assigned to the cluster *j* that minimises the semantic distance of the point from its centroid. The semantic distance is derived from the kernel in Equation (3) as follows:

$$L_w(c_l, x) = (x - c_l) Sem_l^{dissim} Sem_l^{dissim^T} (x - c_l)^T \tag{8}$$

Hence, every time the algorithm computes S_j, its semantic matrix must be computed by means of these new weights. SLAC clusters keypoints according to the degree of relevance and thus generates visual words that are semantically related. Consequently, the visual words obtained are improved in contrast to those obtained using traditional models. As mentioned previously, noisy keypoints from the background of an image affect the quality of the generated visual words. Too many generated visual words are not useful to represent visual content, they degrade the classification power. Thus, these non-informative visual words should be eliminated.

Algorithm 1: The visual words construction algorithm using SLAC

```
Input:  Visual content
Output: Visual words for each visual content
1. Keypoints detection using the DoG algorithm, i=0;
   While (i <=n) // n is a number of images in the collection
        keypoints = DoG(Image[i]); i++;
2. Initialise k centroids $c_1, c_2, ..., c_k$;
3. Initialise weights: $w_{ij} = \frac{1}{D}$, for each centroid $c_j$, $j = 1,..., k$ and for each term $i=1,..., D$;
4. Compute $P$; then compute $P^{dist}$;
5. Compute $Sem^{dissim}$ for each cluster $j$ (Equation (7));
6. For each centroid $c_j$, and for each point x, set:
   $S_j = \{x \mid j = \arg\min_l L_w(c_l, x)\}$,
   where $L_w(c_l, x) = (x - c_l) Sem_l^{dissim} Sem_l^{dissim^T} (x - c_l)^T$
7. Compute new weights:
   for each centroid $c_j$, and for each term $i$:
        Compute Equation (5).
8. For each centroid $c_j$:
   Recompute $Sem_l^{dissim}$ matrix using new weights $w_{ij}$;
9. For each point x:
   Recompute $S_j = \{x \mid j = \arg\min_l L_w(c_l, x)\}$,
10. Compute new centroids:
   $c_j = \frac{\sum x 1_{S_j}(x)}{\sum 1_{S_j}(x)}$ for each $j = 1, ..., k$
```

5.3.1.3 Non-informative Visual Word Elimination

Non-informative visual words are usually the local visual content patterns that are considered not to be useful for retrieval and classification tasks. They are relatively "safe" to remove (Yang & Wilbur 1996) in the sense that their removal does not cause a significant loss of accuracy but rather significantly improves the classification accuracy and computation efficiency of categorisation. Using an analogy with processing text-based documents, for image processing, there exist unimportant visual words, so-called *non-informative visual words* (ψ).These visual words need to be eliminated in order to improve the accuracy of the classification results and to reduce the size of visual word feature space and computation cost. In this thesis, a statistical model is utilised to automatically discover ψ, to eliminate them strengthening the discrimination power. Yang *et al.* (2007b) evaluate several techniques usually used in feature selection for machine learning and text retrieval, e.g. Document frequency, Chi-square statistics, and Mutual information. In contrast to the method of (Yang et al. 2007b), non-informative visual words in this thesis are identified based upon a document frequency and upon a statistical correlation of visual words with all concepts in the collection. In addition, the visual words are normalised in order to compensate for discrepancies in the size of the images.

DEFINITION 2 Non-informative Visual Words (ψ)

A visual word $v \in V,\ V = \{v_1, v_2, ..., v_n\},\ \ n \geq 1$ is uninformative if it:

1. usually appears in many visual content in the collection, thus, it has a high document frequency (*DF*). Since it occurs in several images, it cannot be used to represent any particular image or object
2. and has a small statistical correlation with all classification categories.

From these definitions, ψ *ca*n be extracted from the visual word feature space using a Chi-square statistical model. Having created ϖ in the previous step, ϖ will be quantized into a Boolean vector space model to express each visual content vector. Assume that the appearance of the visual word *i* (ϖ_i) is independent of any concepts $C,\ C \in Z,\ \ Z = \{C_1, C_2, ..., C_n\}$ where $n \geq 1$. Thus, the correlation between ϖ_i and concepts could be expressed in the form of a 2*p contingency table as shown in TABLE 5-2.

DEFINITION 3 The Boolean Vector Space Model

The Boolean vector space model $\mathrm{B} = \{V_i\}_{i=1}^{N}$ contains a collection of *N* visual words. A binary matrix X_{NxM} represents B, where $x_{ij} = 1$ denotes the visual content *i* contains the visual word *j* in the vector space otherwise x_{ij} = 0, where $1 \leq i \leq N$ and $1 \leq j \leq M$.

DEFINITION 4 The 2*p Contingency Table

The 2*p contingency table $T = \{n_{ij}\}_{j=1}^{k}, 1 \leq i \leq 2$ contains the number of instances of visual content containing visual words in each category. A matrix $A_{N \times M}$ represents T, where n_{1j} is the number of visual content instances containing a visual word ϖ_i for the concept C_j; n_{2j} is the number of visual content instances which do not contain visual word ϖ_i in the concept C_j; n_{+j} is the total number of visual content instances in the concept C_j; n_{i+} is the number of visual content in the collection containing the visual wordϖ_i; N is the total number of visual content instances in the training set.

TABLE 5-2: The 2*p contingency table of ϖ_i

	C_1	C_2	…	C_k	Total
ϖ_i -appear	n_{11}	n_{12}	…	n_{1k}	n_{1+}
ϖ_i -not appear	n_{21}	n_{22}	…	n_{2k}	n_{2+}
Total	n_{+1}	n_{+2}	…	n_{+k}	N

where

$$n_{+j} = \sum_{i=1}^{2} n_{ij} \text{ , } n_{i+} = \sum_{j=1}^{k} n_{ij} \qquad (9)$$

$$N = \sum_{i=1}^{2}\sum_{j=1}^{k} n_{ij} = \sum_{i=1}^{2} n_{i+} = \sum_{j=1}^{k} n_{+j} \qquad (10)$$

$$x_{2*p}^{2} = \delta = \sum_{i=1}^{2}\sum_{j=1}^{k} \frac{(Nn_{ij} - n_{i+}n_{+j})^2}{Nn_{i+}n_{+j}} \qquad (11)$$

To measure the independence of each ϖ from all the concepts, the Chi-square statistic (δ) is deployed given in Equation (11). Having calculated the degree of independence, the δ values are sorted by descending order. The δ value indicates the degree of correlation between ϖ and its concepts; the smaller the δ value, the weaker the correlation. These visual words satisfy the second condition of the definition of ψ. However, there exists a problem concerning terms that appear in a small number of documents leading to them to have a small δ value. These terms sometimes could be the feature words. In such a case, the δ value is weighted using Equation (12) (Hao &

Hao 2008).

$$x_{weighted}^{2} = \frac{x_{2*p}^{2}}{DF_{r}} \tag{12}$$

where DF_r denotes the document frequency of the visual word *r*. This model balances the strength of the dependent relationship between a visual word, its concept, and its document frequency. As a result, those ϖ that have δ values less than a threshold (chosen experimentally) are designated as non-informative visual words and are removed because they have a high *DF* and small correlations with all the categories. Obviously, ψ identified in this manner are collection specific which means by changing the training collection one can obtain a different ordered list. The remaining ϖ are informative and are useful for the categorisation task.

5.3.1.4 Visual Words Disambiguation

In existing systems, researchers disambiguate different multiple word senses by combining multiple visual words into a larger unit, a so-called “visual phrase” (Yuan et al. 2007b; Zheng et al. 2009) or “visual sentence” (Tirilly et al. 2008). However, this method has some limitations since the visual phrase is usually constructed using the Frequency Itemsets Mining (FIM) algorithm which is purely based on frequent word collocation patterns without taking into account the term weighting and spatial information. The latter information is crucial in order to discover the semantic similarity between words. Other researchers tried to restructure visual words as a hierarchical model (Jiang & Ngo 2009; Sivic et al. 2008; Wang et al. 2009) in order to disambiguate word senses more explicitly and effectively. These methods convert an unstructured visual words model into a hierarchical structure model using a well-known clustering algorithm e.g. Agglomerative clustering algorithm, Hierarchical Spatial Markov model and Hierarchical Latent Dirichlet Allocation algorithm. Nevertheless, hierarchical models generated from these algorithms have some limitations. First, they are binary hierarchical models that are not always efficient in representing visual content data. In practice, types of relationships among concepts are more diverse. Second, there is no multiple-relationship between parents and a child node. The multiple-relationship means a child node can have more than one parent. For example, a Heptathlon event has a relationship with the Field and Track event since it combines these two events together as one sport for women. As such, the generated hierarchical model used by the existing frameworks cannot represent the semantic information of visual content properly.

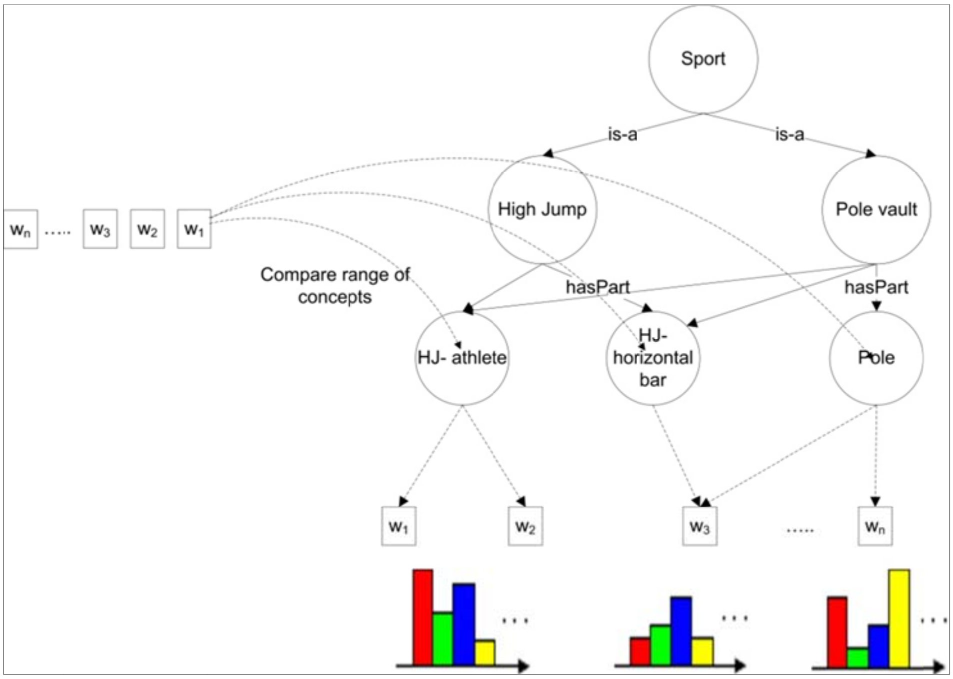


Figure 5-7 Example of a structural ontology model and the different kinds of relationships between concepts for three sport genres. To disambiguate word senses, each visual word is compared to the concept range and assigned to concept(s) in the ontology model

Rather than combining multiple visual words to disambiguate word senses or using a binary tree model, this thesis proposes to transform visual word vector space model to a structural ontology model in order to resolve their limitations. In addition, the proposed method can enhance the image annotation, classification and retrieval performance of the system. Since the ontology model is usually domain specific e.g. natural scene or sports, the structure of concepts and the relationships among concepts for each application differ for each knowledge domain. Furthermore, it is impossible to exploit standard clustering algorithms or expect human beings to generate a general ontology model for every application. Therefore, a pre-designed ontology model, so-called semantic template, for sports domain is needed to enable the system to retrieve information semantically and precisely. Typically in text documents, word sense disambiguation can be done using external knowledge e.g. WordNet. However, WordNet cannot be used in this way for visual words as they do not provide any linguistic information. Therefore, an alternative method is to use mathematic calculations. From the training phase, the semantic concept of each individual object is presented properly and, then, each concept will be used to disambiguate the informative visual words and to assign the concept(s) for each visual word under a pre-designed ontology model which is improved from Wu (2009). The different senses of a visual word can be disambiguated using a concept range from Equation (2). If a visual word is inside the range of any concept, the concept is assigned to the visual word; otherwise the visual word does not respond to any concept and is discarded.

$$c(v) = c_i, \text{ if } \lambda(|v - c_i|,\ r_i) = 1 \quad \text{otherwise}\, discard, \tag{13}$$

$$\lambda(x,b) = \begin{cases} 1, & x < b, \\ 0, & \text{otherwise.} \end{cases} \tag{14}$$

where r_i is the range of object i and ν is a visual word and c_i is a centroid of concept i. This method allows the visual word to be assigned to multiple concepts since the range of concepts may overlap each other. Hence, this method is more practical than the existing systems (Jiang & Ngo 2009; Wang et al. 2009) which lack a multiple-parent relationship. This can handle the *polysemy* problem of a visual word. For example, a visual word can belong to a horizontal bar concept and a pole concept since both objects are similar. Therefore, the use of the concept range technique allows multiple assignments of a visual word to be possible. Since the range of concepts in an ontology model are generated from different views of objects that comprise a sports scene in the training phase using Equation (1), the visual diversity of objects causes the semantics of visual content to be better represented using different visual words. Consequently, the range of concepts is *invariant to the visual appearance* of an object.

Furthermore, this model can be used to detect key objects and classify visual content at a higher-level conceptualisation. The detection of key objects in visual content is related to the frequency of visual words which represent it. If the frequency of related visual words, $f(v_i)$ of a particular object (i.e. athlete) is higher than a threshold (chosen experimentally), this means the visual content contains that object. However, direct use of $f(v_i)$ may be unfair to each instance of visual content in the collection due to scaling differences. Hence, $f(v_i)$ is normalised in order to compensate for discrepancies in the frequency of the visual words. Equation (15) shows the normalisation formula where *N* is a number of instances of visual content.

$$\eta_i = f(v_i) / \sum_{j=1}^{N} f(v_j) \tag{15}$$

To classify the high-level semantics of visual content, the simplest way is by using the detected object information. If the key objects are detected, the detected object information will be used for a scene interpretation based upon *reasoning rules* and *a Bayesian network model*. A reasoning rule is a classification rule that determines the sport genre based on the detected objects. For instance, if an athlete, pole, and horizontal bar are detected in an image, this indicates that the visual content is relevant to a pole vault event. Therefore, the system can classify the visual content using only low-level features and image processing techniques. Here is an example of a reasoning rule for a pole vault event interpretation.

Annotate image as a "pole vault" event if all of the following conditions hold:

$\forall x, \forall y, \forall z \mid Athlete\,(x) \wedge Bar\,(y) \wedge Pole\,(z) \Rightarrow$ pole vault

Meaning: annotate an image represents a "pole vault" if the image contains the objects such as athlete, horizontal bar, and pole.

All detected objects and annotations are counted in order to be possibly used for uncertainty management (Chapter 6) where the system uses probabilities derived from previous data to classify an image. However, objects detected using a number of

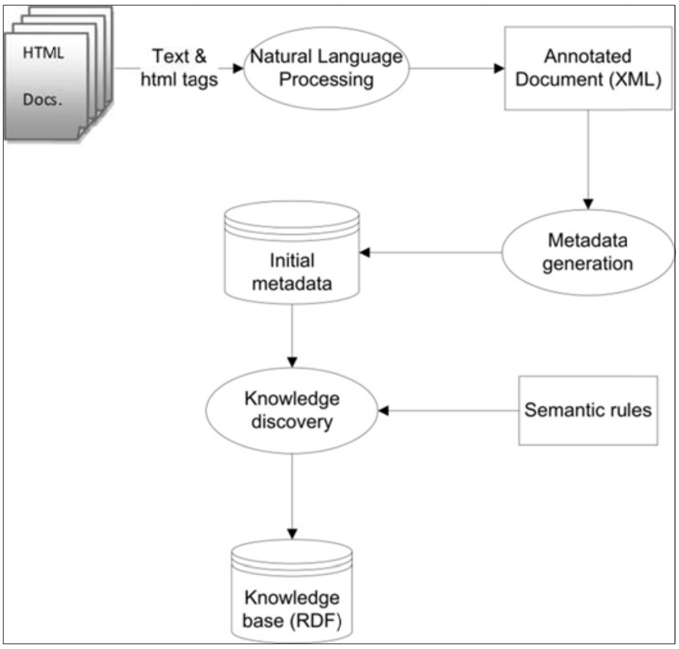


Figure 5-8 Metadata generation process

visual words of each concept has some disadvantages. Since a visual word can belong to multiple similar concepts, it is possible that the system will detect two similar objects in an image that semantically are not supposed to occur together e.g. a pole and a javelin. When the range of concepts for two similar objects overlap, the visual words extracted could belong to both concepts. Then, the system interprets that both objects are found in an image. This uncertainty can be resolved using a semantic reasoning rule. For example, if the system detects an athlete, a bar, a pole and a javelin object in the same image, the reasoning rule will discard a javelin object because there is not any sport genre that uses a bar and a javelin equipment together. In other words, there is very little chance that a bar and a javelin will appear together in an image based upon their semantic relations. In contrast, there is a high probability that a bar and a pole can appear together in the same image e.g. the pole vault event. Based upon this fact, this uncertainty can be handled through this reasoning rule.

Another uncertainty is that the underlying objects are not always detected due to background noise in the image. This uncertainty means the system cannot classify an image properly. To overcome this problem, an uncertainty management technique is proposed (Chapter 6, p.91). Text captions in addition to visual features also are useful for image annotation and classification. They are a vital source for metadata to aid image classification and retrieval. Therefore, techniques for linguistic analysis are provided.

5.3.2 Semantic Linguistic Analysis

Figure 5-8 illustrates the main processes of the semantic linguistic analysis processes executed by the Linguistic Analysis module (Figure 4-1, p.44). There exist text descriptions accompanying some images that can be useful for image classification. Therefore, the main function of this section is to process and analyse text captions to annotate images. First, textual information will be parsed from HTML documents in order to find the implicit meaning hidden in the passage. Second, the relationships among keywords are stored in the KB for later retrieval. The purpose of this process is to identify the information for ontology instances. The output of this step is an RDF file that stores the semantic metadata. HTML files are processed to extract the important textual information (e.g. date, time, place, person name, and event) and to create the semantic metadata and store this metadata in an RDF file.

5.3.2.1 Natural Language Processing

First, a Natural Language Processing tool (NLP) is used for the initial metadata generation because it is easier to work with the information generated by NLP than with the raw document text. However, NLP innovation is out of the research scope here. Therefore, an established NLP framework, ESpotter, is deployed rather than implementing a new text engineering tool. ESpotter (Zhu et al. 2005) provides a function for a Named Entity Recognition (NER) task e.g. person name <mentions-person>, location<mentions-location>, date<mentions-date>, and other proper nouns <mentions-pn> and generates an initial version of the semantic metadata in an XML file format. This annotated document is then processed and the initial metadata extracted. In addition, Espotter provides the position of each detected NER (i.e. <instance content="United States" **pos="27"** />). It also has a further use e.g. word disambiguation as explained in the next section.

5.3.2.2 Metadata Generation and Ambiguous Interpretation

Metadata generation processes the XML files output from the previous step. The metadata output from this step is called the "*initial metadata*" and is stored in a relational database (RDBMS). However, an initial semantic metadata could match many ontology entities. In other words, in some cases, this metadata could be ambiguous.

Consider the following example: "*Brian from USA performs 100m freestyle men in Bejing, China*"

In this example, it is easy for humans to identify that USA refers to the nationality of the athlete and China refers to the host country for the sports event. For a computer system, however, USA and China are ambiguous. A country could refer to the nationality of an athlete or the host country of the sport event, therefore, a system needs to disambiguate and find the most appropriate sense for these terms. To disambiguate such a case, the information in an XML file generated from Espotter is exploited. In a natural language statement, the nationality is usually mentioned after person name and after a preposition such as *of, from* etc. So, this fact is used to create the rules to classify the word sense for the nationality of athlete versus that of the host country for the event. Furthermore, this data needs to be transformed by *generalising* it to higher-level concepts. WordNet can be used for this purpose. For example, *city* can be generalised to higher-level concepts, like country. Algorithm 2 shows the detection rule for nationality and host country.

Algorithm 2: The algorithm for nationality and host country detection

Input: position of the detected name entity.
Output: nationality or host country

```
1. For each detected location name entity $\{l_i\}$ and $l \in \{L\}$ and $i \geq 0$
{
   Generalisation $\{l_i\}$ to higher level using WordNet
   1.1 IF position of location $\{l_i\}$ follows article *of* or *from* etc. THEN
          Location = nationality
   1.2 ELSE
          Location =  host country
}; end for
```

For synonymy and polysemy problems, an external lexical reference system, WordNet, is deployed to solve this problem. WordNet is a semantic network database for English developed at Princeton University. The basic building-block in WordNet is a *synset*. A synset is a set of synonyms denoting the same concept, paired with a description of the synset. The synsets are interconnected with different relational links such as hypernymy (is-a-kind-of), meronymy (is-a-part-of), antonymy (is-an-opposite-of), and others. With the aid of synsets in WordNet, it is possible to relate two words together e.g. "*high jump*" and "*field event*" i.e. a high jump is a kind of a field event, representing a *hypernym relationship*. Therefore, a system can consider this image to be relevant to field event even though the "field event" word does not appear in the text caption. In other words, the system considers the "field event" term as a synonym for the term "high jump".

```xml
<?xml version="1.0" encoding="utf-8" standalone="yes"?>
<ESpotter-Processed-Documents corpusSize="56">
<Document id="0">
<has-directory>E:\NLPProject\HTMLs\img1.html</has-directory>
<has-url>E:\NLPProject\HTMLs\img1.html</has-url>
<has-document-size>56</has-document-size>
<mentions-location>
        <instance content="Atlanta" pos="10" />
        <instance content="United States" pos="27" />
</mentions-location>
<mentions-date>
        <instance content="3 August 1996" pos="11" />
</mentions-date>
<pn>
        <instance content="XXVI Olympiad" pos="17" />
        <instance content="Tennis" pos="19" />
        <instance content="Celebrates" pos="29" />
        <instance content="Victory" pos="31" />
</pn>
<mentions-person>
        <instance content="Andre AGASSI" pos="23" alias="andre" />
</mentions-person>
</Document>
</ESpotter-Processed-Documents>
```

Figure 5-9 Example of a XML file generated by ESpotter

5.3.2.3 Knowledge Discovery

Knowledge discovery aims at extracting implicit information from a corpus of documents (Fayyad 1996). In this thesis, knowledge discovery refers to a process to find the missing information of an image using the previously extracted metadata. The missing metadata can be categorised into two groups, *ambiguous* and *unambiguous*. Unambiguous missing metadata is metadata that has a 1:1 relationship with other metadata e.g. Olympic games event and date. This is because this metadata can be uniquely identified by year (only one Olympic Games event is held in any one year). *Semantic rules* are applied to handle this kind of missing metadata. For instance, if the date in the image is detected as “20 September 2000”, this picture could have a relationship with Sydney Games which has occurred in year 2000, at Sydney (host city), and in Australia (host country). An example of a semantic rule is shown in Figure 5-10. This rule is used for adding relevant metadata to an image based on a recorded date of an image. This means when a new image has been entered to the system without any text caption, the system is able to recognise the Olympic Games event and the venue in an image automatically based upon the recorded date of an image. This clearly enhances the retrieval performance. The ambiguous missing metadata is metadata that has a 1:M relationship with other metadata. In other words, the ambiguous metadata can be matched with several ontology entities. Therefore, a data mining technique is deployed to handle the ambiguous missing metadata.

