Multi-Modal Ontology that Enhances Visual Content Retrieval System

Stefan Poslad
School of Electrical and Electronic Engineering and Computer Science,
Queen Mary University of London, E1 4NS, United Kingdom
Email: stefan@eecs.qmul.ac.uk

Kraisak Kesorn
Computer Science and Information Technology Department,
Science Faculty, Naresuan University, Phitsanulok, 65000, Thailand
Email: kraisakk@nu.ac.th

***Abstract*—This paper presents an ontology-based system for multi-modal image annotation. A novel technique is proposed to represent the semantics of visual content by restructuring visual word vectors to an ontology model from computing a distance between the visual word features and concept features. The second index relies on textual description which is processed to extract and recognise concepts, properties, or instances in an ontology. The two indexes are unified in to a single indexing model used to enhance the image retrieval efficiency. As a result, it is possible to retrieve images with a query using words that do not appear in the caption. The constructed KB was evaluated how well it fits that knowledge domain regarding to its relevance for the application. The results show that the metadata in the presented KB could be exploited efficiently and, thus, it enhances the retrieval performance.**

***Index Terms*—multimodal information, knowledge base, ontology model, image retrieval system**

I. Introduction

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]. For an image retrieval system (IMR), ontologies are used for reducing the semantic gap, the gap between the user perception and the low-level feature abstraction from the visual content, by storing the knowledge structures for summarising, discovering, classifying, browsing and retrieving, and annotating images. Ontology-based frameworks are proposed for IMR in numerous collections [2]-[4]. These frameworks have validated the assumption that ontologies could help improve information retrieval effectiveness by making it possible to find relevant documents that are syntactically not similar to the query terms.

Existing works on IMR have been done based only on single-modality information either textual information or visual features. Consequently, those works suffer from several limitations. For example, the 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. This is because the extracted visual features themselves cannot be used to represent the content of images effectively. Text and image are two distinct types of information from different modalities, as they represent ‘things’ in different ways. However, there are some invariant and implicit connections between textual and visual information [5]. 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.

The combination of textual information with image features information has been proposed to improve image search results. Wang et al. [6] supported multi-modal, text and visual information, in the *canine* domain. Binary histogram is used to represent each of the image features and transformed into ontology model using a hierarchical SVM classification [7] and incorporate with the aforementioned textual description ontology. The proposed method is able to increase classification accuracy and retrieval performance. Khalid et al.[8] proposed multimodality ontology framework for sport domain. Textual descriptions and surrounding text are extracted and then are 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, Label Me Annotation Toolbox [9]. Nonetheless, using global features e.g. colour and edge information cannot represent the semantic of images effectively. For example, the same images have different brightness, size, and camera angle, so called visual heterogeneity problem and this is a well-recognised problem among researchers in image-processing area.

From the analysis of the existing solutions, some limitations still exist as they are unable to handle visual

Manuscript received July 25, 2013; revised September 10, 2013.

doi: 10.12720/ijoee.1.4.284-290

heterogeneity. For instance, when surveyed systems map lower-level features onto a higher level object conceptualisation, an extracted feature may possibly belong to multiple concept of objects leading to visual heterogeneity (one visual appearance has multiple meanings). Therefore, an image representation model should support this requirement. Solution to this problem is vital to achieve a good quality knowledge base for image retrieval system and are, as such, the main focus in this paper.

