Elsevier Editorial System(tm) for Information Fusion
Manuscript Draft

Manuscript Number: IF13K03

Title: A Multi-Modal Incompleteness Ontology Model (MMIO) to Enhance Information Fusion for Image Retrieval

Article Type: Regular Paper Contributions

Keywords: Multi-modal Ontology; Knowledge base; Incomplete Ontology; Visual and Textual Information fusion

Corresponding Author: Dr. Kraisak Kesorn, Ph.D.

Corresponding Author's Institution: Naresuan University

First Author: Stefan Poslad, Ph.D.

Order of Authors: Stefan Poslad, Ph.D.; Kraisak Kesorn, Ph.D.

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 easy task to extract all data from documents or images and to automatically process these and transform them into Ontology model, because of ambiguity of terms or 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 Ontology-based system for image annotation based upon deriving two 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 to an Ontology model from computing a distance between the visual word features and concept features, so called concept range. The second index relies on a textual description which is processed to extract and recognise concepts, properties, or instances, defined in an Ontology. The two indexes are unified 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 exploited 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.

Information Fusion

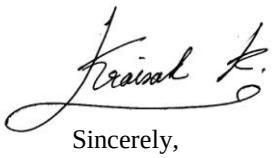
Dear Information fusion Review Coordinator

I would like to submit the attached manuscript, "A Multi-Modal Incompleteness Ontology Model (MMIO) to Enhance Information Fusion for Image Retrieval," for consideration for possible publication in the Research Articles section of the Information Fusion journal.

There are 3 documents submitted together as following:

1) A cover letter
2) Highlights
3) A Manuscript
4) 14 Figures
5) 2 Related self-publications

This paper has not been published or accepted for publication. It is not under consideration at other journals or elsewhere.



Sincerely,

K. Kesorn (Ph.D)

Computer Science and IT department,
Naresuan University
Phitsanulok, 65000
THAILAND
Email: kraisakk@nu.ac.th

Highlights

- Textual and Visual information are integrated into a single Ontology model.
- Bag-of-visual words vector space model is restructured to a hierarchical model using concept range.
- LSI is used to enhance the retrieval performance from an incomplete Ontology model.

A Multi-Modal Incompleteness Ontology Model (MMIO) to Enhance Information Fusion for Image Retrieval

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

[a]School of Electrical and Electronic Engineering and Computer Science,
Queen Mary University of London, E1 4NS, United Kingdom

[b]Computer Science and Information Technology Department, Science Faculty, Naresuan University,
Phitsanulok, 65000, Thailand

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 easy task to extract all data from documents or images and to automatically process these and transform them into Ontology model, because of ambiguity of terms or 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 Ontology-based system for image annotation based upon deriving two 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 to an Ontology model from computing a distance between the visual word features and concept features, so called *concept range*. The second index relies on a textual description which is processed to extract and recognise concepts, properties, or instances, defined in an Ontology. The two indexes are unified 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 exploited efficiently. An additional experiment demonstrates that LSI can

[*]Corresponding author. Tel. +6655-963267

Email address: stefan@eecs.qmul.ac.uk, kraisakk@nu.ac.th (K Kesorn)

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.

Keywords: Multi-modal Ontology, Knowledge base, Incomplete Ontology, Visual and Textual Information fusion

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 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 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.

Gaservic [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 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 contain the phase "London 2012". Ontologies can 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 done based upon only single-modality information, either textual information or visual features. Consequently, those works suffer 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 used to represent 'things' in different ways. However, there are some invariant and implicit connections between them [8]. Often, the textual information surrounding images includes descriptions of images generated by humans. These image captions, these 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.

Combining text and image features has been proposed in order to improve the image search results by several researchers [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 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 to build a complete KB due to several factors. For example:

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 the systems using 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 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 is manipulated and processed by standard information retrieval methods. Image data contains large numbers of unstructured and dynamic visual features. How to establish a good knowledge representation model to represent visual content is very important for IMR [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 et al. [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 support only text-based query. Schreiber et al. [2] presented the 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 higher level conceptualisation. Thus, the system is able to interpret the image content without using text descriptions. However, for the case where some objects needed for interpretation are absent, this leads to a system failure to annotate the image. Wang et al. [20] introduced *m*-PCNN, a multi-channel Pulse Coupled Neural Network, for medical data fusion to diagnose diseases. Similar to work in [21,22] tried to integrate the information from two or more images into a single one to improve image analysis, e.g. in 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 their 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] and incorporates a textual description Ontology. The proposed method is able to increase classification accuracy and retrieval performance. Khalid et al.[15] proposed multimodality Ontology framework for the sport domain. Textual descriptions and 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, Label Me Annotation Toolbox [26]. 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, so called visual heterogeneity problem and this is a well-recognised problem among researchers in the image-processing area. The most similar work to our idea is the work of Chen et al. [27] who, bridge the semantic gap by integrating text and visual features. Their main research motivation is 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 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 (SIFT descriptors) are investigated and used to solve such a 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 concept 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 [28]. 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 because of 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 [29]. This is because of the different languages 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 the 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 [29]. 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 [30]. 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 [31–35].

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 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 [34]. 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/sub-class 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 resolve the limitations of the current state of the art, is introduced.

3. Method

3.1. Knowledge-based Design for Information Fusion

Knowledge representation process starts with Ontology designing and modeling. As such, it relies on a common understanding of how humans understand, acquire and represent the 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*. Ontology is composed of concepts and their relationships. To produce a formal Ontology, an Ontology representation language is selected in order to formalise an Ontology conceptualisation and produce a hierarchy of concepts (organisation of concepts into a "subclass_of" relation). 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 the pole vault event. As such, they are ambiguous. This brings another challenge to the system in order to annotate an image properly and to deal with a polysemy issue. After surveying data at the Olympic organisation website, three main classes of Ontology have been defined, a Sport Domain, Visual Features and 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 Descriptions Ontology* provides the annotation structure template for the (sports) domain.

