

Semantic Restructuring of Natural Language Image Captions to Enhance Image Retrieval

Kraisak Kesorn

School of Electrical and Electronic Engineering and Computer Science,
Queen Mary University of London, Mile End Rd, London, E1 4NS, United Kingdom
Email: kraisak.kesorn@elec.qmul.ac.uk

Stefan Poslad

School of Electrical and Electronic Engineering and Computer Science,
Queen Mary University of London, Mile End Rd, London, E1 4NS, United Kingdom
Email: stefan.poslad@elec.qmul.ac.uk

Abstract—The rapid growth in the volume of visual information can make the task of finding and accessing visual information of interest, overwhelming for users. Semantic analysis of image captions can be used in conjunction with image retrieval systems (IMR) to retrieve selected images more precisely. To do this, we first exploit a Natural Language Processing (NLP) framework in order to extract concepts from image captions. Next, an ontology-based framework is deployed in order to resolve natural language ambiguities. The novelty of the proposed framework is that the combination of LSI with the Ontology framework enables the combined framework to tolerate ambiguities and variations in the Ontology. A key feature is that the system can 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 use of LSI based NLP combined with an ontological framework significantly enhances image retrieval.

Index Terms—image retrieval, latent semantic indexing, natural language processing, knowledge base, semantic model, Ontology.

I. INTRODUCTION

Documents on the Internet are often composed of several kinds of multimedia information when accessed for personal, entertainment, business, and scientific purposes. There are many specific content domains, of interest to different communities of users. One such content domain that has strong universal appeal is the sports domain [1]. Multimedia content can benefit from a multimedia specific retrieval system, rather than using a traditional text-based retrieval system. Concerning image retrieval systems, there are several schemes to retrieve images from content collections such as: Content-Based Retrieval (CBR) [2], [3], [4]; automatic classification of objects and scenes [5], [6], [7], [8], [9]; relevance

feedback from users [24], [25], and image and region labelling [26], [27], [28].

However, there are often considerable differences between human users' high level interpretation of the semantics of visual information and the low-level visual features that can be automatically extracted, creating the so called 'semantic gap' between these. CBR and current text-based retrieval approaches are still far from being able to support semantic-based access. Consequently, the field of Semantic-Based Visual Information Retrieval (SBVIR) has been established and has become a notable research theme for multimedia information retrieval [6] [7]. While numerous ways of performing this task have been proposed and implemented, this field is still very much in its infancy, and most of the associated problems remain unresolved.

One promising approach to enhance visual image retrieval has been to supplement image content analysis with textual annotations associated with the image. In fact, even when content-based techniques are applied, textual information surrounding images should not be disregarded since it often includes some form of human generated description of the image. Image captions can be exploited to help create a knowledge base for the semantic representation of images.

The rest of this paper is organized as follows. In Section II, state-of-the-art frameworks are surveyed and their limitations are analysed. In Section III, the design of the proposed framework infrastructure is described. Section IV describes and discusses the application of the framework within the sports domain and experimental results. Section V presents our conclusions and discusses further work.

II. RELATED WORK

A. Problem Requirements

Before analysing existing solutions, the formal requirements for image retrieval systems are identified. These requirements will serve as a basis for discussion in later sections. Generally, there are several factors inhibiting IMR systems from retrieving images:

1) *Ambiguity of natural language descriptions*: synonyms, polysemy and word inflection can generate false positives with algorithms that are designed to perform exact key word matches. An example of a polysemy problem is the query "Find all images of athletes from the USA". The query implies a distinction between the term USA being used to relate to the nationality of the athlete rather than relating to the location of the sporting event. Keyword matches will not be able to distinguish between these two different relations.

2) *Indirectly relevant concepts*: there are many cases where specific concepts are not mentioned directly in the text captions but which can be inferred semantically and can be used as part of semantic searches. An Ontology-based framework can provide the background knowledge needed to automatically expand terms in captions into other relevant concepts.

3) *Metadata incompleteness*: image captions may not supply all the required information in order to represent the semantics of the image. These uncertainties lower the precision for which the semantics of images can be defined. IMR frameworks may be able to generate the missing information through semantic inferences of knowledge contained in the Ontology framework.

4) *Heterogeneous Ontology Commitment and subjective use of concepts in applications*: although the use of a domain Ontology model has the potential to improve image retrieval within that domain, different communities of users will likely use different conceptualisations and different concept dependencies or commitments [10], e.g., "A football is a type of ball that is kicked". The concept "football" commits to the concept "kicked" in order to define it. Another use may define a different commitment such as a "ball that can be headed". Ontological commitments reflect the subjectivity of different uses of the Ontology in applications. Ontological commitments need to evolve dynamically through usage and experience. They can seldom be completely fixed in advance of usage and through limited usage. A framework is needed that can handle multiple ontological commitments and which allows commitments to be maintained during operation. Typically heterogeneous ontology commitments are handled through a process of either aligning or merging. Noy and Musen [40] define merging as the creation of a single coherent ontology that includes the information from all the sources.

They define alignment as a process in which the sources must be made consistent and coherent with one another but kept separate.

B. Surveyed Systems

Several approaches have been suggested to improve image search relevance and precision using image captions [12] [13] [14], [15]. This section focuses on a discussion of Ontology-based frameworks.

Wang and colleagues [16] have proposed a data-driven approach for image retrieval that uses Web images and their surrounding textual annotations as a source of training data to bridge the semantic gap. Using the WordNet thesaurus, the system is able to resolve synonym problems. Their framework supports automatic metadata extraction from Web pages using a vision-based web page analysis technique. However, this data-driven framework cannot perform semantic searches because the system does not store the semantic relationships in its knowledge base.

Schreiber et al [17] have explored the use of knowledge contained in Ontologies to index and search collections of images. Their proposed system can support semantic relations in annotations. However, their indexing process is done manually. Thus, it is difficult to scale this up, to support a high volume of images.

The Multimedia Thesaurus (MMT) [18] defines a facility for expressing explicit semantic relationships. The LSI algorithm has been used in order to solve NL ambiguities and to support image content subjectivity. However, the system does not satisfactorily address incompleteness e.g., some media data might not be classified into any concept in a knowledge-based model. It is not clear how the system deals with this situation. The MediaNet [19] framework deals with the uncertainty problem by trying to predict image concepts when textual information is not supplied. Experiments with MediaNet have shown that its classifiers and the summarized annotation of images using WordNet result in improved accuracy. However, heterogeneous ontology commitments are not supported by MediaNet.

Zheng et al. [32] have proposed an Ontology-based image retrieval framework. However, this knowledge base is encoded in XML which cannot capture the semantic relationships between concepts and their ontological commitments.

Karanastasi et al. [33] have proposed an approach for semantic processing of NL queries. The OntoNL framework exploits WordNet [30] and takes into account a number of parameters regarding the characteristics of the Ontologies and the types of users. This framework focuses on query processing and query expansion rather on knowledge acquisition based on image captions.