Add *x* to *M* (metadata) if all of the following conditions hold

- $Event\,(y), Photo\ - Date\,(x)$

$\forall x, \exists y, \forall z \mid Photo\,(z) \wedge Photo\ - Date\,(x) \wedge\ has - PhotoDate\ (z, x) \Rightarrow Event\,(y)$

Add Event(y) to metadata of a given image if an image contains the *Photo-Date* entity that happens during the given event, and then this event is considered

Figure 5-10 Example of semantic rule for sports event detection

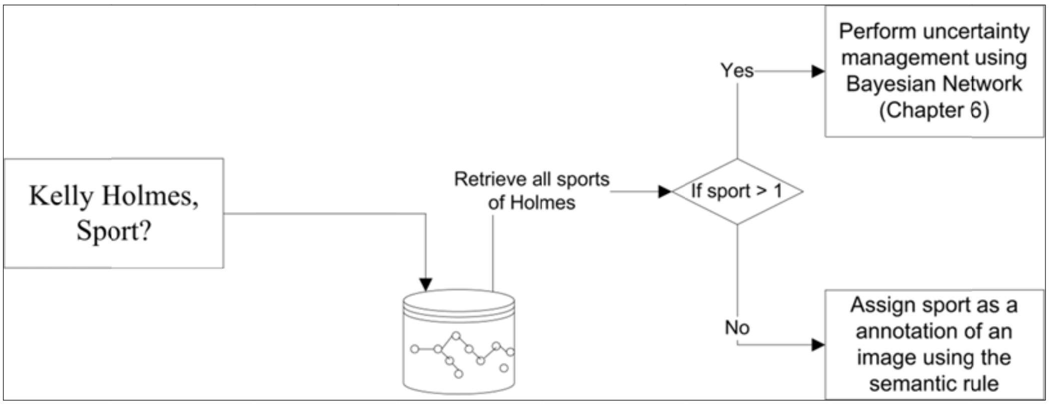

Kelly Holmes, Sport?
Retrieve all sports of Holmes
If sport > 1
Yes
Perform uncertainty management using Bayesian Network (Chapter 6)
No
Assign sport as a annotation of an image using the semantic rule

Figure 5-11 Example of the knowledge discovery process in order to find the missing metadata

For instance, the system tries to find the *Sport* entity (sport name that is relevant to an image) based upon the detected athlete name. If an image contains an athlete name “Kelly Holmes” but no sport genre addressed in the text caption, the system will try to find the related sport genre of this athlete using previous information stored as metadata in the KB. In this example, Holmes usually participated in running events at two different distances, 800m and 1500m. Therefore, the system has to determine which distance that image should be relevant. This can be determined based upon the probability of the previous data. More detail to handle this uncertainty will be described in Chapter 6 (section 6.2, p.99). A knowledge discovery algorithm is shown in Figure 5-11. The advantage of this method is that it is scalable because it works based upon the extracted metadata in the KB. Therefore, the more extracted metadata is stored in the KB, the easier it is to find the missing metadata. Nonetheless, the drawback of this technique is that the uncertainty may occur. To cope with the uncertainty, the Bayesian network will be applied.

5.4 Knowledge-based Representation

Knowledge Representation (KR) is a set of ontological commitments, in the sense that one concept generally refers to and is understood through its relationships with other concepts and through its use (Poslad 2009). In addition, a KR enables efficient machine-readable and machine-understandable computation about knowledge. Semantic metadata generated by the metadata generation step is stored in a relational database. To be able to use this data in a semantic context, it is mapped to an ontology whose data is given a well-defined meaning by representing it using RDF. RDF is an XML extension which enables the system to read and parse the syntax to extract concepts (machine-readable) and to act on the meaning of the concepts (machine-understandable). RDF is selected as a KR rather than RDFS or OWL because of the following reasons:

- The system does not need to use the category relationship which is an additional function of RDFS.
- Cardinality, transitive, and inverse constraints which are available in OWL have not been exploited by the presented system.
- At the time of implementation, the query language for OWL is not mature enough and is quite complex to deploy e.g. SparqlOwl[13] and KAON2[14] .

Therefore, RDF was selected as a knowledge representation language for the metadata extracted from the image description. The data model behind RDF is a directed graph, which consists of nodes and directed arcs linking pairs of nodes. This KR technique allows data from different sources to be shared, exchanged and integrated and enables applications to use data in different contexts. In the next section, the methods that export metadata from the relational database to the semantic model (RDF) are described.

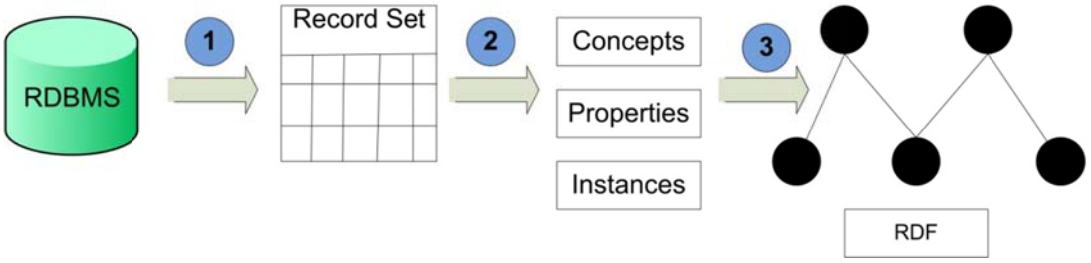


Figure 5-12 Transformation process from relational database to RDF

[13] SPARQL/OWL task force (http://code.google.com/p/owl1-1/wiki/SparqlOwl)

[14] KAON (http://kaon2.semanticweb.org)

5.4.1 The Initial Metadata and Ontology Transformation

To export data from a relational database into RDF, the relational database model has to be mapped to the graph-based RDF data model. There are two approaches for exporting this data from relational databases: in a RDF-direct or an indirect mapping[15].

- *Direct mapping*: this method directly maps a RDBMS schema to RDF. This generic approach can be useful in many cases, but sometimes it may lead to difficulties in synchronising to changes in the database structures, to difficulties in installation, and to use by inexperienced users.
- *Indirect mapping*: uses the application logic to access data. Some content management systems provide APIs and an application logic as a source of information to be exported in RDF.

In this research a simple and fast approach is preferred*: a direct mapping* scheme is used to map data from a RDBMS database to RDF using a JDBC connector API. The mapping process is shown in Figure 5-12.

To transform information in a relational database to RDF, three steps are needed:

1) The initial metadata is retrieved using the SQL select command and the record sets returned from the query are grouped by column.
2) The Jena API is deployed to create ontology concepts, properties and instances.
3) The grouped record set metadata are assigned to the ontology instances created in step 2). First, the record set is selected from a database. Second, the record set is grouped according to the GROUP BY column. Then the class instances are created and assigned an URI or a blank node identifier. Finally, the instances' properties are created and assigned property values and written to a RDF file.

Figure 5-13 shows an example of a generated RDF file. First an RDF file is automatically created by the mapping process. To process the RDF file, a program extracts the knowledge from a RDF file using available tools such as SPARQL or

[15] From Online Community Data to RDF (http://www.w3.org/2007/03/RdfRDB/papers/sioc)

RDQL. These tools also provide APIs for a programming language e.g. Java-based Jena, to access an RDF file.

The generated metadata from ESpotter may be incomplete or incorrect. It is not expected that the quality of the generated metadata reaches the quality of manually created metadata. Some keywords are recognised and they may also be annotated with inappropriate tags because of an error of ESpotter. Therefore, manual correction and annotation of metadata are also supported.

```
<rdf:RDF
   xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:sport="http://protege.stanford.edu/sport#"
  xml:base="http://protege.stanford.edu/sport.rdf">
  <rdf:Description rdf:about="sport#img45">
    <sport:AboutSport rdf:parseType="Resource">
      <sport:hasSportEvent>Track</sport:hasSportEvent>
      <sport:hasSportDetail>100M men</sport:hasSportDetail>
      <sport:hasSportName>Running</sport:hasSportName>
    </sport:AboutSport>
    <sport:hasVisualFeatures rdf:parseType="Resource">
      <sport:hasImageSize>29</sport:hasImageSize>
      <sport:hasImageDimension>290 x 430</sport:hasPhotoDimension>
      <sport:hasImageFormat>JPEG</sport:hasPhotoFormat>
      <sport:hasImagePath>image/img45.jpg</sport:hasPhotoPath>
      < sport:hasSIFTdescriptors>SIFT/img1.csv</sport: hasSIFTdescriptors>
    </sport:hasImageFeatures>
    <sport:hasEvent rdf:parseType="Resource">
      <sport:hasOpenDate>17/09/1988</sport:hasOpenDate>
      <sport:hasClolsingDate>02/10/1988</sport:hasClolsingDate>
      <sport:hasEventYear>1988</sport:hasEventYear>
      <sport:hasFormalName>Games of the XXIV Olympiad</sport: hasFormalName>
      <sport:hasShortName>SEOUL 1988</sport: hasShortName>
    </sport:hasEvent>
……
  </rdf:Description>
```

Figure 5-13 Example of a generated RDF file from the mapping process

5.5 Knowledge-based Evaluation

An ontology as a body of knowledge can be built in many different ways. Therefore, how well the created ontology fits that knowledge domain should be evaluated. Ontology evaluation is an important issue that must be addressed before it will be adopted in the proposed system or application. Various approaches can be used for ontology evaluation. Typically, since an ontology is a fairly complex structure, often the evaluation focuses on different levels of the ontology (Brank et al. 2005). First, the *Lexical or data level* focuses on concepts, relationships and instances in an ontology. It involves comparison of the terms in concepts relationships and instances in an ontology with a corpus of documents (Brewster et al. 2004). Second, the *Taxonomy or hierarchy level* evaluates the degree of structural fit between an ontology and a corpus of documents. This method involves the evaluation of the structural design of an ontology which is usually performed manually by experts. Third, the *Application level* evaluates how the results of the application are affected by the use of the ontology.

As human evaluation is by its very nature subjective, this thesis mainly focuses on an automatic evaluation. Thus, only the Lexical and the Application levels will be deployed to evaluate the presented ontology. For the Application level, if the domain ontology is designed properly and fits the sport domain, the retrieval results should be significantly improved. This will be studied in Chapter 7 (p.107). This section will evaluate the presented ontology using the Lexical or data level, a so-called *Data-driven evaluation*. The main idea of this method is to find similar strings between two sets of data as proposed by Maedche *et al.* (2002). One set of all strings are taken from concepts, relationships and instances of the ontology and used to compare to another set of all strings from a corpus of documents. The similarity is measured through a statistical computation or through the concepts of *precision* and *recall* as known in information retrieval system. In this context, however, precision refers to the percentage of the ontology lexical entries (string used as concepts identifiers and instances) that also appear in a corpus of documents, relative to the total number of strings in ontology. Recall is the percentage of lexical entries in a corpus documents that also appear as concept identifiers, relationships and instances in the ontology, relative to the total number of the lexical entries in a corpus.

5.5.1 Ontology Evaluation using Precision and Recall

An evaluation scheme using precision and recall is designed as follows. It begins with term extraction from text captions of all images in the collections. Here all text captions are processed using the Espotter in order to enhance term extraction (tokenising) such as “Kelly Holmes” should count as one term rather than “Kelly” and “Holmes”. All detected terms are put into the same set, $\mathcal{N}$. The undetected terms from the use of Espotter are processed by removing stop words and put into another set, $\mathcal{R}$. Then, $\mathcal{N}$ and $\mathcal{R}$ are merged together to create a larger set of terms from a corpus documents. Having performed these steps, it obtains totally 22,589 terms excluding stop words from 1,500 text captions in the corpus. Next, all terms in the ontology are counted including terms of concepts, relations, and instances which obtains totally 18,586 terms. To find how many terms in the ontology appear in the corpus, the string matching technique is applied. However, rather than perform simple string matching, WordNet is also exploited for acquiring synonyms and hypernyms of each term from the ontology. This is because terms appear in text caption may differ from terms that appear in the ontology but that are semantically relevant. Having performed string matching, the total number of terms appear in the corpus is for example 18,143 terms. Finally, the precision and recall can be computed as follows:

$$precision = \frac{no.of\ terms\ in\ the\ ontology\ that\ appear\ in\ the\ corpus}{no.of\ terms\ in\ the\ ontology} \qquad (16)$$

$$= \frac{18{,}143}{18{,}586} = 0.9762$$

$$recall = \frac{no.of\ terms\ in\ the\ corpus\ that\ appear\ in\ the\ ontology}{no.of\ terms\ in\ the\ corpus} \qquad (17)$$

$$= \frac{18{,}143}{22{,}589} = 0.8032$$

In this context, the precision value represents how much the metadata in the ontology could be utilised by the system e.g. for matching and querying. This is because almost all of terms in the ontology are the terms that also appear in the corpus. Therefore, the system can use them for matching and querying the terms in the corpus efficiently. From the calculation above, the metadata in the presented ontology could be exploited by the system up to 97.62%. The recall value presents how much metadata in the ontology covers the body of knowledge in the corpus. The result shows that the metadata in the presented ontology is able to capture about 80.32% of the knowledge in the corpus. Another ontology evaluation technique is performed in order to measure the similarity between the metadata in the presented ontology and the content of the corpus using Latent Semantic Analysis (LSA).

5.5.2 Ontology Evaluation using Latent Semantic Analysis

This approach uses a vector space representation of the terms in both the corpus and the presented ontology. Latent Semantic Analysis (also known as Latent Semantic Indexing) is a method in which Singular Value Decomposition (SVD) is used to form semantic generalisations from textual passages. It uses the characteristic that certain words appear in similar contexts to establish relationships between the meanings of the words. LSA compares the *meanings or concepts behind the words* between two set of data. This permits an overall measure of the "fit" between the presented ontology and the corpus of documents.

LSA begins with tokenising data in the corpus, removing stopwords, and calculating the term frequency. Each term is given a weight depending on its appearing frequency in the same document (Local weight) and in the corpus (Global weight), and the number of documents in which it appears (Normalisation). LSA represents terms information in the form of a matrix (vector space model) and performs statistical computations to measure the similarity between two metrics. Figure 5-14 shows an example of a vector space model of keywords (K) appeared in the Ontology (O) and the corpus (C) where N is a weighted term.

Keywords	Ontology ($\mathcal{O}$)	Corpus ($\mathcal{C}$)
K_1	N_{11}	N_{12}
K_2	N_{21}	N_{22}
...	…	…
K_m	N_{m1}	N_{m2}

Figure 5-14 Example of a matrix shows a number of words appeared in an ontology and the corpus

To measure the similarity between the ontology and the corpus, the cosine similarity measurement is deployed in this evaluation. This produces a similarity score (Equation (1)) in the range [0, 1]. Next, the similarity between $\vec{V}(\mathcal{O})$ and $\vec{V}(\mathcal{C})$ is computed as follows:

$$sim(\mathcal{O}, \mathcal{C}) = \frac{\vec{V}(\mathcal{O}) \cdot \vec{V}(\mathcal{C})}{|\vec{V}(\mathcal{O})||\vec{V}(\mathcal{C})|}$$

$$= \frac{(N_{11} \times N_{12}) + (N_{21} \times N_{22}) + \cdots + (N_{m1} \times N_{m2})}{\sqrt{(N_{11}^2 + N_{21}^2 + \cdots + N_{m1}^2) \times (N_{12}^2 + N_{22}^2 + \cdots + N_{m2}^2)}} = 0.8493$$

The similarity between the presented ontology and the corpus is 0.85 which is higher than the recall value in the previous section (p.84). This is because LSI gives weights to terms in the corpus. Consequently, the content in the corpus is represented more efficiently and thus the similarity between both set of data is better measured. This method provides the figure which reflects the coverage of the presented ontology of the corpus.

5.6 Knowledge-based Image Retrieval Processes

In the previous section, the processes to produce semantic metadata have been revealed. This section describes how to exploit the generated semantic metadata, and how to implement semantic queries that meet the requirements for a *Ranked result,* and for *Ontology incompleteness*.

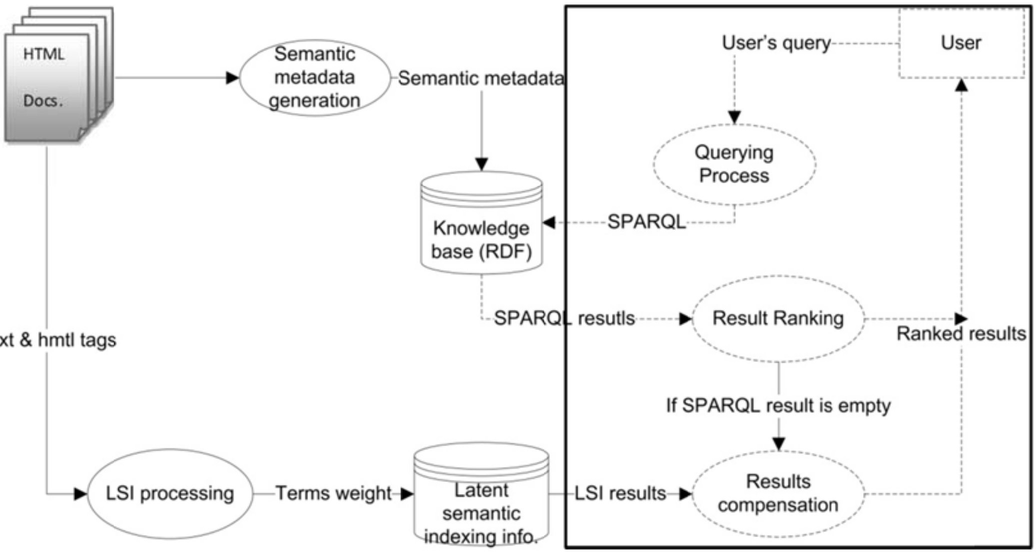


Figure 5-15 Knowledge-based image retrieval processes

Figure 5-15 shows the processes needed for image retrieval. This section presents an approach towards image retrieval based upon extracted knowledge, and visual and textual features. The main contribution of this section is the enhancement of traditional CBIR systems. Ontology-based information retrieval can be seen as an evolution of classic keyword-based retrieval techniques, where the keyword-based index is replaced by a semantic knowledge base. Here, the simple query approach is used for searching information in the system. Users can input two types of querying, Textual or Visual query. Therefore, the Query processing part needs to handle these two types of query properly.

Algorithm 3: Multiple word sense disambiguation

```
Input: A user query
Output: A list of similarity words $\{W\}$
1. The user's query process to remove stopwords and to stem these in order to get a
   set of query keywords $\{Q\}$;
2. Look up all remaining words $\{K\}$ in WordNet and assign senses $\{Q\}$ to all words;
3. For each keyword pairs $\{k_i,q_i\}$
   where $k \in K$ and $q \in C$ {
       3.1 Compute similarity between $\{k_i,q_i\}$ unless $k=c$;
       3.2 Assign score to $k_i$;}
4. Sort $\{K\}$ according to similarity score;
4  Select the concept $(c_i)$ which has the highest similarity value of word sense =
   $\{W\}$.
```

5.6.1.1 Textual Query

This section deals with textual based queries expressed using natural language. It starts with the "*stop words removing process*". The query keywords from users are examined by tokenising them and then stop words are eliminated. The remaining words are assumed to be the important keywords for searching images and these keywords will be disambiguated using WordNet. To find an appropriate word sense for a user query, it is proposed to use an algorithm to disambiguate multiple word senses in a user query as shown in Algorithm 3. Hence, the system can perform the semantic search on RDF with the appropriate word sense detected from a user query. The Query processing translates the user query to *a SPARQL query* and searches for any relevant information within the KB. Finally, images annotated with these instances are retrieved. These images are ranked and presented to the user.

5.6.1.2 Visual Query

When a user inputs an image as a query, the Query process also transforms visual data into high level semantics similar to the technique presented in section 5.3.1 (p.58) in order to generate visual words. The created visual words will be processed in order to annotate an image query based upon relevant concepts using the ontology model as shown in Figure 5-7. Next, all annotations will be used as keywords for searching information in the ontology. A search engine performs a semantic search. This technique offers the potential for enhancing traditional CBIR systems because the search engine can perform conceptual searching using keywords (interpreted from visual data) and can expand these keywords to other relevant concepts. These keywords are used for searching rather than just simple low-level feature matching. As a consequence, more relevant documents can be recognised and retrieved.

5.7 Summary

In this chapter, some graphical ontology tools for development were introduced and analysed. From this analysis, these tools still have some problems in handling a large ontology base. Then the Sport Ontology used for this research is given. Three main ontologies are created, a Sport Domain, Visual Features and an Image Annotation ontology. The Sport Domain Ontology is used for the vocabulary and background knowledge of an image's subject domain description whereas the Image Annotation Ontology is designed for storing the annotations of images in the collection. The Visual Features Ontology mainly stores the extracted low-level features, image size, image format etc. The designed ontology is evaluated using a data-driven scheme that based on the modified precision and recall technique as well as on cosine similarity measurements in order to study how it fits and covers information in the knowledge domain.

This chapter introduces two major components of the presented framework, Visual analysis and Linguistic analysis. In the visual analysis component, an ontology model is used to enhance both image classification and traditional CBIR of visual content. In the linguistic analysis component, the techniques to extract metadata from text captions and to transform these metadata into the knowledge base model are described. To this end, the key contributions of this chapter can be summarised as follows:

First, a technique is presented to improve the quality of the generated visual words using the semantic local adaptive clustering (SLAC) algorithm. Unlike other clustering algorithms, SLAC is an extension of traditional clustering that captures the relevance of the representative keypoints by exploiting a *semantic matrix*. As a result, relevant keypoints can be clustered together and, consequently, the visual words generated, seem more robust and more efficient in representing the semantics of visual data.

Second, a technique to detect the domain specific *non-informative visual words* that add no value when representing visual content and which degrades the categorisation capability, is presented. A Chi-square statistical model is utilised to identify meaningless visual words that have two main characteristics: a high document frequency (*DF*) and a small statistical correlation with the concepts in the collection. These visual words are discarded in order to enhance the discrimination power. To normalise those visual words that appear in a small number of concepts, visual words are weighted before eliminating them.