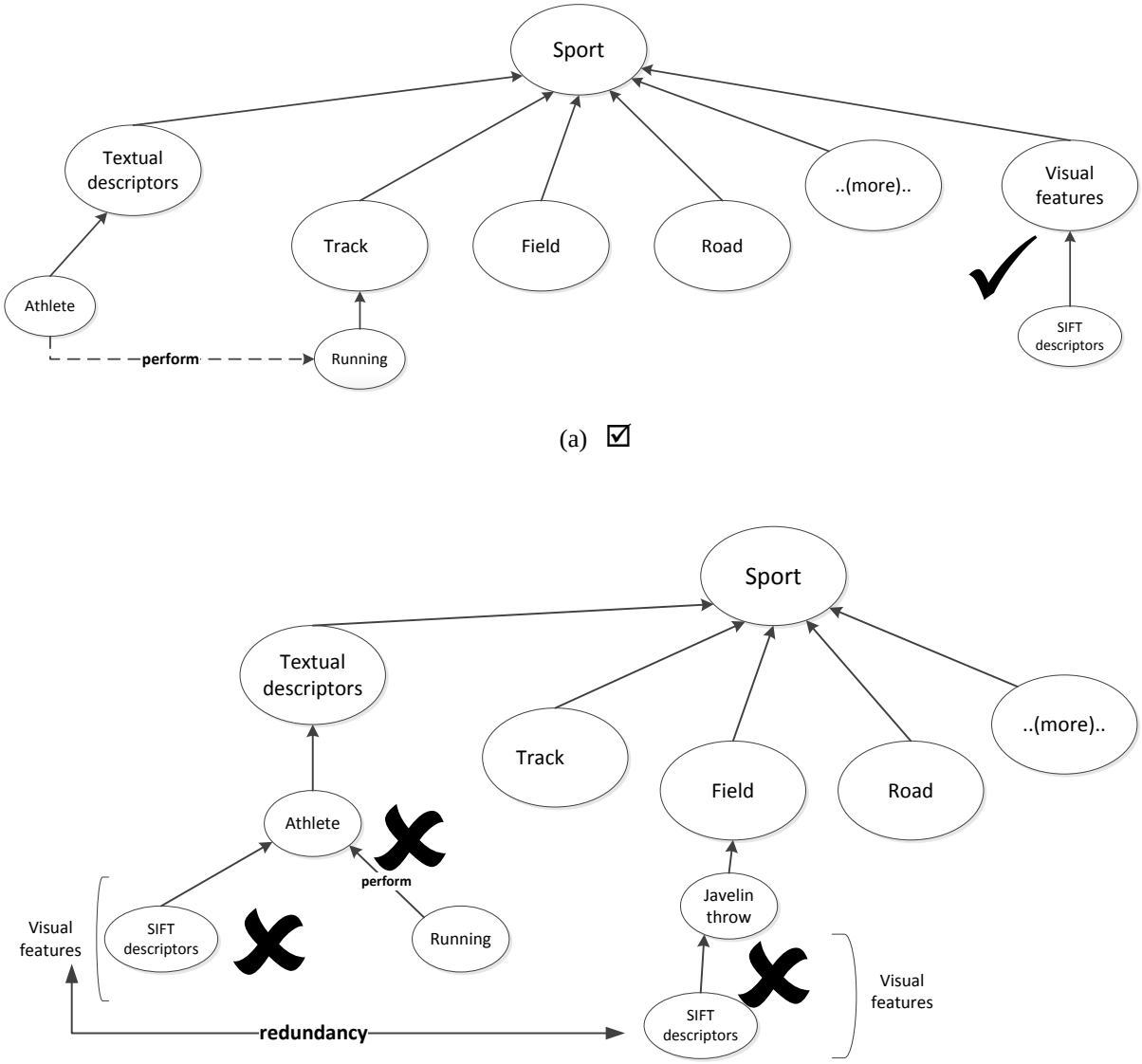


Fig. 1 Ontology design choices.

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 have been decomposed into sub-classes. 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, Visual feature Ontology is designed as main concepts (Fig. 1(a)) rather than as subclasses of a concept in Textual descriptors or Sport domain Ontology (Fig. 1(b)). 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.1(b). Second, the Ontology

structure more efficiently facilitates sport image retrieval (Fig. 1(a)). For instance, "*athletes who perform Track event*", the structure in Fig. 1(b) may make it more difficult to find an answer. For Fig. 1(a), the system easily find 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. 1(b) can make it is impossible to answer this simple query because there no link between *Running* and *Track* event. The Ontology structure in Fig. 1(b) 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, the upper level concepts are a generalisation of the lower level concepts. However, in Fig. 1(b), the *Athletes* concept is not a generalisation of *Running* class as they do not share any common properties. Therefore, the structure and relationships in Fig.1(a) are preferred to represent 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* [36]. 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 [37]. 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 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 LSI [38]. 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 LSI vector space model. A LSI model provides term frequency information which can be used for an implicit semantic search when this information is not contained in the Ontology KB. More details about how LSI is applied are explained in section 4.2.4

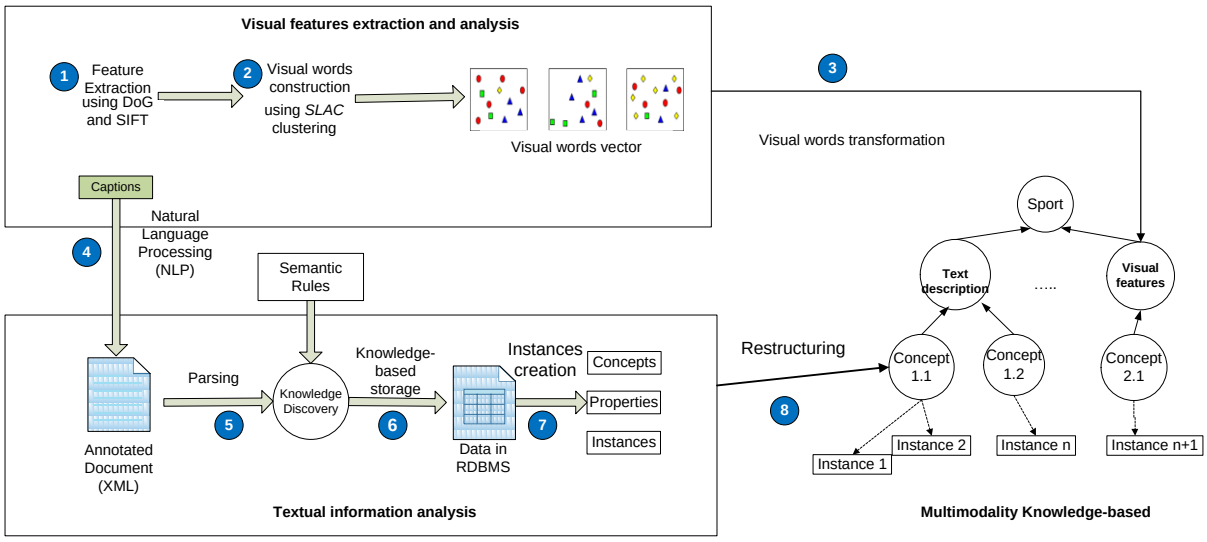


Fig. 2 Knowledge acquisition processes.

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 integrated 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

Higher level representations are computed from lower level ones. Two main processes for visual analysis are defined. First, a low-level image processing extracts useful visual features from images and interpret 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 on the primitive objects and specific prior knowledge (knowledge about a sports event which is not explicitly addressed in the data) 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 [39]. 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, step 1-3): low-level feature extraction, visual word construction and restructuring a vector space model into a hierarchical structure. These are described in more detail 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 [40] 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 to 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 Semantic Local Adaptive Clustering (SLAC) algorithm [41] to cluster the vectors. SLAC improves the conventional clustering techniques by capturing relevant local features within 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 and, thus, SLAC generates semantically related visual words. Consequently, quality of the obtained visual words 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, [42–44] 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 [44], Agglomerative clustering algorithm [45], and hierarchical Latent Dirichlet Allocation algorithm, or hLDA, [46]. 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 (Multimodal Information Fusion to support Image Retrieval using an Incomplete Ontology Model) 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 also apply 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 proposed 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 [47]. Finally, 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 of several 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 between the centroid (ν) of a visual word (ϖ) and the centroid of concept (c_i), the so called *concept range*. The concept range equation [44] [47] is shown as following:

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

The main benefit of the concept range technique 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 Equation (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.

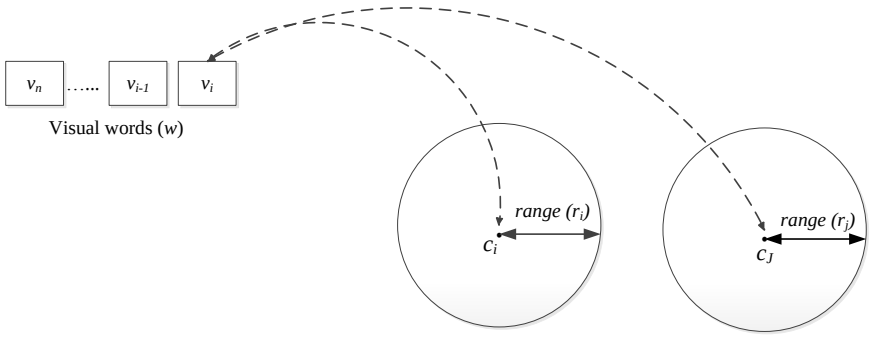


Fig. 3 Visual words are assigned to concept(s).

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 Equation (2), where *N* is a number of instances of visual content.

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

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 [48] 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 enables 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 map to OWL using a JDBC connector API (Fig. 4). This generic approach is useful in numerous cases, but sometimes may is complicated by: difficulties in synchronising it to changes in the database structures, to difficulties in installation and to 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 the Structured Query Language (SQL).
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 OWL file.

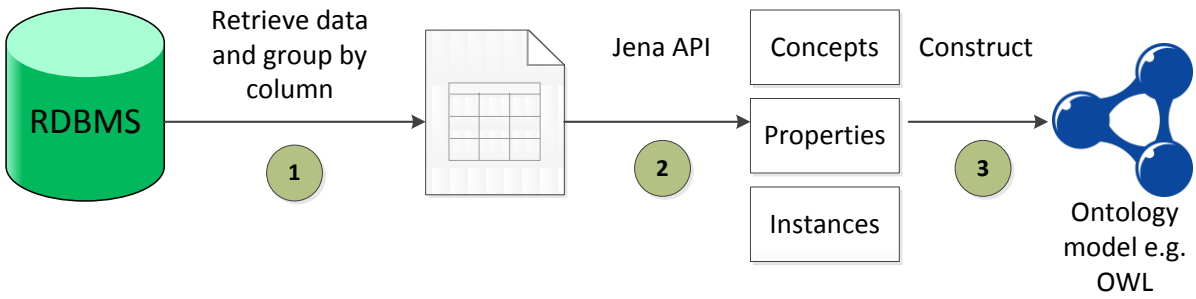


Fig. 4 Transformation process from RDBMS to Ontology model (OWL).