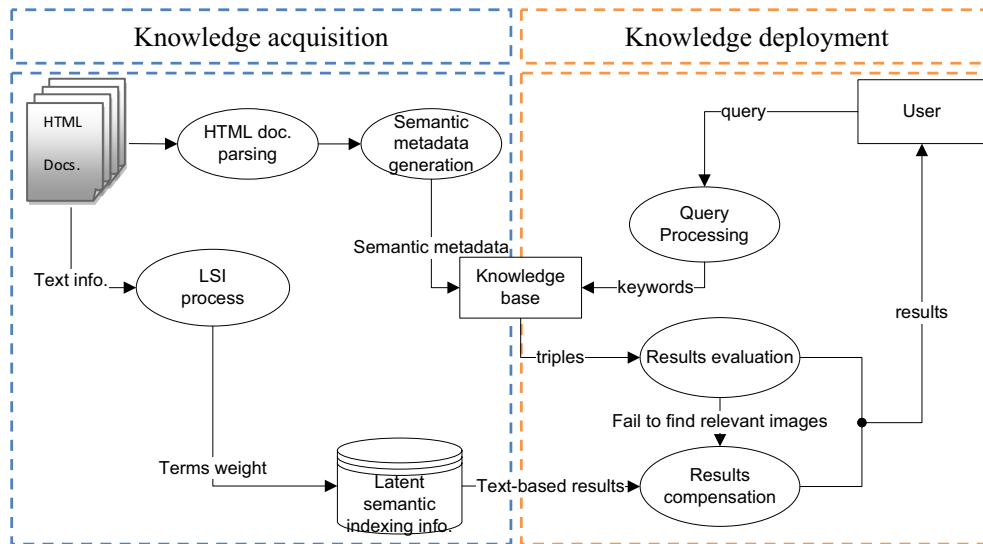


Figure 1. High-level architecture to support knowledge-based searches

MPEG-7 is a standard for multimedia content description that can be used for knowledge-based multimedia retrieval and filtering. The use of XML MPEG-7 is oriented to video rather than images. Several frameworks, [34], [35] and [36], have proposed the use of semantic based models as extensions to MPEG-7. This can create interoperability problems because there are several ways to assign semantics to the MPEG-7 content structures [37], [38].

A. Limitations of Surveyed Systems

Many of the surveyed systems address the issue of natural language ambiguity. This is because this was one of the main motivations for using Ontologies in the first place. However, there are still some issues which are not adequately addressed as follows.

Firstly, heterogeneous ontology commitments are not handled by most systems. It is difficult to build a complete and appropriate Ontology covering different applications in a domain, in one step. Thus, an Ontology-based IMR framework should support heterogeneous Ontologies and semantic metadata.

Secondly, indirectly relevant concepts are often not used for image retrieval. Although some specific terms or concepts are not mentioned directly in the image captions, it is possible that they are still semantically relevant to images.

Thirdly, the issue of missing metadata concepts in text captions is often not addressed. There are several causes for incomplete image captions. The image captions may not provide all the information required to instantiate the Ontology model. Some text captions may not match any Ontology concepts in the semantic model and some images may come without text captions.

The framework presented here addresses these weaknesses. This represents the main novelty and contribution of this paper.

III. PROPOSED FRAMEWORK

Fig. 1 presents a high-level architecture for the proposed framework. In this section, the methods for representing, discovering, storing, and querying the knowledge-base, are described.

Knowledge acquisition consists of several sub-processes. Web documents in HTML format are processed in order to extract their text information. A *semantic metadata generator* uses NLP in order to extract metadata from the text captions. Later, this metadata can be expanded into other relevant metadata and stored as part of the knowledge base, in RDF format. Latent Semantic Indexing (LSI) is performed in parallel with the semantic metadata generation process in order to find the interrelationships between terms and images using a vector space model. A vector space model is an algebraic model for representing text documents as vectors of identifiers such as index terms. If a term occurs in the document, its value in the vector is non-zero. This is widely used in traditional information retrieval system and is able to compensate for NLP of captions which may generate incomplete metadata.

Knowledge deployment applies the Ontology model in order to support semantic queries on image captions. Again there are several sub-processes involved: eliminating stop words within captions, automatically formulating queries to be represented as SPARQL queries¹ and compensating for missing concepts in the image caption. The SPARQL query performs a semantic search on the RDF file and returns results to a user. To ensure that the results are relevant to the query, a statistical computation, in the form of a cosine similarity measurement, is performed.

¹ SPARQL query, See <http://www.w3.org/TR/rdf-sparql-query>

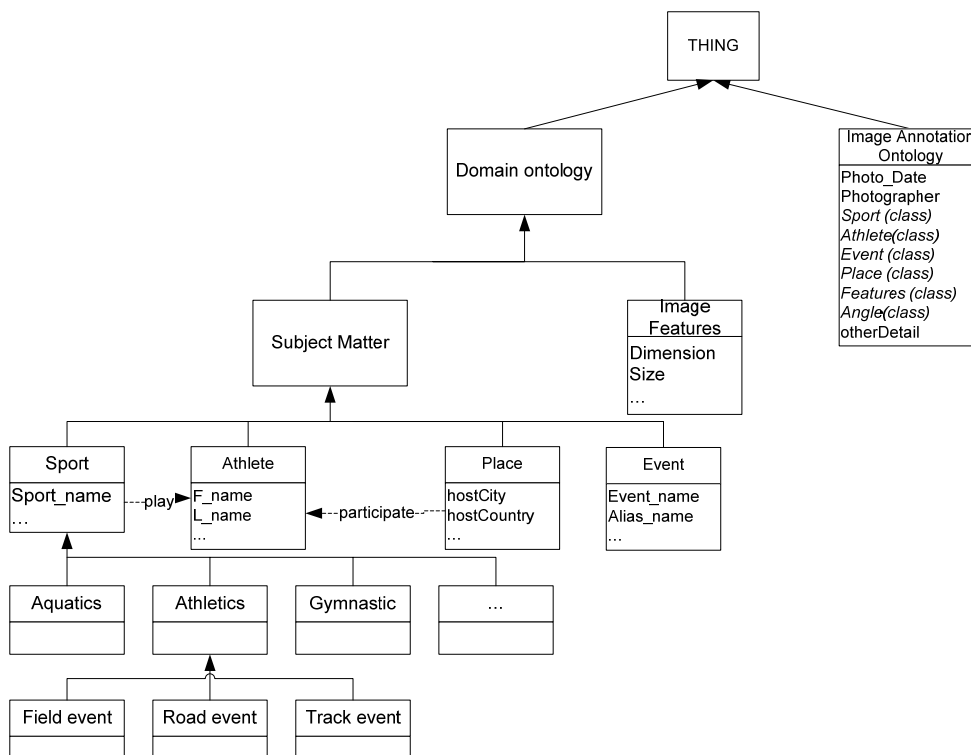


Figure 2. Graph-based semantic model

In some cases, an Ontology-based search may fail to find relevant images because there is no relevant metadata in the Ontology model. In this case, the system will activate the LSI component to compute the similarity of queries and indexed concepts using a vector space model. Then, results from LSI can be used to supplement the results from the Ontology-based search. We call this process results compensation. This is described in more detail in subsequent sections.

A. Knowledge-based Representation

The core component of the system is a knowledge based (actually an Ontology-based) model. The use of Ontologies to describe image documents provides a means to define well structured concepts and conceptual relationships which can subsequently ease the task of annotation and retrieval. To define semantic annotations for sport images, it is partitioned into two sub-Ontologies: a Domain Ontology and an Image annotation Ontology as proposed in [16].

