Information Fusion 20 (2014) 225–241

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

A Multi-Modal Incompleteness Ontology model (MMIO) to enhance information fusion for image retrieval

Stefan Poslad [b], Kraisak Kesorn [a,*]

[a] Computer Science and Information Technology Department, Science Faculty, Naresuan University, Phitsanulok 65000, Thailand
[b] School of Electrical and Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, United Kingdom

ARTICLE INFO

Article history:
Received 7 October 2013
Received in revised form 3 February 2014
Accepted 20 February 2014
Available online 7 March 2014

Keywords:
Multi-Modal Ontology
Knowledge base
Incomplete Ontology
Visual and textual information fusion

ABSTRACT

A significant effort by researchers has advanced the ability of computers to understand, index and annotate images. This entails automatic domain specific knowledge-base (KB) construction and metadata extraction from visual information and any associated textual information. However, it is challenging to fuse visual and textual information and build a complete domain-specific KB for image annotation due to several factors such as: the ambiguity of natural language to describe image features; the semantic gap when using image features to represent visual content and the incompleteness of the metadata in the KB. Typically the KB is based upon a domain specific Ontology. However, it is not an easy task to extract the data needed from annotations and images, and then to automatically process these and transform them into an integrated Ontology model, because of the ambiguity of terms and because of image processing algorithm errors. As such, it is difficult to construct a complete KB covering a specific domain of knowledge. This paper presents a Multi-Modal Incompleteness Ontology-based (MMIO) system for image retrieval based upon fusing two derived indices. The first index exploits low-level features extracted from images. A novel technique is proposed to represent the semantics of visual content, by restructuring visual word vectors into an Ontology model by computing the distance between the visual word features and concept features, the so called *concept range*. The second index relies on a textual description which is processed to extract and recognise the concepts, properties, or instances that are defined in an Ontology. The two indexes are fused into a single indexing model, which is used to enhance the image retrieval efficiency. Nonetheless, this rich index may not be sufficient to find the desired images. Therefore, a Latent Semantic Indexing (LSI) algorithm is exploited to search for similar words to those used in a query. As a result, it is possible to retrieve images with a query using (similar) words that do not appear in the caption. The efficiency of the KB is validated experimentally with respect to three criteria, correctness, multimodality, and robustness. The results show that the multi-modal metadata in the proposed KB could be utilised efficiently. An additional experiment demonstrates that LSI can handle an incomplete KB effectively. Using LSI, the system can still retrieve relevant images when information in the Ontology is missing, leading to an enhanced retrieval performance.

1. Problem statement

The main benefit of using knowledge representation models for Image Retrieval System (IMR), is that they are able to reduce the semantic gap, *the gap between the user perception and the low-level feature abstraction from the visual content,* providing relations between these low-level and high-level concepts can be identified, enhancing concept-based retrieval. Typically an Ontology model is used for knowledge representation, which represents physical things in this world using a hierarchical model expressed in the form of classes and relationships to support human decision-making, learning, reasoning and explanation.

An Ontology provides a useful way for formalising the semantics of the represented information. In principle, an Ontology can actually be the semantic representation for an information system in a concrete and useful manner [1]. Ontologies are used by an IMR for reducing the semantic gap by storing the knowledge structures for summarising, discovering, classifying, browsing, retrieving and annotating images. Ontology-based frameworks are proposed for IMR in numerous collections. Ontologies for manual image annotation and semantic retrieval for collections of animal pictures have

* Corresponding author. Tel.: +66 55 963267.
E-mail addresses: stefan@eecs.qmul.ac.uk (S. Poslad), kraisakk@nu.ac.th (K. Kesorn).

http://dx.doi.org/10.1016/j.inffus.2014.02.003
1566-2535/

been presented in [2]. An Ontology for considering art images has been presented in [3]. In [4], Ontologies also have been applied successfully for handling museum collections. These frameworks have validated the hypothesis that Ontologies can help to improve information retrieval effectiveness by making it possible to find semantically similar documents.

Gasevic et al. [5] has summarised the benefits of using Ontologies in IR and IMR systems as follows. Firstly, the Ontology structure can be exploited to measure the semantic similarity. For example, the term list (''Michael Phelps'', ''Swimming'', ''Gold medal'') has no syntactic similarity to the term list (''London'', ''2012'') although the two lists are semantically relevant. This is because Michael Phelps was the Swimming (Gold medal) winner at the London 2012 Olympic Games. The similarity can be obtained using the relationship between concepts, e.g., Michael Phelps-<*is_champion_of*>-Free_style_swimming-<*participate_in*>-London 2012. Secondly, semantic annotations that may not be explicitly mentioned in a caption can be identified using knowledge stored in an Ontology. For example, if many entities, locations and athletes related to the London 2012 Olympic Games appear in a text caption, and also the time context is London 2012, the annotation system can infer that the London 2012 Olympic Games itself is relevant to an image. Hence, this could be added to the semantic annotation although the text caption does not explicitly contain the time context ''London 2012''. Ontologies may also be used to enable query expansion [6,7] but this will not be described in detail here. These are some of the potential uses of Ontologies in IMR.

Existing work on IMR tends to be based upon only single-modality information, either textual information or visual features. Consequently, such work suffers from several limitations. For example, an IMR system is not able to describe the high-level semantics of images, based only on any distinctive low-level visual features when text descriptions of images are not supplied, because the extracted visual features themselves cannot be used to represent the content of images effectively. Text and image are two different modalities that can be fused to represent 'things' more completely. However, there are some invariant and implicit connections between them that complicates the information fusion of the two [8]. Often, the textual information surrounding images includes descriptions of images generated by humans. These image captions should not be disregarded, as they can aid image interpretation. Nonetheless, exploiting only text information for visual content interpretation can suffer from the ambiguity of the text descriptions used because they are written using natural language, which may be ambiguous and imprecise. As such, using single-modality information is not adequate to enhance the interpretation power for IMR. Multimodality information should be utilised to facilitate image interpretation, classification and retrieval.

Fusing text and image features has been proposed by several researchers in order to improve the image search results [9–15]. These approaches focus on improving the retrieval performance in order to get more accurate results. However, there are some challenges in integrating visual and textual metadata in a knowledge base for IMR. Firstly, the KB model should be designed to interlink both visual and textual metadata together, in order to facilitate the image classification and retrieval performance. Secondly, automatic KB construction and metadata extraction from text captions are very challenging tasks from which to build a complete KB due to several factors:

(1) Those text captions may be ambiguous because they are written using natural language.
(2) Standard Ontology languages such as the W3C Resource Description Framework (RDF) or Web Ontology Language (OWL) cannot directly represent some semantic aspects, e.g. uncertainty and gradual truth, value) because the latter hard-wires a specific logic, description logic, into the Ontology representation.

These are the 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. In this paper, Ontology incompleteness refers to an absence of some semantic metadata and also to 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 systems that use information directly from text captions. This is because of the inadequate coverage of annotations by a KB [16]. As such, IMR 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 knowledge-base for use by an image retrieval system and are, as such, the main focus in this paper.

2. State of the art

The current survey focuses on a discussion of Ontology-based frameworks for IMR that use a shared or a standard encoding, i.e., MPEG-7.

2.1. Ontology-based frameworks for IMR

Numerous techniques have been introduced to resolve the semantic gap problem in the past decade. Early IMR approaches were based on low-level features which fail to capture the underlying conceptual associations in images. 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. 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 are 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 [9–13]. In part through the emergence of the Semantic Web [17], an Ontology model has become common to represent visual content, enabling an IMR system to perform semantic retrieval.

Tansley [18] proposed a method to bridge the semantic gap using Web images and their surrounding text, file name and alterative tags. Using the WordNet Thesaurus [19], the system can solve the NL vagueness problem of text captions. Unfortunately, this framework exploits only textual information and supports only text-based queries. Schreiber et al. [2] presented a method to index and search image collections using Ontologies. The system uses the terminology from WordNet for annotations. The main limitation of this framework is that the knowledge-base is designed as a closed KB, because if a query concept is outside the scope of KB, the system cannot find any relevant images for users. Dasiopoulou et al. [17] presented a framework comprising two main modules, a semantic analysis module and a retrieval module. The domain Ontology provides the conceptualisation and vocabulary for structuring content annotations. 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. The system exploits the low-level features of an image and matches

the extracted low-level features to a 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 failure to annotate the image. Wang and Ma [20] introduced *m*-PCNN, a multi-channel Pulse Coupled Neural Network, for medical data fusion to diagnose diseases. Similar to work in [21,22], *m*-PCNN attempted to integrate the information from two or more images into a single one to improve image analysis for molecular biology and medical image analysis. However, these methods rely only on visual data. Thus, the efficiency of image analysis can decrease when objects are viewed using different camera angles, rotations and sizes. Kesorn et al. [23,24] proposed a method to utilise local low-level features, e.g., Scale-Invariant Feature Transform (SIFT) descriptors, to represent visual content in order to aid image interpretation. However, the main drawback of this technique is the computation cost in order to analyse images in the collection. The interpretation processes performed are based upon vector space model and cannot describe visual content as explicitly as a hierarchical (knowledge-based) model that models conceptual structures and relationships. In addition, the framework exploits only single modality information (visual features) for image interpretation. As a result, the framework can support only a visual-based query and fails to find relevant images using a text-based query.

Multimodality information fusion for image retrieval is emerging as a new and promising research area in recent years. It aims to integrate multi-modal information, e.g., text and visual features, to obtain more accurate retrieval results. The fusion between textual information and image features has been proposed to improve the image search results, for example [9–13]. These approaches only focus on improving the retrieval performance to get more accurate results. However, they ignore the Ontology coverage (completeness) problem. Wang et al. [14] supported multi-modal, text and visual information, in the *canine* domain. A binary histogram is used to represent each of the image features. It is transformed into an Ontology model using a hierarchical Support Vector Machines (SVM) classification [25,26] and incorporates a textual description Ontology. The proposed method is able to increase classification accuracy and retrieval performance. Khalid et al. [15] proposed a multimodality ontology framework for the sport domain. Textual descriptions and any surrounding text are extracted and are then manually mapped to concepts in a domain knowledge base. For visual content, low-level features, e.g., colour layout, dominant colour and edge histogram descriptors, are extracted. These visual features are then classified into categories using a SVM classification technique and a framework called the Label Me Annotation Toolbox [27]. Global features, e.g., colour and edge information, cannot represent the semantics of images effectively. For example, the same images have a different brightness, size, and camera angle. This is a well-recognised problem, the so called visual heterogeneity problem, for researchers in the image-processing area. The work most similar to our idea is the work of Chen et al. [28] who bridge the semantic gap by integrating text and visual features. Their main research motivation is that the lack of feature integration can increase the semantic gap in CBIR. Their framework deploys global low-level features, e.g. colour, shape, texture and edge, to construct a visual thesaurus which can infer the visual meaning behind a textual query. Their experiments show that the proposed method significantly improves retrieval performance and can understand user intention. The main drawback of this method is that global features are not reliably transformed under changing image illumination, rotation, and size. Thus, local features, i.e., SIFT descriptors, are investigated and used to solve this problem.