Third, a method to restructure visual word vector space model into an ontology-based model in order to disambiguate visual word senses is given. Unlike a hierarchical model described in the several state of the art ontology frameworks, the ontology model in this framework is better able to capture the knowledge of sport domain in the real word e.g. it does not use a binary tree and can support multi-parent relationships. This technique and the concept range measurement method are very useful for image classification and retrieval tasks that do not only rely on visual similarity but rather on conceptual similarity. In other words, the technique is able to resolve the visual heterogeneity problem.

Fourth, a method to restructure the unstructured textual information e.g., natural language to form the semantic metadata in the ontology model is presented. Three main steps to transform textual information into the ontology model are proposed. In addition, the metadata can be extended using semantic rules and stored in the semantic model. The main advantage of this approach is that it can find indirectly relevant concepts not explicitly mentioned in the surrounding text by exploiting semantic relations stored in the ontology whereas most of state of the art systems only perform match text snippets to ontology entities.

Since there are inherent uncertainties in the visual and textual analysis processes, e.g. object recognition errors and incompleteness metadata in the knowledge base, these need to be handled in order to enhance the stability of the system. In the next chapter, a method to cope with these uncertainties is introduced.

Chapter 6

Handling Uncertainties in Visual Classification and in the KB

Basically, there are three main causes for the *visual content interpretation or classification uncertainty* (Cullen et al. 1992). First, uncertainty can arise from using an *incomplete image* as an input. The image may not contain sufficient information to make a classification. For example, some key objects may be out of frame due to the camera angle. Thus, the system does not have adequate data to classify the content of an image efficiently. Second, uncertainty may be caused by the *ambiguity of an object*. For example, an object could appear in several sport events. Hence, it is difficult for the system to reach the conclusion that an image is relevant to a particular event. Finally, object *recognition errors* may occur. An object recognition algorithm might not be able to detect some objects of interest in an image due to noise or due to the poor quality of an image. A more robust and reliable system needs a method that can handle the uncertainty and ambiguity in image classification. Furthermore, because of a problem of the incompleteness of semantic metadata in the KB (section 1.1.3, p.5), so the system should not rely only on the metadata in the KB. In addition, it should be designed as an open KB (section 5.1.2, p.51). A system which relies only on information in ontology will return an empty answer to users when there is no relevant information stored in the knowledge base. This is called *KB uncertainty*. In this chapter, therefore, some solutions have been proposed to allow the presented system to handle these uncertainties.

6.1 Uncertainty and Ambiguity of Image Classification

An intuitive proposal for uncertainty management of visual content interpretation is to determine how likely a scene in an image will occur if some objects cannot be detected due to incomplete image, object recognition uncertainty and object ambiguity. To handle these uncertainties, probability theory seems to be the prevailing method for dealing with uncertainty. Andrea *et al.* (2004) proposed a method to transform low-level visual features to high-level descriptors for videos. Objects are extracted from videos using spatial and temporal features and used for interpreting the visual data. This framework, however, ignores some uncertainty problems that may occur. For example, the framework may not be able to extract the underlying objects from a video frame due to an error in an object recognition algorithm or because of background noise. Thus, the system cannot abstract the meaning contained in the visual content.

Cullent *et al.* (1992) presented a method to cope with the uncertainty of image interpretation using constraint satisfaction and failure analysis of the results from informed-backtracking. By using this technique, the framework is able to generate partial solutions and infer values for them, and also overcome errors in object recognition. A Bayesian network has been used to reason about the uncertainty of low-level features and to handle the uncertainty arising from the inseparability between objects that have similar features. Uncertainty management in visual information has also been proposed in (Marengoni et al. 2003). This research applied a Bayesian network and utility theory to reason about satellite images. A Bayesian network has been exploited to model object knowledge in a hierarchical structure and to make decisions in an aerial image interpretation system.

One problem of the proposed framework is that it cannot classify images properly when key objects are absent. For instance, a pole object in a pole vault image (Figure 6-1) is missing. The system could for example classify the content assigning a “high jump” event for the image. In this case, it can be difficult for a computer system to recognise that this image concerns a pole vault event because a key visual object is missing. Usually, humans use their past experiences to interpret visual content when it is ambiguous. Likewise, a computer system can be designed to interpret the meaning of an image based upon the previous data.

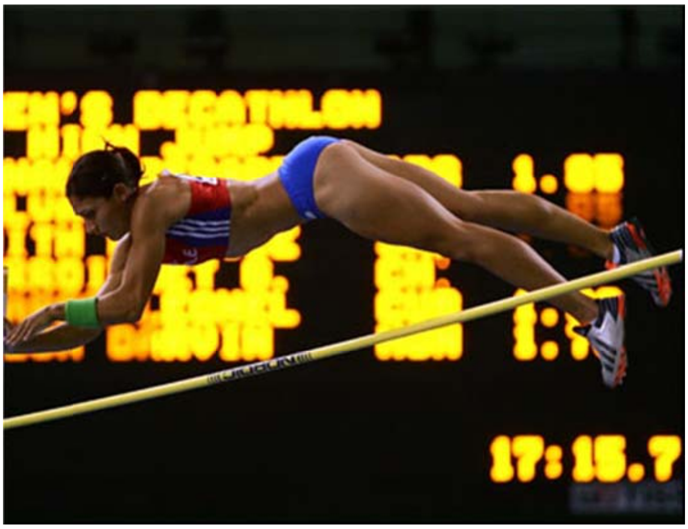


Figure 6-1 Example of a pole vault image which a pole is missing; it is difficult for the system to interpret that this image is relevant to a pole vault event or a high jump event

Clearly, an image classification based solely on visual data is not sufficient to categorise an image in this case. Thus, the system needs additional information to aid the classification process e.g. text captions. Therefore, it is proposed that a probability model, e.g., a Bayesian network, integrates both visual data and text captions to better determine how likely the image represents a specific sports event. To do this, the frequency of occurrence of the objects in images and the frequency of keywords from text captions have been counted and a Bayesian network has been used to model the frequency of occurrence. The system interprets images based on the probability of objects from previous data. This can enhance the categorisation ability of the proposed system and can handle the uncertainty.

6.1.1 Bayesian Network Deployment

A Bayesian network is a directed acyclic graph (DAG) which defines a factorisation of a joint probability distribution over the variables that are represented by the nodes of the DAG and the directed links represent the dependencies between the nodes

(Krebs et al. 1998). It encodes the probabilistic relationships amongst the variables of interest. Each node in a Bayesian network has to specify the prior and the conditional probabilities. The conditional probability of a node represents the conditional belief that a child node is caused by parent nodes. When used in conjunction with statistical techniques, the graphical model has several advantages for data analysis. One, because the model encodes the dependencies among all variables, it readily handles situations where some data entries are missing. Two, because the model has both causal and probabilistic semantics, it is an ideal representation for combining prior knowledge and data.

DEFINITION 5 A Bayesian network (Jensen & Nielsen 2007) $\mathcal{N} = (\mathcal{X}, \mathcal{G}, \mathcal{P})$ consists of

- a DAG $\mathcal{G} = (V, E)$ with nodes $V = \{v_1, \dots v_n\}$ and directed links E
- a set of discrete random variables, $\mathcal{X}$, represented by the nodes of $\mathcal{G}$
- a set of conditional probability distributions, $\mathcal{P}$, containing one distribution, $P(X_v|X_{pa(v)})$, for each random variable $X_v \in \mathcal{X}$ and Pa(v) refers to parent nodes of v.

The conditional probability of a node B having value b was caused by a node A having value a can be described in the expression:

$$P(B = b \mid A = a) = \frac{P(B = b)P(A = a \mid B = b)}{P(A = a)} \tag{18}$$

Equation (18) is also known as *Bayes' rule.* $P(B=b|A=a)$ is known as the “posterior probability” which indicates the probability of B after considering the effect of A on a. The term $P(B=b)$ is called the “prior probability” which refers the probability of B given b alone. The term $P(A=a|B=b)$ is called the “likelihood” which gives the probability of A assuming B on b is true. The term $P(A=a)$ is called the marginal likelihood which is a normalising or scaling factor. To solve a Bayesian network $\mathcal{N} = (\mathcal{X}, \mathcal{G}, \mathcal{P})$, it needs to compute all posterior marginals given a set of evidence ε, i.e., $P(X|\varepsilon)$ for all $X \in \mathcal{X}$. For example, the key objects for the high jump event are Athlete (A) and Horizontal bar (B).

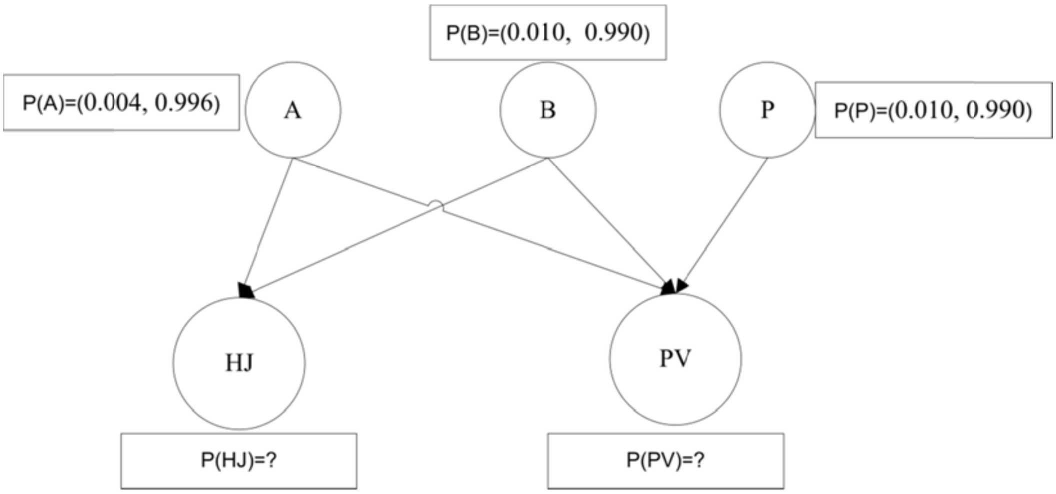


Figure 6-2 Bayesian network for the reduced sport image ambiguous interpretation; A = Athlete; B=Horizontal bar; P=Pole; HJ= High jump event; and PV= Pole vault event

In other words, an athlete and a horizontal bar characterise the high jump event. A pole object combined with an athlete and horizontal bar object can identify the pole vault event. From this background knowledge, a Bayesian network for image classification can be constructed using the DAG $\mathcal{G}$ shown in Figure 6-2. The graph consists of five nodes: Athlete (A), Horizontal bar (B), Pole (P), High Jump event (HJ) and Pole Vault event (PV). This graph is used for image classification to differentiate between a high jump and a pole vault event.

For the quantitative modelling, the probability assessments *P(A)*, *P(B)*, *P(P)*, *P(HJ|A,B)*, and *P(PV|A,B,P)* are needed. For probabilistic reasoning about data in a Bayesian network, the states or possible values of variables (nodes) need to be identified. The possible states of variables can be either "*yes*" or "*no*". Yes means that object appears in an image otherwise is no. TABLE 6-1 can be drawn to show all possible states of all variables in Figure 6-2.

TABLE 6-1: Variables and their possible states for the high jump and pole vault sport event interpretation. yes = an object appears in an image, no = an object does not appear in an image

Variables	**Possible states**
Athlete (A)	no, yes
Horizontal Bar (B)	no, yes
Pole (P)	no, yes
High jump (HJ)	no, yes
Pole vault (PV)	no, yes

TABLE 6-2: Conditional probability distributions for High jump given Athlete and Bar

		High Jump Event	
Athlete	Bar	no	yes
no	no	1.00	0.00
no	yes	0.50	0.50
yes	no	0.90	0.10
yes	yes	0.01	0.99

TABLE 6-3: Conditional probability distributions for Pole vault given Athlete, Horizontal Bar and Pole

			Pole Vault Event	
Athlete	Bar	Pole	no	yes
no	no	no	1.00	0.00
no	no	yes	0.00	1.00
no	yes	no	0.50	0.50
no	yes	yes	0.00	1.00
yes	no	no	0.90	0.10
yes	no	yes	0.00	1.00
yes	yes	no	0.80	0.20
yes	yes	yes	0.00	1.00

The training collection of the proposed system, randomly collected from the Olympic website, contains 500 images for five sport genres, pole vault, high jump, long jump, javelin throw, and hammer throw (100 images for each category). Among these, only 2 images do not contain an athlete object. Therefore, respecting the order of states in TABLE 6-1, the probability distribution for Athlete is *P(Athlete)=(0.004, 0.996)*. Similarly, 99% of high jump and pole vault images contain the horizontal bar object:

P(Bar)=(0.010, 0.990) and 1 out of 100 of pole vault images do not contain the pole object. Therefore, the probability distribution for Pole is *P(Pole)=(0.010, 0.990)*. Next, the initial conditional probability distributions of all variables for the high jump event and the pole vault event need to be specified based on the observed data in training set. The initial conditional probability of the high jump and the pole vault event are shown in TABLE 6-2 and TABLE 6-3 respectively.

From the probabilities specified in TABLE 6-2 and TABLE 6-3, the *P(HJ)* can be computed using Equation (18) as:

$$P(HJ_{=no}, HJ_{=yes}) = \sum_i P(HJ \mid A, B)P(A)P(B) = (0.0208, 0.9792)$$

$P(PV_{=no}, PV_{=yes})$ is computed similarly to *P(HJ)* obtaining (0.0081, 0.9919)

For *P(HJ)*= (0.0208, 0.9792), this indicates that if a new image enters the system and the Athlete and the Bar object can be detected, the system will classify an image as a high jump event with 97.92% confidence. Similar to *P(PV)*, if the three objects, Athlete, Bar and Pole, are recognised, there is a chance about 99.19% that an image could be a pole vault image. When a new image enters the system, the system detects key objects in an image. For instance, if an Athlete and a Bar object are recognised in an image but not a Pole object. According to the probability data in TABLE 6-2 and TABLE 6-3, the probability of $P(HJ_{=yes}|A_{=yes}, B_{=yes})$ and $P(PV_{=yes}|A_{=yes}, B_{=yes}, P_{=no})$ are computed and compared.

$$P\left(HJ_{=yes}\middle|A_{=yes}, B_{=yes}\right) = 0.99$$

$$P\left(PV_{=yes}\middle|A_{=yes}, B_{=yes}, P_{=no}\right) = 0.20$$

These probabilities suggest that an image is more likely to be a high jump image than a pole vault. Thus, the system classifies and annotates that image as a high jump with a certain degree of confidence using the probability measurement, i.e. 99% confidence in this example.

To enhance the categorisation power if the text caption is supplied, the probability of relevant text captions is added to the conditional probability table of a Bayesian network. It is assumed that the text captions are reliable because they are generated by humans and 99% of images in the training collection have text captions which mentions the "*sport name*" or relevant concepts of the sport i.e. 800m or 4x100m explicitly, *P(SportName)=(0.01, 0.99).*

TABLE 6-4: The conditional probability after adding the sport name probability. Sport name = *yes* means a sport name is explicitly mentioned in a text caption otherwise it is set to no.

			High Jump	
Athlete	Bar	Sport name	no	yes
no	no	no	1.0	0.0
no	no	yes	0.0	1.0
no	yes	no	0.5	0.5
no	yes	yes	0.0	1.0
yes	no	no	0.9	0.1
yes	no	yes	0.0	1.0
yes	yes	no	0.9	0.1
yes	yes	yes	0.0	1.0

Having added sport name probability from relevant text captions to the conditional probability table, TABLE 6-4 shows a new conditional probability table. As a further illustration, the following scenario is provided. If the object recognition algorithm can detect only an object "Bar" in an image, the system will interpret that an image is relevant to the "high jump" with 50% confidence ($P(HJ_{=yes}|A_{=no},B_{=yes},S_{=no}) = 0.5$). This is because a Bar object can appear in both a pole vault and a high jump event. When the probability of sport name in text captions is integrated, the degree of confidence is increased from 50% to 100% ($P\left(HJ_{=yes}\middle|A_{=no},B_{=yes},S_{=yes}\right) = 1.0$) as shown in TABLE 6-4. This example illustrates the fact that text captions can enhance the image classification process by increasing the degree of confidence for image interpretation.

In summary, the integration of visual and textual data from text captions to classify visual content using Bayesian network leads to the system being able to handle the uncertainty more effectively when the key objects of an image for classification are missing or ambiguous.

6.2 Handling the Ambiguity of Missing Metadata in the Knowledge Discovery Process

A Bayesian network is also applied to manage the uncertainty of the Knowledge discovery process (section 5.3.2.3 p.78). For instance, the system tries to find a sport name for an image through the detected athlete name. Typically, an athlete performs only one sport in the Olympic Games in the same year. However, he could perform a different sport in previous years. Therefore, the probability from the previous data is used to calculate, for example, how likely a sport appears in an image. For example, "Kelly Holmes" is a UK athlete who usually performs 800m running in Olympic Games between 2000-2008. However, she performed 1500m running in 2004 Olympic Games. Therefore, the probability distribution that an image is more likely to be Kelly performing 800m running is two third (67%) or *P(800m) = (0.33, 0.67).* If the text caption of an image mentions a distance of a track event explicitly such as 400m, 800m or 1500m, the probability distribution of a distance of a track event will be increased.

TABLE 6-5: The conditional probability of the 800m Track event for Kelly Holmes. Kelly= *yes* means the word "Kelly" appear in a text caption otherwise it is set to no. Sport name = *yes* indicates that the "800m" distance is explicitly mentioned in a text caption otherwise it is set to no.

		800m	
Kelly (K)	Sport name (S)	no	yes
no	no	1.00	0.00
no	yes	0.00	1.00
yes	no	0.33	0.67
yes	yes	0.00	1.00

From the observed data in the training set, all text captions of Kelly's images usually mentioned her name, *P(Kelly) = (0.00, 1.00)* and 99% of text caption of Kelly's images mention the distance of a track event that she performs, *P(SportName) = (0.01, 0.99*). Using the Bayes' rule from Equation (18), it obtains:

$$P(800m_{=no}, 800m_{=yes}) = 0.003, 0.997$$

This means the system will annotate an image of Kelly Holmes as the Track event, 800m with 99.7% confidence. A Bayesian network is applied to other similar tasks e.g. when the system found that two sport types related to one athlete. The computational process is similar to the above examples.

6.3 Handling the Incompleteness of the KB

As mentioned previously (in section 3.1 p.26), building an ontology to cover everything in the real world domain is very challenging. A system that relies only on information in the KB, will not return any answer to users when there is no relevant information stored in the KB. In addition, the KB in this thesis is designed as an open KB. Thus, it should re-try to find answers for an *unknown* query using another technique if it fails to find answers in the KB. However, there are some requirements for this second method as follows:

- It should still provide the capability for a semantic search based on textual information in an image caption, e.g. to find the indirectly relevant concepts which are not identified explicitly in the document text, i.e. to improve the retrieval performance compared to a traditional text-based search (String matching).
- It should be able to handle the problem of synonymy and ambiguity of words.

To handle these, a Latent Semantic Indexing (LSI) technique is deployed to cope with the uncertainty of the KB.

LSI is a well-known technique used for text-based information retrieval (Deerwester et al. 1990). Textual information is represented by low-dimensional vectors that can be matched against user queries in the LSI "semantic" space. LSI tries to search for something that is closer to representing the underlying semantics of a document (Praks et al. 2003). Cascia *et al.* (1998) proposed a system which combines textual and visual statistics in a unified vector for content-based search of a WWW image database. LSI is applied for indexing information from HTML tags with different weights e.g. a *title* tag has higher weight than a *H1* tag. After LSI computation, the LSI vector space is integrated with a visual statistic vector space (colour histogram and dominant orientation histogram). Later, users are included in the search loop, to provide relevance feedback, in order to achieve a higher retrieval performance. However, the results show that the retrieval performance is low because HTML tags cannot represent image content correctly, e.g., in some cases, web pages and image galleries may contain several images and/or very little text. This is similar to work by Westerveld *et al.* (2000), in which image features and words from collateral text are combined into a single semantic space. This method supports only QBE in order to retrieve relevant images. However, it is not clear how the proposed system deals with uncertainty, e.g. when there is no text caption accompanying an image. This means the system will perform indexing using only the visual data. In such a case, the retrieval performance may become worse than using text-based retrieval alone.

6.3.1 LSI Deployment

The LSI vector space model is used as a backup method when the system cannot find the desired information in the ontology model. In addition, the LSI technique can be improved by adding a NLP function before indexing the textual information. This can reduce noise words and detect *named entities* more correctly e.g. name of person, event, or places even when multiple word names may contain whitespace. The details of the use of LSI are described below.

LSI is able to handle language vagueness e.g. synonyms, and to present more relevant images in response to a user query through using a more sophisticated statistical calculation. The results can be ranked by descending order according to the relevance value. This means LSI can fulfil the *NL vagueness and the Ranked result requirements*. The LSI computational algorithm is illustrated in Algorithm 4. The LSI process starts with a *Textual information filtering* process to extract textual information from HTML documents. All HTML tags are filtered out and the remaining text will be used to find keywords in the next step.

Then, a *Tokenising step* will process the remaining text from the previous step. Textual information will be tokenised into several words. Unlike other tokenising processes, NLP is also applied to this step in order to enhance the named entities recognition. The words that appeared in the caption are the potential keywords for the image. However, some words from the Tokenising step are not useful and important e.g. a, an, the, without, before. Therefore, they are removed from the set of words by a *Stop words removing step*. The rest of the text, after removing the stop words, is assumed to be the keywords that characterise the image and are stemmed from the original form of each word.

Algorithm 4: LSI computational algorithm

```
Input: Text information
Output: LSI indexing model
1. Textual information filtering $\{P\}$;
2. Tokenising data in $\{P\}$ and Natural Language Processing (NLP)$\{T\}$;
3. Remove stopwords from $\{T\}$-> remaining words $\{K\}$
4. Term frequency calculation
   for each term $k_i$ and $k_i \in K$ and A is a matrix
   if $k_i \notin A$ (found new term) {
      add $k_i \Rightarrow \{A\}$ and $f_i = 1$;
   }
   else {
     $f_i = f_i + 1$;}
5. Term weighting computation
```

Next, *Term frequency calculation step* is performed. The numbers of terms that appear together with the image are counted. These frequencies are used to determine the degree of importance of that term to the image. Term frequency in each document is represented in a matrix in which rows represent keywords and columns represent image using image IDs. Figure 6-3 shows an example of matrix A with term frequency in each image. This matrix is also called a vector space model.

A=		Img1	Img2	img3	img4	img5
	Track	2	0	3	0	0
	Andre KURT	1	3	0	1	1
	United States	0	0	0	1	1
	100m	0	0	0	1	2

Figure 6-3 Example of a term frequency matrix

Having created the term frequency matrix, a *Term-image relationship computation process* computes and assigns a weight to each term using the following equation:

$$W_{ij} = L_{ij} G_i N_j \tag{19}$$

where W_{ij} = Weight of each term, L_{ij} is the local weight for term i in document (image caption) j, G_i is the global weight for term i, and N_j is the normalisation factor for document (image caption) j. There are several formulae for local, global weight, and normalisation factors. In this framework, the selected formulae are based on the survey and recommendation given in (Chisholm & Kolda 1999).