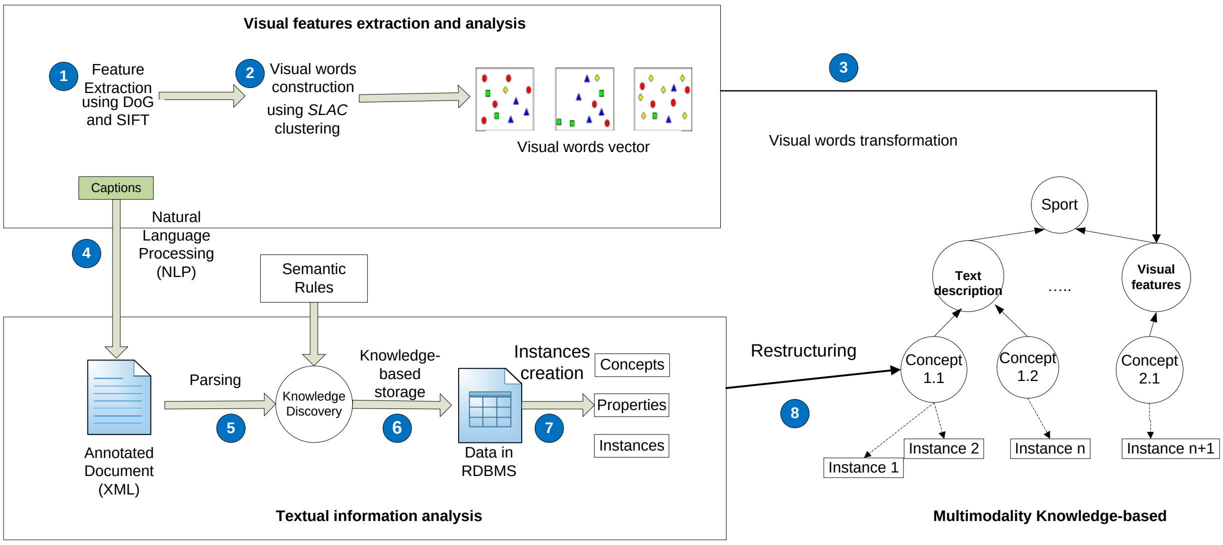

Figure 1. Knowledge-based acquisition technique.

II. Theoretical Framework for Multi-Modal Information Indexing

The main focus of this research is to extract knowledge from *visual data* and *text captions* and to store this extracted information in a unified KB. In order to acquire knowledge from both information sources, first, the low-level features are extracted and processed using a bag-of-visual words (BVW) technique in order to detect objects in images. Later, the extracted visual information is mapped to higher-level semantic conceptualisation based on the ontology model. Second, image captions are analysed and integrated with visual information in the unified knowledge-base in order to enhance the image interpretation and retrieval. Fig. 1 illustrates the main processes of the knowledge-acquisition framework.

A. *Visual Features Extraction and Analysis*

This section explains a process to compute higher level representations from lower level ones. There are two main visual content analysis and interpretation processes: 1) a low-level image processing extracts useful visual features from images and transforms them into primitive objects (person, tree, ball, horizontal bar etc.). Low-level image processing comprises several steps and is often called "image analysis". 2) *Higher level semantic interpretations* are identified based on the primitive objects and specific prior knowledge relevant (the facts that are not explicit in the data e.g. knowledge about a sports event) for the interpretation. This information is integrated to enhance image classification power. Here, we present a novel idea to represent a higher level conceptualisation of visual data derived from lower level features.

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 [10]. However, major limitations of existing BVW models include: they do not preserve the semantics during visual word construction. Hence, this paper proposes a method to generate a new representation model which resolves the above difficulties and enhances image retrieval efficiency. There are three main steps to perform the visual analysis (Fig. 1 process 1-3) which are described in the following sections.

- Low-level features extraction: images are processed (Fig. 1 process 1) to extract the "keypoints", the salient patches, in an image using the Difference of Gaussian (DoG) detector. This process will not explain further due to space limitation. Please read [11] for more details.
- Visual words construction: to construct a visual word, a clustering algorithm is deployed. However, a simple clustering method e.g., *k*-mean performs clustering over all the vectors. Each cluster is considered as a visual word that represents a specific local pattern shared by the keypoints in that cluster. The major drawback of *k*-mean algorithm 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 SLAC algorithm [12] to cluster the vectors. SLAC is a subspace clustering which is an extension of

traditional clustering by capturing local feature relevance within cluster. It can find the semantically similar keypoints and cluster them into the same group using a similarity matrix (Fig. 1 process 2).