From the analysis of the existing solutions, some limitations still exist as follows:

- They 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. They ignore the incomplete Ontology problem or the Ontology evolution or maintenance problem.
- They are unable to handle visual heterogeneity. For instance, when the surveyed systems map lower-level features onto a higher level object conceptualisation, an extracted feature may possibly belong to multiple concepts of objects leading to visual heterogeneity (one visual appearance has multiple meanings). Therefore, an image representation model should support this requirement.

2.2. MPEG-7 versus Ontology-based systems

MPEG-7 is a standard specification for describing multimedia documents. 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 [29]. 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 represent 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. Although, MPEG-7 has become a standard for multimedia search and retrieval, the MPEG-7 structure is not a preferred choice for the KB in this paper for several reasons.

First, syntactic interoperability although Semantic Web as proposed by World-Wide Web Consortium (W3C) is an important framework for developing Internet-based information systems, the combination of the use of Semantic Web and MPEG-7 can cause a lack of syntactic interoperability [30]. This is because different languages are used, e.g. XML, MPEG-7 Description Definition Language (DDL), Resource Description Framework (RDF), RDF Schema (RDFS) and the Web Ontology Language (OWL). However, because MPEG-7 DDL merely adopts an XML Schema, i.e., it represents structure in the form of schema by defining syntactic elements. However, the MPEG-7 DDL lacks particular media-based data types. Unlike RDF Schema or Ontology-based modelling, the MPEG-7 DDL lacks support for the definition of semantic relations, e.g. Beckham-plays-football. Combining a language syntax with a schemata semantics for MPEG-7 is still an open issue.

Second, semantic interoperability MPEG-7 DDL design is based upon XML Schema rather than upon RDF Schema. As a syntax-oriented language, MPEG-7 DDL provides weak or light-weight semantic support, supporting only named attributes and unnamed hierarchical relationships [30]. Therefore, this DDL cannot facilitate reasoning services efficiently, especially in subsumption-based reasoning on concept and relationship 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 are provided MPEG-7 is an XML-based schema that expresses syntax level aspects. No formal semantics are provided so that applications can interpret the meaning of

image descriptions [31]. Finally, the goal of MPEG-7 in this semantics and interoperability context is still questionable. One needs to question if its objective is to be an exchange format or if it is a machine understandable document that can be processed for multimedia descriptions. In contrast to MPEG-7, Knowledge-based models built using Ontologies are a very widespread R&D topic, not only in AI, but also in other disciplines of computing. An Ontology-based KB provides a number of useful features for knowledge representation in general. This paper summarises the most important of these features based on the surveys from [32–36].

First, vocabulary, an Ontology provides the names for referring to concepts or notions of a domain of interest. Ontologies provide *logical statements* that describe not only what the concepts are, but also how they can be related to each other. It is not only the vocabulary that quantifies an Ontology, but the conceptualisations that the concepts in the vocabulary are intended to capture [35]. In addition, Ontologies are usually designed to specify concepts with *unambiguous meanings,* with semantics that are independent of readers and any context.

Second, taxonomy a *taxonomy* is a hierarchical categorisation or entity classification with class/subclass relations [5]. When the taxonomy is used together with the vocabulary of an Ontology, they provide a *conceptual framework* for analysis, discussion, and information retrieval in a domain.

Third, knowledge sharing and reuse the main objectives of Ontologies are *knowledge sharing* and *knowledge reuse* by applications [5]. This is because Ontologies provide a consensual narration of the concepts and relationships in the domain and this information can be shared and reused among applications.

Hence, an, Ontology-based KB is the preferred choice for storing multimodality information in this research. In the next section, our Ontology design to overcome the limitations of current state of the art methods, is introduced.

3. Method

3.1. Knowledge-based design for information fusion

Knowledge representation process starts with Ontology design and modelling. As such, it relies on a common understanding of how humans understand, acquire and represent knowledge. In order to manage facts and ideas, humans usually transform their knowledge into a structured model: Similar things are grouped together regarding certain common properties or characteristics, which define an abstraction that is used 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 "subclass_of" relations). The process of conceptualisation of a domain has four main phases: define the concept taxonomy, define a set of relations connected among concepts, define the constraints for the values in a relation and define axioms on the relations and concepts. Typically, a knowledge model that is created requires a subsequent refinement process in order to validate and rectify it. This process is repeated until the system has achieved the desired level of performance.

In this research, a knowledge-base for Sport is constructed. The structures and relationships for the sports-domain Ontology 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 20 genres for Athletics sports as this is sufficient to bring a number of challenges to the proposed system. First, the visual appearance of events for different types of athletics is quite similar. It is very challenging for a system to categorise properly based upon the extracted low-level features. Second, some objects that appear in images are shared between two or more types of athletics, e.g. a horizontal bar can appear in the high jump and in 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 the polysemy issue. After surveying data at the Olympic organisation website, three main classes of Ontology have been defined, a Sport Domain, a Visual Features and a Text Descriptors Ontology. 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 and image size. This Ontology is designed to support image interpretation. The *Text Descriptors Ontology* provides the annotation structure template for the (sports) domain.

The hierarchy in the KB is designed as follows: Level i of the hierarchical model is a generalisation of the concepts of level $i-1$. Some object classes are decomposed into subclasses. Edges in a concept graph represent several types of relationships between classes e.g. is-a, participate-in, and part-of. The Ontology structure is designed as follows. First, redundant concepts are minimised. For example, the Visual feature Ontology is designed as main concepts (see Fig. 1a) rather than as subclasses of a concept in Textual descriptors or Sport domain Ontology (see Fig. 1b). This is because subclassing the visual feature concept from other classes can lead to concept redundancy For example, there are two concepts, SIFT and Colour in Fig. 1b. Second, the Ontology structure more efficiently facilitates sport image retrieval (Fig. 1a. For instance, "*athletes who perform Track event*", the structure in Fig. 1b may make it more difficult to find an answer. For Fig. 1a, the system easily finds the answers to this query by following the route *Athletes-<perform>-Running-<is-a>Track* and then, all athletes who perform all sports which are under the Track concept will be retrieved. In contrast, the structure in Fig. 1b can make it impossible to answer this simple query because there no link between *Running* and *Track* event. The Ontology structure in Fig. 1b does not model the concepts and relationships in the real world and application satisfactorily. Thus, this structure is less effective and scalable when the Ontology covers a large number of sports. The search performance will be degraded. Typically, upper level concepts are a generalisation of lower level concepts. However, in Fig. 1b, the *Athletes* concept is not a generalisation of the *Running* class as they do not share any common properties. Therefore, the structure and relationships in Fig. 1a are preferred in order to represent the sports domain information.

3.2. A KB design to tackle the ontology incompleteness problem

Ontology incompleteness refers to the absence of some semantic metadata and also to relationships between concepts that cannot be represented in an Ontology. This type of KB architecture is called *a closed KB*. 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 KB model implies that any data that is not present in the KB is *false*, while an *open KB* model states the data that is not presented in the KB is *unknown* or *unresolvable* [37]. 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 [38]. 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 advantage of this approach is that it can find information

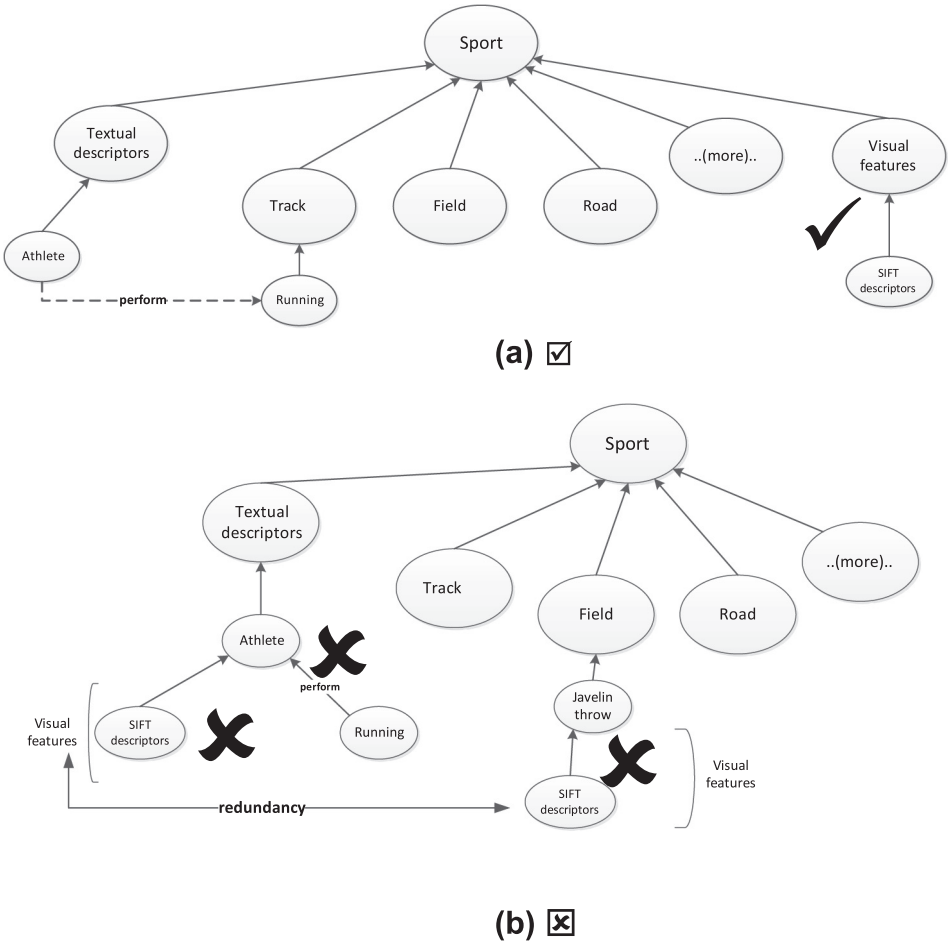

Fig. 1. Ontology design choices.

or images not included in the training set; annotation keywords come from a web-based search engine. In our study, the proposed system is designed to overcome the Ontology incompleteness problem using Latent Semantic Indexing (LSI) [39]. LSI was selected to handle this problem because it enables a semantic search based on textual information in image captions. LSI is able to discover the relevant information which does not explicitly appear in the document text and can enhance the retrieval efficiency compared to a conventional text-based search (string matching). In addition, it is able to handle the problem of synonymy and ambiguity of words. An unresolvable query will be forwarded to a LSI module in order to perform a second search on the LSI vector space. 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. LSI begins by constructing a term-document matrix, A, to identify the occurrences of the m unique terms within a collection of n documents. In a matrix A, each term is represented by a row, and each document is represented by a column, with each matrix cell, a_{ij}, initially representing the number of times the associated term appears in the indicated document, tf_{ij}. This matrix is usually very large and very sparse. Once a term-document matrix is constructed, local and global weighting functions can be applied to it to condition the data. The weighting functions transform each cell, a_{ij} of A, to be the product of a local term weight, L_{ij}, which describes the relative frequency of a term in a document, and a global weight, G_i, which describes the relative frequency of the term within the entire collection of document. More details about how LSI is applied are given in Section 4.2.4.