4.2. Multi-modal 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 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 [49]. Therefore, a huge effort [50–53] 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.

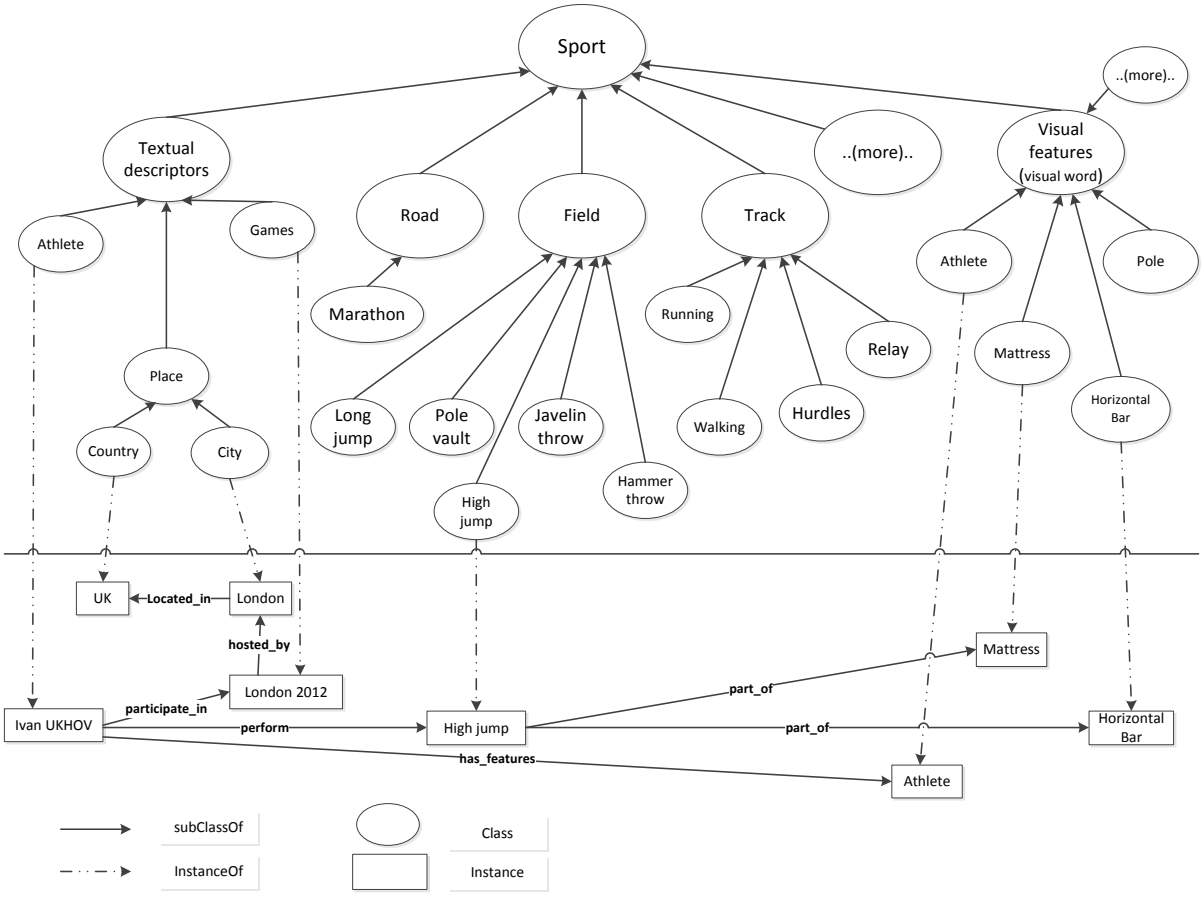


Fig. 5 Multi-modal Ontology structure.

4.2.1. Sports 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 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 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 domain Ontology to cover everything in a real world 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 [38]. LSI tries to search for things that are closer to representing the underlying semantics of a document [54].

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 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.

$A=$

	Img1	Img2	img3	img4	img5
swimming	1	1	1	0	1
Michael Phelps	0	1	1	1	1
United States	1	1	0	1	1
free style	0	0	1	0	1

Fig. 6 Example of a term frequency matrix.

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

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

where W_{ij}= Weight of each term, L_{ij} is the local weight for term i in document (image caption) j, G_i is the global weight for term i, and N_j is the normalisation factor for document (image caption) j. There are several formulae for local, global weight, and normalisation factors. In this framework, the selected formulae are based on the survey and recommendation given by [55].

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.

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

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 s 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 normalisation factor compensates for differences in document lengths. Equation (5) shows the normalisation formula.

$$N_j = \frac{1}{\sqrt{\sum_{i=0}^{m} \left(G_i L_{ij}\right)^2}} \quad (5)$$

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 Equation (6) in order to reduce the matrix dimensions.

$$A = USV^T \quad (6)$$

Having computed the SVD, the columns in *U* is reduced to *k* and the values in matrix *S* are sorted in ascending order. In other words, the matrix U indicates the semantically importance of concepts. This is from where the term "latent semantics" comes from. Unimportant concepts are regarded as "semantic noise". To reduce the dimensions of the matrix, only the largest *k* singular values will be selected. Therefore, the decomposition can be approximated as:

$$A \approx U_k S_k V_k^T \quad (7)$$

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.

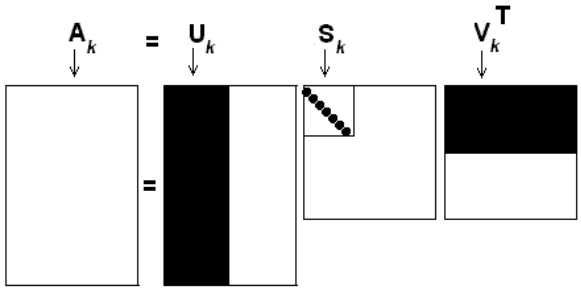


Fig. 7 Rank-*k* SVD approximation for matrix dimensional reduction.

The main advantage of LSI is the ability to infer indirect concepts when those concept 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 8,000 images (400 images from each group).

5.1.1. 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 [56]. 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 the Ontology structure designing that 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 compare the retrieval performances with commercial search engines.

5.1.2. 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 [57].
- *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 and 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 be able to find all relevant images based upon any distinctive low-level visual features. Therefore, both modalities should be used 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 contains 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 everything in a domain in one phase of development. When the information in the knowledge-based is incomplete, retrieval performances (precision and recall) are 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*).

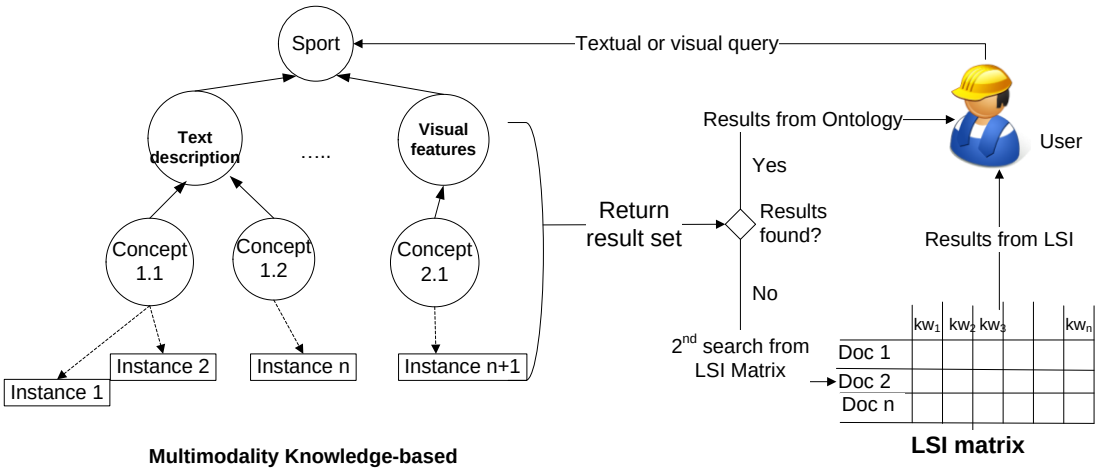


Fig. 8 Image retrieval mechanism to support the presented KB design.