- Visual word restructuring: in existing systems, researchers tried to restructure visual words as a hierarchical model [13] in order to disambiguate word senses more explicitly and effectively. Nevertheless, hierarchical models generated from these algorithms have some limitations. First, they are binary hierarchical model*s* that are not always efficient in representing visual content data. In practice, types of relationships among concepts are more diverse. Second, there is no multiple-inheritance between parents and a child node. The multiple-inheritance means a child node can have more than one parent. For example, a Heptathlon event has a relationship with the Field and Track event since it combines these two events together as one sport for women. As such, the generated hierarchical model used by the existing frameworks cannot represent the semantic information of visual content properly. This paper proposes transforming a visual word vector space model into a structural ontology model in order to resolve the aforementioned limitations.

In this paper, we proposed the mathematics-based technique to categorise visual words to concept(s) in an ontology model (Fig. 1 process 3). Typically, visual content may contain noise from its background, the visual objects of interest are manually separated from the background. The objects in the visual content are the extracted keypoints with respect to the local appearance of those objects. These keypoints are considered relevant because they are from the same object. This method can eliminate noise from the background. Then, only the keypoints of objects will be further processed to generate visual words [14]. The linkage between the visual words and high level semantics for an object category can be obtained, which serves to connect low level features to high level semantic objects. It is noted that when performed manually, such object separation is not an efficient method for a large-scale multimedia system. However, this method is applied for training only in order to allow the system to learn the sets of visual words. As a consequence, a set of visual words (ϖ) and $\{\varpi_i \in C_i\}$ are obtained for each object category C_i. Different visual words represent different views of different parts of an object. Having obtained 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 of a visual word (ϖ)'s centroid(ν) to the concept's centroid (c_i) and can be calculated using the following formula [13]:

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

The concept range is useful for the visual word sense disambiguation and image classification. From the training phase, the semantic concept of each individual object is presented properly and, then, each concept will be used to disambiguate the informative visual words and to assign the concept(s) for each visual word under a pre-designed ontology model which is improved from Wu [14]. The different senses of a visual word can be disambiguated using a concept range from (1). If a visual word is inside the range of any concept, it is assigned to that concept; otherwise the visual word does not respond to any concept and is discarded.

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

Furthermore, this model can be used to detect key objects and classify visual content at a higher-level conceptualisation. The detection of key objects in visual content is related to the frequency of visual words which represent it. If the frequency of related visual words, $f(v_i)$ of a particular object (i.e. athlete) is higher than a threshold (chosen experimentally), this means the visual content contains that object. However, direct use of $f(v_i)$ may be unfair to each instance of visual content in the collection due to scaling differences. Hence, $f(v_i)$ is normalised in order to compensate for discrepancies in the frequency of the visual words. Equation (2) shows the normalisation formula where N is a number of instances of visual content.

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

B. Textual Information Extraction and Analysis

Fig. 1 process 4-8 shows the main processes of the textual information analysis processes executed by the Textual Information Analysis module (TIA) (Fig. 1 process 5-7). 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, a Natural Language Processing tool (NLP) is used for the initial metadata generation (Fig. 1 process 4). An established NLP framework, ESpotter, is deployed rather than implementing a new text engineering tool. ESpotter [15] provides a function for a Named Entity Recognition (NER) task e.g. person name, location, date, and other proper nouns. ESpotter generates an initial semantic metadata in an XML file format.

The knowledge representation (KR) that we choose is a set of ontological commitments, in the sense that one concept generally refers to and is understood through its relationships with other concepts and through its use. In addition, a KR enables efficient machine-readable and machine-understandable computation about knowledge. Semantic metadata generated by the metadata generation process is stored in a relational database-RDBMS (Fig. 1 process 5). To be able to use this data in a semantic context, it is mapped to an ontology the data of which is given a well-defined meaning by representing it using OWL. 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 a simple and fast approach is preferred*:* a direct mapping scheme is used to map data from a RDBMS database to OWL using a JDBC connector API. This method directly maps a RDBMS schema to OWL. This generic approach can be useful in many cases, but sometimes it may lead to difficulties in synchronising to changes in the database structures, to difficulties in installation, and to use by inexperienced users. To transform information in a relational database to OWL, three steps are needed:

- The initial metadata is retrieved using the SQL select command and the record sets returned from the query are grouped by column.
- The Jena API (http://jena.sourceforge.net) is deployed to create ontology concepts, properties and instances. Jena provides off-the-shelf methods for creating ontology classes, properties, and assigning instances to these classes and properties.
- Then the class instances are created and assigned an URI or a blank node identifier. Later, the instances' properties are created and assigned property values and written to OWL file.