4. Theoretical framework for multi-modal information indexing

4.1. Automatic knowledge acquisition

Automatic Knowledge Acquisition extracts knowledge from both *visual data* and *text captions* and stores this extracted information in a knowledge-based model. In order to acquire knowledge from both these information sources, firstly, the low-level features are extracted and processed using a bag-of-visual words (BVW) technique in order to recognise objects in images. Later, the extracted visual information is mapped to a higher-level semantic conceptualisation based on the Ontology model. Second, image captions are analysed and fused with visual information in the unified knowledge-base in order to enhance the image interpretation and retrieval. Fig. 2 illustrates the main processes of the knowledge-acquisition framework.

4.1.1. Visual feature extraction and analysis

Two main processes for visual analysis are defined to compute higher level representations from lower level ones. First, low-level image processing extracts useful visual features from images and interprets them as primitive objects. Low-level image processing comprises several steps and is often called "image analysis". Second, *higher level semantic interpretations* are identified based upon the primitive objects and upon specific prior knowledge (knowledge about a sports event, which is not explicitly addressed in

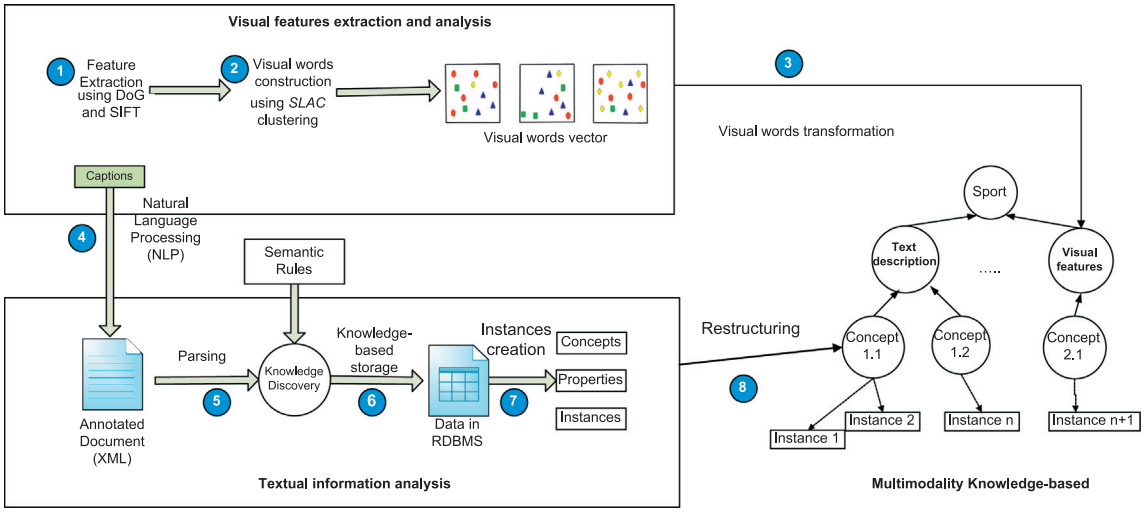

Fig. 2. The knowledge acquisition processes.

the data) that is relevant for the interpretation. This information is integrated to enhance image classification. Here, a novel idea to represent a higher level conceptualisation of visual data derived from lower level features is presented.

This paper 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 [40]. However, major limitations of existing BVW models include that they do not preserve the semantics during visual word construction. Hence, this paper presents a new representation model that resolves the above difficulties and that enhances image retrieval efficiency. We already proposed the idea to extract knowledge from visual data in [23]. However, we extend our idea to support multimodality information to the framework presented in this paper. There are three main steps to perform the visual analysis (Fig. 2, steps 1–3): low-level feature extraction, visual word construction and restructuring a vector space model into a hierarchical structure. These steps are described in more detail in the following sections.

4.1.1.1. Low-level feature extraction. Local patches of images are extracted and they are considered as the candidates for the main visual words. In this paper, the Difference of Gaussian (DoG) detector [41] is used for the automatic detection of keypoints from images (Fig. 2, step 1). The patches are represented as numerical vectors that represent feature descriptors. Each patch is converted into a 128 dimensional vector using a SIFT descriptor. Hence, each image is represented as a set of vectors of the same dimension, independent of the vector order.

4.1.1.2. Visual word construction. Similar SIFT descriptors are assigned into the same group exploiting clustering algorithm. Each cluster represents the shared local pattern of keypoints, a so called "visual word". Those visual words are used later to produce a BVW model represented in the form of a vector. To construct a visual word, a clustering algorithm is deployed e.g. *k*-mean algorithm. The major drawback of *k*-mean is that it appears to be unaware of the spatial location of keypoints. As a result, semantic information between the low level features and the high level semantics of objects in the visual content is lost. As such, we propose to use the Semantic Local Adaptive Clustering (SLAC) algorithm [42] to cluster the vectors. SLAC improves the conventional clustering techniques by capturing relevant local features within a cluster. Hence, it can find the semantically similar keypoints and cluster them into the same group using a similarity matrix (Fig. 2, step 2).

Keypoints are clustered based on the degree of relevance. Thus, SLAC generates semantically related visual words. Consequently, the quality of the visual words obtained is enhanced compared to those obtained using conventional models [24]. The number of the clusters is the bag of visual words' size (BVW). The BVW representation is similar to the bag-of-words representation technique in text document as both can be converted into a vector space model.

4.1.1.3. Visual word vector space model restructuring. Existing systems [43–45] have disambiguated word senses by restructuring visual words into a hierarchical model. These methods deployed state of the art clustering algorithms e.g. Hierarchical Spatial Markov model [45], Agglomerative clustering algorithm [46], and hierarchical Latent Dirichlet Allocation algorithm, or hLDA [47]. Nevertheless, there are some drawbacks of these algorithms which affect the image interpretation efficiency. First, the generated hierarchical model is usually a binary tree. This is not always an effective representation model for visual content data. Typically, the number of relationships between concepts is more than a binary one. Second, there is no multiple-inheritance between parents and a child node such that a child node inherits properties from several parent nodes. 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 (Heptathlon is_subclassOfField_sport and Track_sport). As such, the semantic information of images is not effectively represented using a hierarchical model in existing frameworks. Instead of integrating several visual words in order to distinguish word senses or using a binary tree model, this paper proposes restructuring a conventional vector space model of visual word into a hierarchical model to overcome the aforementioned limitations. Furthermore, the MMIO framework can improve the quality of the image annotation, classification and retrieval efficiency of IMR. We proposed an idea for visual feature restructuring in [23] and apply it here.

Visual words extracted from visual content are different from words extracted from a text document. Typically, the word sense can be disambiguated through exploiting WordNet. Unfortunately, WordNet is not applicable in this case because linguistic information is not supported by visual words. Therefore, we propose a technique called *concept range* to classify visual words into concept(s) in an Ontology-based structural model (Fig. 2, step 3).

First, the objects of interest are manually extracted from other objects and background and then they are processed to extract the keypoints. The idea behind this process is to eliminate noise from the background. As a consequence, all extracted keypoints from the same object are considered semantically relevant and, thus, visual words are constructed from these semantically keypoints [48]. Thus, the linkage between low-level features and high-level conceptualisations of each object category is preserved. However, manual object separation is done for training purpose only and this allows the system to learn and recognise such an object effectively. Thus, we obtained a bag of visual words (ϖ) and $\{\varpi_i \in C_i\}$ for each object category C_i which substitutes various views for several parts of an object. Having obtained a bag of visual words, the concept range of each object will be calculated. The range (r_i) of a concept i is the maximum distance between the centroid (v) of a visual word (ϖ) and the centroid of concept (c_i), the so called *concept range*. The concept range [45,48] is defined as follows:

$$r_i = \max |v - c_i|, \quad v \in \varpi \tag{1}$$

The main benefit of the concept range is to distinguish the senses of visual word and to increase the image classification power. The different senses of a visual word can be disambiguated using a concept range, see Eq. (1). A visual word can be considered as multiple concepts if its centroid is in the range of those concepts. Therefore, this technique can represent facts more effectively than existing techniques when words can have multiple meanings. Some visual words will be discarded if they do not match to any concept in the structural Ontology model, as shown in Fig. 3. As a result, this method 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 visually similar. In other words, the visual appearance of an object is *invariant* using the proposed method.

The detection of key objects in an image is related to the frequency of visual words that 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 image contains that object. Because of scaling differences of images in the collection, using $f(v_i)$ directly could be erroneous. Hence, $f(v_i)$ is normalised in order to compensate for discrepancies in the frequency of the visual words. The normalisation formula is shown in Eq. (2), where N is a number of instances of visual content.

$$\eta_i = f(v_i) \Big/ \sum_{j=1}^{N} f(v_j) \tag{2}$$

4.1.2. Textual information extraction and analysis

Fig. 2, step 4–8 shows the main processes of the textual information analysis processes executed by the Textual Information Analysis (TIA) module. There exist text descriptions accompanying some images that can be useful for image classification. The main function of TIA 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. HTML files are processed to extract the important textual information (e.g. date, time, place, person name, and event) and to temporary store this metadata in the tables of a relational database. Second, this metadata is transformed into 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 OWL file that stores the semantic metadata.

4.1.2.1. Document processing. First, a Natural Language Processing tool (NLP) is used for the initial metadata generation (Fig. 2, step 4). As NLP innovation is not the focus here, an established NLP framework, ESpotter, is deployed rather than implementing a new NLP tool. ESpotter [49] provides a function for a Named Entity Recognition (NER) task, e.g. person name, location, date, and other proper nouns. ESpotter produces the semantic metadata in the form of XML format. Then, this annotated document is processed and the metadata extracted.