6. Experimental Results and Discussion

Four experiments are conducted to evaluate the performance of the MMIO model to validate the proposed hypotheses.This section presents an approach towards image retrieval based upon extracted knowledge, and visual and textual features (Fig. 8). If the domain Ontology is designed properly, i.e. in this case to fit the sport domain, the retrieval results should be significantly improved. It starts when the query keywords from users are examined by tokenising them and, then, stop words are eliminated. The remaining words are assumed to be the important keywords for searching images and these keywords will be disambiguated using WordNet. To find an appropriate word sense for a user query, it is proposed 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 where p f**{
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 in 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 searching, average precision has been used. More details of the experiments is explained in the next section.

6.1.1. 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 implement another five state-of-the-art techniques, LSI [38], Original Bag-of-Words (OBW) [39], LIRE (www.semanticmetadata.net/lire/), Lucene, and Visual-Features-Ontology (VFO), with similar experiment settings and then compared the results with the proposed MMIO method. VFO contains only visual features in the form of an Ontology. VFO is built based on 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.

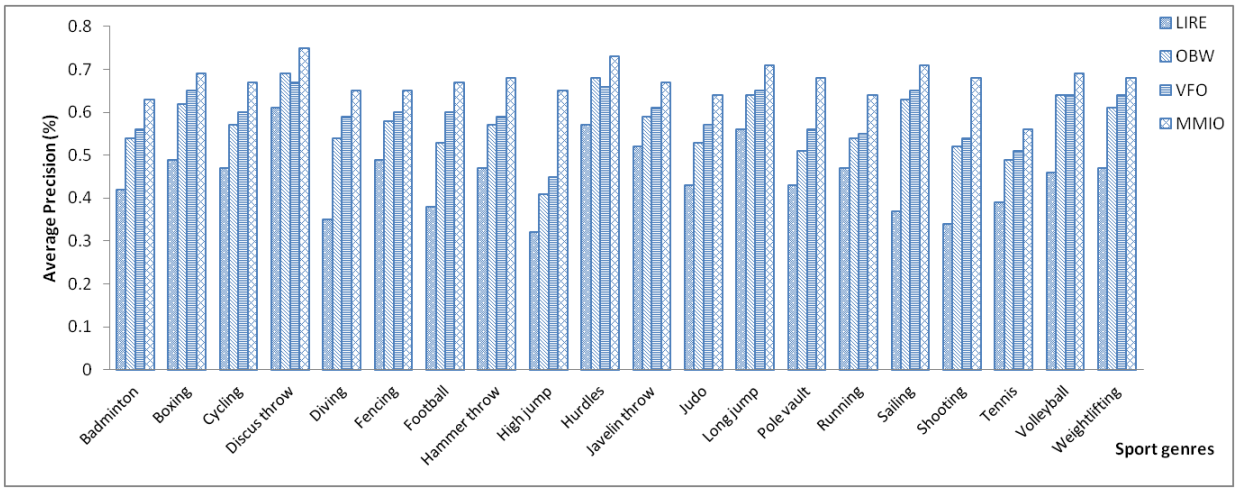


Fig. 9 Precision graph comparison for visual queries.

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. This is 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 properly, e.g. concepts in this kind of tree cannot overlap or support multiple inheritances. 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 acquires the lowest precision compared with the other techniques.

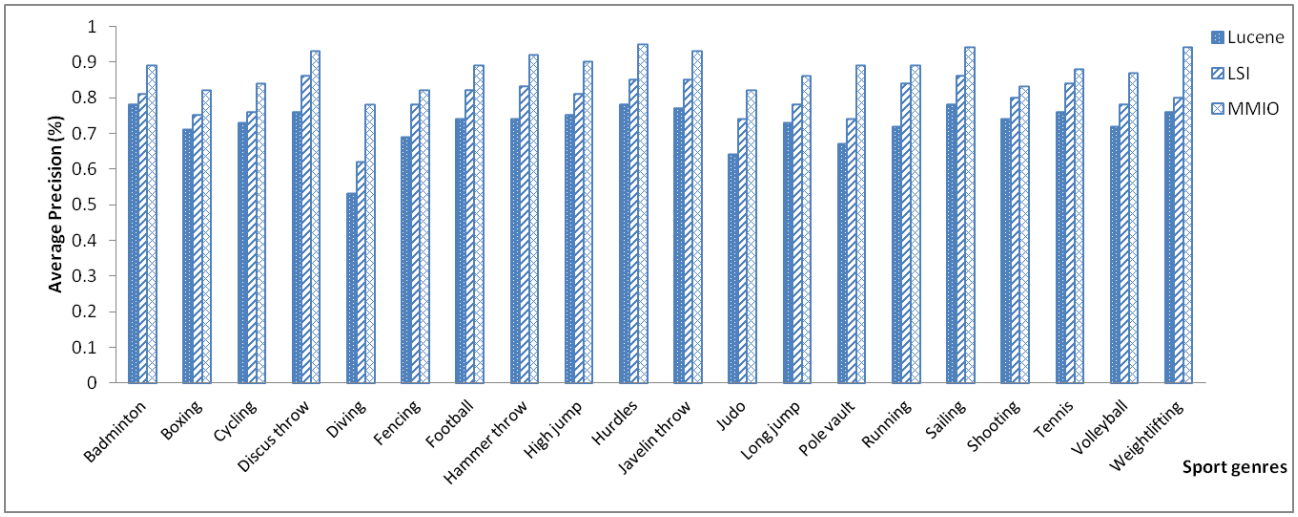


Fig. 10 Precision graph comparison for textual queries across sports genres.

When a textual query is submitted, all keywords will be extended to find other similar keywords in WordNet and 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 under 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 and 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 Fig. 9 and Fig. 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.

6.1.2. Experiment II: Semantic Search Study

In this experiment, testing the hypothesis investigates that if the structure of Ontology 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 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 the Olympic games in Great Britain" and "Great Britain athletes participate the Olympic games in Thailand". The search engine of Lucene will retrieve 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 which are not expressed in the query. As a result, only relevant images regarding 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.

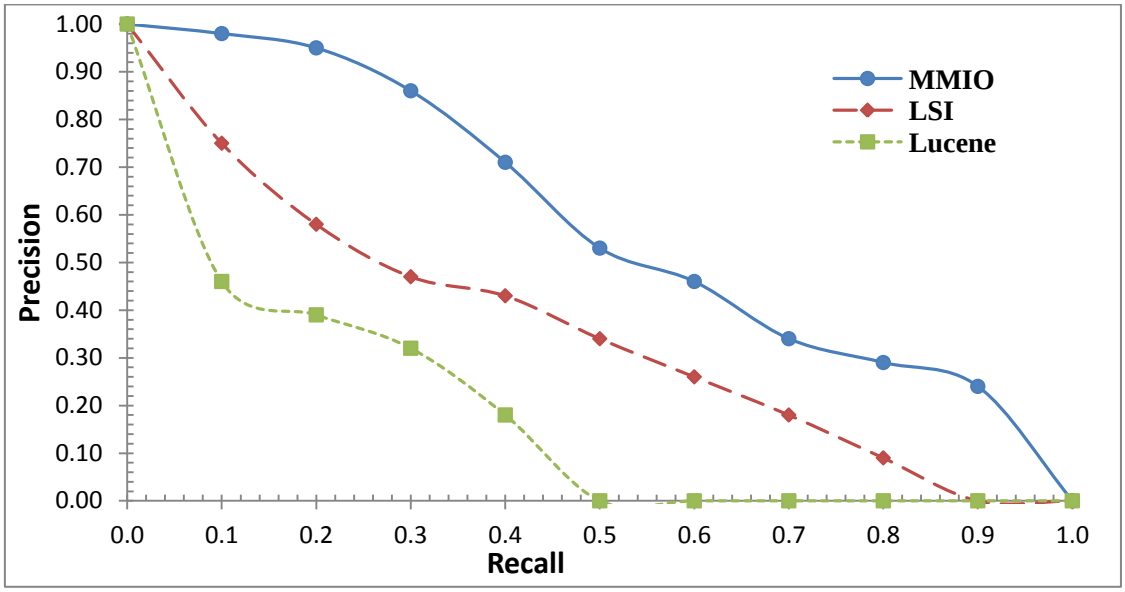


Fig. 11 Retrieval results of the query for MMIO, LSI and Lucene.

Since MMIO contain multi-modal information, it can handle both types of queries more effectively than other state of the art techniques which focus only on use of 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 aVisual-feature Ontology. However, creating a complete Ontology is a challenging task. In the next section, the issue of Ontology incompleteness is studied and analysed.