The *Domain Ontology* describes the vocabulary and background knowledge for the subject domain of images. This defines two main classes ‘Subject_matter’, and ‘Image_features’ with a numbers of related properties. The Subject_matter class is sub-classed into the various types of domain

concepts that are needed to describe a sport image, such as Athlete, Event, Place, and Sport. The Image Feature class represents metadata about image features e.g. image format (e.g., jpg), size, and the resolution of an image.

The *Image annotation Ontology* is designed to store the annotations of images in the sports domain. It corresponds to three main aspects of the image such as: what the image depicts; how, when, and where the image is recorded; how the image is stored. 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. Fig. 2 presents a graph-based semantic model for the proposed framework.

B. Knowledge-based Discovery from Text Captions

The main focus of this process is to extract knowledge from image captions and to store this extracted knowledge in a semantic model. Firstly, the image captions are parsed from HTML files and then a NLP framework processes those text captions and generates outputs as XML files. We deploy an established NLP framework called ESpotter [29] rather than to develop our own.

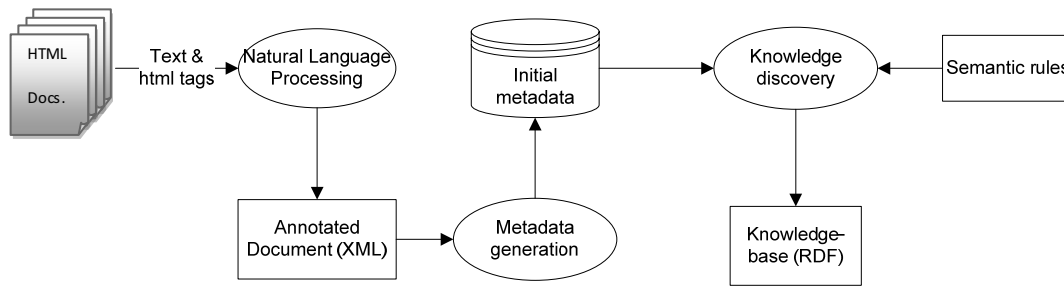


Figure 3. Sub-processes for semantic metadata generation

Another popular framework for NLP is the GATE framework². Although the GATE framework includes some useful inbuilt functions compared to ESpotter, e.g., for exporting metadata to relational databases such as Oracle and PostgreSQL, the downside of GATE is that it is more complicated to use, to configure, and to export metadata in XML format. ESpotter supports more efficient lexicon and patterns recognition and exports its results in XML format. These XML documents will then be parsed to form the initial metadata using a ‘metadata generation’ process and stored in a relational database (RDBMS).

In practice, semantic metadata is often ambiguous. For example the sentences, “Athlete David from the United Kingdom” and “Games of the XXX Olympiad will be held in the United Kingdom” uses ‘United Kingdom’ with respect to two different semantic relations: as a nationality for an athlete and as a host country. This ambiguity for the term United Kingdom cannot be resolved using syntactic keyword matches alone. Therefore, a Knowledge discovery step is required to disambiguate the metadata. To find the most appropriate ontology entity, a statistical technique TF-IDF [38] is applied to do this. The TF-IDF (term frequency–inverse document frequency) weight is a weight often used in information retrieval and text mining. This weight is a statistical measure used to evaluate the most similar concept in the Ontology to the word being considered.

Following disambiguation, knowledge discovery performs further metadata processing in order to find any implicit relationships amongst the Ontology concepts in the metadata. To do this, *semantic rules* are applied to expand the metadata to other relevant concepts. Consequently, new metadata may be associated with an image. For instance, if the date in the text caption of an image is detected as “20 September 2000”, this picture is considered to have a relationship with the *Sydney Olympic Games* which took place in year 2000 in Sydney (the host city), and Australia (the host country). This serves to handle the *metadata incompleteness requirement* which is addressed in section II. Simple rules are used mainly for expanding metadata to other relevant concepts

which may be absent in text captions e.g., subtype of sport (field event, track event, aquatic event etc.), host city, host country, and name of the Olympics games. An example of a simple semantic rule which is expressed in first-order logic is shown below. In this paper, we apply our framework to the Olympics Games domain only. This enables us to use a one-to-one mapping between the time, place and event. This does not take into account other sport events in the current of implementation. An example semantic rule is shown below:

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

- $Event(x), PhotoDate(x)$

$$\forall x, \exists y \mid Photo(x) \wedge PhotoDate(x) \wedge happensDuring(x) \Rightarrow Event(x)$$

Add $Event(x)$ to metadata of a given image if an image contains $PhotoDate$ that happens during the given event.

The overall aim is to generate semantic concepts from caption instances that align with concepts in the Ontology model. However, the Ontology model may not capture all the Ontological commitments and capture application specific uses of terms used in the captions. Hence, some concepts identified in the captions may not be matched with any particular Ontology concept but may still be important for the meaning of an image. For example, the NLP framework extracts the concepts ‘victory’ or ‘celebration’. These concepts are associated with the winner or results of a sports event. Although the Ontology does not currently contain these concepts, they are not discarded, they are assigned to the ‘*otherDetails*’ entity in the Ontology model. Further work will investigate more advanced ways of expanding the Ontology through use.

It is not assumed that the quality of the generated metadata reaches the same quality as the manually created metadata. Therefore, manual correction and annotation of metadata are also supported. Fig. 3

² GATE framework, See <http://www.gate.ac.uk>

illustrates the sub-processes for the semantic metadata generation process.

LSI is currently exploited in this framework (Fig. 1) to solve the subjective use of Ontology and to support the heterogeneous Ontology commitments problem. After textual information is parsed from 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. The term weight in each document is represented in the matrix format where rows represent keywords and columns represent the image (identified by an image ID). Fig. 4 shows an example of the LSI matrix and the term weights (*TW*).

Keywords	img1	img2	img3	img4	img5
Olympiad	0.577	0.707	0.707	0	0
Tennis	0.577	0	0.707	0	0
AGASSI	0.577	0.707	0	0.577	0.577
USA	0	0	0	0.577	0.577
Celebrate	0	0	0	0.577	0.577

Figure 4. LSI matrix after assigning weights to each term

Each term will be assigned a weight to show the importance of that term to the image. Term weight (*TW*) is a product of local weight (L_{ij}), global weight (G_i), and normalization factor multiplication (N_j). Various term weighting formulas have been proposed by several researchers. Erica et al [20] undertook experiments in order to evaluate and compare several term weighting schemes. Therefore, we selected some formulas they suggested. To compute the local term weight, the Square root scheme (SQRT) [20] formula was selected. Equation (1) gives the SQRT formula.

$$L_{ij} = \begin{cases} \sqrt{f_{ij} - 0.5} + 1 & \text{if } f_{ij} > 0 \\ 0 & \text{if } f_{ij} = 0 \end{cases} \quad (1)$$

where f_{ij} is the frequency of term i in image j . The global weights (G_i) are functions of how many times each term appears in the entire collection. The Inverse Document Frequency (IDFB) [20] method is deployed for the global term weighting computation. Equation (2) gives the IDFB formula.

$$G_i = \log \left(\frac{N}{n_i} \right) \quad (2)$$

where N is the number of images in the collection and n_i is the number of images in which term i appears. The normalization factor compensates for discrepancies in the lengths of the documents. Equation (3) shows normalization formula.

$$N_j = \frac{1}{\sqrt{\sum_{i=0}^m (G_i L_{ij})^2}} \quad (3)$$

The term-image matrix information is stored in the form of a table in a RDBMS. This information is useful when a IMR system fails to find relevant images due to heterogeneous ontology conceptualisations and commitments. This issue will be described in detail in section D.