4.1.2.2. Knowledge-based representation for textual information. The knowledge representation (KR) in this paper is a set of ontological commitments. In addition, KR supports efficient machine-readable and machine-understandable computation about knowledge. XML metadata is parsed and stored in a Relational Database Management System or RDBMS (Fig. 2, step 5). Thereafter, this data is restructured into an Ontology, represented in OWL, in order to be used in a semantic context. To export initial metadata from a relational database into OWL, the relational database model has to be mapped to the concepts in a knowledge-based model. In this research, data from a RDBMS database is directly mapped to OWL using a JDBC connector API (Fig. 4). This generic approach is useful in numerous cases, but sometimes is complicated by difficulties: in synchronising it to changes, in the database structures, and in installation and use by inexperienced users. The mapping process is shown in Fig. 4. To transform information in a relational database to OWL, three steps are performed.

(1) The data in the RDBMS is retrieved (Select and Group by column) using Structured Query Language (SQL) queries.
(2) The Jena API (http://jena.apache.org) is deployed to construct Ontology concepts, properties and instances. Jena provides off-the-shelf methods for creating Ontology classes, properties, and assigning instances to these classes and properties.
(3) A URI or a node identifier is assigned to those generated instances. Later, the instances' properties are created and assigned property values and written to an OWL file.

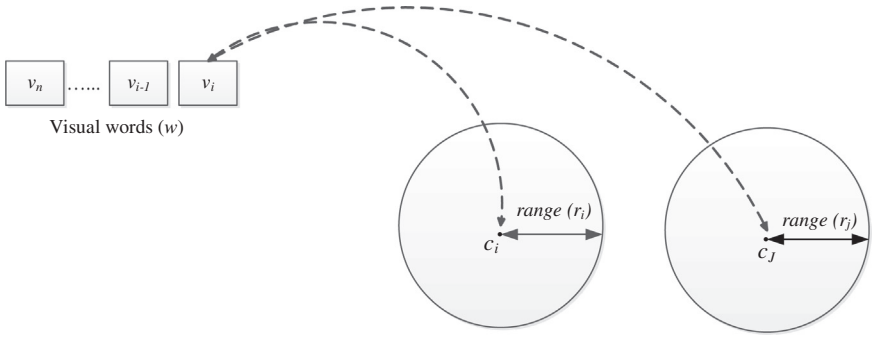

Fig. 3. Visual words are assigned to concept(s).

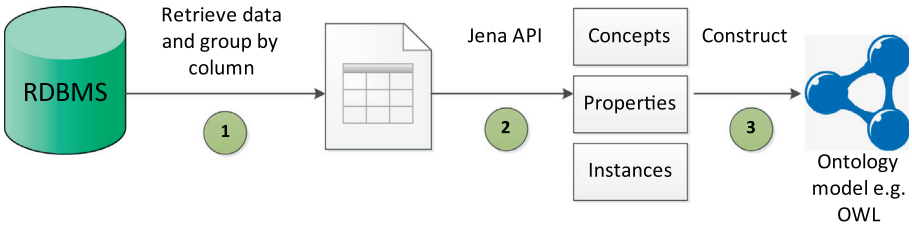

Fig. 4. Transformation process from an RDBMS data structure to an Ontology model (OWL).

4.2. Multi-Modal ontology structure

Since both unstructured textual and visual information are transformed into hierarchical structures, they are fused into a unified knowledge model, so called Multi-Modality Incompleteness Ontology (MMIO), to enhance the retrieval performance. The main motivation for the enhancement is because MPEG-7 does not support machine processable semantic annotation of image subject matter [50]. Therefore, a huge effort [51–54] has been undertaken to transform or integrate MPEG-7 and Ontolgoies to resolve such a problem. In this section, we briefly describe the structure of multi-modality Ontology. Fig. 5 shows the structure of presented Multi-Modal Ontology.

4.2.1. Sport event ontology

The Sports event Ontology provides the vocabulary and background knowledge for sports, e.g. sport name, genre, and discipline. Classes and relationships defined in the sports Ontology are extracted from the Olympic website (http://www.olympic.org), which provides standard descriptions and defines relationships for various aspects of sports. Fig. 5 illustrates the structure of the presented Ontology. An ellipse represents pre-defined classes in the Ontology and a rectangle indicates Ontology instances. Some parts of the Ontology are omitted due to space limitations.

4.2.2. Textual description ontology

The Textual Description Ontology encapsulates image annotations. This Ontology contains concepts used in annotations of sports images, e.g. athlete, game, host city and host country. In other words, it provides metadata to answer three types of queries associated with what, when, and where.

4.2.3. Visual features ontology

This Ontology represents the metadata about the image itself e.g. format (jpg, bmp), size, resolution of a picture, and concerning the content in terms of low-level features (visual words). This

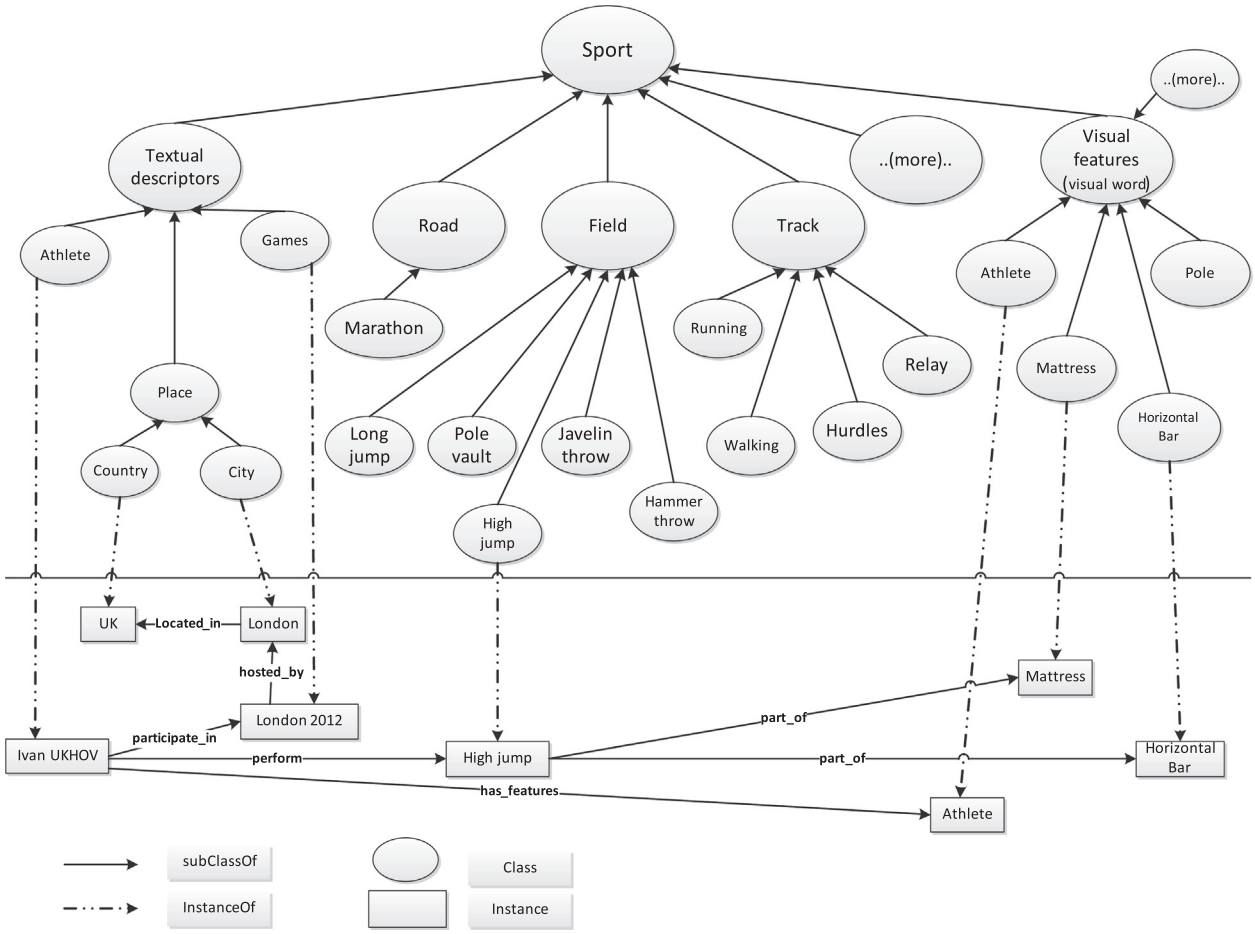

Fig. 5. Multi-Modal Ontology structure.

Ontology is constructed using the method described in Section 4.1.1 and incorporates the textual description Ontology in order to enhance the retrieval performance.

4.2.4. Handling the incompleteness of the KB using Latent Semantic Indexing

As mentioned previously, building a real world domain Ontology to cover everything in that domain is very challenging. A system that relies only on information in the KB, will not return any answer to users' queries when there is no relevant information stored in the KB. Thus, the system should be designed to cope with this uncertainty. For example, it can try to answer such an *unresolvable* query in a different way when it fails to find answers in the KB. However, there are some challenges in answering unresolvable queries. The system should still provide a capability to support a semantic search based on textual information in an image caption. In addition, it should be able to handle the synonym problem and ambiguities in some words. To handle these challenges, Latent Semantic Indexing (LSI) is proposed as a technique to cope with the uncertainty in the KB. LSI is a well-known technique used to enhance text-based information retrieval [39]. LSI tries to search for things that are closer to representing the underlying semantics of a document [55].

The LSI vector space model is used as a backup method when the system cannot find the desired information in the Ontology model. Here, we modify the LSI technique 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. 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 in descending order according to the relevance value. The LSI process starts with a 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 tokenisation 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 'noise' words from the tokenisation step are not useful or important, e.g. a, an, the, without and before. Therefore, they are removed from the set of words. The remaining text, after removing the stop words, is assumed to be the keywords that characterise the image and that are stemmed from the original form of each word. Next, the term frequency is calculated, i.e., the numbers of occurrences of each term that appear together with the image are counted. The Degree of Importance between the term and the image can be derived from these frequencies and represented in the form of a matrix. Rows of a matrix represent keywords whereas columns refer to image instance that contains those keywords using image IDs. Fig. 6 shows an example of matrix A with a term frequency in each image. This matrix is also referred to as a vector space model.

Having created the term frequency matrix, a *Term-image relationship computation process* computes and assigns a weight to each term using Eq. (3):

$$W_{ij} = L_{ij} G_i N_j \tag{3}$$

where W_{ij} is 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 by [56].