6.1.3. 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 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 presented KB model, so called *open KB*, can handle *unresolvable* data better than a general KB, so called *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.

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 London 2012 Games

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

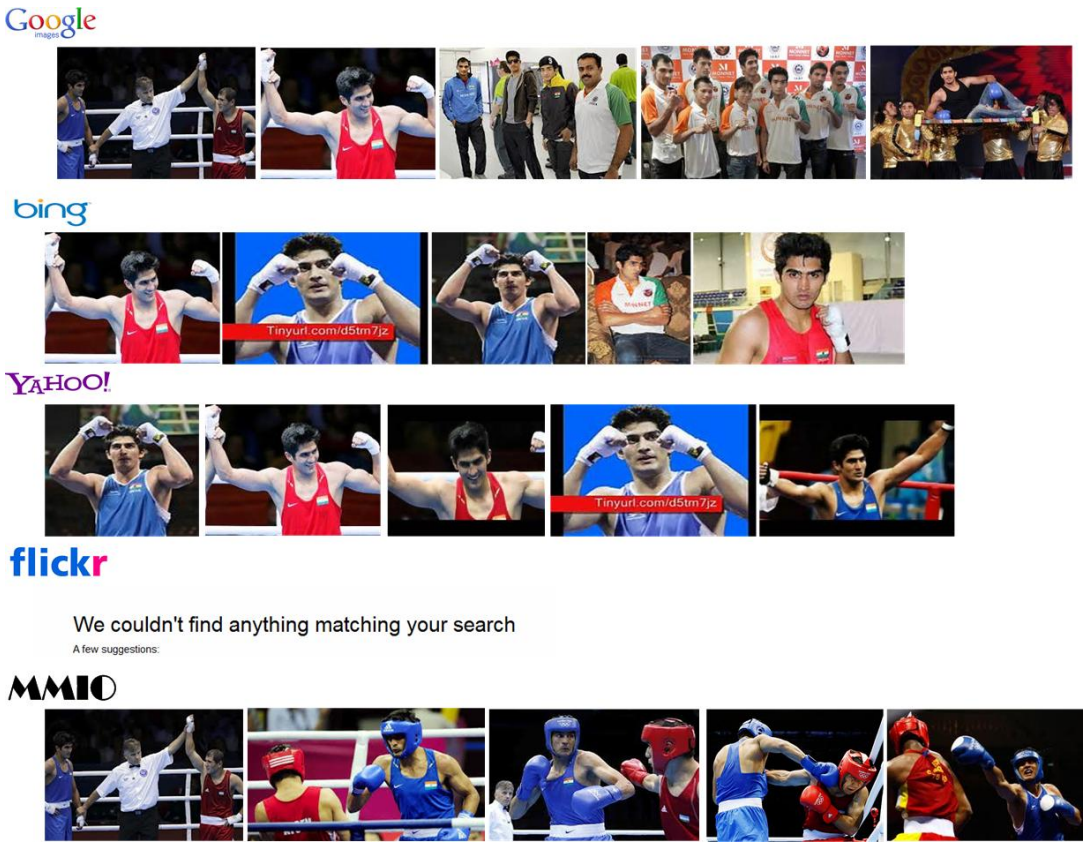
Fig. 12 Example of LSI results when the KB does not contain the London 2012 information.

6.1.4. Experiment IV: Retrieval Performance Comparison against State-of-the-Art Search Engines

An additional experiment was conducted to compare the 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.

Four of the search engines can find images that Vijender Singh (*athlete*) is performing the Boxing (*sport*) competition in London 2012 (*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 “performs Boxing competition”, instead they show mostly Vijender Singh’s face. This is because these search engines use 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.



*The experiment conducted on 05/09/2013

Fig. 13 Top five results for the example query “Vijender Singh performs the Boxing competition in London 2012” obtained from five search engines.

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 recognized 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 recognize 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 of experiment.

*The experiment conducted on 05/09/2013

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

Fig. 14 shows the results of two search engines using a visual query (QBE). The query (Q) is an image of "Vijender Singh performs the Boxing competition in London 2012". 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 will use 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 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 the Textual Description Ontology.

To summarise this evaluation, firstly, using text-based query retrieval state-of-the art search engines, they will miss several relevant images that do not have the terms in the captions matched to the query and, thus, precision is reduced. MMIO is able to infer the implicit semantic behind the query, select only the semantically relevant images, and yield better results. Secondly, the results from using visual-based query 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.

7. Concluding Remarks

This research promotes the design of a knowledge-based framework (MMIO) that provides a semantic-based solution to multimodal image retrieval. The MMIO KB only relies on visual similarity but also on conceptual similarity to improve image retrieval. In addition to the visual analysis component, this paper also proposes a technique for acquiring knowledge from text captions to improve the knowledge representation used for knowledge-based visual content retrieval. The main innovation for textual information analysis is to extract the essential metadata from text captions and transform the unstructured metadata into semantic concepts automatically. Three main steps are defined to complete this task. Both textual and visual information are

encoded into a unified KB model in the form of OWL to facilitate semantic retrieval. This enables MMIO to retrieve image correctly and robustly. In addition, the MMIO KB is designed as an open KB rather than as a closed KB. The advantage of this is that the system can handle unknown queries more effectively and robustly than existing image search engines.

Further work is to extend MMIO to support personalized 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, 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: Work. 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, 2nd 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.
[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. Work. 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 Appl. 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. Informatics Sustain. Res. Innov., 2011: pp. 382–393.
[16] G. Nagypál, Possibly Imperfect Ontologies for Effective Information Retrieval, University Karlsruhe (TH), Germany, 2007.
[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.
[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. Appl. 38 (2011) 11472–11481.
[25] H. Wang, S. Liu, L.-T. Chia, Does ontology help in image retrieval?: a comparison between keyword, text ontology and multi-modality ontology approaches, in: Proc. 14th Annu. ACM Int. Conf. Multimed., 2006: pp. 109–112.
[26] 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.
[27] 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.
[28] 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.
[29] 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.
[30] 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.
[31] T.R. Gruber, A Translation Approach to Portable Ontology Specifications, Knowl. Acquis. 5 (1993) 199–220.
[32] 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.
[33] E.N. Guarino, R. Poli, N. Guarino, Formal Ontology, Conceptual Analysis and Knowledge Representation, Int. J. Hum.-Comput. Stud. 43 (1995) 625–640.
[34] B. Chandrasekaran, J.R. Josephson, V.R. Benjamins, What Are Ontologies, and Why Do We Need Them?, IEEE Intell. Syst. 14 (1999) 20–26.
[35] D. McGuinness, Ontologies Come of Age, in: D. Fensel, J. Hendler, H. Lieberman, W. Wahlster (Eds.), Semantic Web Why What, MIT Press, Boston, MA, 2003.
[36] S. Poslad, Ubiquitous Computing Smart Devices, Environments and Interactions, John Willey & Sons, London, United Kingdom, 2009.
[37] A. Llorente, S. Rüger, Using Second Order Statistics to Enhance Automated Image Annotation, in: Proc. 31th Eur. Conf. IR Res., 2009: pp. 570–577.
[38] 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.
[39] G. Csurka, C.R. Dance, L. Fan, J. Willamowski, C. Bray, Visual Categorization with Bags of Keypoints, in: Int. Work. Stat. Learn. Comput. Vis., 2004: pp. 1–22.
[40] D.G. Lowe, Distinctive Image Features from Scale-Invariant Keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.
[41] 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: pp. 87–105.

[42] 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.
[43] 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.
[44] 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.
[45] B. Walter, K. Bala, M. Kulkarni, K. Pingali, Fast Agglomerative Clustering for Rendering, in: IEEE Symp. Interact. Ray Tracing, 2008: pp. 81–86.
[46] D.M. Blei, A.Y. Ng, M.I. Jordan, Latent Dirichlet Allocation, J. Mach. Learn. Res. 3 (2003) 993–1022.
[47] L. Wu, S.C.H. Hoi, N. Yu, Semantic-Preserving Bag-of-Words Models for Efficient Image Annotation, in: Proc. 1st ACM Work. Large-Scale Multimed. Retr. Min., 2009: pp. 19–26.
[48] 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.
[49] 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.
[50] 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: pp. 113–122.
[51] S. Dasiopoulou, V. Tzouvaras, I. Kompatsiaris, M.G. Strintzis, Enquiring MPEG-7 Based Multimedia Ontologies, Multimed. Tools Appl. 46 (2010) 331–370.
[52] 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, 2006: pp. 291–294.
[53] 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: pp. 237–250.
[54] P. Praks, S. Snasel, J. Dvorský, Latent Semantic Indexing for Image Retrieval Systems, in: Proc. SIAM Conf. Appl. Linear Algebra, 2003: pp. 1–8.
[55] 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.
[56] J. Brank, M. Grobelnik, D. Mladenić, A Survey of Ontology Evaluation Techniques, in: Proc. Conf. Data Min. Data Warehouses, 2005: pp. 166–169.
[57] G. Guida, G. Mauri, Evaluating Performance and Quality of Knowledge-Based Systems: Foundation and Methodology, IEEE Trans. Knowl. Data Eng. 5 (1993) 204–224.