C. Knowledge-based Storage

The semantic metadata generated by the metadata generation process is stored in a RDBMS. MySQL is used in this framework. To be able to use this data in a semantic context, it is mapped to the Ontology to give 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 RDF³, the Resource Description Framework, which can be represented as a directed graph consisting of nodes and directed arcs linking pairs of nodes (Fig. 1). RDF was chosen as it represents a compromise between supporting named relationships between concepts, being efficient to parse and supporting a standard query language. Although, other representations such as OWL⁴, the Web Ontology Language, are more semantically expressive, OWL is far more complex to process and query. RDF data structures must be mapped to RDBMS structures and vice versa. There are two approaches for mapping metadata in RDBMS form to RDF form, using direct or indirect mappings [21].

- *Direct mapping*: is a direct mapping from the relation database schema to the RDF. This generic approach can be useful in many cases, but sometimes it may lead to difficulties in reflecting changes in the database structure.
- *Indirect mapping*: uses application logic to access data. Some content management systems provide APIs and the application logic as a source of information to be exported into RDF form.

In this research, a *direct mapping* scheme is used to map the stored data from a MySQL database to the RDF/XML format using a JDBC connector. The mapping architecture is shown in Fig. 5.

³ The Resource Description Framework (RDF), See <http://www.w3.org/TR/REC-rdf-syntax/>. RDF adds support for named associations between concepts to XML.

⁴ 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.

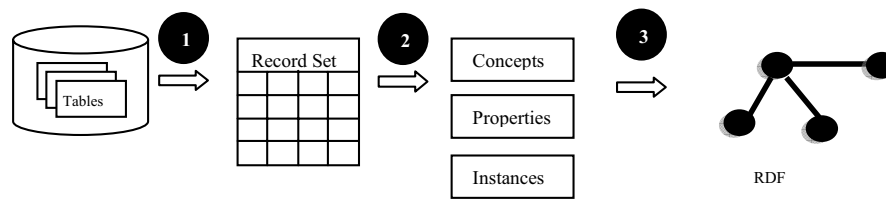


Figure 5. RDBMS to RDF mapping

Our mapping process is inspired by Berlin [31] and consists of the following steps:

- 1) Selection of a record set from the database using SQL: The initial metadata is retrieved from the MySQL tables using the SQL select command. The record sets returned from the query are grouped by the 'imageID' column.
- 2) Creation of instances and identifiers: The Jena API⁵ (Application Programming Interface) is deployed to create the Ontology concept instances and their properties, corresponding to the Ontology model given in Fig. 2.
- 3) Mapping of the grouped record set metadata to properties of instances: The grouped record set metadata is assigned to the ontology entities created in step 2.

The contribution of this process is to automatically apply an Ontology-based model to restructure semantic concepts in the natural language captions into a semantic or hierarchical structure which contain the relationships between the metadata and associated concepts.

D. Knowledge-based Image Retrieval

The query keywords from users are first tokenized. Then stop words are eliminated. Stop words represent the frequently occurring, insignificant words that appear in a text document. Common stop words include: *a, an, the, in, of, on, are, be, if, into, which* etc. These words do not provide a significant meaning to the documents or images in this research. Therefore, they should be removed in order to reduce 'noise' and to reduce the computation time.

The SPARQL query language is a W3C recommendation for querying data from RDF documents which form part of the the knowledge base. A query returns a list of instance tuples that satisfies the query. If the tuples only comprise domain concepts, the images that are annotated with these instances and can simply be retrieved. When a user inputs the query into the system, stop words are

removed and then the query is transformed into a SPARQL query automatically. For example, "Find all images about a particular field event *X*". Then, this user query can be formulated as:

```
SELECT ?photograph
WHERE { ?photo sport:Sport ?subType.
        ?subType sport:hasSportType ?athletics
        ?athletics sport:subAthletics ?fieldEvent }
```

If the proposed IMR system fails to find relevant images using the semantic model, LSI is activated to compute the similarity of concepts between the query and indexing terms using *cosine similarity*. The results of LSI will then be used instead of the knowledge-base search results. In other words, the performance of the framework will degrade better when Ontology conceptual variations occur, by compensating with the LSI results. Equation (4) defines the *cosine similarity* formula. Let $\{P_i\}_{i=1}^N$ be the set of all images in the collection, query (Q_i) and image (P_{ij}) have t terms and their associated weights are WQ_{ik} and WP_{ijk} respectively, for $k=1$ to t . The similarity between the query and image is measured using the following inner product:

$$sim(Q_i, P_{ij}) = \frac{\sum_{k=1}^t WQ_{ik} \times WP_{ijk}}{\sqrt{\sum_{k=1}^t (WQ_{ik})^2} \times \sqrt{\sum_{j=1}^m (WP_{ijk})^2}} \quad (4)$$

IV. IMPLEMENTATION AND EVALUATION

For the purpose of experimental evaluation, a collection of sport images from the Olympic organization website⁶ was assembled. Our framework was tested by selecting sample queries, and compared to the Lucene⁷, full-featured text search engine. In the current stage of implementation,

⁵ Jena library, See <http://jena.sourceforge.net>

⁶ <http://www.olympic.org>

⁷ <http://lucene.apache.org>

we only deal with textual information to build the knowledge base. Therefore, it is sufficient to compare our framework only with the text-based search part. Four hundred images were used for testing. The test environment was implemented using Java 1.6 and Jena APIs version 2.5; MySQL version 5.0 was used for storing the initial metadata; and Lucene version 2.3.2 was used for a comparative evaluation.

A. Hypotheses Evaluation

To evaluate the retrieval performance of our framework, some hypotheses were established against the requirements of the IMR given in section II. The main hypotheses evaluated are the following:

Hypothesis 1 (H1): The ontology model can be combined with the NL system to help overcome the NL ambiguity problem. Our framework enhances the correct interpretation of query keywords depending on their meaning rather than on their syntax matching between search terms and terms in the text captions.

Hypothesis 2 (H2): The semantic model can find the indirectly relevant concepts which are not identified explicitly in the document text. As discussed in section II-A, sometimes, some specific concepts are not mentioned directly in the text captions but they might be semantically relevant.

Hypothesis 3 (H3): Ontology-based searching provides an acceptable level of information retrieval performance even when Ontology variations are present. The retrieval performance of the ontology technique is expected to be not worse than the keyword-based search even though there are heterogeneous ontology conceptualisations and commitments within a domain.

B. Retrieval Performance Measurement

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 documents (as specified in a user query) in the document collection. Let B denote the retrieved documents which the system returns for the user query.