Local weight calculates the frequency of each term appearing in a document and in a query. To compute the local term weighting, the Logarithms scheme is used to adjust the weight of a term in documents that have different lengths. Local weight can compute using the following formula (Eq. (4)).

$$L_{ij} = \begin{cases} \log f_{ij} + 1 & \text{if } f_{ij} > 0 \\ 0 & \text{if } f_{ij} = 0 \end{cases} \tag{4}$$

where f_{ij} is the frequency of term i in image caption j. The global weight computes the frequency of each term appearing in the entire collection. The main idea of global weight is that a term appears less in the collection, the more important the weight it is. To compute a global term weighting, the Inverse Document Frequency (IDF) method is deployed. The global weight formula is illustrated in Eq. (5).

$$G_i = \log\left(\frac{N}{n_i}\right) \tag{5}$$

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 differences in document lengths. Eq. (6) shows the normalisation formula.

$$N_j = \frac{1}{\sqrt{\sum_{i=0}^{m} (G_i L_{ij})^2}} \tag{6}$$

The reasons to normalise each term are as follows. Firstly, long documents tend to gain a higher score since these tend to have more words and more occurrences of each word. These tend to have a high score simply because they are longer, not necessarily because they are more relevant. Secondly, 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 the formula in Eq. (7) in order to reduce the matrix dimensions.

$$A = USV^T \tag{7}$$

Having computed the SVD, the columns in U are reduced to k and the values in matrix S are sorted in ascending order. In other words, the matrix U indicates the semantic importance of concepts. This is from where the term "latent semantics" comes from. Unimportant concepts are regarded as "semantic noise". To reduce

$A =$

	Img1	Img2	img3	img4	img5
Keyword1	1	1	1	0	1
Keyword2	0	1	1	1	1
Keyword3	1	1	0	1	1
Keyword4	0	0	1	0	1

Fig. 6. Example of a term frequency matrix.

the dimensions of the matrix, only the largest k singular values will be selected. Therefore, the decomposition can be approximated as Eq. (8):

$$A \approx U_k S_k V_k^T \tag{8}$$

where U_k contains the top k most important concept vectors sorted by ascending order, S_k has singular values and is diagonal and V_k^T has rows that are the right singular vectors. The SVD approximation, the so-called *Rank-k approximation*, can be created by selecting only the first k columns of U and the first k rows of S and V^T as illustrated in Fig. 7.

The main advantage of LSI is the ability to infer indirect concepts when those concepts are not explicitly mentioned in the text captions. For example, consider the image caption, "*Jessica Ennis wins the gold medal in London 2012*". In this example, the image caption contains an athlete name "Jessica Ennis" but no sport genre is explicitly addressed in the caption. However, LSI can find the missing sport genre of Ennis using relationships between terms in the vector space model. Based upon historical data, Ennis usually participates in running events and therefore "Ennis" and "running" have a strong relationship. Thus, the LSI algorithm can introduce a new term, *running*, to the image automatically. When a user submits a query for "running" images, LSI will consider this as a relevant image and it can be returned to the user. Therefore, the system does not rely solely on metadata in the Ontology and can handle the Ontology incompleteness problem efficiently. This architecture of KB can be considered as a type of open KB.

5. Experimental settings, evaluation protocols and research hypotheses

The experimental method proposed to evaluate the MMIO model is as follows. We constructed an image collection in order to evaluate the MMIO model because sports images are not readily available in such standard test collections. Therefore, a new test collection needs to be created. It was decided that images obtained from the Google image search engine could be used to form a sports domain image collection to develop and test the Ontology. The image collection contains 20,000 images of twenty sport genres (badminton, boxing, cycling, discus throw, diving, fencing, football, hammer throw, high jump, hurdles, javelin, Judo, long jump, pole vault, running, sailing, shooting, tennis, volleyball, and weightlifting). It is divided into two groups, a training set containing 12,000 images (600 image from each group) and a test or evaluation set with 8000 images (400 images from each group).

5.1. Evaluation protocols

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. 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 [57]. First, the *Application level* evaluates the performance of the Ontology to perform some specific tasks, e.g. searching data. Second, the *Taxonomy or hierarchy level* evaluates how well the Ontology structure design fits a corpus of documents. This method involves the evaluation of the structural design of an Ontology, which is usually performed manually by experts. As human evaluation is by its very nature subjective, this paper mainly focuses on computerised evaluations. Thus, the Application level will be deployed to evaluate the presented Ontology. We conducted four experiments: Experiments I and II determine the retrieval performance of the system, Experiment III evaluates the usefulness of LSI when a query fails, because of a lack of known data, and Experiment IV compares the retrieval performances with commercial search engines.

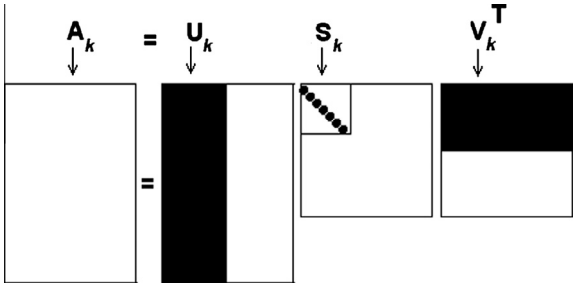

Fig. 7. Rank-k SVD approximation used for matrix dimensional reduction.

5.2. Research hypotheses

To evaluate the framework efficiency, several aspects need to be studied. In this paper, we evaluate the quality of the proposed KB with respect to three aspects:

- *Correctness* represents the ability of a KB to provide the right information to a user within its actual coverage [58].
- *Multimodality* refers to the capability of a KB that can store multi-modal information and deal with a variety of forms (mode) of queries.
- *Robustness* refers to the capability of a KB to handle an unknown query that can find relevant data for a user.

From these aspects of evaluation, three hypotheses are established.

The first hypothesis focuses on Ontology design and structure. The classes and relationships in this paper represent information derived from the Olympic website. If the structure of Ontology is designed appropriately, it is more able to find the right answer to a user query.

Hypothesis 1. The Structure and relationships of the proposed Ontology allows the system to perform a semantic search, e.g. to find semantically related information, which is not explicitly mentioned in image caption. As a consequence, its retrieval performance is significantly improved (*Correctness evaluation*).

Almost all domain Ontologies usually contain single-modality information to capture image content even though text and visual features can represent image content in different ways. When text descriptions of images are not supplied, an IMR system should also be able to find all relevant images based upon any distinctive low-level visual features. Therefore, both text and image modalities should be fused in a unified model to enhance the retrieval performance. In addition, the KB should handle both visual and textual queries.

Hypothesis 2. A Multi-Modal Ontology fuses both textual and low-level image feature information as a unified model. As a consequence, the system can handle both forms (mode) of queries (*Multi-modal evaluation*).

Most existing IR systems do not support Ontology incompleteness [16] (see Section 1). It is very challenging to construct an Ontology that can cover the complete domain, in one phase of development. When the information in the knowledge-based is incomplete, the retrieval performance (in terms of precision and recall) is affected. This leads to the following hypothesis:

Hypothesis 3. The use of LSI can enhance an IR system to handle an incomplete KB. LSI allows the search mechanism to find semantically relevant data, even though the data in a KB is absent (*Robustness evaluation*).

6. Experimental results and discussions

Four experiments are conducted to evaluate the performance of the MMIO model to validate the proposed hypotheses. This section presents an image retrieval approach based upon the extracted knowledge, and the visual and textual features (Fig. 8). If the domain Ontology is designed suitably, i.e. in this case to fit the sport domain, the retrieval results should be significantly improved. Image retrieval starts when the query keywords from users are examined, by tokenising them and then eliminating the stop words. 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 that an algorithm to disambiguate multiple word senses in a user query is used, as shown in Algorithm 1. Hence, the system can perform a semantic search whose terms are represented in OWL with the appropriate word sense, derived from a user query. Query processing translates the user query into a SPARQL (Protocol And RDF Query Language, see www.w3.org/TR/rdf-sparql-query) query and searches for any relevant information within the KB. The retrieved images are ranked and presented to the user.

Algorithm 1: Word sense disambiguation

Input: A user query
Output: A list of similarity words{W}
10: Remove stopwords from user query and to stem them and get a set of query keywords {Q};
20: Look up all remaining words {K} in WordNet and assign senses {Q} to all words;
30: For each word in {K}
40: Look up for all its senses in WordNet{Q}
50: where $k \in K$ and $q \in Q$ {
60: Compute similarity between $\{k_i, q_i\}$;
70: Assign similarity score to k_i;}
80: Sort {K} according to similarity score;
90: Select the concept (c_i) which has the highest similarity value of word sense = {W}.

To evaluate the KB model at an application level, the retrieval performance of the proposed framework is compared with other state of the art techniques, e.g. Lucene (http://lucene.apache.org). Lucene is an open-source framework providing Java-based indexing and search technology for text-based information. To evaluate the performance of image searches, the average precision has been used. More details of the experiments are explained in the next section.

6.1. Experiment I: correctness and multi-modal KB evaluation

To evaluate the annotation efficiency, we compare the retrieval performance of the proposed (Multi-Modal Incompleteness Ontology (MMIO) method with state of the art techniques, using precision and recall metrics. The experiments are conducted repeatedly, 10 times, and are used to compute average precision. We implemented five state of the art techniques for use in our comparative evaluation. These are LSI [39], Original Bag-of-Words (OBW) [40], LIRE (www.semanticmetadata.net/lire/), Lucene, and Visual-Features-Ontology (VFO). We used similar experiment settings and then compared their results with those of our proposed MMIO method. VFO contains only visual features in the form of an Ontology. VFO is based upon visual words generated by the simple k-mean algorithm. Then, these visual words are transformed into a hierarchical model using the Agglomerative Clustering algorithm. LIRE is a Java-based framework for photos and images retrieval based on their *colour* and *texture* characteristics. The comparative evaluation experiment is divided into two parts: Firstly, the retrieval performance from text-based queries is examined. The MMIO method is compared to those frameworks (LSI, Lucene and MMIO) that support only text queries. Secondly, query-by-examples are performed using images as queries for the remaining comparison frameworks that exploit low-level image features for indexing and retrieval.

OBW represents image data using a classical vector space model. However, this model ignores the high-level semantic information among those visual words. This affects the retrieval performance. In Fig. 9, OBW obtains a lower precision than MMIO because more inaccurate images are retrieved. OBW retrieves all images that have similar SIFT descriptors. Unfortunately, those similar images are not semantically relevant. In contrast, MMIO preserves the semantic information during its visual words construction in the training phase. In addition, in contrast to VFO, MMIO exploits the use of a hierarchical, Ontology model, which can express visual content through concepts and relationships.