C. *Multi-Modality Ontology Structure*

Since both unstructured textual and visual information is transformed into hierarchical structures, they are merged into a unified knowledge model, so called Multi-Modality Ontology (MMO), to enhance the retrieval performance and as an alternative to the MPEG-7 standard. This is because MPEG-7 does not support machine processable semantic annotation of image subject matter [16]. In this section, we briefly describe the structure of multi-modality ontology. Fig. 2 shows the structure of presented multi-modal ontology. Ellipse represents pre-defined classes in ontology and rectangle indicates to ontology instances. Some parts of the ontology are omitted due to the space limitation.

- Sport event ontology provides the vocabulary and background knowledge of sport e.g. sport name, genre, and disciplinary. Classes and relationships in ontology are extracted from the Olympic website (www.olympic.org) which provides standard descriptions and relationships in various aspects of sports.
- Textual description ontology encapsulates high-level semantic sport and image annotations. It provides metadata to answer three types of queries, what, when, and where.
- Visual features ontology represents the metadata about the representation of an image as a whole e.g. format (jpg, bmp), size, resolution of a picture, and low-level features (visual words). This ontology is constructed using the method describe in Section II-A and incorporated with the textual description ontology in order to enhance the retrieval performance.

III. Experimental Settings, Evaluation Protocols, and Research Hypotheses

This section introduces and discusses the experimental method for the presented framework evaluation. Since sport image collections are not readily available in standard test collections, a new test collection needs to be designed. It was decided that images from the Olympic organization website and the Google image search engine should form the basis for the test collection to provide domains where a test ontology is developed. 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 throw, Judo, Long jump, Pole vault, Running, Sailing, Shooting, Tennis, Volleyball, and Weightlifting). The image collection is divided into two groups, a training set containing 12,000 images (600 image from each group) and a testing set with 8,000 images (400 images from each group).

A. *Research Hypotheses*

Almost all ontology usually contains single-modality information to capture image content. Text and visual features represent image content in different ways. Therefore, our main hypothesis is "Both modalities could be used in a unified model to enhance the retrieval performance, In addition, the system should handle both forms of queries".

B. *Experimental Results and Discussions*

Guided by the proposed framework, four experiments are conducted to evaluate the performance of the proposed system in the variety aspects and validate all hypotheses addressed previously.

To evaluate the annotation efficiency, we compare the retrieval performance of the proposed method (Multi-Modal Ontology-MMO) with state of the art techniques by adopting precision measurement because we focus on how many retrieved images are correct. The experiments are conducted repeatedly about 10 times and used to compute average precision. We implement another five techniques, LSI, Original Bag-of-Words (OBW) [10], LIRE (www.semanticmetadata.net/lire), Lucene (http://lucene.apache.org), and Visual-Features-Ontology (VFO), with the similar experiment settings and compare the results with the proposed method. VFO contains visual features (without textual information) in the form of ontology.