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

$$Precision = \frac{|A \cap B|}{|B|} \quad (5)$$

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

$$Recall = \frac{|A \cap B|}{|A|} \quad (6)$$

Using precision-recall pairs, a so-called precision-recall diagram, can be drawn that shows the precision values at different recall levels. In this paper, the retrieval performance is reported using the 11-point Interpolated Average Precision graph [22]. The interpolated precision P_{interp} at a certain recall level is r defined as the highest precision found for any recall level $r' \geq r$:

$$P_{interp}(r) = \max(r'), \quad r' \geq r \quad (7)$$

In order to evaluate the retrieval performance of two systems, the F score is also employed [23]. The F score is the harmonic mean of recall and precision, a single measure that combines recall and precision. The value of F score is between $[0, 1]$. An F score of 0 means no relevant documents have been retrieved, and the F score of 1 means all retrieved documents are relevant. The harmonic mean F assumes a high value only when both precision and recall are high.

Therefore, the determination of the maximum value for F can be interpreted as an attempt to find the best possible compromise between recall and precision. Equation (8) shows the F score formula.

$$F \text{ score} = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (8)$$

C. Selecting the User Queries to Evaluate

To evaluate the main hypotheses, the results of a classical full-text search engine (Lucene) have been compared with the results from of our framework. Test queries in this experiment have been selected to evaluate the hypotheses as follows.

Sample Query 1 (Q1): Find all images of any athletes with a specific nationality who compete at a specific location e.g., USA athletes in Australia. This query is used for NL ambiguity testing. In a data-driven IMR, the system cannot distinguish between 'USA athletes in Australia' versus 'Australia athletes in USA'. The 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.

TABLE I.
 QUERIES AND CONCEPTS WHICH ARE SPECIFIED AS THE CONSTRAINT IN A QUERY FOR SEARCHING

Query	Concepts				
	Athlete	Sport	Event	Place	Image_features
Q1: Find all images of any athletes with a specific nationality	✓	-	-	✓	-
Q2: Find all images of field sport.	-	✓	-	-	-
Q3: Find all images of a specific sport in a specific Olympics Games e.g., swimming in Sydney 2000	-	✓	✓	-	-
Q4: Find all images of a specific host city e.g., Barcelona	-	-	-	✓	-
Q5: Find all images of a specific Olympics Games	-	-	✓	-	-
Q6: Find all images which has a specific file size e.g., < 3 MB	-	-	-	-	✓
Q7: Find all images of a specific athlete participating in a specific host country e.g., Agassi in USA	✓	-	✓	-	-
Q8: Find all images of track event which has a specific file size	-	✓	-	-	✓
Q9: Find all images related to a specific host country	-	-	-	✓	-
Q10: Find all images a specific event e.g., opening ceremony	-	-	✓	-	-
Q11: Find all images of an athlete in a specific Olympic Games participating in a specific host city e.g., Popov in the XXV Olympiad, Moscow	✓	-	✓	✓	-

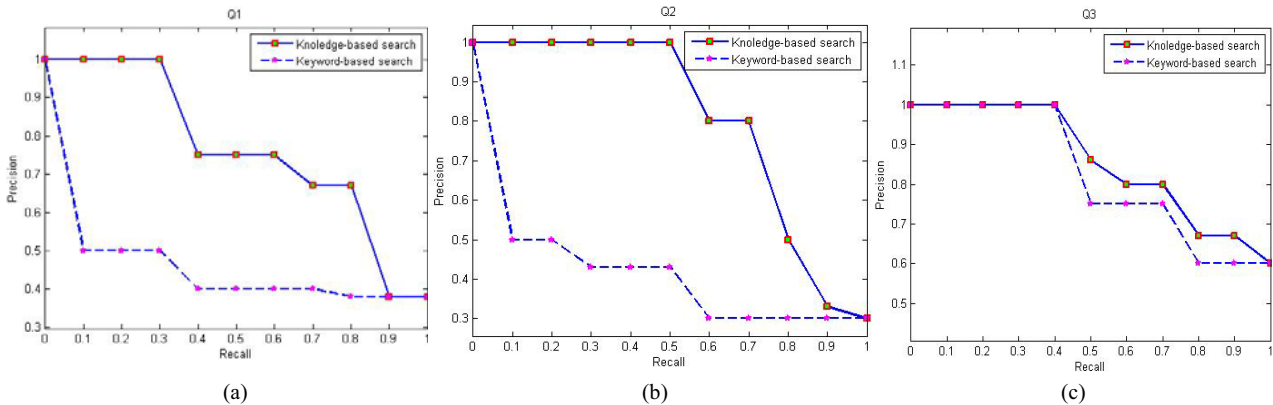


Figure 6. Experimental results comparison for the knowledge-based search and the keyword-based search

Sample Query 2 (Q2): Find all images related to the sub-concept of sport type, e.g., a field event. A field event usually refers to all the kinds of sports that athletes perform in the field e.g., hammer throw. This query aims to test the H2 hypothesis. Image descriptions in the collection usually only mention the sport name. They do not refer to the sub-concept of the particular sport. Therefore, an IMR framework should recognize these sub-concepts automatically.

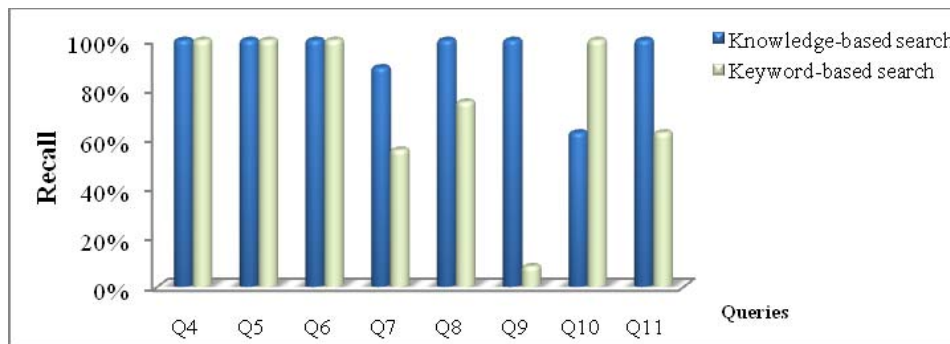
Sample Query 3 (Q3): Find all images of a specific sport type in a particular sport event. This query aims to test the H3 hypothesis. To do this, some information of images in Ontology is deliberately deleted, resulting in incompleteness in the ontology, e.g., we delete information about swimming images for the Sydney 2000 Olympics games. The query then fails to find relevant images, using the semantic model alone. Our framework will then in turn be triggered to use the LSI results.

In this experiment, we select the various sample queries which traverse each concept in the semantic model (Fig. 2). Table I shows the sample queries against the concepts which are traversed by the knowledge-based search algorithm. Here ‘✓’ means that the concept is visited by the knowledge-based search and ‘-’ means that the concept is not visited by the search algorithm. For example, Q1, *Find all images of any athletes with a specific nationality* e.g., images of USA athletes participating in Australia. This query is about athletes and places for the Olympics Games. Hence, the search traverses or visits the ‘Athlete’ and ‘Place’ concepts. If a visited concept consists of sub-concepts e.g., the ‘Sport’ concept, the knowledge-based search engine also visits related sub-concepts for concepts contained in the query. For example, Q2, *Find all images about field event*. As shown in Fig. 2, the ‘Field event’ is the sub-concept of the ‘Sport’ concept. Thus, the ‘Sport’ concepts are visited in order to retrieve information for the ‘Field event’ concept.