Local weights are functions of how many times each term appears in a document. The local weight is computed according to the terms in the given document or the query. To compute the local term weighting, the Logarithms scheme (LOGA) is used in this research. Logarithms are used to adjust within-document frequency because a term that appears ten times in a document is not necessarily ten times as important as a term that appears once in that document. Equation (20) shows the LOGA formula.

$$L_{ij} = \begin{cases} \log f_{ij} + 1 & \text{if } f_{ij} > 0 \\ \\ 0 & \text{if } f_{ij} = 0 \end{cases} \tag{20}$$

where f_{ij} is the frequency of term i in document (image caption) j. The global weights are functions of how many times each term appears in the entire collection. Global weighting tries to give a discrimination value to each term. This technique is based on the idea that the less frequently a term appears in the whole collection, the more discriminating it is. To compute a global term weighting, the Inverse Document Frequency (IDF) method is deployed. IDF formula is illustrated in Equation (21).

$$G_i = \log\left(\frac{N}{n_i}\right) \tag{21}$$

where N is the number of documents in the collection and n_i is the number of documents in which term i appears. The normalisation factor compensates for discrepancies in the lengths of the documents and has to be done after the local and global weighting. Equation (22) shows the normalisation formula.

$$N_j = \frac{1}{\sqrt{\sum_{i=0}^{m} \left(G_i L_{ij}\right)^2}} \tag{22}$$

The reason to normalise each term is because:

1. Long documents tend to gain a higher score since these tend to have more words and more same word occurrences. These tend to score high simply because they are longer, not necessarily because they are more relevant.
2. Repetition does not necessarily mean that terms are more relevant. The same terms might be repeated within different contexts or topics.

The entire collection is represented by a matrix *A*. Next, the Singular Value Decomposition (SVD) of the matrix A is computed using formula in Equation (23) in order to reduce the matrix dimensions.

$$A = USV^T \tag{23}$$

Having computed SVD, it obtains so-called *concept vectors* (left-singular vectors – the columns in U), which can be interpreted as individual (semantic) topics present in the collection. An important property of SVD is that concept vectors are ordered according to their "significance", which is defined by values of the singular values k_i stored in ascending order in the diagonal matrix *S*. Informally, the concept significance says which concepts are semantically important and which are not. This is where the "latent semantics" comes from. Unimportant concepts are regarded as "semantic noise".

To reduce the matrix dimensions, only the largest k singular values will be selected. Therefore, the decomposition can be approximated as:

$$A \approx U_k S_k V_k^T \tag{24}$$

where U_k contains the first k most important concept vectors, S_k contains the respective singular values and $S_k V_k^T$ contains the document vectors represented using the first k concept vectors. In other words, by SVD the original m-dimensional vectors are projected into a vector space of dimension k ($k <= m$). The SVD approximation (so-called *Rank-k approximation*) can be created by selecting only first k columns of U and first k rows of S and V^T .

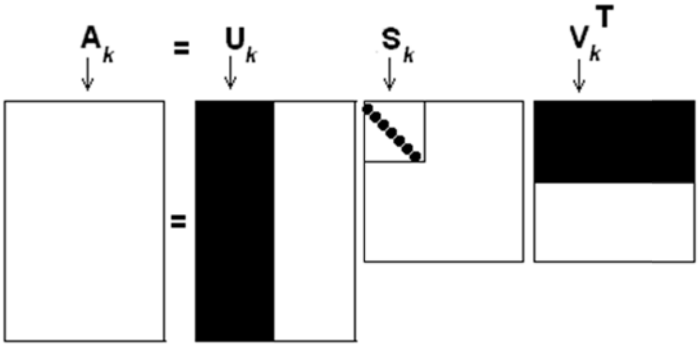


Figure 6-4 Rank-*k* approximation for matrix dimensional reduction

For the retrieval of images, the document vectors and the query vector are compared using a similarity measure e.g. the *cosine similarity* measure.

6.4 Integration of LSI with the SBIR

The results from a SPARQL query might be not adequate due to incomplete semantic metadata in the ontology. Moreover, even if the ontology could theoretically provide adequate coverage, it may still be the case that for a specific image no semantic metadata was generated due to object recognition or natural language processing errors. In such cases, the metadata query would clearly not return any result to a user. Therefore, the SPARQL query results can be compensated with the LSI search results to solve this problem. This serves to solve the *ontology incompleteness requirement.*

It is notable that the word "*compensation*" is used instead of "*combination*" because the results from SPARQL are not combined with the LSI-based results. Result compensation refers to a mechanism to activate LSI to search for information from the corpus of documents. Hence, the LSI results are used instead of SPARQL results. This is because the combination of the results from both approaches often decreases the retrieval performance. LSI-based searches work based upon statistic calculation. It is possible that LSI-based searches can obtain some irrelevant images. These irrelevant results can reduce the precision and recall when they are combined with the ontology-based search results. Therefore, the LSI results are used to compensate the SPARQL results for the case when the ontology-based search fails to find any relevant information.

6.5 Summary

This chapter illustrated the techniques to deal with uncertainties in both visual and textual information. First, a Bayesian network is exploited to handle the uncertainty of the Knowledge discovery process and the image classification. As an open KB, it is possible that terms in a query is unknown by the KB. Therefore, an LSI technique is introduced to cope with the incompleteness of the KB.

The uncertainty of visual classification is handled using a Bayesian network based upon the probabilities of visual data and text captions. The classification performed is based on the degree of confidence calculated from the conditional probability. In addition, the use of a LSI model to enhance the open KB leads to better reliability for IMR through compensating the retrieval results from the ontology model for the case when metadata in the KB is incomplete. The LSI algorithm has been modified by adding an NLP function in order to improve the word tokenising function and the name entity recognition performance. Consequently, words in the image caption, e.g. name of person, event, place etc, can be more correctly identified.

The main assumption underlying the need to apply LSI technique to SBIR system is that even though ontologies in practice are often incomplete, the system should not fail, returning no results to users. Therefore, the system needs a backup approach to aid retrieval mechanism when the system fails to find relevant information. Hence, an innovative idea was proposed to integrate a LSI technique with SBIR in order to solve the *ontology incompleteness* problem (Section 3.1). In next chapter, the experimental results are given and discussed.

Chapter 7

Experimental Results Evaluation

The ideas and algorithms previously presented in Chapter 5 and 6 were implemented in a prototype system. The system's user interface allows non-expert users to more easily interact with the system and exploits an ontology and the generated metadata in order to make the search process more semantic. This chapter begins by describing the implementation of the prototype system and then describes the experiments and experimental results. The chapter ends with an evaluation of the system performance.

7.1 System Implementation

The user interface of the prototype system and the workflow of the end-user are described. Users are able to specify queries in a simple natural language by typing these into a text field as this is more intuitive and natural for users rather than needing to know the syntax to write a structured query or having to fill out a form. The system should be simple and easy to use so the user interface of the prototype system is designed to be similar to the Google search engine. In addition, results should be ranked according to the degree of relevance to the user query to enable users to focus on the top ranked, more relevant results to the query rather than on the less relevant results.

7.1.1 Implementation of the Semantic Search Process

The semantic search process is one of the most important components of the system. For a text query, the system takes a NL user query, parses it (`QueryParsing`), transforming the query to a SPARQL query and performs a semantic search (`SemanticSearch`), and returns a list of ranked results to the user interface

(`ResultsRanking`). In addition, if the ontology KB search (SBIR) cannot find any relevant documents using the ontology model, the LSI search engine is activated (`LSISearch`).

For a visual query (QBE), the system extracts low-level features (SIFT) and generates visual words (`VisualWordGeneration`). Later, it maps low-level features to high-level semantic annotations through matching visual words to concepts in the ontology model (`ConceptMatching`). Then it uses those high-level semantics annotations as keywords for searching relevant images in the KB (`SemanticSearch`). Finally, the results are ranked and returned to user via the user interface (`ResultsRanking`). Figure 7-1 shows the sub-processes for the semantic search process.

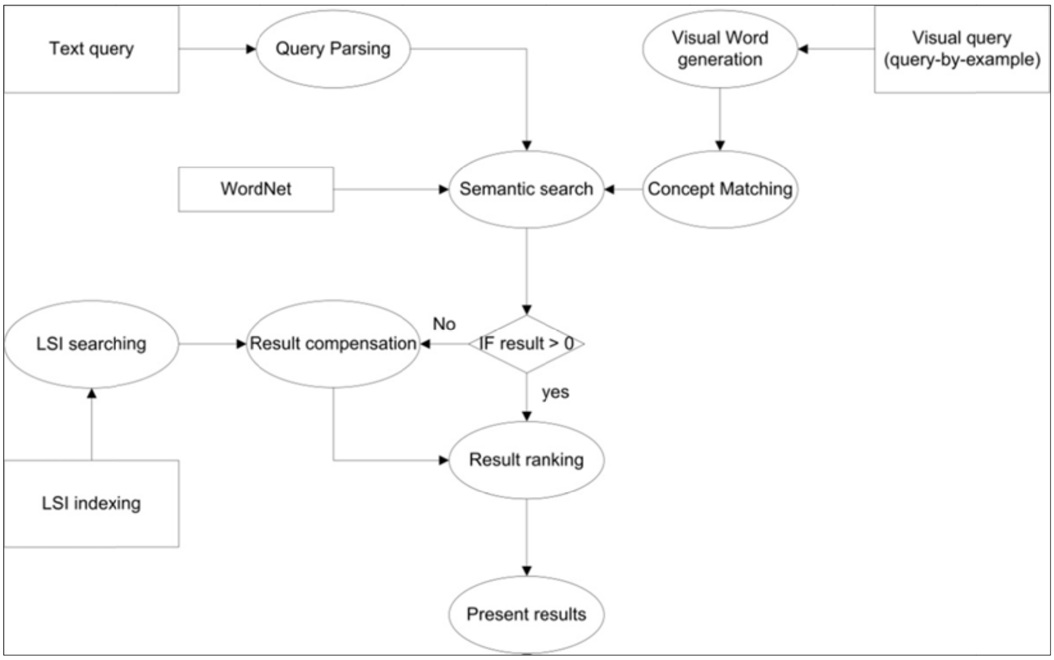


Figure 7-1 Semantic search processes

7.1.1.1 Implementation of Query Parsing Module for the Text-based Query

Users can enter a query keyword via a text field in a similar way to the Google search engine Web page and press a “Submit Query” button. A query string will be stored in a variable using the method `getQueryKeyword()`. Before eliminating stop words from the query, it needs to separate the query into individual words using the `StringTokenizer()` function. To eliminate stop words from the query, a list of stop words is read from a text file by the `FileReader()` method and stored in a buffer variable in order to be further processed. All special characters (e.g. ! @ # $ % ^ & * + ; : [ ]) are deleted from query keywords and, then, all words are compared with words in the stop word list. If any words from the query are not in the stop word list, then they become the feature keywords to be used in the `SemanticSearch()`function. Algorithm 5 gives the query parsing algorithm.

Algorithm 5: Query parsing algorithm (`QueryParsing`)

```
     Input: A list of words from a user query
     Output: Feature keywords
  1. Get a list of words from a user query and store it a variable
     InputKeyword =  getQueryKeyword();
  2. Separate words in the user query
     Str= StringTokenizer(InputKeyword);
  3. Read stop word list from text file
     StopWords = FileReader("stopword.txt");
  4. While (Str.hasMoreTokens) {
     Delete all special character e.g. ! @ # $ % ^ & * + ; : [ ]
         IF str.nextToken <> null and !StopWords.contain(str.nextToken){
           //Add str.nextToken as a feature keyword for the semantic search
           FeatureList.add(str.nextToken);
         } //end if
     }//end while
```

7.1.1.2 Implementation of the Semantic Search Module

After the list of feature keywords are obtained, they are disambiguated and expanded to other relevant concepts using WordNet. Therefore, a set of relevant keywords is obtained. All relevant keywords are transformed to form a SPARQL query using the `SparqlTransform()` module. To perform a semantic search, the information in the knowledge base which is stored in RDF format needs to be read and stored in memory using the `FileManager()` function. Then, the SPARQL query can be executed against the knowledge base. All relevant information is retrieved using the `QueryExecutionFactory()`.Thus, the results can be ranked according to the degree of relevance and the results presented to users using the `ResultRanking()`.

The semantic search algorithm is illustrated in Algorithm 6.

```
Algorithm 6: Semantic search algorithm (SemanticSearch)
Input: A list of feature keywords
Output: All relevant images
  1. Disambiguate word senses and expansion using WordNet
         Perform Algorithm 3 to disambiguate word sense to get a list similarity
     words {W};
  2. SPARQL transformation
        Sparql = SparqlTransform(W);
  3. Read information from the knowledge base
     in = FileManager.get().open("sport.rdf");
  4. //Allocate space in memory to store the knowledge model
     Model  model = ModelFactory.createDefaultModel();
     //read RDF file into model
     model.read (in," ");
  5. Execute the SPARQL query against the knowledge base
     QueryExecution qe = QueryExecutionFactory.create(Sparql, model);
     //Store results in a variable "results"
     ResultSet results = qe.execSelect();
  6. Result ranking
     If  results <> null
        //Perform similarity calculation
        scores = sim(W[i],results[i]);
        //Rank the result by descending order
        RankedResult = ResultRanking(scores);
     Else
        LSISearch(W);
```

7.1.1.3 Implementation of the LSI Search Algorithm

If a semantic search fails to find any relevant information in the knowledge base, the LSI search engine will be activated to search all relevant data from the LSI document indexing database using `LSISearch`() in order to supplement the failed semantic search. The LSI search method is described in Chapter 6 (section 6.3.1, p.101) and is shown in Algorithm 7. The algorithms so far deal with NL text type queries. In this prototype, however, users can also query using visual data, a so-called visual query using a QBE method. Users start by browsing through images in order to locate an image which will be used as a query. Then, the system will extract SIFT low-level features and generate visual words as described in Chapter 5 (section 5.3.1 p.58). The generated visual words are further processed in order to translate visual content into high-level semantics. Therefore, in the subsequent sections, the algorithms for dealing with a visual query will be illustrated.

Algorithm 7: LSI search algorithm (`LSISearch`)

```
Input: A list of feature keywords {W}
Output: All relevant images which have relevant score more than a threshold
  1. Calculate term weight of all feature keywords as show in Equation (19)
       1.1 Local weight calculation as shown in Equation (20)
       1.2 Global weight calculation as shown in Equation (21)
       1.3 Normalization calculation as shown in Equation (22)
  2. Read all data from LSI database (store all keywords with their weights) {L}
       L= exec('Select * from LSI');
  3. Compute the similarity between {$W_i$, $L_i$}
       scores = sim(Wi, Li)
  4. Rank the result by descending order
       RankedResult = ResultRanking(scores);
  5. Select all results which have score more than zero and present those results to
     a user
       FinalResults = exec('Select * from RankedResult where score >= a
     threshold');
      Display(FinalResults);
```

7.1.1.4 Implementation of the Visual Word Construction Algorithm

An image selected by the user will be computed to find the keypoints in an image using the SIFT descriptor (`KeypointsDetection`). Then, all detected keypoints are clustered using the SLAC algorithm to form visual words (each cluster is a visual word).

Algorithm 8: Visual word construction algorithm (`VisualWordGeneration`)

```
Input: A query image (I)
Output: Sets of visual words
Detect keypoints {KP}
    KP = DoG(I); //DoG is used for keypoints detection
    SIFT= FeatureRepresentation (KP); // convert keypoints to the SIFT descriptors
Visual word construction {VW} using SLAC clustering algorithm
    VW = SLAC(SIFT); //as shown in Algorithm 1
```

7.1.1.5 Implementation of the Concept Matching Algorithm

As described in Chapter 5, the generated visual words will be matched to the concepts in the ontology model using concept range measurement method (Figure 5-7) in order to classify the query image. Before matching the concepts with visual words, the non-informative visual words need to be detected and removed to enhance the classifying ability and to reduce the computational cost. Having matched visual words with concepts, the query image is classified and a concept label of the query image will be used as a keyword for the semantic search process to find the relevant information in the KB.

Algorithm 9: Concept matching algorithm (`ConceptMatching`)

Input: Visual words

Output: High level semantic conceptualisation of an image

```
1. Non-informative visual word removal.
   1.1 Chi-square computation as shown in Equation (11)
   1.2 Weight the Chi-square value as shown in Equation (12)
   1.3 If Chi-square value (W) >= threshold Then
           FVW= W; // It is a feature visual word (FVW);
       Else
           NIF = W; //It is a non-information visual word (NIF);
2. Concept matching compares visual words in {FVW} with the centroid of each
   concept (obtain from training phase) in the ontology layer as show in Equation
   (2).
3. High level semantic annotation is performed based on the frequency of visual
   words in each concept.
   3.1 Normalise the visual word frequency as shown in Equation (15)
   3.2 For i=0 and i<= a number of images in the collection
           WHILE j <= a number of detected objects
               IF object (j) has a visual word frequency >= a threshold (chosen
                  experimentally)
                  The image contain object (i);
           j=j+1; i=i+1;
   3.3 Perform image classification using a Bayesian Network based upon all
       detected objects in image (i).
4. The classification label will be used as a keyword will extended to other
   relevant concepts using WordNet. Having been extended, all relevant
   keywords {K} are obtained.
5. All relevant keywords {K} are used for the semantic search process
```

7.2 Evaluation of Experimental Results

This section introduces and discusses the experimental results using sport image collections. The evaluation was conducted with the presented framework and implemented based upon the ideas introduced in this thesis.

7.2.1 Hypotheses to Evaluate

To evaluate the retrieval performance of the framework, some hypotheses were established in relation to the IMR requirements given in Chapter 3 (TABLE 3-1 and TABLE 3-2). The main hypotheses come from the survey of state of the art frameworks and the research objectives. They are described as follows.

7.2.1.1 Hypotheses for Visual Analysis Technique Evaluation

Hypothesis 1 (H1): *The SLAC clustering algorithm can generate better quality visual words compared to a conventional clustering algorithm e.g. k-mean algorithm since it takes the term weight and keypoints distance into account.*

A major drawback of the *k*-mean algorithm is that it ignores the spatial location of keypoints. This leads to a loss of semantic information between low level features and high level semantic objects in an image. In contrast, SLAC clusters information based on term weighting and spatial information (the distance between keypoints) in the form of a similarity matrix. The algorithm could improve the quality of visual words by preserving the semantic information of objects in the image. When the quality of visual words has been improved, they can represent the content of images effectively and enhance the sport image categorisation power.

Hypothesis 2 (H2): *Non-informative visual word elimination does not degrade the classification power. In contrast, it enhances the visual words approach to represent image content more concisely.*

Since some visual words generated from keypoints may contain a lot of noise, resulting in low-quality visual words that are not useful to represent the image, these should be detected and eliminated. Non-informative visual words are safe to remove. They improve the classification accuracy and also require less memory space. The remaining visual words are informative and distinctive, enhancing the categorisation power when representing images.

Hypothesis 3 (H3): *Restructuring the visual words space model using an ontology model can resolve the visual heterogeneity problem more effectively. Consequently, the retrieval performance is increased significantly.*

Combining multiple visual words to disambiguate word senses in state of the art systems is based on statistical computation which do not represent the actual semantic relationships between visual words. An ontology is an efficient model to resolve semantic type problems but is usually applied to textual information. This thesis exploits the ontology model to disambiguate word senses for visual words. However, one needs to evaluate its performance to perform such a task and compare this against competitor techniques.

7.2.1.2 Hypotheses for Knowledge Model Evaluation

Hypothesis 4 (H4): *The ontology model can handle the NL ambiguity problem.*

As was discussed in Chapter 3, Natural language may be ambiguous. An ontology-based image retrieval system should handle the vagueness in natural language effectively, e.g. ambiguity in user queries. The IMR system should interpret query keywords depending on their meaning rather than on relying on syntax matches between search terms and text captions terms.

Hypothesis 5 (H5): *The semantic model can find the indirectly relevant concepts which are not identified explicitly in an image description or caption.*

Indirectly relevant concepts are often not used for image retrieval. Even although some specific terms or concepts are not mentioned directly in image captions, it is possible that they are still semantically relevant to images. As listed in the IMR requirements (TABLE 3-1), the image retrieval system should be able to propose related concepts in an image even when they are not part of text captions.

Hypothesis 6 (H6): *An ontology-based search integrated with LSI can be useful even when an ontology-based knowledge model is incomplete. LSI enables the system to deal with unknown terms in queries.*

Ontology incompleteness is not handled by most existing systems. It is often difficult to build a complete and appropriate ontology covering different applications in a domain in one phase of development. Often an effective ontology requires some deployment and a community of practice. When the information in the KB is incomplete, precision and recall are affected. An intuitive solution is to supplement the results from ontology KB search engine (SBIR) with the results of a LSI-based search engine. LSI is used to handle metadata in a query that is unknown to the KB by using a statistical calculation to index metadata extracted from the corpus documents. This method allows the search mechanism to search for information beyond the data stored in the KB.

7.2.2 Retrieval Performance Measurements

The two classical measures used to evaluate the performance of information retrieval systems are *precision* and *recall*. Precision is defined as the number of relevant documents retrieved, divided by the total number of documents retrieved by that search. Recall is defined as the number of relevant documents retrieved, divided by the total number of existing relevant documents (which should have been retrieved). Let *A* denote all relevant images (as specified in a user query) in the image collection. Let *B* denote the retrieved images which the system returns for the user query.

- *Precision* is defined as the portion of relevant images in the retrieved image set, i.e.

$$Precision = \frac{|A \cap B|}{|B|} \tag{25}$$

- *Recall* is defined as the portion of relevant images that were returned by the system and all relevant images in the collection, i.e.

$$Recall = \frac{|A \cap B|}{|A|} \tag{26}$$

Using precision-recall pairs, a so-called precision-recall diagram, can be drawn that shows the precision values at different recall levels. Here, the retrieval performance is reported using the 11-point Interpolated Average Precision graph (Manning et al. 2008). The interpolated precision P_{interp} at a certain recall level is *r* defined as the highest precision found for any recall level $r' \geq r$:

$$P_{interp}(r) = max(r'\ ), \quad r' \geq r \tag{27}$$

Cosine similarity is used to measure the similarity between the queried and the stored visual content in a collection.

7.3 Evaluation Methodology

In this section, the methodology to evaluate the system will be described, including the selection of the test collection and user queries.