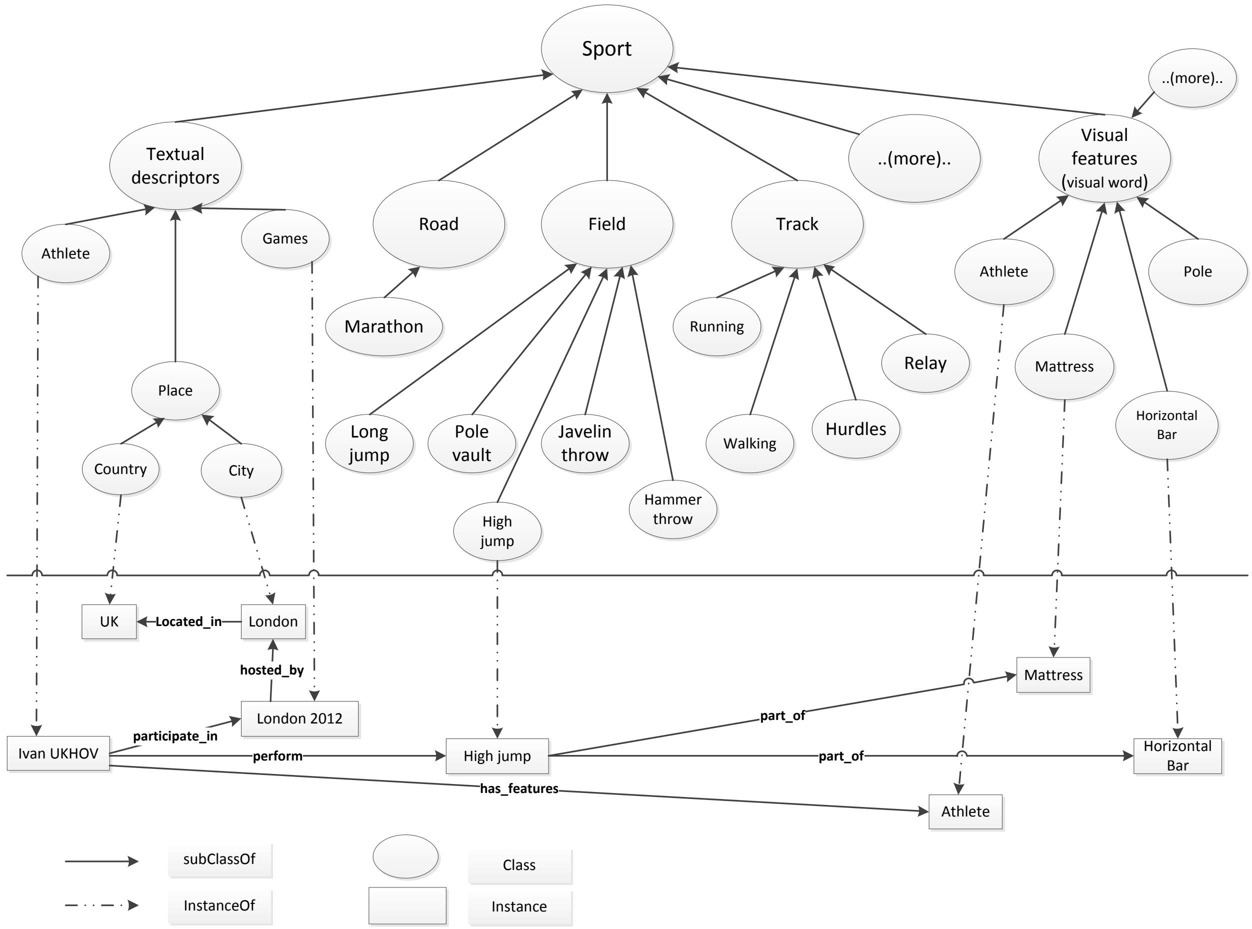

Figure 2. Multi-modal ontology structure.

VFO is constructed from visual words generated using simple *k*-mean algorithm and they are restructured from vector space model into hierarchical model using Agglomerative clustering algorithm. LIRE is java-based framework for photos and images retrieval based on their colour and texture characteristics. This experiment is divided into two parts: Firstly, the retrieval performance from text-based queries is examined. The proposed method is compared to those frameworks (LSI, Lucene and MMO) that support only text queries. Secondly, query-by-examples are performed using images as queries for the other frameworks that exploit low-level features for indexing and retrieval. OBW represents image data using classical vector space model. However, this model ignores the high-level semantic information among those visual words and this affects the retrieval performance. In Fig. 3, it obtains lower precision than MMO because more inaccurate images are retrieved. OBW retrieves all images that have similar SIFT descriptors. Unfortunately, those similar images are not semantically relevant to the query. In contrast, MMO preserve semantic information during visual words construction in training phase. In addition, MMO exploits hierarchical model which can express visual content more explicitly and efficiency than vector space model. The structure of ontology can reason that what images are relevant to the query. As a result, several relevance images are recognised. MMO is also superior to VFO because a binary tree cannot represent image content properly e.g. some concepts in this kind of tree cannot overlap nor have multiple inheritances. Therefore, its retrieval performance is not much different from OBW. LIRE retrieves images based on local visual features e.g. colour and texture. This is possible that LIRE retrieves visually similar images but, however, some of those images are not always semantically relevant to the query. As a result, LIRE obtains the lowest precision compared to other techniques.