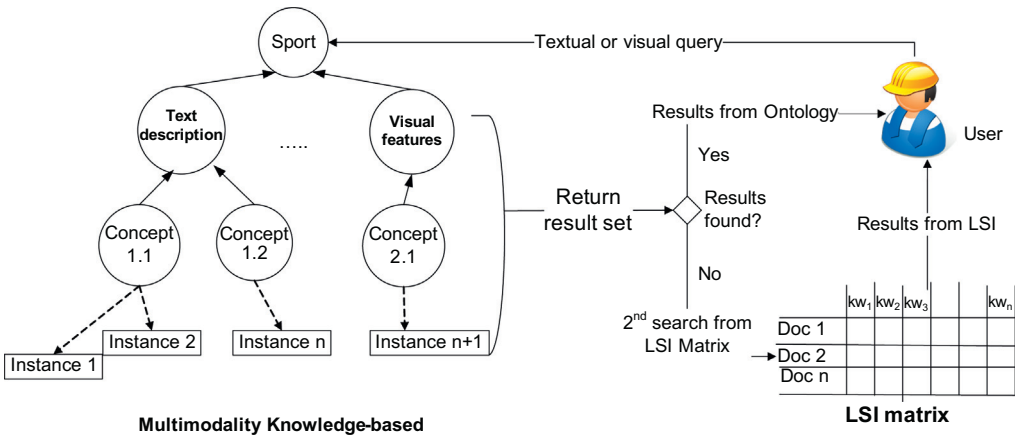

Fig. 8. Image retrieval based upon the extracted knowledge, and visual and textual features.

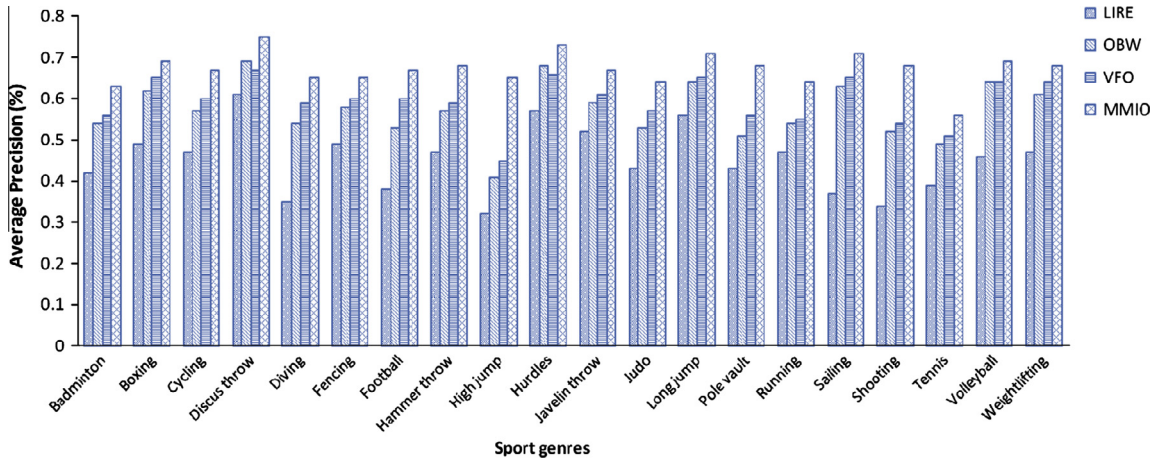

Fig. 9. Precision graph comparison for visual queries.

This is represented more efficiently than a pure vector space model. The structure explicitly defined by an Ontology can be used to reason about which images are more relevant to a query. VFO represents visual information in the form of a binary tree structure. Unfortunately, a binary tree cannot represent image content adequately, e.g. concepts in this kind of tree cannot overlap and cannot support multiple inheritance. Therefore, some irrelevant images are retrieved. Thus, VFO obtains a lower precision than MMIO. Because of the less effective representation model of VFO, its retrieval performance is similar to that of OBW. LIRE retrieves images based upon local visual features e.g. colour and texture. Hence, LIRE tends to retrieve visually similar images. However, some of those visually similar images are not always semantically relevant to the query. As a result, LIRE has the lowest precision compared with the other techniques.

When a textual query is submitted, all keywords are extended in order to find other similar keywords in WordNet and then matched with the metadata in an Ontology. Fig. 10 demonstrates that MMIO significantly improves the retrieval performance of Lucene and LSI. This result is expected because the MMIO leverages the structure of an Ontology to help to filter out irrelevant results by performing a conceptual search. MMIO exploits Ontology annotations and relationships to retrieve relevant data rather than just performing simple string matching. As a result, all images based on the same concept are considered as relevant images, even though there are no keywords in a query that also appear in the associated image captions.

The Ontology structure and relationships defined by MMIO are very useful to aid query reasoning. This can be used to more effectively differentiate similar queries that have semantic differences. An example of query reasoning is explained in the next section. The experimental results shown in Figs. 9 and 10 indicate that MMIO can dynamically handle both forms of queries, while other techniques can support only one particular type of a query. Therefore, the first hypothesis (H1) is verified.

6.2. Experiment II: semantic search study

In this experiment, testing the hypothesis investigates that if the domain Ontology structure is designed appropriately. It is able to better disambiguate the vagueness in a user query without needing to exploit any external knowledge, e.g. WordNet. To evaluate this hypothesis, an illustrative query "*Thailand athletes who participate at the Olympic games in Great Britain*" is used to evaluate the retrieval performance. The proposed framework has been compared with other two techniques, that of LSI and Lucene. Fig. 11 shows that MMIO is superior to Lucene and LSI. Lucene is not able to differentiate a query "Thailand athletes participate at the Olympic games in Great Britain" and "Great Britain athletes participate at the Olympic games in Thailand". The search engine of Lucene

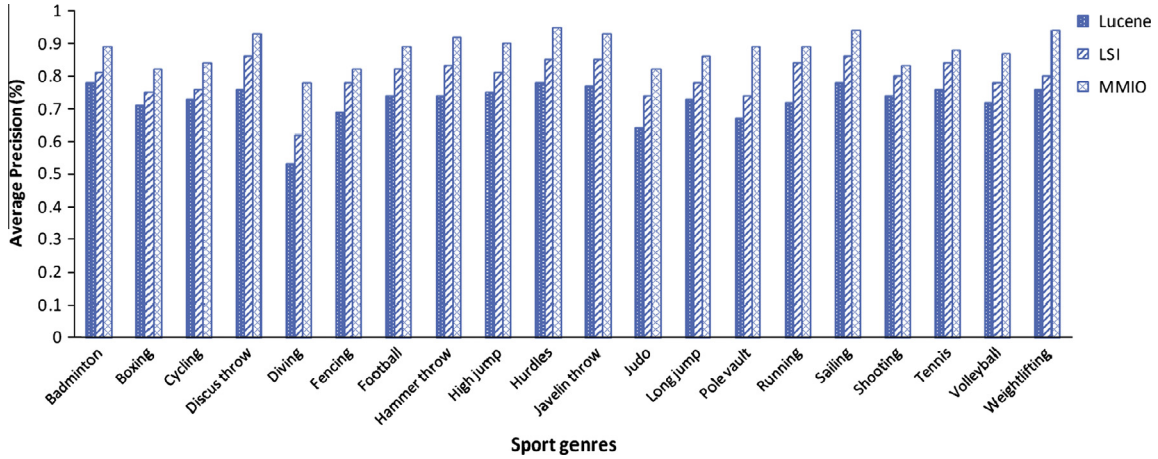

Fig. 10. Comparison of the precision of state of the art methods for textual queries across sports genres.

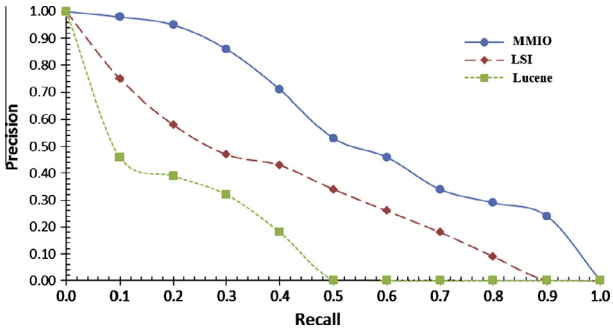

Fig. 11. Retrieval results of the query for MMIO, LSI and Lucene.

retrieves all documents containing words "*Thailand, athlete, Olympic games and Great Britain*". Therefore, it retrieves a high number of irrelevant images including Great Britain athletes who participate in sport events in Thailand which leads to a low precision and recall. Lucene performs searches based upon a string matching technique. When a string used in a query does not match the data in the KB, it returns no result (the precision is zero) whereas LSI can find more relevant images than Lucene by using a semantic indexing mechanism. Thus, it obtains a higher precision.

In contrast, the opposite is possible for a MMIO search using a SPARQL query. In a SPARQL query, the relationship between two concepts can be explicitly specified. In this experiment, the SPARQL query ignores the "Great Britain athletes"-<*participate*>-"Olympic games"-<*located*>-"Thailand" relationship and other relationships, not expressed in the query. As a result, only relevant images regarding the concepts and relationships in the SPARQL query are retrieved. This mechanism significantly improves precision and recall compared to Lucene and LSI. Hence, the second hypothesis (H2) is successfully validated.

Since MMIO contains multi-modal information, it can handle both types of queries more effectively than other state of the art techniques which focus only on using a single media type. The lack of text descriptions for images can affect the text-based query method and text Ontology. Thus, the MMIO framework can more effectively handle this problem through its use of a Visual-feature Ontology. However, creating a complete Ontology is a challenging task. In the next section, the issue of Ontology incompleteness is studied and analysed.

6.3. Experiment III: knowledge-based incompleteness evaluation

The hypothesis evaluated in this section is 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 better deal with unknown query terms and metadata. To assess this hypothesis, some metadata for images in the Ontology is removed, leading to missing information in the Ontology as unknown data is present in text captions. For instance, metadata for boxing images for the London 2012 Olympic Games was deleted, and then a query "*Boxing in London 2012*" is submitted to the system. Fig. 12 illustrates an example of the search results using LSI when the information concerning London 2012 is deleted. LSI is able to recognise all relevant images for London 2012 even though an image caption does not contain the word "London", e.g. the third image in Fig. 12. This is because the system uses

10 August 2012, Australian light heavyweight Damien Hooper celebrates at the end of a round in his fight with Marcus Browne of the USA.

Team USA's Marcus Browne takes a standing count as Australia's world number two Damien Hooper pulls out a fantastic last round to beat him 13-11 in their light heavyweight bout in the London 2012 Olympics Games

Nicola Adams (in blue) jabs Chungneijang Mery Kom Hmangte of India during their 51kg fly women's boxing match in the Olympic Park

Fig. 12. Example of LSI results when the KB does not contain the London 2012 information.