7.3.1 Selecting a Test Collection

Standard test collections exist that provide a "golden standard" to evaluate the retrieval performance of IMRs. However, sport images are not readily available in standard test collections. Therefore, a new test collection needs to be designed. It was decided that images from the Olympic organisation website[16] and the Google image search engine[17] should form the basis for the test collection to provide domains where a test ontology[18] is developed. The image collection contains 2,000 images of five sport genres (high jump, long jump, javelin throw, hammer throw and pole vault). The image collection is divided into two groups, a training set containing 500 images (100 image from each group) and a testing set with 1,500 images (300 images from each group).

Once a test collection has been defined, the set of all relevant images (needed for recall calculation) needs to be determined. To do this use is made of the Olympic website that has categorised the sport genres with respect to their concepts. This categorisation enables the relevant images for a specific topic to be identified without scanning all images in an entire collection.

[16] Olympic Organisation (http://www.olympic.org)

[17] Google image search (http://images.google.co.uk)

[18] Since Olympic website constantly changes, it is important to note that I accessed and collected data from Olympic website in March, 2007.

7.3.2 Selecting the User Queries to Evaluate

Besides the test collection issue, another major issue is the selection of user queries. Here are the main criteria used to define the user queries:

- User queries should test all hypotheses in Section 7.2.1 e.g. using NL query or using synonym words (indirect words) in the query.
- User queries should relate to many entities in the ontology in order to investigate the retrieval performance.
- User queries should be able to test the reliability and stability of the system. For example, querying information which is not contained in the ontology in order to study how the presented system behaves when it cannot find the relevant data in its knowledge base.

From the requirements above, these are the user queries which will be used to test and evaluate the hypotheses mentioned previously:

Query 1 (Q1): Find all images of any athletes with a specific nationality who compete at a specific host country e.g. USA athletes participate the Olympic game in Australia. This query is used for NL ambiguity testing. In a data-driven IMR, the system cannot distinguish between “USA athletes participate in the Olympic games in Australia” versus “Australia athletes participate in the Olympic Games in USA”. This query challenges the presented system to disambiguate the user query which is in the form of NL. Therefore, a knowledge-based IMR is expected to return images which contain USA athletes which participate in Australia rather than images of Australia athletes which participate in USA.

Query 2 (Q2): Find all images of a sport genre by using its indirectly relevant keywords e.g. “field event”. A field event usually refers to all the kinds of sports that athletes perform in the field e.g. pole vault, javelin throw and hammer throw. This query aims to test the H5 hypothesis. Image descriptions in the collection usually only mention the sport name. They do not refer to the super-concept (a higher-level concept in the ontology model) of the particular sport. Therefore, the proposed framework should recognise these indirect concepts automatically.

Query 3 (Q3): Find all images for which information is not contained in the ontology. This query aims to test the H6 hypothesis. Although it is difficult to build an ontology to cover all information in a domain of interest, the proposed IMR should tolerate the incompleteness of ontologies because it is designed as an open KB system. Therefore, it is able to find relevant images using an alternative search scheme.

7.4 Evaluation Results

This section presents the results of the evaluation. A collection of sport images from the Olympic organisation website and the Google image search engine was assembled for testing. This collection is different from the set used for training. The evaluation has two main parts: the *visual content representation* and the *image search engine*. Visual content representation evaluation aims to evaluate the image representation technique using BVW described in Chapter 5 whereas the retrieval performance is tested by the image search engine part using precision and recall.

7.4.1 Visual Content Representation Evaluation

All Images in the collection are processed to extract keypoints. Images from a training set produces a total of 328,653 keypoints and images from the testing set generates 853,276 keypoints by the DoG algorithm. Later, these keypoints are converted into SIFT descriptors and a SLAC clustering algorithm is utilised for clustering the visual words based on the extracted keypoints.

7.4.1.1 Evaluation of Categorisation Algorithms

First, the performance of classification algorithms needs to be investigated since different classifiers suit to different data. Typically, a *k*-mean clustering algorithm is used to generate visual words in the existing system. The major drawback of *k*-mean is that users need to specify the number of visual words (*k* value) which will be produced by a *k*-mean algorithm. Choosing the right vocabulary size involves a trade-off between discrimination and generalisation. Using a small vocabulary, a visual-word feature is not very discriminatory because dissimilar keypoints can map to the same visual word. As the vocabulary size increases, the feature becomes more discriminatory but also becomes less generalisable and tolerant of noise, since similar keypoints can map to different visual words. Using a large vocabulary increases the cost associated with clustering keypoints, computing visual-word features, and running supervised classifiers. Accordingly, it is not clear how many visual words should optimally represent the visual content. In contrast to a *k*-mean clustering algorithm, the SLAC clustering algorithm does not need to specify the number of visual words. It clusters keypoints based on their weight and distance. Therefore, the number of visual words produced by the SLAC is likely to be suitable to represent visual content.

To evaluate the classifiers, the average precision value was calculated for three different classifier algorithms[19]: Naïve Bayes, SVM-Linear, and SVM-RBF. These are used to cluster the same set of visual words produced by SLAC but using different term weighting schemes. Since term weighting is a key technique in information retrieval, its use was investigated for visual-word feature representation. Two major factors that affect the term weighting are *tf* (term frequency) and *idf* (inverse document frequency). A third factor is *normalisation*, which converts the feature into a unit-length vector to eliminate the difference between short and long documents (small and big images' size). Some popular term weighting schemes in IR are then applied to the visual-word feature vectors. These schemes are summarised in TABLE 7-1.

TABLE 7-1: Weighting schemes for visual-word feature (t_i)

Name	Factors	Value for t_i
BIN	*binary*	1 if t_i is present, 0 if not
TFY	*Tf*	tf_i
TFN	*tf, normalisation*	$\frac{tf_i}{\sum_i tf_i}$
TFI	*tf, idf*	$tf_i \cdot \log(N / n_i)$
TIN	*tf, idf,* *normalisation*	$\frac{tf_i \cdot \log(N / n_i)}{\sum_i tf_i \cdot \log(N / n_i)}$

[19] Three classification algorithms are exploited using the Weka framework (www.cs.waikato.ac.nz/ml/weka)

TABLE 7-2: A comparison of the classification performance based on using three classifiers for five sport genres and using visual words (including non-informative visual words)

Classifiers	Weighting schemes	Average Precision (AP)				
		High jump	Long jump	Pole vault	Javelin throw	Hammer throw
Naïve Bayes	BIN	0.451	0.581	0.581	0.547	0.528
	TFY	0.589	0.667	0.667	0.638	0.539
	TFN	0.585	0.656	0.666	0.641	0.592
	TFI	0.589	0.638	0.475	0.574	0.494
	TIN	0.589	0.664	0.573	0.617	0.591
SVM-Linear	BIN	0.567	0.586	0.475	0.572	0.486
	TFY	0.592	0.594	0.479	0.588	0.503
	TFN	***0.631***	***0.681***	***0.692***	***0.643***	0.594
	TFI	0.547	0.615	0.634	0.597	0.552
	TIN	0.612	0.668	0.685	0.628	0.564
SVM-RBF	BIN	0.534	0.631	0.594	0.582	0.498
	TFY	0.363	0.635	0.596	0.507	0.376
	TFN	0.363	0.462	0.671	0.581	***0.596***
	TFI	0.544	0.647	0.592	0.597	0.512
	TIN	0.624	0.677	0.422	0.574	0.486

The classification results from the five classifiers and the highest precision values for each sport category are printed in bold-italic face to highlight the best precision values. Based on the classification results in TABLE 7-2, the SVM-Linear classifier with the TFN weighting scheme is a good choice as it has produced the best performance in these experiments (except for the hammer throw event in which the SVM-RBF with TFN weight scheme provides a slightly better classification performance than SVM-Linear). Thus, these techniques will be applied to further experiments in subsequent sections.

Typically, SVM-RBF often obtains better results than SVM-linear. In this experiment, however, SVM-linear provides a better result than the SVM-RBF classification. From the data analysis, when the features in a vector space are very large, SVM-linear often obtains better results than SVM-RBF since mapping to a high dimensional space does not improve the performance. This is the main reason why the results of SVM-linear in this experiment are superior to the SVM-RBF results. This difference has also been reported by Hsu *et al.* (2010).

7.4.1.2 Evaluation of the SLAC Clustering Algorithm

The H1 hypothesis states that the SLAC clustering algorithm can generate better quality visual words compared to a simple clustering algorithm (e.g. *k*-mean algorithm) since it takes both the weight and keypoints distance into account. To evaluate this hypothesis, the results for image classification using visual words generated from the SLAC and the *k*-mean clustering algorithm are compared. First, the extracted keypoints with the SLAC algorithm are clustered to obtain visual words. Then the *k* value in the *k*-mean algorithm is set to be equivalent to the number of clusters generated from SLAC. Therefore, both algorithms generate an equal number of visual words. Both algorithms are compared using three different categorisation algorithms in TABLE 7-2 with the TFN weighting scheme. The results of this comparison are shown in TABLE 7-3.

As the numerical values in TABLE 7-3 shown, the classification results using visual words generated from SLAC is clearly superior to those using the visual words generated from the *k*-mean algorithm. This figure indicates that SLAC produces better quality of visual words that preserve the semantic information between the keypoints, and higher level semantic information for visual content, because it clusters relevant keypoints using a similarity matrix calculated from term weighting which represents the inter-semantic relationships among terms and spatial information (distance between keypoints). As a result, the generated visual words represent the visual content more effectively and enhance the categorisation power. In contrast, *k*-mean ignores the spatial information between keypoints and this leads to a loss of semantic information between low level features and high level semantic objects in the visual content. Consequently, the visual keywords cannot represent the semantic image properly. Thus, the H1 hypothesis is validated.

TABLE 7-3: The comparison of classification results between SLAC and k-mean clustering algorithm using three different classifiers with TFN weighting scheme

	High jump		Long jump		Pole vault	
Classifiers	SLAC	*K*-mean	SLAC	*K*-mean	SLAC	*K*-mean
Naïve Bayes	***0.385***	0.363	***0.462***	0.417	***0.666***	0.592
SVM-Linear	***0.631***	0.563	***0.681***	0.605	***0.692***	0.624
SVM-RBF	***0.585***	0.548	***0.656***	0.586	***0.671***	0.617

TABLE 7-3 (Cont.)

	Javelin throw		Hammer throw	
Classifiers	SLAC	*K*-mean	SLAC	*K*-mean
Naïve Bayes	***0.598***	0.571	***0.563***	0.504
SVM-Linear	***0.672***	0.605	***0.641***	0.557
SVM-RBF	***0.631***	0.582	***0.611***	0.549

TABLE 7-4: The proportion of visual words (ϖ) using SLAC clustering before and after the non-informative visual word {ψ} removal

	{ϖ}	{ψ}	{ε}={ϖ}-{ψ}	Proportion (ψ/ϖ)
keypoints {γ}	5,978	843	5,135	14.10%

{ϖ} = visual words; {ψ} = non-informative visual words detected using the Chi-square model {ϖ}-{ψ} = informative visual words {ε}

7.4.1.3 Evaluation of Classification Performance after the Non-Informative Visual Word Removal

The proportion of visual words (ϖ) before and after removing non-information visual words {ψ} are compared. The number of visual words shown in TABLE 7-4 is from all categories and generated by a SLAC clustering algorithm. From TABLE 7-4, the number of ϖ generated after removing uninformative visual words is reduced by 14%, compared to the original number of ϖ. Consequently, this reduces the vector space model by several dimensions for the classification task. However, this elimination may affect the classification performance (TABLE 7-2). Thereafter, the effect of ψ removal on the classification performance needs to be studied. The classification results between visual words with uninformative visual words {ϖ+ψ}, visual words without uninformative visual words {ϖ-ψ} using the SVM-Linear algorithm with TFN weighting scheme are compared. The experiments are conducted using 20 differents queries and, then, all the precision value are averaged, to generate the so-called Average Precision (AP).

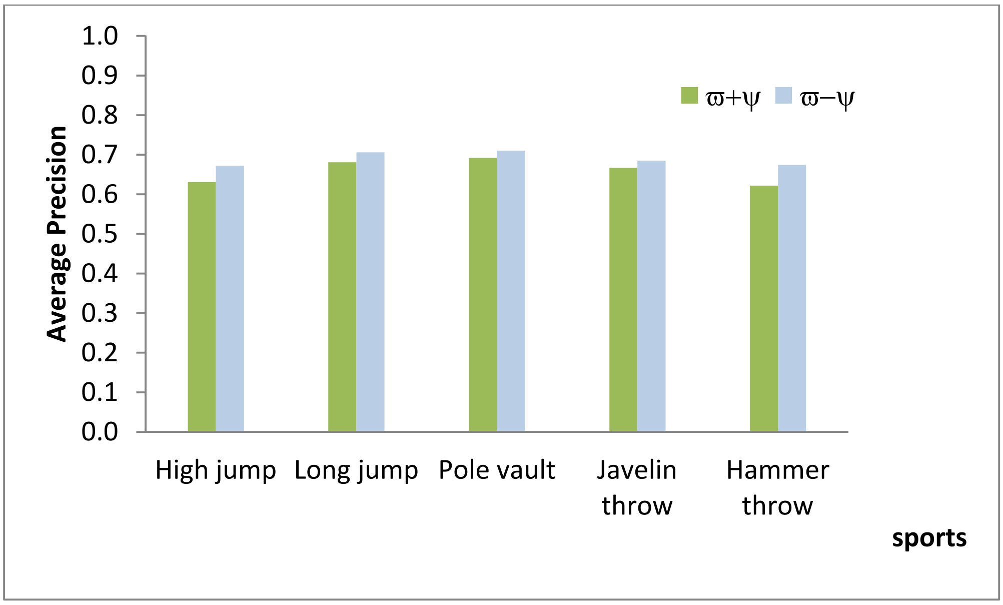


Figure 7-2 Classification performance comparison of the visual words (ϖ) before and after ψ removal using a SVM-Linear with TFN weighting scheme

Figure 7-2 shows that the classification accuracy is influenced by eliminating ψ. For example, in the pole vault event, the classification accuracy is increased from 69% to 71% when ψ is removed and the improvement of classification accuracy trends similarly in all other categories. This consistent improvement suggests that ψ removal does not cause a loss in classification accuracy but it is able to improve the classification performance. Although a large number of ϖ makes a feature become more discriminative, ϖ also makes the feature vector less generalisable and might contain more noise. These are major factors that degrade the classification performance of visual words before ψ removal. A large number of ϖ also increases the cost associated with clustering keypoints in computing visual-word features and in generating supervised classifiers. Hence, the H2 hypothesis is verified.

7.4.1.4 Statistical Significance Testing

The classification performance shown in Figure 7-2 seems to be not significantly improved because it is a very small difference (3% on average) between two methods. Testing for statistical significance is a mathematical calculation to evaluate whether or not this difference is significant. In several cases, differences between two methods are small but statistically significant due to the large sample size. When a sample size is large, very small differences will be detected as significant. This thesis contains 1,500 images for testing. Thus, 3% classification improvement could be significant. There are five steps for the statistical significance calculation.

1) State the research hypothesis: a research hypothesis states that the expected relationship between two variables (i.e. remove and non-remove ψ). Therefore, the research hypothesis in this evaluation is that "*the classification performance can be improved when non-informative visual words (ψ) are removed*".
2) State the Null hypothesis: a null hypothesis usually states that there is no relationship between two variables. Here, the null hypothesis is "*removing non-informative visual words (ψ) will provide the same classification performance as non-removing ψ*".
3) Select a probability of error level (alpha level): in social sciences, the alpha (*p*) = 0.05 is widely used among researchers. This means that a probability of 5% of making an error is accepted as the error level.
4) Interpret the results: If a T-test result is higher than the critical value from a probability table, the difference is significant and the null hypothesis will be rejected.

A T-test is calculated by comparing the average value on some variable obtained for two groups. This experiment containing 1,500 images for testing collection which are used to construct visual words (ϖ) and then remove non-informative visual words (ψ). Then, the image classification are performed based upon two groups (5 sport types) of data, ϖ+ψ and ϖ-ψ (Figure 7-2). To perform a T-test, the results are transformed into a table as shown in TABLE 7-5.

TABLE 7-5: T-test calculation

	ϖ+ψ	ϖ-ψ
Number of observations	5	5
Mean ($\bar{X}$)	0.6586	0.6894
Variance (*var*)	0.0009	0.0003

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{var_1}{n} + \frac{var_2}{n}}} \tag{28}$$

$$t = \frac{0.6894 - 0.6586}{\sqrt{\frac{0.0003}{5} + \frac{0.0009}{5}}} = 1.733$$

Degree of freedom (*df*) = number of observations-2 = 10-2=8. From the distribution table with *df*=8 and *p*=0.05, *t* must equal or exceeds 1.645. Therefore, this can conclude that there is a statistically significant probability that a relationship between the two variables exists and that this is not due to chance. The null hypothesis is rejected. This lends support to the research hypothesis and indicates that 3% is a significant improvement of the image categorisation performance when ψ is removed.

7.4.1.5 Evaluation of Image Classification Performance

A Bayesian Network technique was introduced in section 6.1.1 (p.93) in order to enhance the visual content classification and uncertainty management. In this section, the performance for visual classification using a hierarchical Bayesian network will be evaluated and compared with other categorisation frameworks.

In contrast to the Naive Bayesian classifier used in TABLE 7-2, the classification in this section is performed using the detected objects in images whereas the classification in TABLE 7-2 is performed based on the extracted visual words.

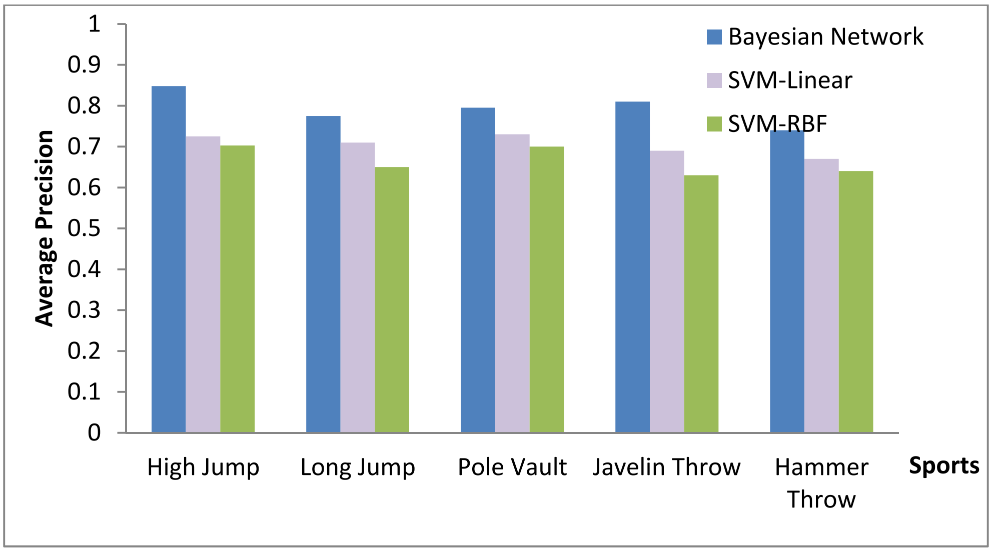


Figure 7-3 Object-based classification performance comparison between a Bayesian Network, SVM-Linear and SVM-RBF

The classification results in Figure 7-3 indicate that the Bayesian network can improve the classification power compared to SVM-Linear and SVM-RBF. This is because the Bayesian network classifies data based on the hierarchical structure. The structural model addresses the relationship between the key objects and all possible concepts of sports explicitly. In addition, it exploits the conditional probability to cope with the uncertainty which may occur during the classification process. This probability aids the classifier in making a decision about which category (sport genre) an image should be in when an uncertainty occurs, e.g. when the underlying objects are missing. As such, this mechanism leads the proposed system to obtain a higher classification performance than the other two methods. Since SVM-Linear and SVM-RBF performs categorisation based purely upon a statistical calculation, they do not have a mechanism to deal with the uncertainty. Therefore, they obtain a lower classification performance than the Bayesian Network.

Among all sports, the hammer throw event obtains the lowest classification accuracy, about 74%. From analysing the classification data, the main cause is that the system cannot detect a hammer object in the hammer throw images. From the observation of images in the test collection, in several cases, a hammer object is very small and is merged into the image background. Hence, the keypoints of a hammer object are very difficult to detect and extract (Figure 7-4) because there are a lot of noises from the background. Consequently, a hammer object cannot be detected because the generated visual words are different from the visual words generated in the training phase. When the system cannot detect a hammer object in an image, it may misclassify a hammer throw image as other sport e.g. a long jump using the probability table. Therefore, this also decreases the classification accuracy of long jump images.

Although the Bayesian network can significantly enhance the categorisation accuracy, its drawback is the computational expensive compared to SVM-Linear and SVM-RBF. This is because it needs to compute the conditional probability of every image that needs to be classified and to update the conditional probability table after finishing each classification. As a result, it spends more time in categorising images compared to the other two methods.

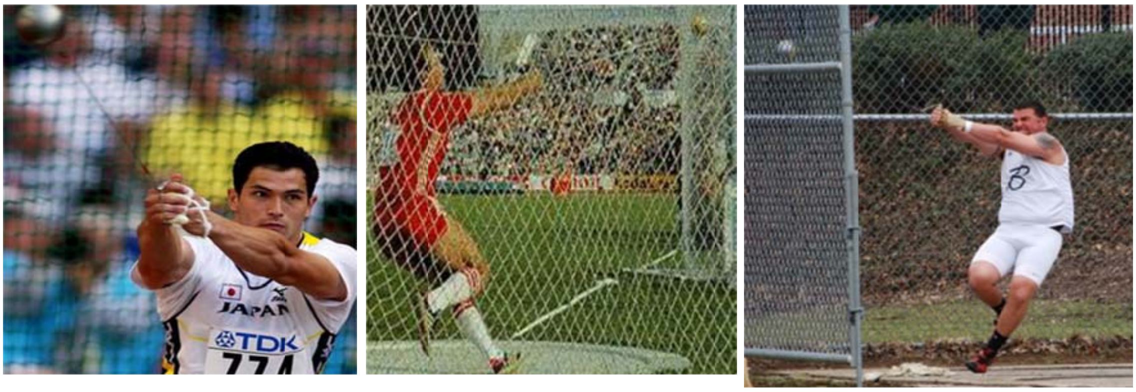


Figure 7-4 Examples of hammer throw images which a hammer object is merged with the background.