Fig. 1 Ontology design choices.
Click here to download high resolution image

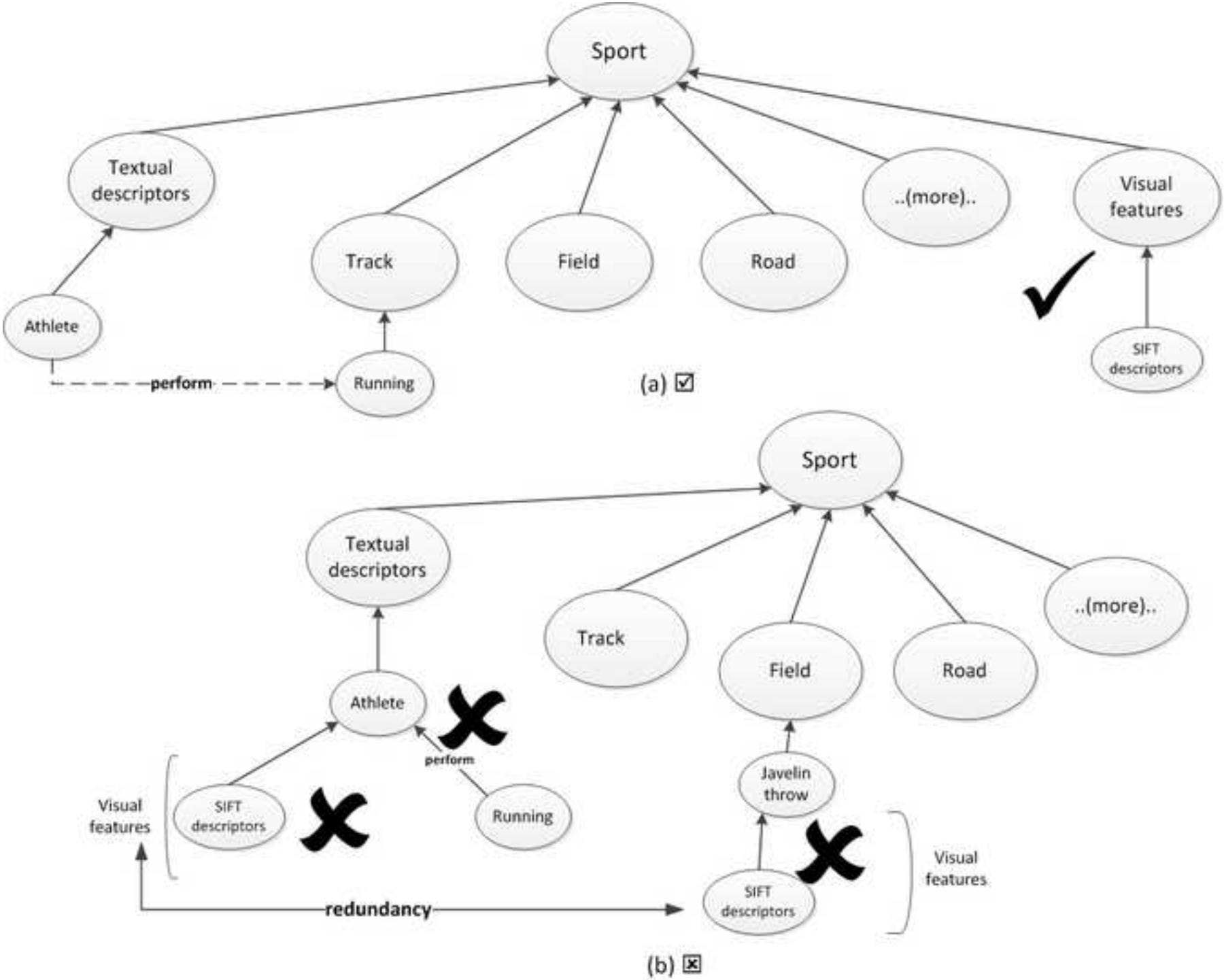


Fig. 2 Knowledge acquisition processes.
Click here to download high resolution image

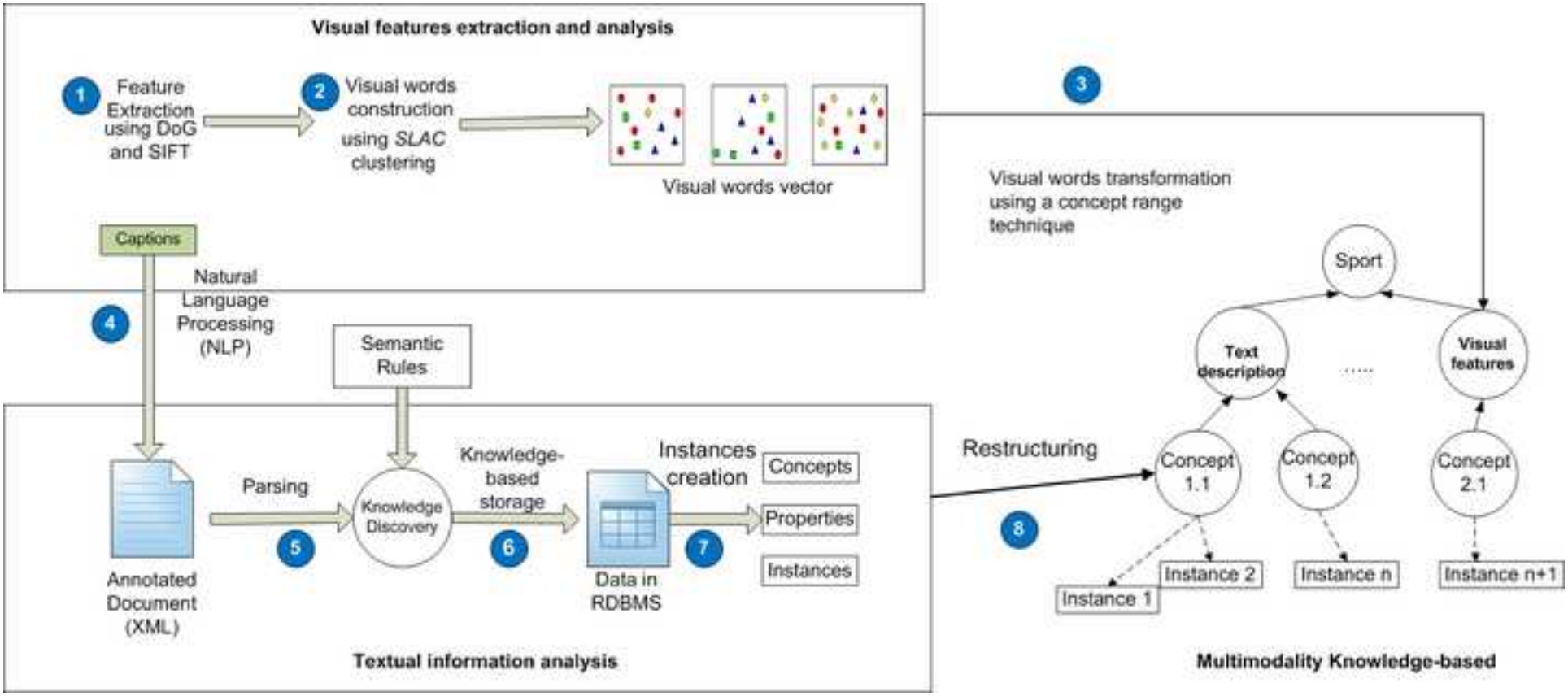


Fig. 3 Visual words are assigned to concept(s).
Click here to download high resolution image