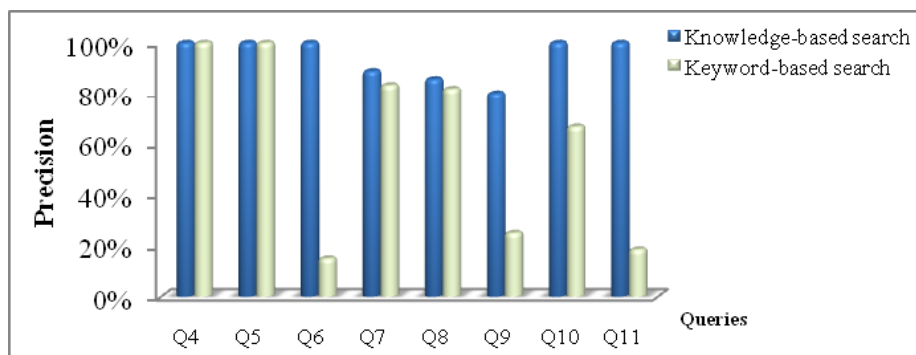
TABLE II
RECALL, PRECISION, AND F SCORE COMPARISON FOR TWO SEARCHING APPROACHES

Types of Queries	Queries	Recall		Precision		F score	
		Knowledge	Keyword	Knowledge	Keyword	Knowledge	Keyword
Simple	Q4	100%	100%	100%	100%	100%	100%
	Q5	100%	100%	100%	100%	100%	100%
	Q6	100%	100%	100%	15%	100%	26%
Complex	Q7	89%	56%	89%	83%	89%	67%
	Q8	100%	75%	86%	82%	92%	78%
	Q9	100%	8%	80%	25%	89%	13%
	Q10	63%	100%	100%	67%	77%	80%
	Q11	100%	63%	100%	19%	100%	29%
Average		94%	75%	94%	61%	93%	62%

Note: the KB-search = the Knowledge-based search, and the KW-search = the Keyword-based search.



(a)



(b)

Figure 7. Precision and recall comparison between the knowledge-based and keyword based techniques

D. Empirical Results and Evaluation

The proposed system was tested with 400 images and text captions taken from the Olympic website. The ontology was created during the knowledge extraction step with 2,550 instances stored in a RDF file. The experimental results were reported in three precision and recall graphs which are shown in Fig. 6 a, b, and c.

For Q1, the H1 hypothesis states that “*The ontology model can handle NL ambiguity*”. As Fig. 6a shows, the knowledge-based search is superior to a keyword-based search. The semantic search

supports the expression of more precise information, leading to more accurate answers. For example, in a keyword-based system, it is not possible to distinguish a query ‘USA athletes participate in Australia’ versus ‘Australia athletes participate in USA’ but it is possible with a semantic (SPARQL) query. All athlete instances and semantic relationships matched to the SPARQL query will be retrieved. The SPARQL query for Q1 is shown as follows:

```

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

```

This SPARQL query will ignore the ‘Australia athletes’ -participate - ‘USA’ relationship and other relationships which are not expressed in the query. This mechanism dramatically improves precision and recall compared to the keyword-based search. The H1 hypothesis is successfully evaluated.

The H2 hypothesis states that “*The semantic model can find the indirectly relevant concepts which are not identified directly in the document text*”. To evaluate this hypothesis, query (Q2) is specified using some keywords which do not appear directly in the text captions. In this example, a user wants to find images about a specific ‘field event’ e.g., the pole vault. As shown in Fig. 6b, the knowledge-base contains semantic relationships with sub-concepts of sport such as field event, track event and road event. The proposed system is thus able to recognize images annotated with a sport name which belong to the field event concept whereas the keyword-based search only recognizes a document as relevant if it contains words such as ‘field’ or ‘event’. This means that the knowledge-base search obtains better precision and recall than the text-based approach. In summary, the semantic model improves the retrieval performance significantly and hence this confirms the H2 hypothesis.

The H3 hypothesis states that “*Ontology-based searching provides an acceptable level of information retrieval performance even when Ontology variations are present*”. Fig. 6c shows that the performance of the knowledge-based search and the keyword-based search is similar. This is because the knowledge-based search could not find any images related to swimming images in the Sydney 2000. This triggers LSI to be activated and the results of LSI are used instead of the results of the knowledge-based search.

LSI computes the similarity of terms in a user query and Ontology model terms in a matrix. Fig. 6c shows that our framework obtains a slightly better performance compared to the keyword-based search because LSI can perform semantic search which result in it finding implicit relationships between keywords and images. Finally, the results are presented in descending order to users. Our proposed framework provides good retrieval performance even the knowledge base may not explicitly contain all the variations of ontological commitments seen in practice, hence H3 is validated.

Additional sample queries were also used for testing and gathering more information (Table I). The sample queries are classified into two categories. The first category (Q4-Q6) is related to a simple query structure containing keywords which are explicitly annotated in Ontology instances and in text captions e.g., sport name or athlete name.

In the second category (Q7-Q11), the specified query keywords do not appear explicitly in the image captions. Therefore, the search engine needs to find the implicit concepts based upon the semantic relationships among Ontology classes in the semantic model. The comparison metrics used for the knowledge-based search and the keyword-based search are precision, recall, and the F score. The queries are grouped into two categories. The average precision, recall, and F score of each query are presented in each table’s cell. The last record of the table shows the average precision, recall, and F score values for all queries. The corresponding numerical values are reported in Table II.

In Fig. 7a, the graph illustrates that the recall for the knowledge-based search and the keyword-based search are similar for the query in the simple category (Q4-Q6). This is because the keywords in the query explicitly appear in Ontology labels and text captions. Therefore, both search approaches can find relevant information easily and obtain a similar performance. However, when we examine complex queries (Q7-Q11), the recall for the knowledge-based technique outperforms the recall for the keyword-based technique which means more relevant images in a repository are retrieved compared to the keyword-based search. This is because the knowledge-based search is able to recognize images which have semantic relations with the query terms. For example, Q9, “*Find all images related to sport events in Spain*”. The knowledge-based search recognizes that images of Barcelona (host city) have a semantic relationship with Spain which is their host country whereas the keyword-based technique could not. Therefore, more relevant images are retrieved leading to a significantly improve recall.

However, the keyword-based approach obtains a better recall in Q10, “*Find all images about opening ceremony*”. This is because it retrieves all the images which their text captions contain “opening” or “ceremony” word. Thus, all the relevant documents about the opening ceremony are retrieved in addition to some other irrelevant images e.g., images about ‘medal ceremony’ or ‘closing ceremony’. Therefore, the keyword-based search obtains a higher recall but lower precision than the knowledge-based search.