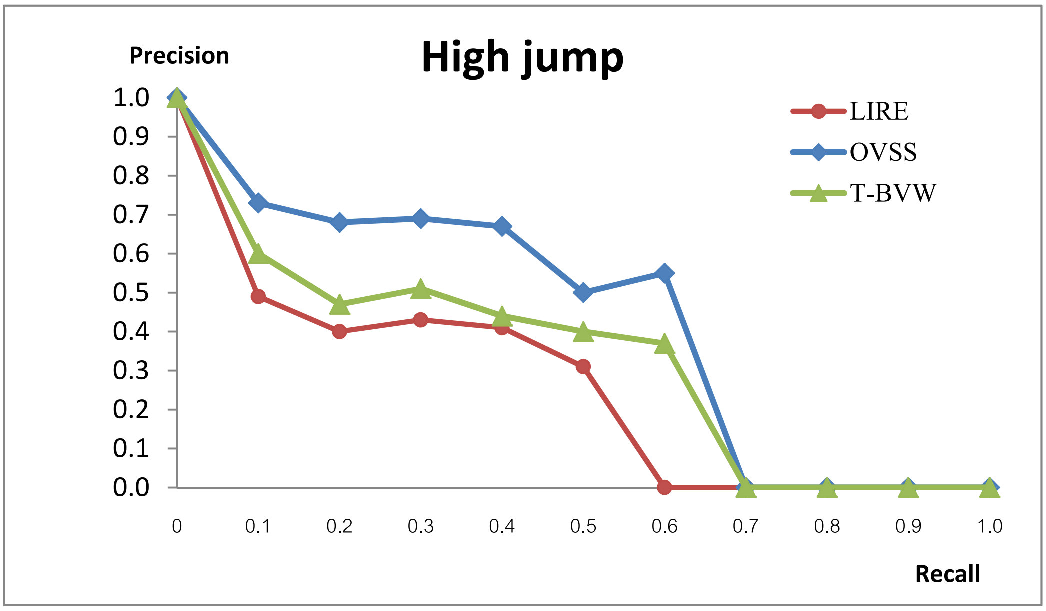
7.4.1.6 Evaluation of Image Retrieval Performance using the Visual Query

H3 states that restructuring the visual words space model to an ontology model can resolve the visual heterogeneity problem more effectively. Consequently, the retrieval performance is significantly increased. To evaluate this hypothesis, the retrieval performance (precision-recall graph) between the proposed method, the so-called Ontology-based Visual Semantic Search, (OVSS), a traditional bag-of-visual word model[20] (T-BVW) and a traditional CBIR (LIRE framework[21]) are compared. The search algorithm for the proposed technique described in section 5.6.1.2 (Visual query) and the experimental results are shown in Figure 7-5. The results show that the retrieval performance is affected by the proposed technique. Since the OVSS technique analyses an image query and interprets it into a high-level conceptualisation, the search engine is able to perform conceptual searching rather than a simple low-level feature matching. As a consequence, more relevant images can be recognised and retrieved. This leads to the OVSS technique obtaining the highest precision and recall compared to other techniques. The content-based search (LIRE) retrieves all images which have similar low-level features. Unfortunately, some of them are not semantically relevant to the image query. As a result LIRE, obtains a lower precision and recall compared to other techniques. T-BVW attains a better retrieval performance compared to LIRE because the use of associative visual words efficiently distinguishes visual content more than the colour and texture features used in LIRE. However, it obtains a lower precision and recall than the OVSS method because the T-BVW model represents visual content based on the feature space model whereas OVSS deploys a hierarchical model which expresses visual content more explicitly than the feature space model. This structured model is able to disambiguate visual word senses more effectively. As a result, this structural model enhances the visual content interpretation and the retrieval performance.

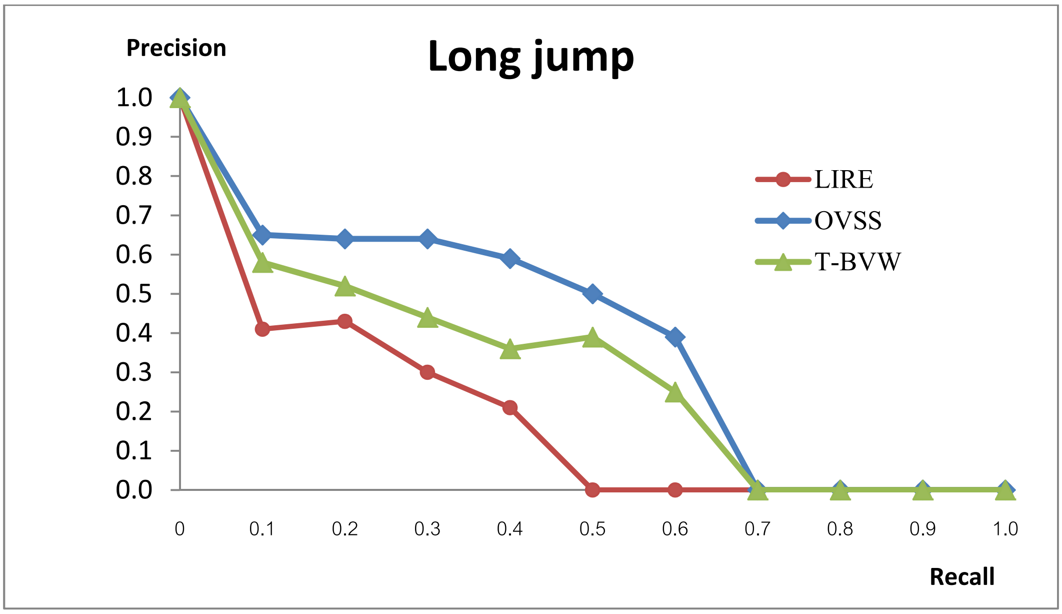
The use of the ontology model incorporated with the BVW model allows the systems to recognise all semantically similar images even when their visual appearance is different. In other words, the proposed technique is highly invariant to visual appearance. Therefore, the H3 hypothesis is successfully evaluated.

[20] A traditional BVW model is constructed from visual words generated by the SLAC algorithm and removes noisy visual words. The remaining visual words are used to produce a “Bag of Visual Words” that is represented as a vector (histogram).

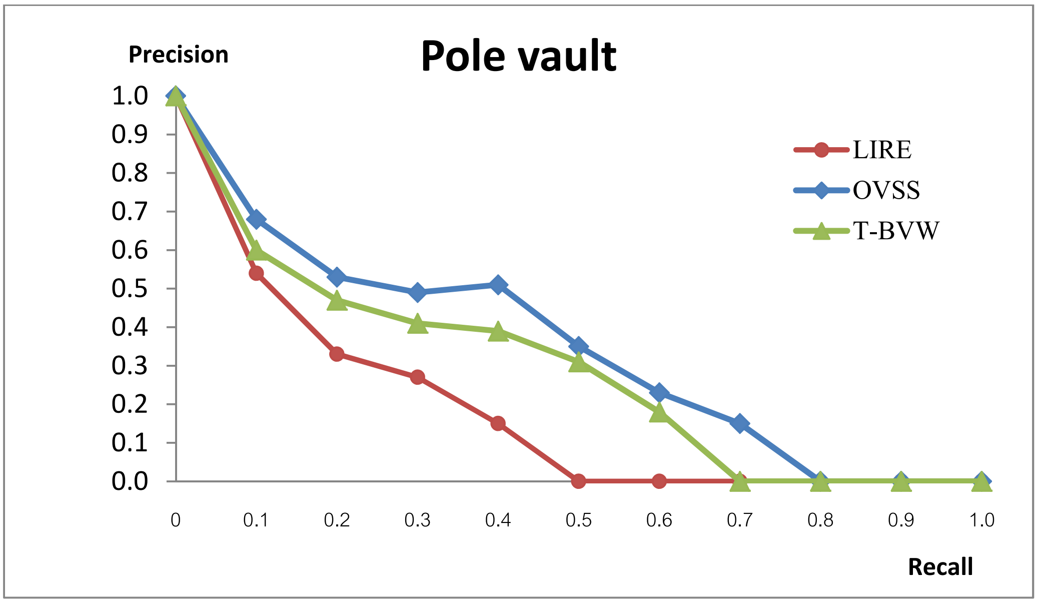
[21] Lucene Image Retrieval framework (http://www.semanticmetadata.net/lire)


High jump
Precision
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Recall
LIRE
OVSS
T-BVW

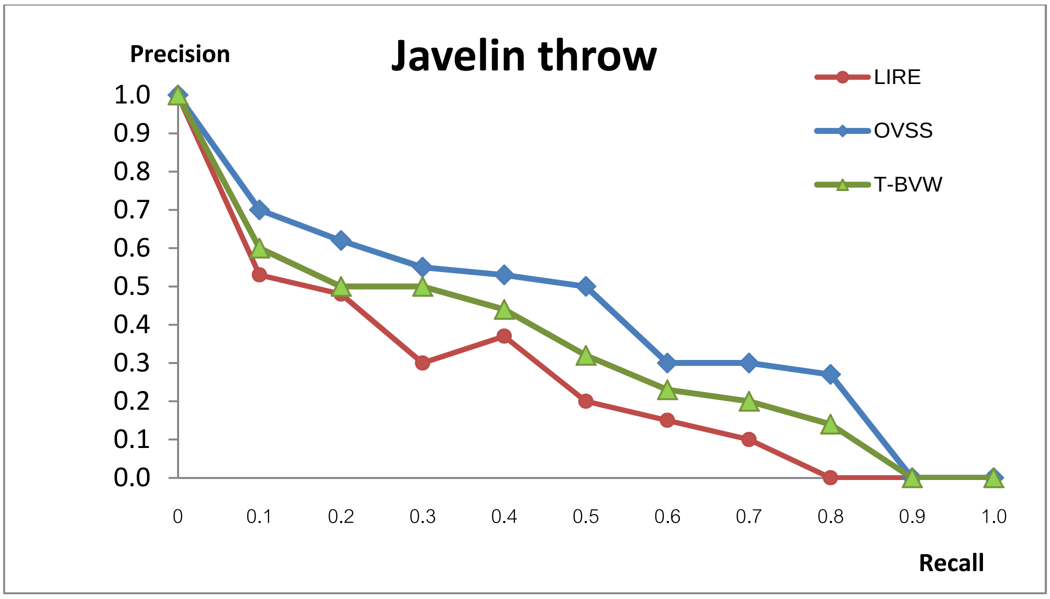
(a)


Long jump
Precision
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Recall
LIRE
OVSS
T-BVW

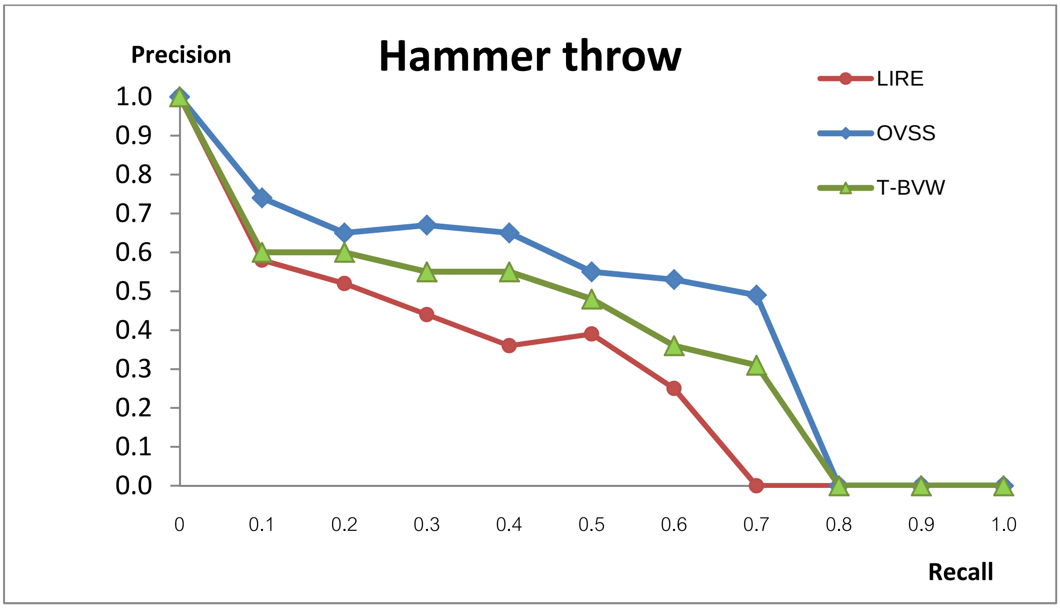
(b)


Pole vault
Precision
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Recall
LIRE
OVSS
T-BVW

(c)


Javelin throw
Precision
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0.0
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Recall
LIRE
OVSS
T-BVW

(d)



(e)

Figure 7-5 Retrieval performance results comparison of OVSS (the presented method), BVW and LIRE (the CBIR framework)

7.4.2 Image Retrieval Performance Evaluation using Text Queries

In this section, the retrieval performance is evaluated using textual queries and then compared the retrieval performance between SBIR and the Lucene[22] full-featured text search engine is compared. The test environment was implemented as described in Chapter 4 (Section 4.2 p.46).

7.4.2.1 Evaluation of the Ability to Disambiguate User Queries Using the Ontology Model

H4 states that the ontology model can handle the NL ambiguity problem. If the structure of ontology is designed appropriately, it is able to disambiguate the vagueness of the user query without exploiting any external knowledge (e.g. WordNet). To evaluate this hypothesis, Q1 is used to evaluate the retrieval performance. Figure 7-6 shows, the SBIR search is superior to Lucene. The SBIR

[22] Lucene full text search engine (http://lucene.apache.org)

supports the expression of more precise information, leading to more accurate results. In Lucene, it is not possible to distinguish a query "USA athletes participate in Australia" versus "Australia athletes participate in USA". The search engine of Lucene will retrieve all documents containing words USA, athlete, and Australia. Therefore, it retrieves a high number of irrelevant images including Australia athletes who participate in the Olympic Games in USA which leads to a low precision and recall.

In contrast, the opposite is possible with the SBIR search using a SPARQL query. In a SPARQL query, the relationship between two concepts can be explicitly specified. Thus, all athlete instances and semantic relationships which are matched to the SPARQL query will be retrieved. The SPARQL query for Q1 is shown in Figure 7-7. This SPARQL query ignores the "Australia athletes" *-participate* - "USA" relationship and other relationships which are not expressed in the query. This mechanism significantly improves precision and recall compared to Lucene. Hence, the H4 hypothesis is successfully validated.

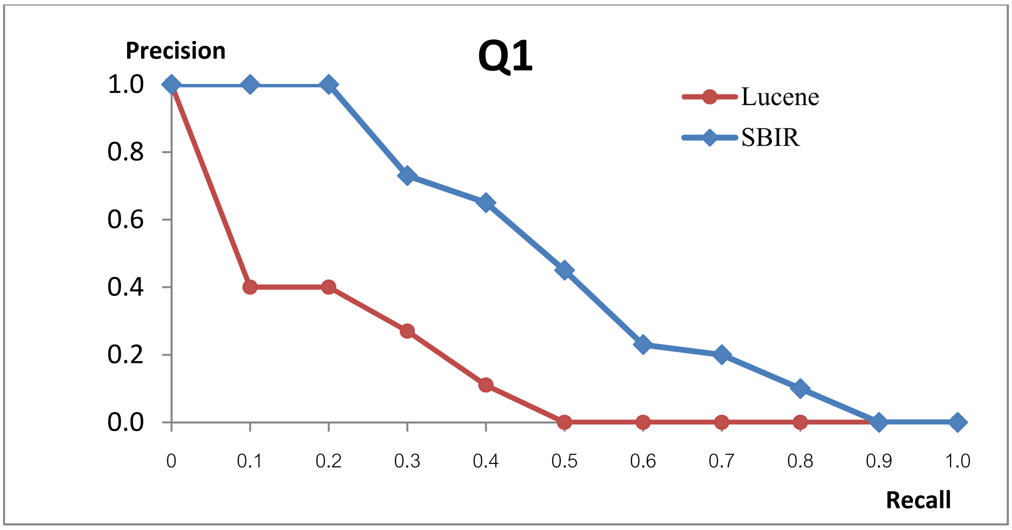


Figure 7-6 Retrieval results of Q1 for SBIR and Lucene text-based search

```
SELECT ?photoID, ?photoPath
WHERE {?photo sport:hasAthlete ?athlete.
       ?athlete sport:hasNationality ?nationality
       ?athlete sport:paticipateIn ?hostCountry
FILTER (regex(?nationality, “USA”,”i”) &&
      regex(?hostCountry, “Australia”,”i”)}
```

Figure 7-7 SPARQL query for Q1 which is able to disambiguate user queries using explicit relationships between concepts

7.4.2.2 Evaluation of Finding the Indirect Relevant Concepts

H5 states that the semantic model can find the indirectly relevant concepts which are not identified explicitly in the text caption. To evaluate this hypothesis, Q2 is specified using some keywords which do not appear directly in a text caption. The experimental results are shown in Figure 7-8.

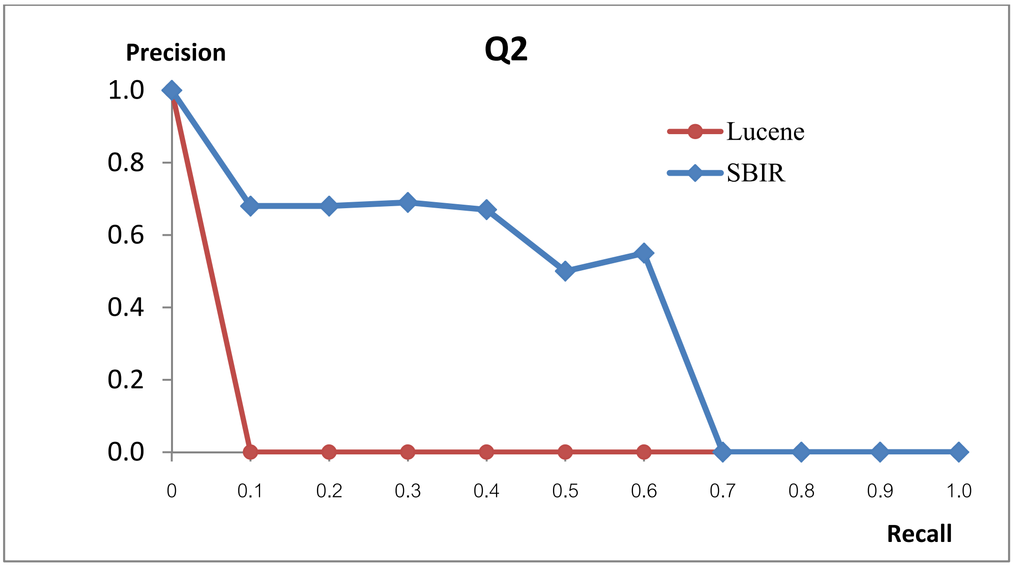


Figure 7-8 Retrieval results using unknown keywords which does not appear in text caption

In this example, a user wants to find images of a “field event” (e.g. the pole vault and high jump). As shown in Figure 7-8, the knowledge-base defines semantic relationships in terms of sub-concepts of the Sport(Athletics) concept such as field event, track event and road event. The proposed system is thus able to recognise all images annotated with a sport name which belong to the "field event" concept although word "field event" does not appear in any text captions whereas Lucene fails to retrieve any relevant documents since the keyword of the query does not appear in any text captions. Therefore, SBIR obtains a higher precision and recall than the text-based approach. In addition, although the ontology model does not contain the "field event" concept, the system is still able to recognise those sports that are relevant to the "field event" sport using WordNet (Algorithm 3). In summary, the semantic model improves the retrieval performance significantly and hence this confirms the H5 hypothesis.

7.4.2.3 Evaluation of Retrieval Results when Information in the KB is Incomplete

H6 states that an ontology-based search integrated with LSI can be useful even when an ontology-based knowledge model is incomplete. LSI enables the system to deal with unknown query and metadata properly. To evaluate this hypothesis, some metadata for images in the ontology is deleted resulting in incompleteness in the ontology when unknown data is present in text captions, e.g. metadata for high jump images for the Beijing 2008 Olympic games was deleted, then a query "*Find all high jump images for Beijing 2008*" is submitted to the system. Figure 7-9 illustrates an example of the search results using LSI when the information concerning Beijing 2008 is deleted. LSI is able to recognise all relevant images for Beijing 2008 even though an image caption does not contain the word "Beijing" e.g. the third image in Figure 7-9. This is because the system uses co-occurrence information from LSI to find all relevant images. In this example, Beijing, 2008, and China have a high co-occurrence value since they usually appear together in other image captions in the collection. Therefore, LSI can recognise that the third image is also relevant to the query even though there is no word "Beijing" in the text captions. This cannot be achieved by using a simple text-based search engine.

Figure 7-10 shows that the retrieval performance of LSI and Lucene search is not that different. This is because SBIR could not find any images related to high jump images for the Beijing 2008. This triggers the LSI to be activated and the results of LSI are used instead of the results of SBIR. LSI computes the similarity of terms in a user query and its indexing data.

However, LSI still obtain a better performance compared to the text-based (Lucene) search because LSI can perform a semantic search which results in it finding implicit relationships between keywords and images using its matrix containing term frequency information. Consequently, the LSI algorithm provides better results than a text-based search.

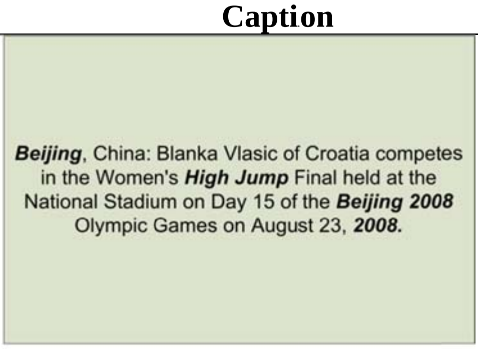
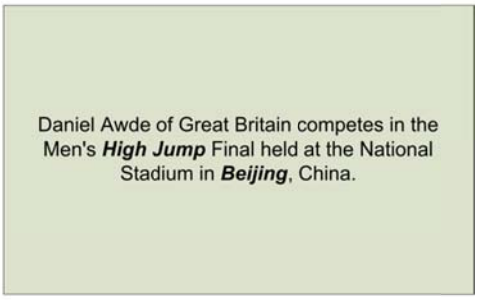
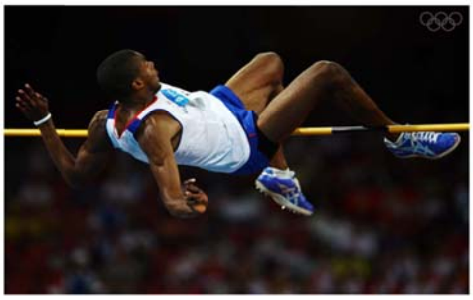
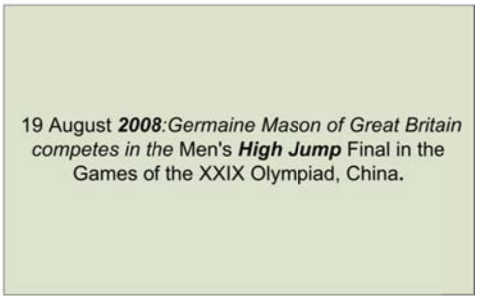
Image	Caption
	 ***Beijing***, China: Blanka Vlasic of Croatia competes in the Women's ***High Jump*** Final held at the National Stadium on Day 15 of the ***Beijing 2008*** Olympic Games on August 23, ***2008.***
	 Daniel Awde of Great Britain competes in the Men's ***High Jump*** Final held at the National Stadium in ***Beijing***, China.
	 19 August ***2008***:*Germaine Mason of Great Britain competes in the* Men's ***High Jump*** Final in the Games of the XXIX Olympiad, China.