When a textual query is entered into the system, all keywords will be extended to find other similar keywords in WordNet and matched with the metadata in ontology. Fig. 4 demonstrates that MMO significantly improves the retrieval performance of Lucene and LSI. This result is not surprise us at all because the structure of ontology in MMO is very helpful to filter out irrelevant results by performing conceptual search. MMO exploits ontology annotations and relationships to retrieve relevant data rather than just perform simple string matching. As a result, all images in the same concept can be recognised as relevance images to the query although there is no keywords in a query appeared in image captions.

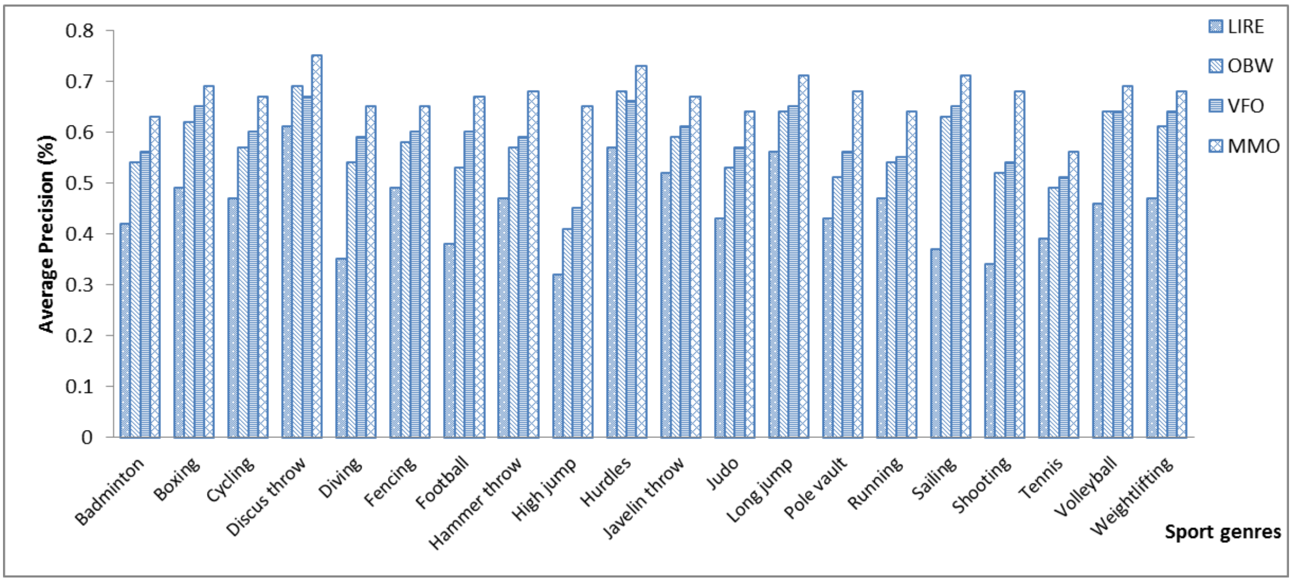

Figure 3. Precision graph comparison for visual queries

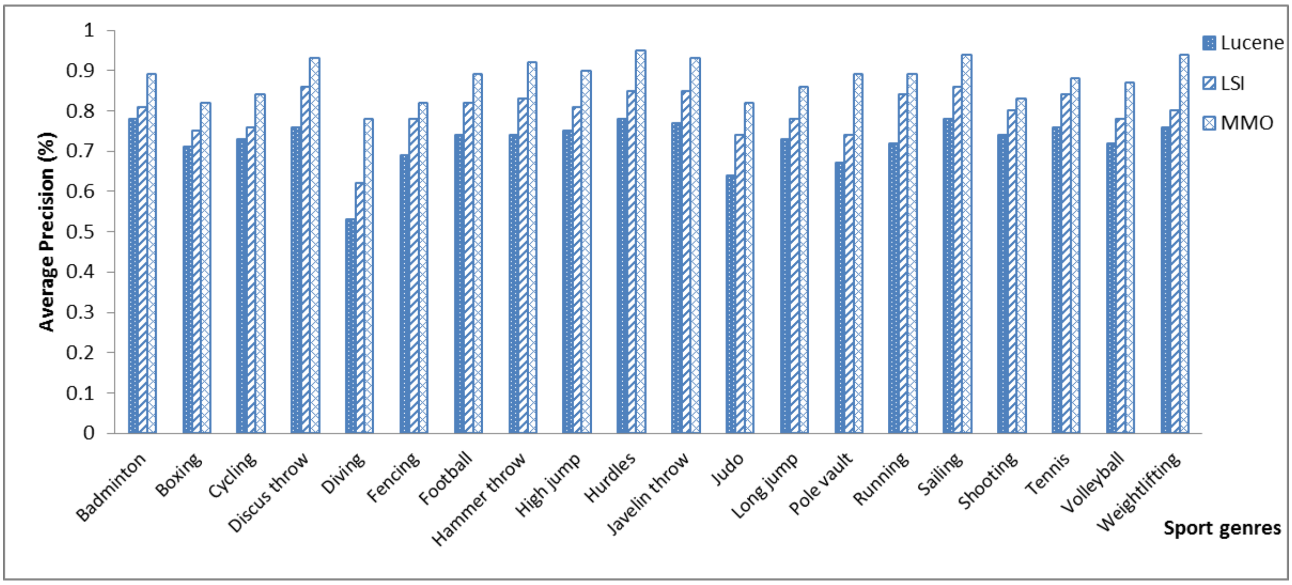

Figure 4. Precision graph comparison for textual queries.

The structure and relationships of MMO are very useful for query reasoning and can differentiate similar queries but have semantic differences effectively. An example of query reasoning is explained in next section. From the experimental results shown in Fig. 3 and Fig. 4 indicate that MMO can dynamically handle both forms of queries while other techniques can support only one particular type of a query. Therefore, the hypothesis in this research is verified.

IV. Conclusion and Future Work

This research introduces an idea to design a knowledge base that provides a semantic-based image retrieval solution for multimodality information. The presented KB is useful for image retrieval so that this not only relies on visual similarity but also on conceptual similarity. Aside from the visual analysis component, this paper also proposes a technique for acquiring knowledge from text captions, for their knowledge representation and for knowledge-based visual content retrieval. Both textual and visual information are encoded into a unified KB model in the form of OWL format to facilitate semantic retrieval. With the proposed framework, the image retrieval system can retrieve image correctively and dynamically.

The framework is now extended to support personalized image retrieval (PIMR) which aims to improve the retrieval process by taking into account the particular interests of individual users. 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/text-based model. Finally, naming differences can vary according to the linguistic representation. The concepts underlying such terms may be used differently by the different users at different levels of granularity and in different situations with divergent interpretations. As such, PIMR that models user profiles should take terminological heterogeneity problem into account.

ACKNOWLEDGMENT

This research has been supported by National Science and Technology Development (NSTDA), Thailand. Project no: SCH-NR2011-851.

REFERENCES

[1] R. Meersman, "Semantic ontology tools in information system design," in *Proc. Foundations of Intelligent Systems, 11th International Symposium*, 1999, pp. 30–45.