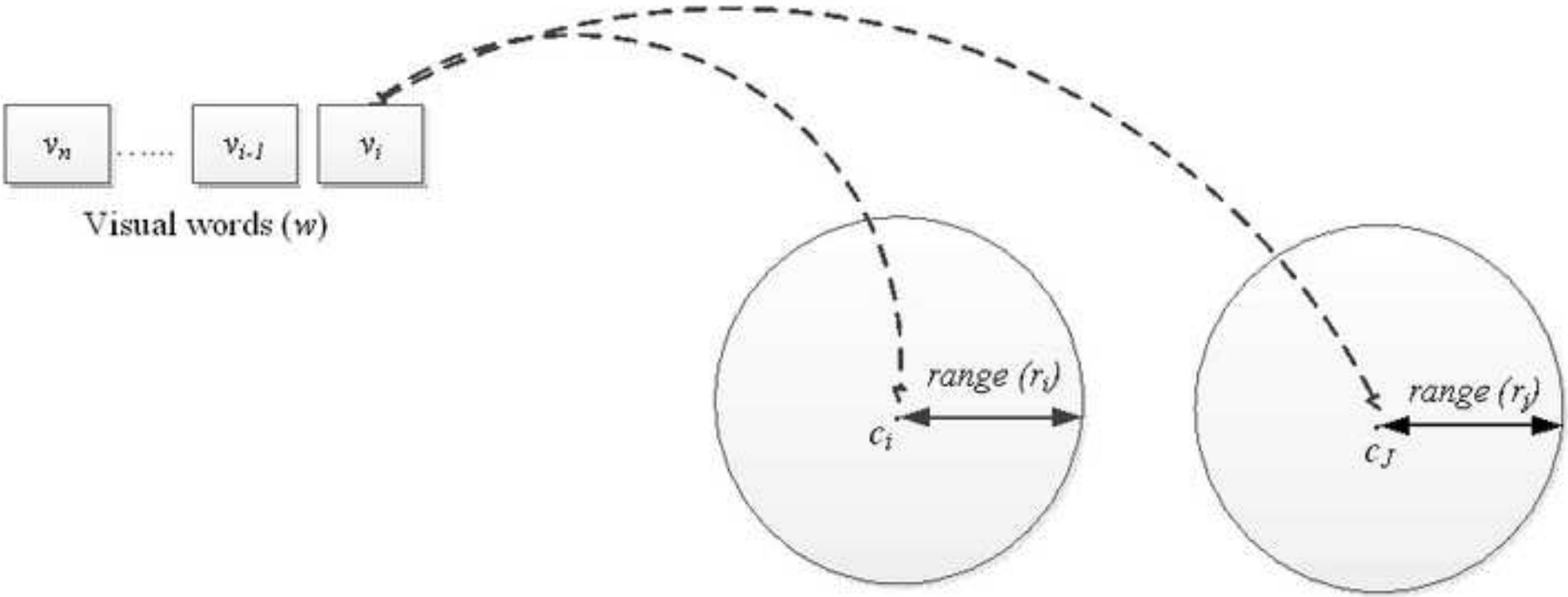


Fig. 4 Transformation process from RDBMS to Ontology model (OWL)
Click here to download high resolution image

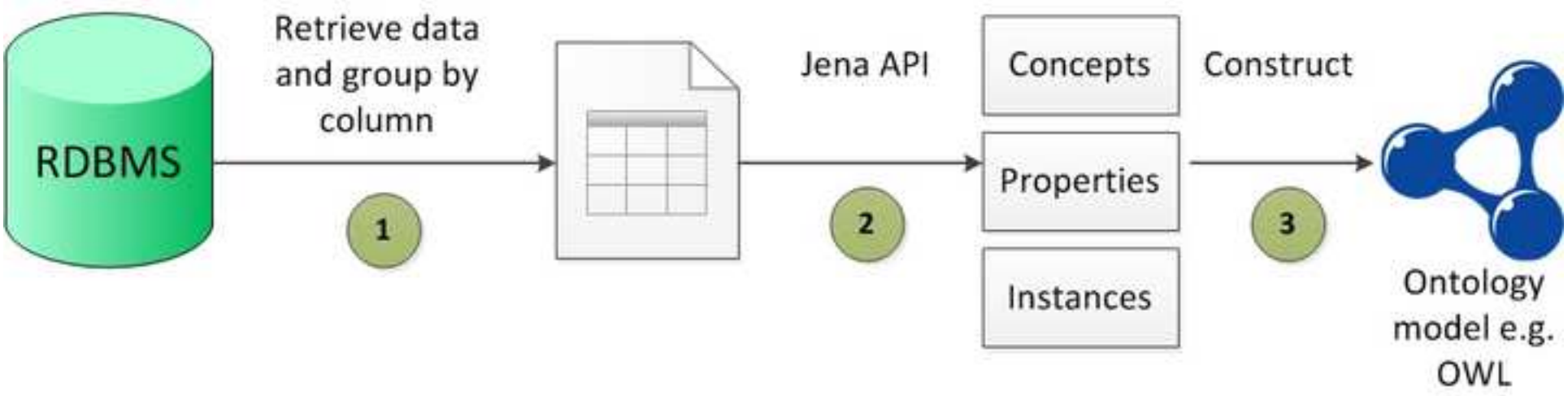


Fig. 5 Multi-modal Ontology structure.

Click here to download high resolution image

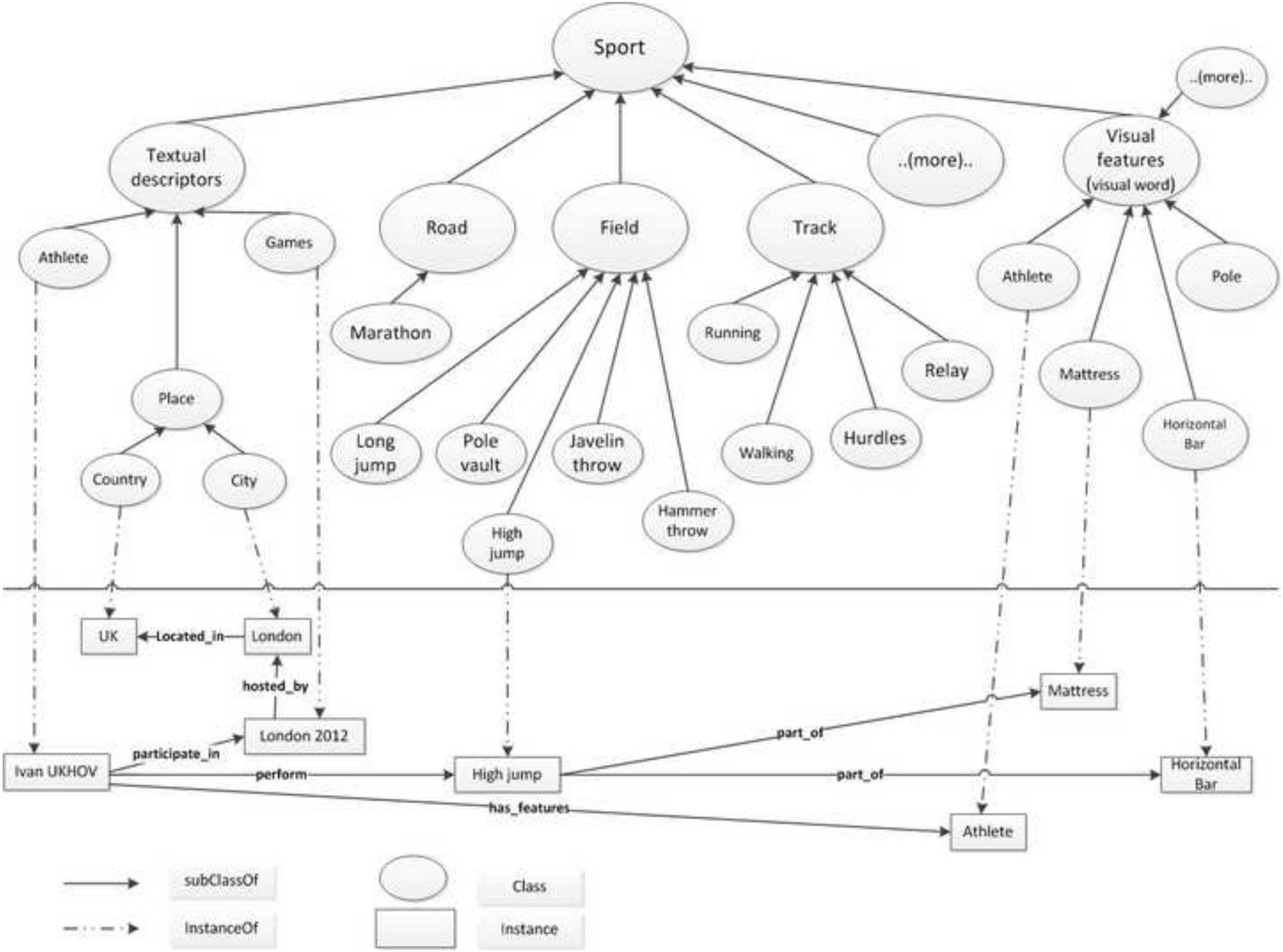


Fig. 6 Example of a term frequency matrix.
Click here to download high resolution image