Figure 7-9 Example of LSI results when the KB does not contain the Beijing 2008 information

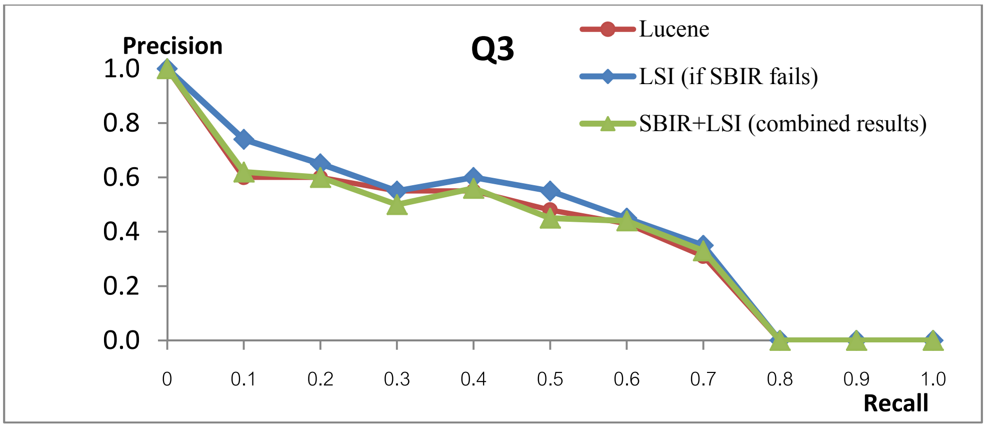


Figure 7-10 Retrieval performance when an ontology does not contain some information required by users

Results compensation from LSI was used instead of combining results from LSI and SBIR because the result combination affects the efficiency of the retrieval results. Additional experiments demonstrate that when the results from SBIR are combined with the results from using LSI, the retrieval performance is significantly reduced in terms of precision and recall. A probable reason for this is that the results from SBIR are significantly better compared to the results of LSI. When they are combined, the average precision and recall are affected. Hence, SBIR should be used as a supplementary technique as and where necessary so that it can provide a high quality of retrieval results instead of combining the results with other search algorithms (e.g. LSI). Whenever SBIR cannot find the relevant information in the KB, the LSI should be activated and its results should be used instead of SBIR search results.

An incomplete KB in this experiment is considered as the same as an open KB model described in section 5.1.2 (p.51) because it can handle *unknown* data better than a closed KB model. A closed KB model is much stricter and provides the answers based only upon the metadata that is contained in its KB. If the desired metadata is not present, it will not try another method. In contrast, an open KB model is more useful because it is able to try a second method for searching relevant information in the corpus of documents if an unknown data is present. Likewise, if the KB in the presented system is incomplete, the search mechanism will try a second search through LSI instead of SBIR. The search mechanism in this framework is much more efficient than the text-based search engine i.e. Lucene as it is able to perform semantic search using the ontology KB model. In addition, it is able to cope with uncertainty of the KB and still perform a semantic search when unknown data is present through using LSI. From the results shown in Figure 7-10, the proposed framework provides a better retrieval performance than other methods even if the KB is incomplete, hence H6 is validated.

7.5 Summary

This chapter evaluated the main ideas proposed in Chapter 5 and 6. Before testing the proposed system, some hypotheses are established. Evaluations of the system are conducted against these hypotheses. Next, the measurements for evaluating the retrieval performance are introduced. The retrieval effectiveness was measured in this work using the classical IR measures--precision, recall, and 11-point Interpolated Average Precision graph. These measurements are used as standard measurements to compare the retrieval performance of the proposed system and some other state of the art frameworks.

Since there is no available standard test collection for sport images, a new test collection needs to be established. Therefore, 2,000 images collected from the Olympic organisation website and the Google image search engine are used as the basis for the test collection (training and testing set). The experimental results are compared Lucene (full text-based search engine) and LIRE (CBIR). Based on the experimental results, it is shown that the proposed framework can enhance both typical CBIR and text-based search engine. Of particular interest is that it can provide a good retrieval performance, even though information in the KB is incomplete, by exploiting a LSI algorithm. LSI governs the term frequency information which is not present in the ontology KB. Another advantage of the proposed framework is that it can classify the visual content when a text caption is not supplied. This uncertainty handling is important to make the system more reliable. In addition, the proposed framework exploits multi-semantic information from multimodal (visual and textual) information that improves the image retrieval performance significantly.

Chapter 8

Conclusions and Future Work

This chapter summarises the research presented in this thesis and discusses directions for future work. This thesis has addressed the core requirements and challenges concerned with acquiring and representing visual knowledge and classifying and retrieving visual data using the extracted knowledge. This thesis has proposed solutions to these challenges through integrating the visual data processing and linguistic processing within the same framework. This chapter contains a summary of the work done in this thesis, the key contributions, achievements, the limitations of the presented framework. Finally, directions for future work are presented.

8.1 Discussion, Achievements and Novelty

8.1.1 Achievements

The proposed framework has been developed as presented in chapters 5 and 6. The semantic model can resolve the vagueness of the natural language, support uncertainty during image interpretation, and represent the visual content used for image retrieval, more effectively. This framework could be a prototype for a semantic search which can handle both textual and visual information. The key contributions of this thesis have been accepted by 5 conferences, a journal and 1 book chapter as shown below.

Kesorn, K., & Poslad, S. (2008). Semantic Representation of Text Captions to Aid Sport Image Retrieval. In *IEEE Proceedings of the International Symposium on Intelligent Signal Processing and Communications Systems* (pp. 1-4).

Kesorn, K., & Poslad, S. (2008). Use of Semantic Enhancements to NLP of Image Captions to Aid Image Retrieval. In *IEEE Proceedings of the 3rd International Workshop on Semantic Media Adaptation and Personalization* (pp. 52-57).

Kesorn, K., & Poslad, S. (2009). Enhanced Sports Image Annotation and Retrieval Based Upon Semantic Analysis of Multimodal Cues. In *ACM SIG Multimedia Proceedings of the 3rd Pacific Rim Symposium on Advances in Image and Video Technology* (pp. 817-828).

Kesorn, K., Liang, Z., & Poslad, S. (2009). Use of Granularity and Coverage in a User Profile Model to Personalise Visual Content Retrieval. In *IEEE Proceedings of the 2nd International Conference on Advances in Human-Oriented and Personalized Mechanisms, Technologies, and Services* (pp. 79-84).

Chimlek S., Kesorn K., Piamsa-nga P., Poslad S., (2010) Semantically Similar Visual Words Discovery to Facilitate Visual Invariance. In *Proceedings of IEEE International Conference on Multimedia & Expo* (To appear).

Kesorn, K., & Poslad, S. (2009). Semantic Restructuring of Natural Language Image Captions to Enhance Image Retrieval. *Journal of Multimedia*, *4*(5), 284-297.

Liang, Z., Kesorn, K., & Poslad, S. (2010). The USHER System to Generate Personalised Spatial Maps for Travellers. In M. Wallace, I. Anagnostopoulos, P. Mylonas, & M. Bielikova (Eds.), *Semantics in Adaptive and Personalised Services: Methods, Tools and Applications*. Springer.

Among some of these conferences and journals, the acceptance rate for the oral presentation is highly competitive. In particular, only 15% of submitted papers were accepted at the IEEE International Conference on Multimedia & Expo (ICME 2010). For the Journal of Multimedia (JMM), only 5 papers were accepted for that issue. The paper presented at the 2nd International Conference on Advances in Human-Oriented and Personalized Mechanisms, Technologies, and Services, was awarded the best student paper award.

8.1.2 Novelty

This research introduces a framework to analyse both visual and textual information and to create a knowledge base for visual content representation that provides a semantic-based classification and retrieval solution for visual content. The novelty in this thesis can be separated into two parts, visual content analysis and textual information analysis.

8.1.2.1 Visual Content Analysis

This thesis has developed a technique for analysing, representing, classifying and retrieving visual data using a visual words model integrated with an ontology model (section 5.3.1, p.58). The main contributions of this part of the framework are as follows:

Semantic Content Visual Analysis and Representation

In this part of the framework, the visual data analysis and representation based on the Bag-of-visual words (BVW) technique was improved. The first key feature is that *the technique is able to discover semantically similar keypoints through applying the semantic local adaptive clustering (SLAC) algorithm*. Unlike other clustering algorithms, SLAC is an extension of traditional clustering, capturing the relevance of representative keypoints through exploiting a *semantic matrix* constructed from term weighting and spatial information (distance between keypoints). This method clusters the relevant keypoints more effectively than the simple *k*-mean clustering algorithm used in existing systems. Furthermore, the term weighting method is improved. Rather than using an inverse document frequency (*idf*) scheme, *a weighting measure based on local adaptive clustering (LAC) is exploited*. The main disadvantage of the *idf* weighting scheme is that it concerns only document frequency without taking the distance between terms into account. Consequently, the generated visual words are more robust and can more effectively represent the semantics of visual data.

To enhance the image classification performance, *an innovative technique to detect non-informative visual words*, which would otherwise ineffectively represent visual content and degrade the categorisation capability, has been proposed. A Chi-square statistical model is utilised to identify meaningless visual words. Two main characteristics of the useless visual words are considered using the Chi-square model: those with a high document frequency (DF); and those that have a small statistical correlation with all the concepts in the collection. These visual words will be discarded in order to enhance the discrimination power. As a result, the classification accuracy is significantly improved.

This contribution has been accepted and published in the proceedings of the IEEE International Conference on Multimedia & Expo 2010.

Semantic Visual Content Classification and Retrieval

Another novel technique is the idea to disambiguate visual words using *the concept range* (Equation (2) and to map these generated visual words to the ontology model. This technique is useful for image classification and for image retrieval so that this not only relies on visual similarity but also on conceptual similarity. In other words, the technique is able to resolve the synonymy (visual heterogeneity) and polysemy problem. The main advantage of the ontology model is it can disambiguate word senses more explicitly and effectively compared to statistical methods, e.g. Frequency Itemset Mining (FIM) as used in the state of the art frameworks. Each visual word can be disambiguated by comparing the range of concepts in the ontology model obtained in the training phase. This method has an advantage over existing methods because it allows a visual word to be assigned to multiple concepts as a range of concepts may overlap each other. Hence, this can effectively resolve the polysemy problem of a visual word effectively. In addition, the knowledge infrastructure uses an ontology for the annotation and interpretation of visual data, identifying the high level semantic concepts and leading to a narrowing of the semantic gap. These annotations can be further used during image retrieval to match the keywords in a user query.

A Bayesian Network is deployed to aid the image categorisation based upon the detected objects and conditional probability information. These can significantly improve the classification accuracy compared to SVM-Linear and SVM-RBF. In addition, the system is able to handle the uncertainty that can arise during classification. This is because a Bayesian network encodes the dependencies among data. It can handle the situation where some data are missing and this information can be used for prediction.

8.1.2.2 Textual Information Analysis

Aside from the visual analysis component, this thesis also proposes a technique for acquiring knowledge from text captions, for their knowledge representation and for knowledge-based visual content retrieval.

Knowledge-based Acquisition from Text Captions and Knowledge Representation

The main innovation for textual information analysis is to extract the essential metadata from text captions and transform the unstructured metadata into semantic concepts. Three main steps are defined in this thesis to complete this task (Figure 5-12). Then, the KB is encoded in RDF format to facilitate semantic retrieval. Since text captions do not provide all information needed by ontology, *two methods are proposed to handle two types of missing metadata*, *unambiguous and ambiguous metadata* (section 5.3.2.3, p.78). The semantic rules are defined in order to deal with unambiguous missing metadata. It is exploited in order to find the relevant information in the KB and to identify any missing information needed in the ontology model. To deal with ambiguous missing metadata, a Bayesian network is also deployed.

This contribution has been published in the ACM SIG Multimedia Proceedings of the 3rd Pacific Rim Symposium on Advances in Image and Video Technology and in the IEEE Proceedings of the International Symposium on Intelligent Signal Processing and Communications Systems.

Knowledge based Retrieval and Uncertainty management

Another key idea is that a hybrid technique that combines natural language and semantic restructuring can be used to enhance semantic retrieval. Since an ontology often cannot effectively cover all information in the domain of knowledge, *this thesis proposed the use of LSI technique* (Chapter 6, p.100) to complement the use of an ontology when the latter is incomplete. LSI is deployed for text caption analysis in parallel with the use of a NLP framework. This thesis improved one particular step of the LSI algorithm which is very important for heightening the indexing quality of LSI, *the tokenising step*. Unlike the existing tokenising process in LSI algorithms, NLP is also applied to this step in order to enhance the tokenising power and to obtain the tokenised words more correctly using the named entities recognition function e.g. the system should get “Great Britain” rather than “Great” and “Britain”. This is a minor step but crucial in raising the indexing quality and the retrieval performance of the LSI search engine. As a result, LSI can more effectively capture the semantically relevant documents and retrieve information with more accuracy than a conventional text-based search engine can. In addition, *the use of LSI as a second search for SBIR can fulfil the open KB issue in which it can handle unknown data more efficiently*.

This contribution has been published in proceedings of the 3rd International Workshop on Semantic Media Adaptation and Personalization (SMAP) and Journal of Multimedia.

8.2 Limitations of the Proposed Framework

In this thesis, techniques for a new image retrieval system are proposed that offer distinct advantages over existing text-based and content-based search engines. Since no single framework works perfectly, the limitations of the proposed framework are discussed as follows:

- First, the SLAC algorithm suffers from being computationally expensive. The running time of one iteration of the SLAC is $O(kND^2)$ whereas the time complexity of the k-mean algorithm is $O(kND)$, where k is the number of clusters, N is the number of visual contents, and D is the number of keypoints. One possible solution is the noisy keypoints could be detected and eliminated before generating visual words. Consequently, D is decreased, thus the computational cost is reduced.
- Second, non-informative visual words detected using Chi-square model are domain specific. Changing the training collection, one can obtain a different ordered list.

8.3 Future work

This section introduces some directions for future work. The issue of information overload for potentially millions of users, with a variety of interests, brings another challenge to the research field of information retrieval, *Personalised Information Retrieval*. Personalised information retrieval aims to improve the retrieval process by taking into account the particular interests of individual users (Vallet et al. 2007). The use of personalisation to relieve the information overload is envisioned as a major research area (Gonzalez 2008). This is because traditional information retrieval tools select the same content for different users in response to the same query, which may be barely related to some users' interests. Thus, users need powerful search tools in order to help them to retrieve images that are of interest to them. The use of a personalised image retrieval (PIMR) system is identified as a key step in order to cope with the variety of users and the continuous growth in the number of multimedia documents in the future. In order to design a PIMR system, the main challenges are summarised in TABLE 8-1.

TABLE 8-1: Summary of the PIMR system challenges

Challenges	Descriptions
Automatic user preference acquisition	Manual user profile creation is not possible for a large scale system with thousands of users; automatic user preference acquisition is more *scalable*
Dynamic capture of users' interests	Users' profiles are usually not static but vary with time and depend on the situation. Therefore, profiles should be automatically modified based on observations of users' actions. In other words, a *dynamic* system is needed to capture users' interest.
Richness of the semantic representation	User preferences should be represented in a richer, more precise, and less ambiguous way than in a keyword/text-based model (Chen & Williams 2008).

Challenges	Descriptions
Terminological heterogeneity	Naming differences can vary according to the linguistic representation. The concepts underlying such terms may be used differently by the different users at different levels of granularity and in different situations with divergent interpretations (Poslad & Zuo 2008). Systems that model user profiles should take this issue into account.

In the next section, an initial idea for the PIMR system is proposed to support those challenges in TABLE 8-1. The objective of this framework is to represent users' interests in a formal way, such that different user models can form customised views of, and can be checked to be valid, with respect to the global ontology based domain model, generating a flexible indexing structure for different users.

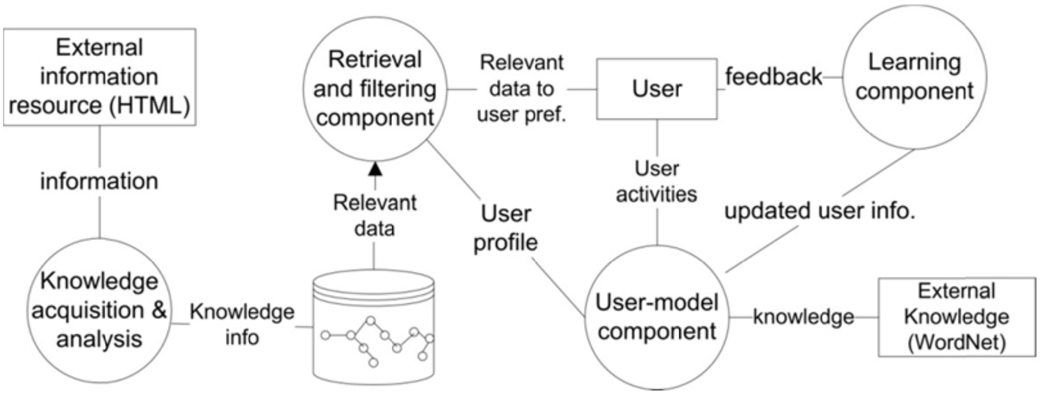

External information resource (HTML)
information
Knowledge acquisition & analysis
Knowledge info
Relevant data
Retrieval and filtering component
Relevant data to user pref.
User profile
User
User activities
feedback
Learning component
updated user info.
User-model component
knowledge
External Knowledge (WordNet)

Figure 8-1 An overview of a framework for PIMR

8.3.1 A Personalised Image Retrieval (PIMR) System

The PIMR system uses four basic components (Figure 8-1): knowledge acquisition and analysis, a user-model, machine learning; and retrieval and filtering. The details of each component are described as follows.

The knowledge acquisition and analysis component obtains or collects data from text-based visual content captions. The textual information is analysed and represented in an ontology representation language e.g. RDF or OWL.

The user-model component implicitly gathers information about users and their interests, and constructs user profiles. A hybrid method is used for acquiring knowledge about users that combines a statistic calculation and an ontology-based knowledge model. Ontology-based users' profiles are constructed from unstructured textual data in image captions accessed by users. An external lexical reference system, WordNet, and domain ontology are exploited in order to achieve a higher degree of automation. Because some users' interests may change over time, a system that relies only on a static user's profile, might become worse when the user changes his/her interests. For example, Bob is a fan of British swimming team but the British swimming team's performance is very poor and they are likely to lose competitive matches. Then, Bob may change his interest to another team which has a much better potential to win the match. Thus, the system should be dynamic, continuously and incrementally refining, extending, and updating during the system operation in order to cope with new facts and new evidence about users' viewing preferences. This requirement led to the development of a learning model with respect to users' profiles:. To create dynamic user profiles, the usage information is collected during users' search sessions. An initial profile for a new user will be created after a user enters a query to the system. Ontologies enable an initial user profile to be matched with existing concepts in the domain ontology and with relationships between these concepts. Building an Ontological model of user's interest may cause inconsistencies if the domain ontology does not contain any of the words that form a given user's preferences (terminological problem). To solve this problem, text captions can be augmented by adding a few semantically similar or related terms after processing with the NLP technique. WordNet is exploited as a lexical reference system in order to find these additional related terms. Hence, the similarity between terms and concepts in the domain ontology are computed to determine the best matched categories with users' preferences.

Static user profiles can constructed based upon information from user activities, referred to as multi-shot or multi-session user profiles which are recorded by the user-model component using a statistical model, e.g. LSI. However, this static multi-shot user model relies on previous usage data. This can result in a failure to filter out irrelevant visual content because users' interests are dynamic and are likely to change over time. Therefore, multi-shot interests are not always reliable and may not always accurately reflect a user's interests. Therefore, a dynamic model is needed to cope with this problem.

The learning component is used to adapt to changes in users' interests. The retrieval process includes a learning process to detect shifts in user's interests and updates in user profiles. Otherwise, inaccuracies occur in profiles that affect the retrieval results. User profiles can be updated implicitly during and after the retrieval process. An example of learning algorithm for adaptive user profile proposed in (Liu et al. 2004) as follows:

$$M(i,j)^t = \frac{N_i^{t-1}}{N_i^t} M(i,j)^{t-1} + \frac{1}{N_i^t} \sum_k VCT(k,j) * CT(k,i) \tag{29}$$

where M^t is the modified user profile at time *t*; N_i^t is the number of instances of visual content which are related to the i-th concept that have been accumulated from time zero to time *t*; the second term on right hand side of (28) is the sum of the weight of the j-th term in the text captions (VCT) that are related to the i-th concept (CT) and obtained between time t-1 and time t divided by N_i^t; and k is a number of related image. This approach allows the system to learn and update users' interests rapidly and makes user profile more *dynamic* than in the previous framework.

The retrieval and filtering component matches the user profile with the information in the knowledge base and decides whether or not the data is relevant to the user. The decision is not binary (i.e. relevant or not relevant) but it is probabilistic (i.e. information receives a relevance rank). User queries need to perform word sense disambiguation and their semantic similarity needs to be measured. To ensure that the results are relevant to the query, a statistical computation, in the form of a *cosine similarity* measurement, is performed. In addition, the results obtained from the cosine similarity measure are further filtered according to the user profile.

A *Personal relevance measure* has been proposed in (Castells et al. 2005) to calculate the similarity between a user preference (u) and a weighted keyword associated with visual content (*p*) in the KB. The personal relevance measure is defined as:

$$prm(u,p) = \frac{u \cdot p}{\|u\|\|p\|} \tag{30}$$

To calculate the similarity between a user preference, query and the visual content, integrating the *cosine similarity* and the *personal relevance measure* a so-called combSum model (Castells et al. 2005) is needed. The combSum model merges the two rankings using a linear combination of the relevance scores.

$$score(d,q,u) = \lambda \cdot prm(u,p) + (1-\lambda)sim(p,q) \tag{31}$$

where $\lambda \in [0,1]$. The searching results are ranked and presented to user.

8.3.2 Benefits of using the PIMR System

The use of an ontological knowledge model can capture users' interests more effectively. The fact that the statistical technique relies solely on numeric data can result in a failure to understand the *meaning* of users' interests. For instance, Bob has accessed a picture of the Beijing Olympics games. Use of a statistical model alone, however, fails to capture the context of the visual content in which he is interested. This feature is not supported by usage mining techniques, but through using a knowledge-based model. A knowledge-based model is able to share concept-based representations for retrieval. Ontologies can define user interests using the same concept space used to describe the visual information content. The main difference from previous frameworks is that text captions are augmented by adding a few semantically similarity or related terms obtained from WordNet. Hence, the presented system tolerates uncertainty for the case when a domain ontology does not contain any of specific words that match users' interests.