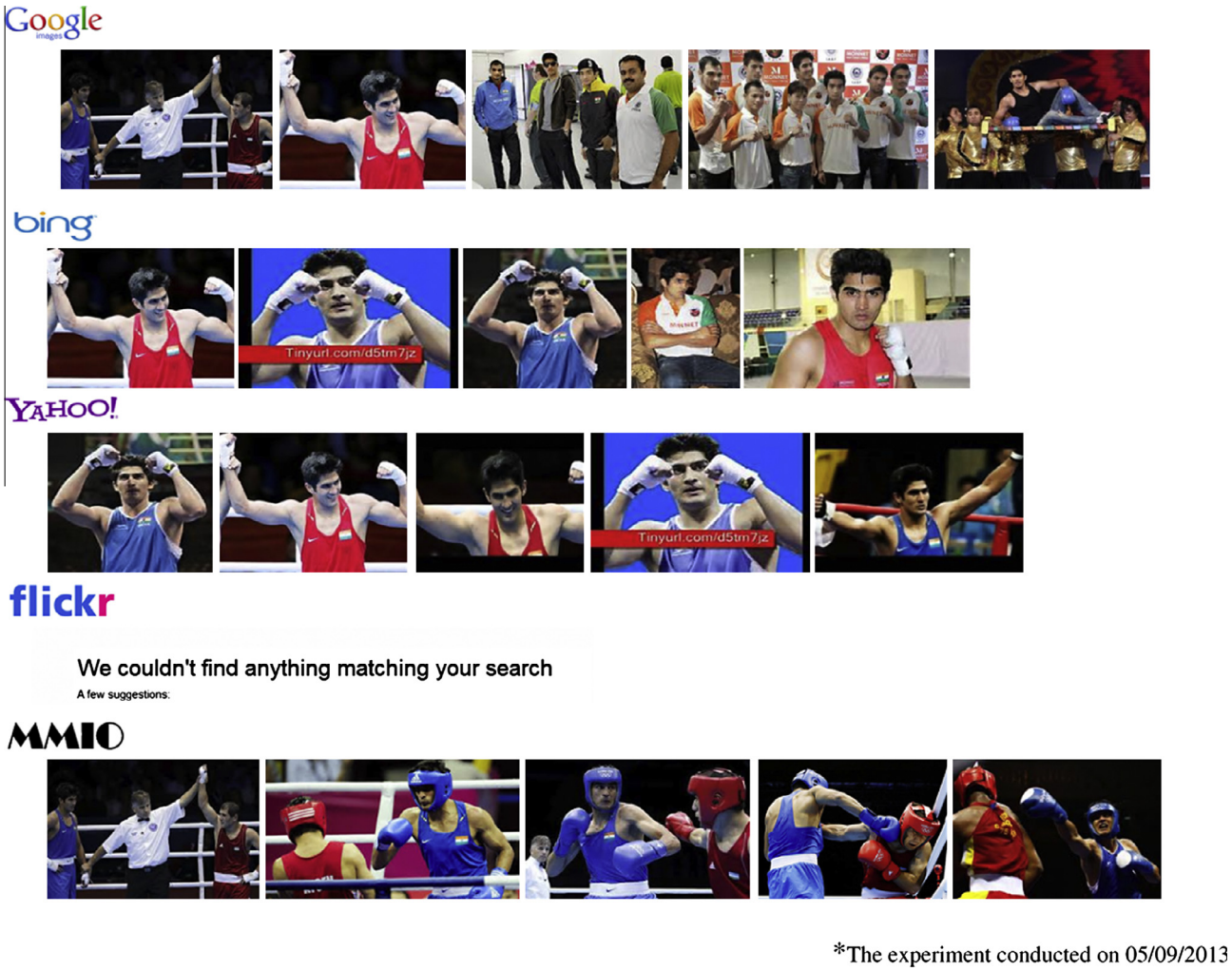

Fig. 13. Top five results for the example query "Vijender Singh performs the Boxing competition in London 2012" obtained from four state of the art search engines and our MMIO one.

co-occurrence information from LSI to find all the relevant images. In this example, Olympic park, and London, 2012 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 no word "London" appears in the text captions. This cannot be achieved by using a simple text-based search engine. The KB model presented, acts as an *open KB*. It can handle *unresolvable* data better than a general KB that behaves as more of a *closed KB*. 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 returns an empty or null result to users. In contrast, an open KB model can use a second method to search for relevant information, within the corpus of documents if unknown data is present. Similarly, if the KB in the MMIO model is incomplete, the search mechanism can try a second search using LSI instead of using SBIR. In other words, the open KB is more robust than a closed KB when an unknown query is encountered. Hence, Hypothesis H3 is validated.

6.4. Experiment IV: image retrieval performance comparison of MMIO against state of the art search engines

An additional experiment was conducted to compare the image retrieval performance against that of four commercial search engines, Google, Bing, Yahoo, and Flickr. Twenty queries have been used to evaluate the system and to compare MMIO against the state of the art search engines. To illustrate this evaluation, we present the results of one query example. We consider only the top 5 search results for each search engine. In this experiment, an illustrative query "Vijender Singh performs in a boxing competition in London 2012" is used to evaluate the retrieval performance of individual search engines. Fig. 13 illustrates the top five images obtained using different commercial search engines.

All these four search engines can find images that Vijender Singh (*athlete*) participates in a Boxing (*sport*) competition at the London 2012 Olympics (*time and place*). However, the ranking results from each search engine are different. Flickr fails to find any image that matches the query because it cannot handle natural language queries. Meanwhile, Google, Bing, and Yahoo can find images of Vijender Singh. Unfortunately, almost all of them are not semantically relevant because they do not exhibit any relevance to "participates in a Boxing competition", instead they show mostly Vijender Singh's face. This is because these search engines use a string matching technique to match image descriptions with keywords in the query. However, they do not understand the actual meaning of the query. In addition, some relevant images are not retrieved because they do not have any terms in captions matched with the query. Thus, those images are discarded.

Fig. 14. Comparison of the top ten ranked results using a visual query obtained from Google and MMIO.

MMIO, in contrast, understands the intention behind the query, and is able to find the relevant images for "Vijender Singh performs the Boxing competition in London 2012" more effectively using terms and relationships (triples) defined in the domain Ontology. In addition, some images can be recognised as relevant images, even although they do not have the query terms appearing in their captions. For example, the last image in the MMIO results (Fig. 13) has a caption "Indian boxers at London 2012". Through LSI, "Vijender Singh" has a strong relationship with "Indian" using a co-occurrence computation. Hence, MMIO can recognise that this image is relevant to the query. As such, MMIO significantly increases the IR precision by ranking images that do not relate to the boxing competition, as less relevant.

Next, a query-by-example (QBE) is performed to study the results between MMIO and Google. Only two image search engines are compared because Yahoo and Bing do not support QBE at the time our experiments were performed.

Fig. 14 shows the results of two search engines using a visual query (QBE). The query (Q) is an image of "Vijender Singh boxes at the London 2012 Olympics". Since the search engines retrieve images based upon the similarities of the visual features, e.g. colour, shape, and texture, all obtained images are visually similar. Nonetheless, some of them are semantically irrelevant, e.g. different domains of interest (the 7th result from Google), different sport genres (the 9th result from Google), different people and different times and places. Those images are ranked highly by Google because colour and shape dominate visual similarities. However, MMIO uses annotations defined in the Text Description Ontology to enable the most similar image to infer other related images. All images that are semantically related with those annotations are ranked higher in the result set. In other words, MMIO not only uses visual features but also textual features to determine the most relevant images. For example, the top five images of MMIO in Fig. 14 are more semantically relevant to the query in term of people (Vijender Singh), sport genre (Boxing), time and place (London 2012) than other images in the result set. This mechanism significantly improves the precision of MMIO, aided through use of a Textual Description Ontology.

7. Conclusions and further work

A Multi-Modal Incompleteness Ontology-based (MMIO) framework has been proposed for image retrieval. This is based upon fusing the low-level visual features extracted from images with the key concepts found in image (text) annotations, representing these in an Ontology. Based upon an experimental validation, we found that Queries that can leverage the fused image features with text concept models (MMIO) have a better recall and precision than (state of-the art) methods that only use one content mode, i.e., either images (LIRE, VFO, OBW) or text (Lucene, LSI), but not both.

The Ontology structure defined by MMIO aids query reasoning and can be used to more effectively differentiate similar queries that have semantic differences. Through exploiting semantically

related information, not explicitly mentioned in image caption, the precision of the images using MMIO is significantly improved compared to state of the art methods for visual queries and for text queries. If the structure of Ontology is designed appropriately, it is able to better disambiguate unrelated similarities (vagueness) in a user query, e.g., athletes representing a first country participating in games held in a second country versus athletes representing the second country participating in games held in the first country, without needing to exploit any external knowledge, e.g., from WordNet. This significantly improves image precision and recall. Similar words to those used in a query, but that do not appear in the caption, can also be exploited as search terms using a Latent Semantic Indexing (LSI) algorithm.

Text-based query retrieval state of the art search engines tend to miss several relevant images that do not have the terms in the captions matched to the query. Thus, their precision is reduced. MMIO is able to infer the implicit semantics behind the query, selecting only the semantically relevant images, and yielding better results. Secondly, the results from using visual-based queries in commercial search engines may not match the user intention due to inefficient content representation of global low-level features. A better representation using Visual Features Ontology in conjunction with a Textual Description Ontology can significantly improve the precision and recall.

Our further work will focus on two areas, query time evaluation and personalisation. A fundamental characteristic of the MMIO framework is that it has an increased response time to return its results to users, since it has to process both textual and visual information for a query. Of interest is also, what the difference is for visual versus text only queries. The measurement of the difference in throughput is left to further work. Further work can also be undertaken to extend MMIO to support personalised image retrieval (PIMR) and to improve the retrieval performance by considering an individual user's interests. Several challenges need to be overcome. Firstly, 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. Secondly, user preferences should be represented in a richer, more precise and less ambiguous way than in a keyword and text-based model. Finally, naming differences of things can vary according to the linguistic representation. The concepts underlying such terms may be used differently by different users at different levels of granularity and in different situations with divergent interpretations. As such, a PIMR that models user profiles should take terminological heterogeneity problem into account.

Acknowledgments

This research has been supported in part by National Science and Technology Development (NSTDA), Thailand. Project No: SCH-NR2011-851. We also thank Ms. Jenny Williams for her proof-reading.