For the precision graph (Fig. 7b), the knowledge-based technique outperforms a keyword-based technique, particularly in Q6-Q11. This is because the knowledge-based search retrieves relevant images more accurately than the keyword-

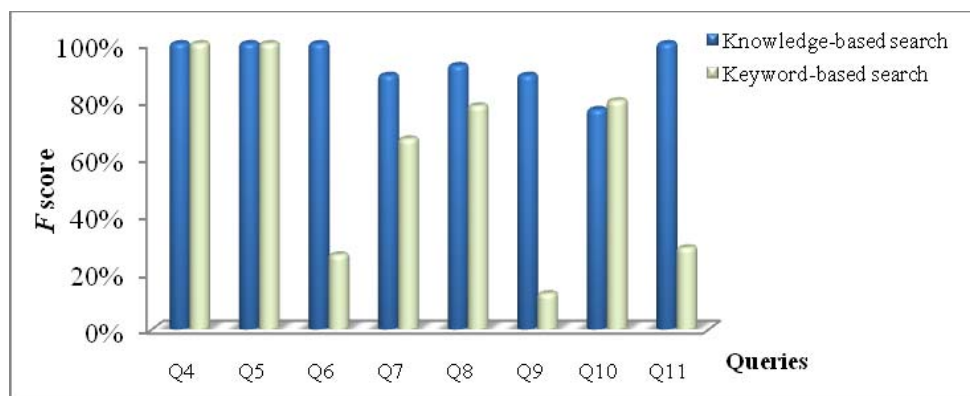


Figure 8. *F* score comparison for the knowledge-based and keyword based techniques

based approach. That said, the keyword-based technique retrieved all images that contain query keywords. Unfortunately, these retrieved documents might not be semantically relevant to the query. As a result, precision is reduced.

Finally, the *F* score of the knowledge-based search is much greater than the keyword-based search especially in Q6-Q11 (Fig. 8) except for Q10. Based on the average from all queries (Table II), 93% of all relevant documents in a collection are recognized by the knowledge-based technique whereas the keyword-based system retrieves only 62% of the relevant documents in a repository.

Hence, almost all documents which are retrieved by the knowledge-based search are relevant to the query, only 7% are irrelevant. In contrast, about 40% of retrieved documents from the keyword-based search are irrelevant to the query.

V. CONCLUSIONS AND FURTHER WORKS

This paper has proposed an IMR framework in order to support several major requirements for image retrieval using text captions: NL ambiguity, indirectly relevant concepts, metadata incompleteness and support for variable Ontology conceptualisations. We describe a framework to utilise semantic concepts within natural language image captions to aid image retrieval. Our innovation is to combine an Ontology-based model to restructure the semantic concepts in natural language captions. A key feature is that the hybrid combination of natural language and semantic restructuring degrades proportionately when a set of Ontology concepts is incomplete as this is compensated by LSI. We conclude that the main hypotheses of the work, that (Ontological) knowledge based techniques can significantly enhance the effectiveness of image retrieval system, have been validated. Although the proposed system is effective and it fulfils the requirements given in section II, some further challenges remain.

Currently, the variable Ontology conceptualisations and commitments used by different users are not well structured semantically, but rather grouped more in an ad hoc manner. We need to investigate how to create a more extensible structure that can support Ontology variations, e.g., whether or not the Ontology model should be partitioned into more specific types of interlinked sport models in order to model sports at a finer level of granularity. Ontology variations may not only be caused by different applications using different Ontology conceptualisations and commitments but also by different viewpoints of the same conceptualisation. These different viewpoints exist because different users may understand the conceptualisations at different levels of granularity and may use a different scope (horizontal coverage) and perspective (vertical coverage) for concepts.

The Ontology model used here is proprietary. The alignment of this ontology model against more standard conceptualisations for multimedia structures as defined in MPEG-7/21 needs to be investigated. Because several possible mapping from MPEG-7/21 XML structures into more semantic structures have been proposed by different researchers, we need to understand how these differ and which should be used as best practice.

In addition, Ontology models need to be maintained. In the real world, knowledge is not static, it often changes over time. Therefore, a more automated approach to Ontology maintenance is needed. A change in conceptualisation often needs to be reflected in the underlying domain Ontology. Consequently, these changes have effects on the performance and validity of the knowledge-based system. If the knowledge base is not updated in a systematic way according to clearly defined policies, the retrieval system may not include some relevant knowledge sources and can deliver incorrect answers to users. The issue of open-world versus closed-world semantics must also be considered. Therefore, we plan to expand this framework to improve the

reliability and consistency of semantic metadata management.

ACKNOWLEDGEMENT

This work was undertaken as a contribution to, and partially funded by, the EU FP7 project, My-e-Director 2012.

REFERENCES

- [1] Poslad, S., Pnevmatikakis, A., Nunes, M. et al. (2009) Directing Your Own Live and Interactive Sports Channel. 10th Int. Workshop on Image Analysis for Multimedia Interactive Services, WIAIMS'09, Special Session on Event, Behaviour Video Analysis for Interactive Multimedia Services, 6-8 May, London, 2009
- [2] Niblack, W., Barber, R., Equitz, W., Flickner, M., Glasman, E., Petkovic, D., and et al., "The QBC Project: Querying Images by Content Using Color, Texture, and Shape," *In Proc. of SPIE Storage and Retrieval for Image and Video Databases*, pp. 173-181, Feb., 1993.
- [3] Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., and Jain, R., "Content-Based Image Retrieval at the End of the Early Years", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 12, pp. 1349-1380, Dec. 2000.
- [4] Smith, J.R., and Chang, S.-F., "VisualSEEk: A Fully Automated Content-Based Image Query System", *In Proc. of ACM Int. Multimedia Conf. and Exhibition (ACM MM-1996)*, pp. 87-98, Nov. 1996.
- [5] Tseng, B.T., Lin, C.-Y., Naphade, M.R., Natsev, A., and Smith, J.R., "Normalized Classifier Fusion for Semantic Visual Concept Detection", *In Proc. of Int. Conf. on Image Processing (ICIP-2003)*, Vol. 2., pp. 535-538, Sep., 2003.
- [6] Paek, S., Sable, C., L., Hatzivassiloglou, V., Jaimes, A., Schiffman, B.,H., Chang, S.-F., and et al., "Integration of Visual and Text based Approaches for the Content Labeling and Classification of Photographs", *In Proc. of ACM SIGIR Workshop on Multimedia Indexing and Retrieval (ACM, SIGIR-1999)*, pp.15-19, Aug., 1999.
- [7] Paek, S., and Chang, S.F., "A Knowledge Engineering Approach for Image Classification based on Probabilistic Reasoning Systems", *In Proc. of IEEE Int. Conf. on Multimedia and Expo (ICME-2000)*, pp. 1133-1136, July-Aug. 2000.
- [8] Naphade, R.M., Kozintsev, I.V., and Huang, T.S., "A Factor Graph Framework for Semantic Indexing", *IEEE Trans. on Circuits and Systems for Video Technology*, pp. 35-39, Jan., 2000.
- [9] Naphade, M.R., and Smith, J.R., "Learning Regional Semantic Concepts from Incomplete Annotation", *In Proc. of Int. Conf. on Image Processing (ICIP-2003)*, Vol. 3., pp. 603-606, Sep., 2003.
- [10] Guarino, N., Carrara, and M., Giaretta, P. "Formalizing ontological commitments", *In Proc. of the 12th Int Conference on Artificial Intelligence*, Vol. 1, pp. 560 - 567, 1994.
- [11] Srihari, R. K., Chopra, R., and Burhans, D., "Use of Collateral Text in Image Interpretation," *In Proc. ARPA Image Understanding Workshop*, pp.897-907, 1994.
- [12] Frankel, C., Swain, M. J., and Athitsos, V., "WebSeer: An Image Search Engine for the World Wide Web," Technical Report TR-96-14, The University of Chicago, Illinois, Aug., 1996.
- [13] Hu J., and Bagga, A., "Categorizing Images in Web documents," *IEEE Trans. On Multimedia*, Vol. 11, Issue 1, pp. 22-30, Jan.-Mar., 2004.
- [14] Song, X., Ching-Yung, L., and Ming-Ting, S., "Autonomous Visual Model Building based on Image Crawling through Internet Search Engines", *In Proc. of the 6th ACM SIGMM international workshop on Multimedia information retrieval (MIR'04)*, pp. 315-322, Oct., 2004.
- [15] Wang, X., Ma, W., and Li, X., "Data-Driven Approach for Bridging the Cognitive Gap in Image Retrieval", *In Proc. of IEEE Int. Conf. on Multimedia and Expo. (ICME'04)*, Vol.3, pp. 2231-2234, Jun., 2004.
- [16] Schreiber, A., Dubbeldam B., Wielemaker J., and Wielinga B.J., "Ontology-based photo annotation", *IEEE Intelligent Systems*, pp. 66-74, 2001.
- [17] Hyvönen, E., Stynman A., and Saarela S., "Ontology-Based Image Retrieval," *In Proc. of the 12th Int. Conf. on World Wide Web (WWW2003)*, 20-24 May 2003 - Poster.
- [18] Tansley, R., "The multimedia thesaurus: Adding a semantic layer to multimedia information," *PhD Thesis University of Southampton*, Southampton, UK, 2000.
- [19] Benítez, A.B., "Multimedia Knowledge: Discovery, Classification, Browsing, and Retrieval," *PhD Thesis*, Columbia University, USA, 2005.
- [20] Chisholm, E., and Kolda T.G., "New Term Weighting Formulas for The Space Method in Information Retrieval", Technical Report Number ORNL-TM-13756, Oak Ridge National Laboratory, Oak Ridge, TN, March 1999.
- [21] Bojars, U., and Breslin G.J., "From Online Community Data to RDF," W3C Workshop on RDF Access to Relational Databases (RdfRDB Workshop), Cambridge, MA, USA, 25-26 October, 2007.