Using an Ontology representation for the domain knowledge and users' profiles allows inference mechanisms to find more relevant information more effectively. For example, Bob can view the visual contents for Freestyle, Backstroke, Butterfly, and Breaststroke swimming. With a knowledge-based model, the system can infer that Bob is interested in Aquatics sport (generalisation). In addition, using *inference* and *reasoning* mechanisms, it is possible to predict the 'Aquatics sport' as a 'general' interest of Bob. This information can be represented in users' profile using a hierarchical structure at an appropriate level of granularity and coverage. As a result, the user model is represented with richer information than conventional methods. Learning dynamic user preferences from only the most recent observation leads to a user model that can adjust more rapidly to a user's changing interest. This makes a user profile more dynamic.

This initial idea has been accepted and awarded as the best student paper at the International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services (CENTRIC 2009) and is included as a section in the book "*Semantics in Adaptive and Personalised Services: Methods, Tools and Applications*".

In addition, PIMR will be combined with other techniques of IMR e.g. CBIR and SBIR. These techniques will work together based upon an ontology-based KB. This means a new framework is able to interpret the meaning of visual content when there is no text captions supplied using CBIR technique. It also semantically retrieves visual content based upon concept of users' queries (SBIR) and users' preferences rather than performing a syntactic search and providing similar results to all users. Figure 8-2 illustrates the main focus (an intersection part of the three main aspects) of the future work of this thesis.

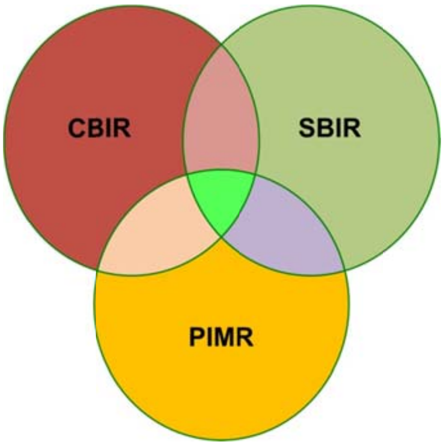


Figure 8-2 Future work of this thesis will combine three main aspects of the IR system

REFERENCES

Alhwarin, F. et al., 2008. Improved SIFT-Features Matching for Object Recognition. In *International Academic Conference on Vision of Computer Science-BSC*. pp. 179-190.

AlSumait, L. & Domeniconi, C., 2008. Text Clustering with Local Semantic Kernels. In M. W. Berry & M. Castellanos, eds. *Survey of Text Mining II: Clustering, Classification, and Retrieval.* London, United Kingdom: Springer-Verlag London Limited, pp. 87-105.

Bach, J. et al., 1996. Virage Image Search Engine: an Open Framework for Image Management. In *International Conference on Storage and Retrieval for Still Image and Video Databases IV*. pp. 87-76.

Barnard, K. et al., 2003. Matching Words and Pictures. *Journal of Machine Learning Research*, 3, 1107-1135.

Benitez, A., Smith, J. & Chang, S., 2000. MediaNet: A Multimedia Information Network for Knowledge Representation. In *Proceedings of the 2000 SPIE Conference on Internet Multimedia Management System*. pp. 1-12.

Berners-Lee, T., Hendler, J. & Lassila, O., 2001. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific American Magazine*, 34-43.

Brank, J., Grobelnik, M. & Mladenić, D., 2005. A Survey of Ontology Evaluation Techniques. In *Proceedings of the 8th Multiple Conferences of Information Scociety*. pp. 166-169.

Brewster, C. et al., 2004. Data Driven Ontology Evaluation. In *Proceedings of International Conference on Language Resources and Evaluation*. pp. 1-4.

Cascia, M.L., Sethi, S. & Sclaroff, S., 1998. Combining Textual and Visual Cues for Content-based Image Retrieval on the World Wide Web. In *IEEE Workshop on Content-based Access of Image and Video Libraries*. pp. 24-28.

Castells, P. et al., 2005. Self-tuning Personalized Information Retrieval in an Ontology-Based Framework. In *OTM Workshops on the Move to Meaningful Internet Systems*. pp. 977-986.

Cavallaro, A. & Ebrahimi, T., 2004. Interaction between High-level and Low-level Image Analysis for Semantic Video Object Extraction. *EURASIP Journal on Applied Signal Processing*, 2004, 786-797.

Chandrasekaran, B., Josephson, J.R. & Benjamins, V.R., 1999. What Are Ontologies, and Why Do We Need Them? *IEEE Intelligent Systems*, 14(1), 20-26.

Chen, S. & Williams, M., 2008. Learning Personalized Ontologies from Text: A Review on an Inherently Transdisciplinary Area. In *Personalized Information Retrieval and Access: Concepts, Methods and Practices*. New York, USA: IGI Global, pp. 1-29.

Chen, X., Hu, X. & Shen, X., 2009. Spatial Weighting for Bag-of-Visual-Words and Its Application in Content-Based Image Retrieval. In *Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*. pp. 867-874.

Chisholm, E. & Kolda, T.G., 1999. *New Term Weighting Formulas For The Vector Space Method In Information Retrieval*, USA: Computer Science and Mathematics Division, Oak Ridge National Laboratory.

Cristianini, N., Shawe-Taylor, J. & Lodhi, H., 2002. Latent Semantic Kernels. *Journal of Intelligent Information Systems*, 18(2), 127-152.

Csurka, G. et al., 2004. Visual Categorization with Bags of Keypoints. In *International Workshop on Statistical Learning in Computer Vision*. pp. 1-22.

Cullen, P., Hull, J. & Srihari, S., 1992. A Constraint Satisfaction Approach to the Resolution of Uncertainty in Image Interpretation. In *Proceedings of the 8th International Conference on Artificial Intelligence for Applications*. pp. 127-133.

Cullum, J.K. & Willoughby, R.A., 2002. *Lanczos Algorithms for Large Symmetric Eigenvalue Computations: Theory*, Boston, USA: SIAM.

Dasiopoulou, S., Doulaverakis, C. & Mezaris, V., 2007. An Ontology-Based Framework for Semantic Image Analysis and Retrieval. In Y. Zhang, ed. *Semantic-Based Visual Information Retrieval*. USA: IRM Press, pp. 269-293.

Datta, R. et al., 2008. Image Retrieval: Ideas, Influences, and Trends of the New Age. *ACM Computing Surveys*, 40(2), 1-60.

Deerwester, S. et al., 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41, 391-407.

Domeniconi, C. et al., 2007. Locally Adaptive Metrics for Clustering High Dimensional Data. *Data Mining and Knowledge Discovery*, 14(1), 63-97.

Duygulu, P. et al., 2002. Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary. In *Proceedings of the 7th European Conference on Computer Vision-Part IV*. pp. 97-112.

Fayyad, U.M., 1996. *Advances in Knowledge Discovery and Data Mining*, Cambridge, MA: MIT Press.

Forsyth, D. & Fleck, M., 1997. Body Plans. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. pp. 678-683.

Gasevic, D., Djuric, D. & Devedzic, V., 2009. *Model Driven Engineering and Ontology Development* 2nd ed., London, United Kingdom: Springer.

Gomez-Perez, A., Corcho, O. & Fernandez-Lopez, M., 2004. *Ontological Engineering: with examples from the areas of Knowledge Management, e-Commerce and the Semantic Web* 1st ed., London, United Kingdom: Springer.

Gonzalez, R., 2008. Exploring Information Management Problems in the Domain of Critical Incidents. In *Personalized Information Retrieval and Access: Concepts, Methods and Practices*. New York, USA: IGI Global, pp. 55-76.

Gruber, T.R., 1993. A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5(2), 199-220.

Guarino, E.N., Poli, R. & Guarino, N., 1995. Formal Ontology, Conceptual Analysis and Knowledge Representation. *International Journal of Human-Computer Studies*, 43(5-6), 625-640.

Hampapur, F.V., 2003. *Towards the Semantic Web: Ontology-Driven Knowledge Management* J. Davies & D. Fensel, eds., New York, USA,Wiley.

Hao, L. & Hao, L., 2008. Automatic Identification of Stop Words in Chinese Text Classification. In *2008 International Conference on Computer Science and Software Engineering*. pp. 718-722.

Harris, C. & Stephens, M., 1988. A Combined Corner and Edge Detector. In *Proceedings of the 4th Alvey Vision Conference*. pp. 147-151.

Haubold, A., Natsev, A. & Naphade, M., 2006. Semantic Multimedia Retrieval using Lexical Query Expansion and Model-Based Reranking. In *IEEE International Conference on Multimedia and Expo*. pp. 1761-1764.

Hendler, J., 2001. Agents and the Semantic Web. *IEEE Intelligent Systems*, 16(2), 30-37.

Hironobu, Y.M., Takahashi, H. & Oka, R., 1999. Image-to-Word Transformation Based on Dividing and Vector Quantizing Images With Words. In *Proceedings of International Workshop on Multimedia Intelligent Storage and Retrieval Management*. pp. 405-409.

Hollink, L. et al., 2003. Semantic Annotation of Image Collections. In *Workshop on Knowledge Markup and Semantic Annotation*. pp. 1-3.

Hsu, C., Chang, C. & Lin, C., 2010. *A Practical Guide to Support Vector Classification*, Taiwan: Department of Computer Science National Taiwan University.

Hu, J. & Bagga, A., 2004. Categorizing Images in Web Documents. *IEEE Multimedia*, 11(1), 22-30.

Hyvönen, E., Styrman, A. & Saarela, S., 2002. *Ontology-Based Image Retrieval*, Helsinki Institue for Information Technology, Finland: HIIT Publications.

Jensen, F.V. & Nielsen, T.D., 2007. *Bayesian Networks and Decision Graphs* 2nd ed. M. Jordan, J. Kleinberg, & B. Scholkopf, eds., New York, USA: Springer.

Jiang, Y. & Ngo, C., 2009. Visual Word Proximity and Linguistics for Semantic Video Indexing and Near-Duplicate Retrieval. *Computer Vision and Image Understanding*, 113(3), 405-414.

Jing, F. et al., 2003. Learning in Region-Based Image Retrieval. In *Proceedings of the IEEE International Symposium on Circuits and Systems*. pp. 206-215.

Joachims, T., 1998. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. *Machine Learning*, 1398, 137-142.

Kalfoglou, Y., 2001. Exploring Ontologies. In *Handbook of Software Engineering and Knowledge Engineering: Fundamentals*. World Scientific Publishing, pp. 863-887.

Ke, Y. & Sukthankar, R., 2004. PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2004*. pp. 513-506.

Khan, L., McLeod, D. & Hovy, E., 2004. Retrieval effectiveness of an Ontology-based Model for Information Selection. *The VLDB Journal The International Journal on Very Large Data Bases*, 13(1), 71-85.

Krebs, B., Burkhardt, M. & Korn, B., 1998. Handling Uncertainty in 3D Object Recognition Using Bayesian Networks. In *Proceedings of the 5th European Conference on Computer Vision-Volume II*. pp. 782-795.

Lee, Y., Lee, K. & Pan, S., 2005. Local and Global Feature Extraction for Face Recognition. In *Proceedings of the 5th International Conference on Audio- and Video-Based Biometric Person Authentication*. pp. 219-228.

Liu, F., Yu, C. & Meng, W., 2004. Personalized Web Search For Improving Retrieval Effectiveness. *IEEE Transaction on Knowledge and Data Engineering*, 16(1), 28-40.

Liu, Y. et al., 2007. A Survey of Content-Based Image Retrieval with High-Level Semantics. *Pattern Recognition*, 40(1), 262-282.

Llorente, A. & Rüger, S., 2009. Using Second Order Statistics to Enhance Automated Image Annotation. In *Proceedings of the 31th European Conference on IR Research*. pp. 570-577.

Lowe, D.G., 1999. Object Recognition from Local Scale-Invariant Features. In *Proceedings of the International Conference on Computer Vision*. pp. 1150-1157.

Lowe, D.G., 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2), 91-110.

Ma, W. & Manjunath, B., 1997. NeTra: a Toolbox for Navigating Large Image Databases. In *Proceedings., International Conference on Image Processing*. pp. 568-571.

Maedche, A. & Staab, S., 2002. Measuring Similarity between Ontologies. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management, Ontologies and the Semantic Web*. pp. 251-263.

Manning, C.D., Raghavan, P. & Schütze, H., 2008. *Introduction to Information Retrieval*, London, United Kingdom: Cambridge University Press.

Marengoni, M. et al., 2003. Decision Making and Uncertainty Management in a 3D Reconstruction System. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(7), 852-858.

McGuinness, D., 2003. Ontologies Come of Age. In D. Fensel et al., eds. *The Semantic Web: Why, What, and How*. Boston, MA: MIT Press.

Meersman, R., 1999. Semantic Ontology Tools in Information System Design. In *Foundations of Intelligent Systems, 11th International Symposium*. pp. 30-45.

Mezaris, V., Kompatsiaris, I. & Strintzis, M., 2003. An Ontology Approach to Object-Based Image Retrieval. In *Proceedings of the International Conference on Image Processing*. pp. 511-514.

Mikolajczyk, K. & Schmid, C., 2003. A Performance Evaluation of Local Descriptors. In Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition. pp. 257-263.

Miller, G.A., 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11), 39-41.

Moller, R. & Neumann, B., 2008. Ontology-Based Reasoning Techniques for Multimedia Interpretation and Retrieval: Theory and Applications. In Y. Kompatsiaris & P. Hobson, eds. *Semantic Multimedia and Ontologies*. London, United Kingdom: Springer, pp. 55-98.

Moravec, H., 1977. Towards Automatic Visual Obstacle Avoidance. In *Proceedings of the 5th International Joint Conference on Artificial Intelligence*. p. 584.

MPEG Requirements Group, 1999. MPEG-7 Context, Objectives and Technical Roadmap V.12. In *ISO/IEC JTC 1/SC 29/WG 11*. International Organization for Standardization, pp. 1-13.

Nack, F., van Ossenbruggen, J. & Hardman, L., 2005. That obscure object of desire: multimedia metadata on the Web, part 2. *IEEE Multimedia*, 12(1), 54-63.

Nagypál, G., 2007. *Possibly Imperfect Ontologies for Effective Information Retrieval*, Germany: University Karlsruhe (TH).

Naphade, M. & Smith, J., 2003. Learning Regional Semantic Concepts from Incomplete Annotation. In *Proceedings of the International Conference on Image Processing*. pp. 603-606.

Natsev, A. et al., 2007. Semantic Concept-Based Query Expansion and Re-ranking for Multimedia Retrieval. In *Proceedings of the 15th International Conference on Multimedia*. pp. 991-1000.

Neches, R. et al., 1991. Enabling Technology for Knowledge Sharing. *AI Magazine*, 12(3), 36-56.

Paek, S. & Chang, S., 2000. A Knowledge Engineering Approach For Image Classification Based On Probabilistic Reasoning Systems. In *IEEE International Conference on Multimedia and Expo*. pp. 1133-1136.

Poslad, S., 2009. *Ubiquitous Computing Smart Devices, Environments and Interactions*, London, United Kingdom: John Willey & Sons.

Poslad, S. & Zuo, L., 2008. An Adaptive Semantic Framework to Support Multiple User Viewpoints over Multiple Databases. *Advances in Semantic Media Adaptation and Personalization*, 93, 261-284.

Praks, P., Snasel, S. & Dvorský, J., 2003. Latent Semantic Indexing for Image Retrieval Systems. In *Proceedings of SIAM Conference on Applied Linear Algebra*. pp. 1-8.

Rui, Y. et al., 1998. Relevance Feedback: A Power Tool for Interactive Content-Based Image Retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 8(5), 644-655.

Rui, Y., Huang, T.S. & Chang, S., 1999. Image Retrieval: Current Techniques, Promising Directions, and Open Issues. *Journal of Visual Communication and Image Representation*, 10(1), 39-62.

Schmid, C. & Mohr, R., 1997. Local Grayvalue Invariants for Image Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(5), 530-535.

Schreiber, A.T. et al., 2001. Ontology-Based Photo Annotation. *IEEE Intelligent Systems*, 16(3), 66-74.

Schreiber, G. et al., 1994. CommonKADS: A Comprehensive Methodology for KBS Development. *IEEE Expert*, 9(6), 28-37.

Sheikholeslami, G., Chang, W. & Zhang, A., 1998. Semantic Clustering and Querying on Heterogeneous Features for Visual Data. In *Proceedings of the sixth ACM international conference on Multimedia*. pp. 3-12.

Sinclair, P.A.S. et al., 2005. Concept Browsing for Multimedia Retrieval in the SCULPTEUR Project. In *The 2nd Annual European Semantic Web Conference*. pp. 28-36.

Sivic, J. et al., 2008. Unsupervised Discovery of Visual Object Class Hierarches. In IEEE Conference on Computer Vision and Pattern Recognition. pp. 1-8.

Smeulders, A.W.M. et al., 2000. Content-Based Image Retrieval at the End of the Early Years. *IEEE Transaction on Pattern Analysis and Matching Intelligent*, 22(12), 1349-1380.

Smith, J.R. & Chang, S., 1997. Visually Searching the Web for Content. *IEEE MultiMedia*, 4(3), 12-20.

Smith, J.R. & Chang, S., 1996. VisualSEEk: A Fully Automated Content-Based Image Query System. In *Proceedings of the 4th ACM international conference on Multimedia*. pp. 87-98.

Song, X., Lin, C. & Sun, M., 2004. Autonomous Visual Model Building Based on Image Crawling through Internet Search Engines. In *Proceedings of the 6th ACM SIGMM International Workshop on Multimedia Information Retrieval*. pp. 315-322.

Swain, M.J., Frankel, C. & Athitsos, V., 1997. WebSeer: An Image Search Engine for the World Wide Web. In *Technical Report*. The University of Chicago.

Szummer, M. & Picard, R.W., 1998. Indoor-Outdoor Image Classification. In *Proceedings of the 1998 International Workshop on Content-Based Access of Image and Video Databases*. pp. 42-51.

Tamura, H., Mori, S. & Yamawaki, T., 1978. Textural Features Corresponding to Visual Perception. *IEEE Transactions on Systems, Man and Cybernetics*, 8(6), 460-473.

Tansley, R., 2000. *The Multimedia Thesaurus: Adding A Semantic Layer to Multimedia Information*, United Kingdom: University of Southampton.

Tirilly, P., Claveau, V. & Gros, P., 2008. Language Modeling for Bag-of-Visual Words Image Categorization. In *Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval*. pp. 249-258.

Tong, S. & Chang, E., 2001. Support Vector Machine Active Learning for Image Retrieval. In *Proceedings of the ninth ACM International conference on Multimedia*. pp. 107-118.

Troncy, R. & Carrive, J., 2004. A Reduced Yet Extensible Audio-Visual Description Language: How to Escape From the MPEG-7 Bottleneck. In *Proceedings of the 4th ACM Symposium on Document Engineering, DocEng'04*. pp. 87-89.

Tseng, B. et al., 2003. Normalized Classifier Fusion for Semantic Visual Concept Detection. In *Proceedings of the International Conference on Image Processing*. pp. 535-538.

Vailaya, A., Jain, A. & Hong, J.Z., 1998. On Image Classification: City vs. Landscape. In *Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries*. pp. 3-8.

Vallet, D. et al., 2007. Personalized Content Retrieval in Context Using Ontological Knowledge. *IEEE Transactions on Circuits and Systems for Video Technology*, 17(3), 336-346.

Wang, J. et al., 1999. Semantics-Sensitive Retrieval for Digital Picture Libraries. *Digital Library Magazine*, 5(11), URL: http://www.dlib.org.

Wang, J., Li, J. & Wiederhold, G., 2001. SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture Libraries. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 23(9), 947-963.

Wang, L., Lu, Z. & Ip, H.H., 2009. Image Categorization Based on a Hierarchical Spatial Markov Model. In *Proceedings of the 13th International Conference on Computer Analysis of Images and Patterns*. pp. 766-773.

Wang, Z. et al., 2003. Content-Based Image Retrieval with Relevance Feedback using Adaptive Processing of Tree-Structure Image Representation. *International Journal of Image and Graphics*, 3(1), 119-143.

Wenyin, L. et al., 2001. Semi-Automatic Image Annotation. In *Conference on Human-Computer Interaction, Interact 2001*. pp. 326-333.

Westerveld, T., 2000. Image Retrieval: Content versus Context. In *Content-Based Multimedia Information Access, RIAO 2000 Conference*. pp. 276-284.

Wong, S.K.M., Ziarko, W. & Wong, P.C.N., 1985. Generalized Vector Spaces Model in Information Retrieval. In *Proceedings of the 8th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. pp. 18-25.

Wu, L., Hoi, S.C. & Yu, N., 2009. Semantic-Preserving Bag-of-Words Models for Efficient Image Annotation. In *Proceedings of the 1st ACM Workshop on Large-Scale Multimedia Retrieval and Mining*. pp. 19-26.

Yang, J. et al., 2007. Evaluating Bag-of-Visual-Words Representations in Scene Classification. In *Proceedings of the International Workshop on Multimedia Information Retrieval*. pp. 197-206.

Yang, S., 2003. *A Probabilistic Approach to Human Motion Detection and Labeling*. California Institute of Technology.

Yang, Y. & Wilbur, J., 1996. Using Corpus Statistics to Remove Redundant Words in Text Categorization. *Journal of the American Society for Information Science*, 47(5), 357-369.

Yuan, J., Wu, Y. & Yang, M., 2007a. Discovery of Collocation Patterns: from Visual Words to Visual Phrases. In *IEEE Conference on Computer Vision and Pattern Recognition*. pp. 1-8.

Yuan, J., Wu, Y. & Yang, M., 2007b. From Frequent Itemsets to Semantically Meaningful Visual Patterns. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. pp. 864-873.

Zhang, Y., 2007. Toward High-Level Visual Information Retrieval. In *Semantic-based Visual Information Retrieval*. London, United Kingdom: IGI Publishing, pp. 1-20.

Zhao, R. & Grosky, W.I., 2002. Narrowing the Semantic Gap - Improved Text-Based Web Document Retrieval Using Visual Features. *IEEE Transactions on Multimedia*, 4(2), 189-200.

Zheng, X. et al., 2004. Locality preserving clustering for Image Database. In *Proceedings of the 12th annual ACM International conference on Multimedia*. pp. 885-891.

Zheng, Y. et al., 2008. Toward a Higher-Level Visual Representation for Object-based Image Retrieval. *The Visual Computer*, 25(1), 13-23.

Zhou, X.S. & Huang, T.S., 2003. Relevance Feedback in Image Retrieval: A Comprehensive Review. *Multimedia Systems*, 8(6), 536-544.

Zhu, J., Uren, V. & Motta, E., 2005. ESpotter: Adaptive Named Entity Recognition for Web Browsing. In *Intelligent IT Tools for Knowledge Management Systems, KMTOOLS 2005*. pp. 518-529.