References

[1] R. Meersman, Semantic ontology tools in information system design, in: Found. Intell. Syst. 11th Int. Symp., 1999: pp. 30–45.
[2] A.T. Schreiber, B. Dubbeldam, J. Wielemaker, B. Wielinga, Ontology-based photo annotation, IEEE Intell. Syst. 16 (2001) 66–74.
[3] L. Hollink, G. Schreiber, J. Wielemaker, B. Wielinga, Semantic annotation of image collections, in: Workshop Knowl. Markup Semantic Annot., 2003, pp. 1–3.
[4] P.A.S. Sinclair, S. Goodall, P.H. Lewis, K. Martinez, M.J. Addis, Concept browsing for multimedia retrieval in the SCULPTEUR Project, in: 2nd Annu. Eur. Semantic Web Conf., 2005: pp. 28–36.
[5] D. Gasevic, D. Djuric, V. Devedzic, Model Driven Engineering and Ontology Development, second ed., Springer, London, United Kingdom, 2009.
[6] A. Haubold, A. Natsev, M. Naphade, Semantic multimedia retrieval using lexical query expansion and model-based reranking, in: IEEE Int. Conf. Multimed. Expo, 2006: pp. 1761–1764.
[7] A. Natsev, A. Haubold, J. Tešić, L. Xie, R. Yan, Semantic concept-based query expansion and re-ranking for multimedia retrieval, in: Proc. 15th Int. Conf. Multimed., 2007, pp. 991–1000.
[8] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, R. Jain, Content-based image retrieval at the end of the early years, IEEE Trans. Pattern Anal. Matching Intell. 22 (2000) 1349–1380.
[9] J.R. Smith, S.-F. Chang, Visually searching the web for content, IEEE Multimed. 4 (1997) 12–20.
[10] M.J. Swain, C. Frankel, V. Athitsos, WebSeer: an image search engine for the world wide web, in: Tech. Rep., The University of Chicago, 1997. <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.44.9234&rep=rep1&type=pdfm, retrieved 07/07/2013>.
[11] R. Zhao, W.I. Grosky, Narrowing the semantic gap – improved text-based web document retrieval using visual features, IEEE Trans. Multimed. 4 (2002) 189–200.
[12] J. Hu, A. Bagga, Categorizing images in web documents, IEEE Multimed. 11 (2004) 22–30.
[13] X. Song, C.-Y. Lin, M.-T. Sun, Autonomous visual model building based on image crawling through internet search engines, in: Proc. 6th ACM SIGMM Int. Workshop Multimed. Inf. Retr., 2004, pp. 315–322.
[14] H. Wang, L.-T. Chia, S. Liu, Image retrieval ++–web image retrieval with an enhanced multi-modality ontology, Multimed. Tools Applic. 39 (2008) 189–215.
[15] Y.I.A.M. Khalid, S.A. Noah, S.N.S. Abdullah, Towards a multimodality ontology image retrieval, in: Proc. 2nd Int. Conf. Vis. Inform. Sustain. Res. Innov., 2011, pp. 382–393.
[16] G. Nagypál, Possibly Imperfect Ontologies for Effective Information Retrieval, University Karlsruhe (TH), Germany, 2007. <http://d-nb.info/986790028/34, retrieved 07/07/2013>.
[17] S. Dasiopoulou, C. Doulaverakis, V. Mezaris, An ontology-based framework for semantic image analysis and retrieval, in: Y.-J. Zhang (Ed.), Semantic-Based Vis. Inf. Retr., IRM Press, USA, 2007, pp. 269–293.
[18] R. Tansley, The Multimedia Thesaurus: Adding A Semantic Layer to Multimedia Information, University of Southampton, 2000. <http://eprints.soton.ac.uk/253833/> (retrieved 07.07.13).
[19] G.A. Miller, WordNet: a lexical database for English, Commun. ACM 38 (1995) 39–41.
[20] Z. Wang, Y. Ma, Medical image fusion using M-PCNN, Inf. Fusion 9 (2008) 176–185.
[21] L. Yang, B.L. Guo, W. Ni, Multimodality medical image fusion based on multiscale geometric analysis of contourlet transform, Neurocomputing 72 (2008) 203–211.
[22] T. Li, Y. Wang, Biological image fusion using a NSCT based variable-weight method, Inf. Fusion 12 (2011) 85–92.
[23] K. Kesorn, S. Poslad, An enhanced bag of visual word vector space model to represent visual content in athletics images, IEEE Trans. Multimed. 14 (2012) 1520–1532.
[24] K. Kesorn, S. Chimlek, S. Poslad, P. Piamsa-nga, Visual content representation using semantically similar visual words, Expert Syst. Applic. 38 (2011) 11472–11481.
[25] J. Shawe-Taylor, S. Sun, A review of optimization methodologies in support vector machines, Neurocomputing 74 (2011) 3609–3618.
[26] Y. Ji, S. Sun, Multitask multiclass support vector machines: model and experiments, Pattern Recogn. 46 (2013) 914–924.
[27] B. Russell, A. Torralba, K. Murphy, W. Freeman, LabelMe: a database and web-based tool for image annotation, Int. J. Comput. Vis. 77 (2008) 157–173.
[28] Y. Chen, H. Sampathkumar, B. Luo, X. Chen, ILike: bridging the semantic gap in vertical image search by integrating text and visual features, IEEE Trans. Knowl. Data Eng. 25 (2013) 2257–2270.
[29] MPEG Requirements Group, MPEG-7 Context, Objectives and Technical Roadmap V. 12, in: ISOIEC JTC 1SC 29WG 11, International Organization for Standardization, 1999, pp. 1–13.
[30] F. Nack, J. van Ossenbruggen, L. Hardman, That obscure object of desire: multimedia metadata on the Web, part 2, IEEE Multimed. 12 (2005) 54–63.
[31] R. Troncy, J. Carrive, A reduced yet extensible audio-visual description language: how to escape from the MPEG-7 bottleneck, in: Proc. 4th ACM Symp. Doc. Eng. DocEng'04, 2004, pp. 87–89.
[32] T.R. Gruber, A translation approach to portable ontology specifications, Knowl. Acquis. 5 (1993) 199–220.
[33] G. Schreiber, B. Wielinga, R. de Hoog, H. Akkermans, W. Van de Velde, CommonKADS: a comprehensive methodology for KBS development, IEEE Expert 9 (1994) 28–37.
[34] E.N. Guarino, R. Poli, N. Guarino, Formal ontology, conceptual analysis and knowledge representation, Int. J. Hum.-Comput. Stud. 43 (1995) 625–640.
[35] B. Chandrasekaran, J.R. Josephson, V.R. Benjamins, What are ontologies, and why do we need them?, IEEE Intell Syst. 14 (1999) 20–26.
[36] D. McGuinness, Ontologies come of age, in: D. Fensel, J. Hendler, H. Lieberman, W. Wahlster (Eds.), Semantic Web Why What How, MIT Press, Boston, MA, 2003.
[37] S. Poslad, Ubiquitous Computing Smart Devices, Environments and Interactions, John Willey & Sons, London, United Kingdom, 2009.
[38] A. Llorente, S. Rüger, Using second order statistics to enhance automated image annotation, in: Proc. 31th Eur. Conf. IR Res., 2009, pp. 570–577.
[39] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, R. Harshman, Indexing by latent semantic analysis, J. Am. Soc. Inf. Sci. 41 (1990) 391–407.

[40] G. Csurka, C.R. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: Int. Workshop Stat. Learn. Comput. Vis., 2004, pp. 1–22.
[41] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.
[42] L. AlSumait, C. Domeniconi, Text clustering with local semantic kernels, in: M.W. Berry, M. Castellanos (Eds.), Surv. Text Min. II Clust. Classif. Retr., Springer-Verlag London Limited, London, United Kingdom, 2008.
[43] J. Sivic, B.C. Russell, A. Zisserman, W.T. Freeman, A.A. Efros, Unsupervised Discovery of Visual Object Class Hierarchies, in: IEEE Conf. Comput. Vis. Pattern Recognit., 2008, pp. 1–8.
[44] Y.-G. Jiang, C.-W. Ngo, Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval, Comput. Vis. Image Underst. 113 (2009) 405–414.
[45] L. Wang, Z. Lu, H.H. Ip, Image Categorization Based on a Hierarchical Spatial Markov Model, in: Proc. 13th Int. Conf. Comput. Anal. Images Patterns, 2009, pp. 766–773.
[46] B. Walter, K. Bala, M. Kulkarni, K. Pingali, Fast agglomerative clustering for rendering, in: IEEE Symp. Interact. Ray Tracing, 2008, pp. 81–86.
[47] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[48] L. Wu, S.C.H. Hoi, N. Yu, Semantic-preserving bag-of-words models for efficient image annotation, in: Proc. 1st ACM Workshop Large-Scale Multimed. Retr. Min., 2009, pp. 19–26.
[49] J. Zhu, V. Uren, E. Motta, ESpotter: adaptive named entity recognition for web browsing, in: Intell. IT Tools Knowl. Manag. Syst. KMTOOLS 2005, 2005, pp. 518–529.
[50] R. Arndt, R. Troncy, S. Staab, L. Hardman, M. Vacura, COMM: Designing a well-founded multimedia ontology for the web, in: Proc. 6th Int. Semantic Web Conf., 2007, pp. 11–15.
[51] S. Dasiopoulou, V. Tzouvaras, I. Kompatsiaris, M.G. Strintzis, Capturing MPEG-7 semantics, in: M.-A. Sicilia, M.D. Lytras (Eds.), Metadata Semant., Springer, US, 2009.
[52] S. Dasiopoulou, V. Tzouvaras, I. Kompatsiaris, M.G. Strintzis, Enquiring MPEG-7 based multimedia ontologies, Multimed. Tools Applic. 46 (2010) 331–370.
[53] M.A. Rahman, M.A. Hossain, I. Kiringa, A.E. Saddik, Ontology-based unification of MPEG-7 semantic descriptions, in: Int. Conf. Electr. Comput. Eng. 2006, pp. 291–294.
[54] L. Bai, S. Lao, W. Zhang, G.J.F. Jones, A.F. Smeaton, Video semantic content analysis framework based on ontology combined MPEG-7, in: N. Boujemaa, M. Detyniecki, A. Nürnberger (Eds.), Adapt. Multimed. Retr. Retr. User Semant., Springer, Berlin Heidelberg, 2008.
[55] P. Praks, S. Snasel, J. Dvorský, Latent semantic indexing for image retrieval systems, in: Proc. SIAM Conf. Appl. Linear, Algebra, 2003, pp. 1–8.
[56] E. Chisholm, T.G. Kolda, New Term Weighting Formulas For The Vector Space Method In Information Retrieval, Computer Science and Mathematics Division, Oak Ridge National Laboratory, USA, 1999.
[57] J. Brank, M. Grobelnik, D. Mladenić, A Survey of Ontology Evaluation Techniques, in: Proc. Conf. Data Min. Data Wareh., 2005, pp. 166–169.
[58] G. Guida, G. Mauri, Evaluating performance and quality of knowledge-based systems: foundation and methodology, IEEE Trans. Knowl. Data Eng. 5 (1993) 204–224.