[2] A. T. Schreiber, B. Dubbeldam, J. Wielemaker, and B. Wielinga, "Ontology-based photo annotation," *IEEE Intell. Syst.*, vol. 16, no. 3, pp. 66–74, 2001.

[3] L. Hollink, G. Schreiber, J. Wielemaker, and B. Wielinga, "Semantic annotation of image collections," in *Workshop on Knowledge Markup and Semantic Annotation*, 2003, pp. 1–3.

[4] P. A. S. Sinclair, S. Goodall, P. H. Lewis, K. Martinez, and M. J. Addis, "Concept browsing for multimedia retrieval in the SCULPTEUR project," in *Proc. the 2nd Annual European Semantic Web Conference*, 2005, pp. 28–36.

[5] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Anal. Matching Intell.*, vol. 22, no. 12, pp. 1349–1380, 2000.

[6] H. Wang, L. T. Chia, and S. Liu, "Image retrieval ++–web image retrieval with an enhanced multi-modality ontology," *Multimed. Tools Appl*, vol. 39, no. 2, pp. 189–215, 2008.

[7] H. Wang, S. Liu, and L. T. Chia, "Does ontology help in image retrieval? A comparison between keyword, text ontology and multi-modality ontology approaches," in *Proc. 14th Annual ACM International Conference on Multimedia*, 2006, pp. 109–112.

[8] Y. I. A. M. Khalid, S. A. Noah, and S. N. S. Abdullah, "Towards a multimodality ontology image retrieval," in *Proc. the 2nd International Conference on Visual informatics: Sustaining* Research *and Innovations*, 2011, pp. 382–393.

[9] B. Russell, A. Torralba, K. Murphy, and W. Freeman, "LabelMe: A database and web-based tool for image annotation," *Int. J. Comput. Vis.*, vol. 77, no. 1, pp. 157–173, 2008.

[10] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Proc. International Workshop on Statistical Learning in Computer Vision*, 2004, pp. 1–22.

[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 91–110, 2004.

[12] L. AlSumait and C. Domeniconi, "Text clustering with local semantic kernels," in *Survey of Text Mining II: Clustering, Classification, and Retrieval*, M. W. Berry and M. Castellanos, Eds. London, United Kingdom: Springer-Verlag London Limited, 2008, pp. 87–105.

[13] L. Wang, Z. Lu, and H. H. Ip, "Image categorization based on a hierarchical spatial markov model," in *Proc. the 13th International Conference on Computer Analysis of Images and Patterns*, 2009, pp. 766–773.

[14] L. Wu, S. C. H. Hoi, and N. Yu, "Semantic-Preserving bag-of-words models for efficient image annotation," in *Proc. the 1st ACM Workshop on Large-Scale Multimedia Retrieval and Mining*, 2009, pp. 19–26.

[15] J. Zhu, V. Uren, and E. Motta, "ESpotter: Adaptive named entity recognition for web browsing," in *Intelligent IT Tools for Knowledge Management Systems, KMTOOLS 2005*, 2005, pp. 518–529.

[16] R. Arndt, R. Troncy, S. Staab, L. Hardman, and M. Vacura, "COMM: Designing a well-founded multimedia ontology for the web," in *Proc. the 6th International Semantic Web Conference*, 2007, pp. 11–15.

Stefan Poslad has a PhD from the University of Newcastle upon Tyne, UK. He is a senior lecturer at the School of Electronic Engineering and Computer Science, Queen Mary, University of London. His research interests 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.

Kraisak Kesorn holds a BSc in Computer Science of Chiang Mai University and an MSc in Information Technology (IT) of King Mongkut's Institute of Technology Ladkrabang from Thailand. He has a PhD in Electronic Engineering and Computer Science from Queen Mary College, University of London, United Kingdom. Dr. Kesorn is currently teaching and researching at the Computer Science and IT department, Faculty of Science, Naresuan University, Thailand, e.g. courses in Information Retrieval for undergraduates and post-graduates. His current research interests include semantic multimedia retrieval, knowledge-based modelling for information retrieval, and cross-language (Thai-English) information retrieval, as well as Semantic Business Intelligence (BI).