- [22] Manning, D. C., Raghavan P., and Schütze H., "Introduction to Information Retrieval," Cambridge University Press, ISBN-10: 0521865719. UK, 2008.
- [23] Khan, L., McLeod D., and Hovy E., "Retrieval effectiveness of an ontology-based model for information selection," *International Journal on Very Large Data Bases (VLDB)* Vol. 13, No. 1, pp 71 - 85, 2004.
- [24] Rui, Y., T.S. Huang, M. Ortega, and S. Mehrotra, "Relevance Feedback: A Powerful Tool for Interactive Content-Based Image Retrieval", *IEEE Trans. on Circuits and Video Technology*, Vol. 8, No. 5, pp. 644-655, Sept., 1998.
- [25] Zhou, X.S., and T.S. Huang, "Relevance Feedback for Image Retrieval: a Comprehensive Review", *Multimedia Systems*, Vol. 8, No. 6, pp. 536-544, Apr., 2003.
- [26] Barnard K., Duygulu P., Forsyth D., DE Freitas N., Blei D.M., and Jordan M.I., "Matching words and pictures", *Journal of Machine Learning Research*, Vol. 3, pp. 1107-1135, Feb., 2003.
- [27] Duygulu, P., Barnard K., DE Freitas J.F.G., and Forsyth D.A., "Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary," *In Proc. of the 7th European Conference on Computer Vision Copenhagen (ECCV02)*, pp. 349-354, Jan., 2002.
- [28] Mori, Y., H. Takahashi, and R. Oka, "Image-to-Word Transformation based on Dividing and Vector Quantizing Images with Words", *In Proc. of Int. Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM-1999)*, pp. 405-409, Jul., 1999.
- [29] Zhu, J., Uren, V., and Motta, E., "ESpotter: Adaptive Named Entity Recognition for Web Browsing", *In Proc. of the Intelligent IT Tools for Knowledge Management Systems (KMTOOLS 2005)*, pp. 518-529, Dec., 2005.
- [30] Miller, G. A. "Nouns in WordNet: a lexical inheritance system", *International Journal of Lexicography*, Vol. 3., No.4, pp. 245 - 264, 1990.
- [31] Berlin, F., Bizer, C., "D2R MAP-Map Language Specification," unpublished.
- [32] Zheng, W., Ouyang, Y., Ford, J., and Makedon., F., S., "Ontology-based Image Retrieval", *In Proc. of the 2nd WSEAS Int. Conf. on Non-linear Analysis, Non-linear Systems and Chaos (WSEAS2003)*, pp. 15-17, Dec., 2003.
- [33] Karanastasi, A., and Christodoulakis, S., "Semantic Processing of Natural Language Queries in the OntoNL Framework", *In Proc. of the IEEE Int. Conf. on Semantic Computing (IEEE ICSC)*, pp. 686-693, Sept., 2007.
- [34] Tsinaraki C., Polydoros P., and Christodoulakis S., "Interoperability support between MPEG-7/21 and OWL in DS-MIRF", *IEEE Trans. on Knowledge and Data Engineering (IEEE-TKDE)*, Vol. 19, No. 2, pp. 219-232, Feb., 2007.
- [35] Hunter J., "Adding Multimedia to the Semantic Web Building an MPEG-7 Ontology", *In Proc. of the 1st Int. Semantic Web Working Symposium (ISWC)*, pp. 261-281, 2001.
- [36] Garcia R., and Celma, O., "Semantic Integration and Retrieval of Multimedia Metadata", *In Proc. of the 5th Int. Workshop on Knowledge Markup and Semantic Annotation*, pp. 69-80, 2005.
- [37] Nack, F., Ossenbruggen, J., and Hardman L., "That Obscure Object of Desire Multimedia Metadata on the Web (Part II)", *IEEE Multimedia*, Vol.12. No. 1, pp. 54-63, 2005.
- [38] Troncy R., and Carrive, J., "A Reduced Yet Extensible Audio-Visual Description Language: How to Escape From the MPEG-7 Bottleneck", *In Proc. of the 4th ACM Symposium on Document Engineering (DocEng'04)*, pp. 87-89, 2004.
- [39] Dill, S., Eiron, N., Gibson, D., Gruhl, D., Guha, R., Jhingran, A., and et al., "SemTag and Seeker: bootstrapping the semantic web via automated semantic annotation", *In Proc. of the 12th international conference on World Wide Web*, pp. 178-186, May, 2003.
- [40] Noy, N., and Musen, M., "PROMPT: Algorithm and tool for automated ontology merging and alignment", *In Proc. of the 17th National Conference on Artificial Intelligence (AAAI'00)*, pp. 450 - 455, 2000.



Kraisak Kesorn has a BSc in Computer Science and a MSc in Information Technology (IT) from Thailand. He is currently a full-time student in the School of Electronic Engineering and Computer Science at Queen Mary (college) University of London, United Kingdom. He has been funded from Thai Government for his PhD. His research interests include semantic multimedia retrieval and Knowledge based model construction. He is working on image retrieval based upon ontology model.



Stefan Poslad has a PhD from the University of Newcastle upon Tyne, UK. He is lecturer in the School of Electronic Engineering and Computer Science, Queen Mary, University of London. His research interest are Ubiquitous Computing (Ubiquitous Computing book out in 2009), intelligent interaction involving the Semantic Web and Software Agents. He has led and been active in several international collaborative projects in these areas and has over 60 related research publications.