$A =$

	Img1	Img2	img3	img4	img5
swimming	1	1	1	0	1
Michael Phelps	0	1	1	1	1
United States	1	1	0	1	1
free style	0	0	1	0	1

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

Click here to download high resolution image

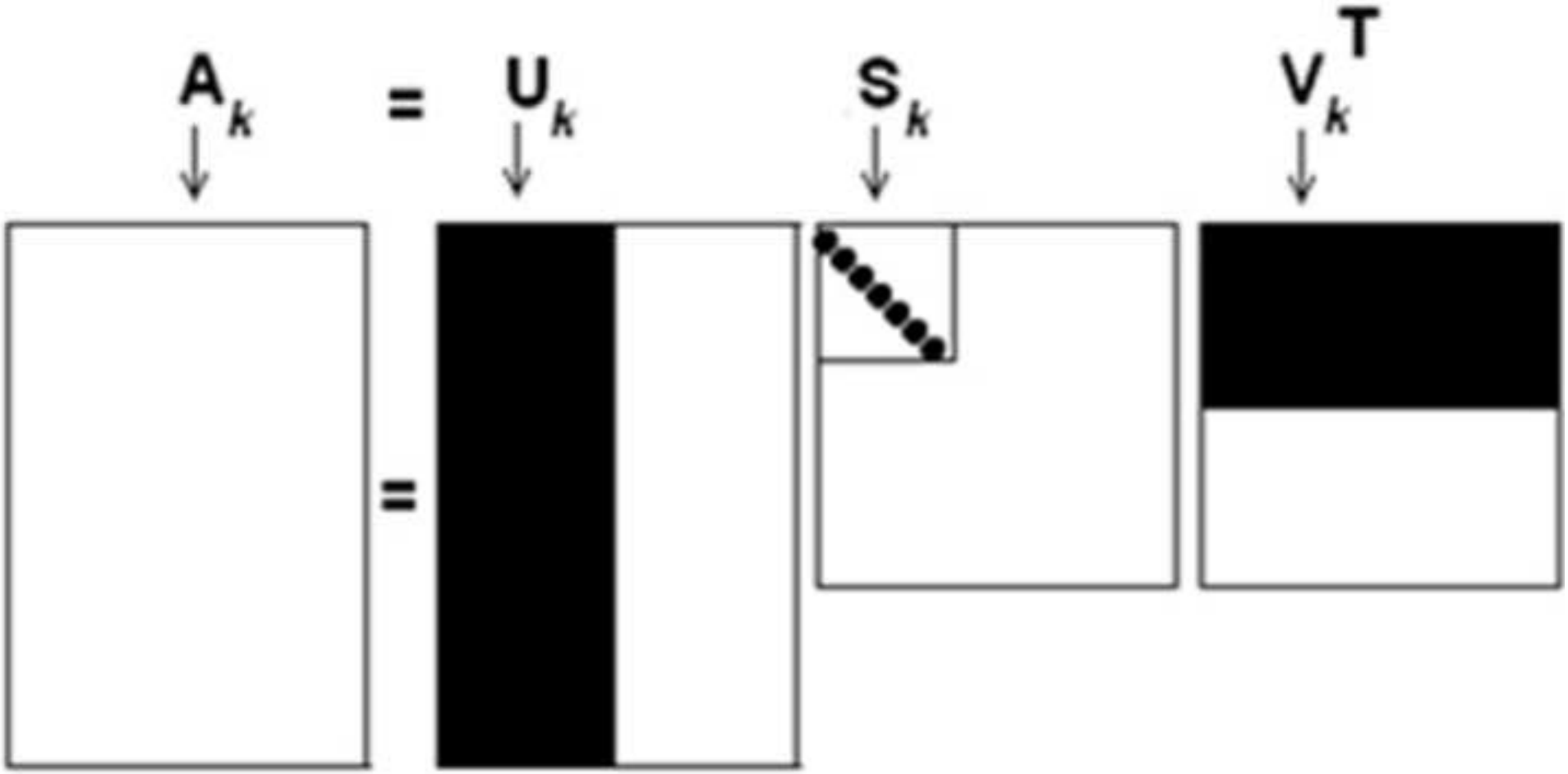


Fig. 8 Image retrieval mechanism to support the presented KB des

Click here to download high resolution image

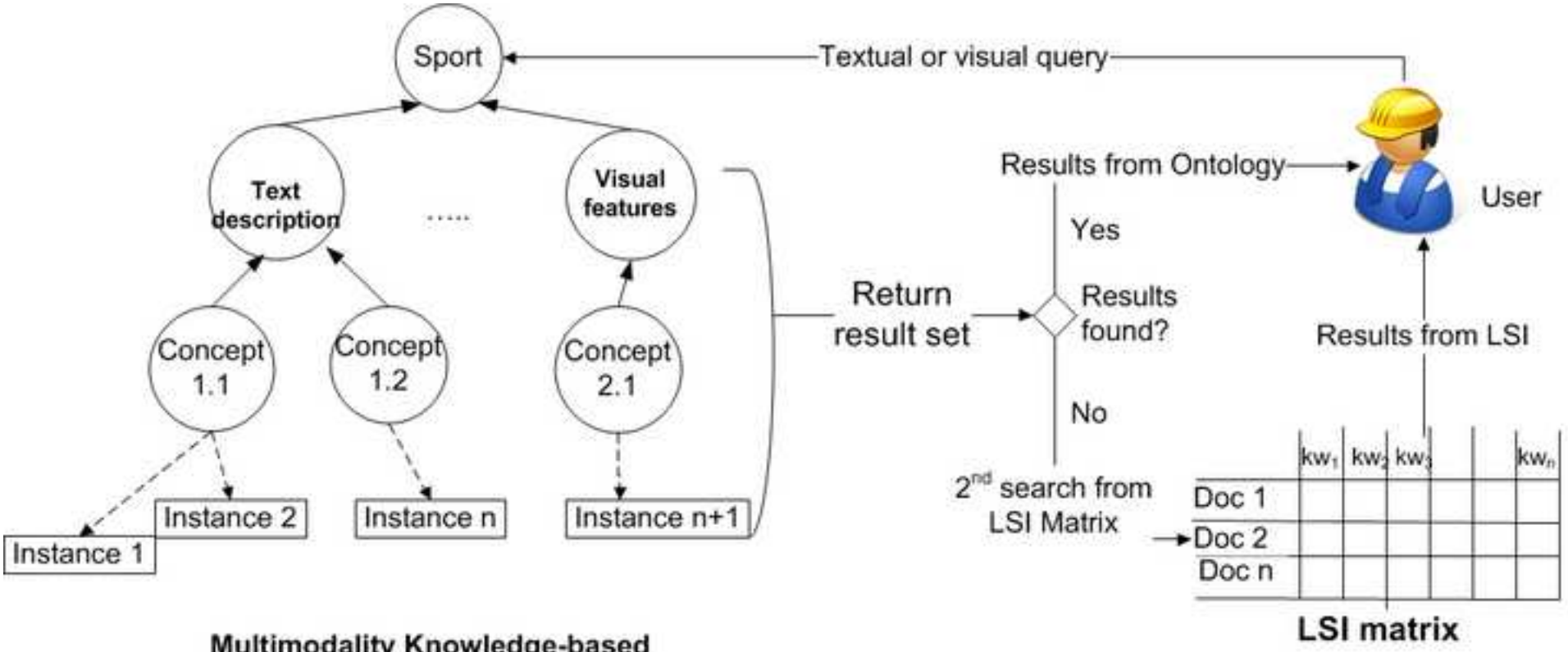


Fig. 12 Example of LSI results when the KB does not contain the
Click here to download high resolution image

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 London 2012 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. 9 Precision graph comparison for visual queries.
Click here to download high resolution image

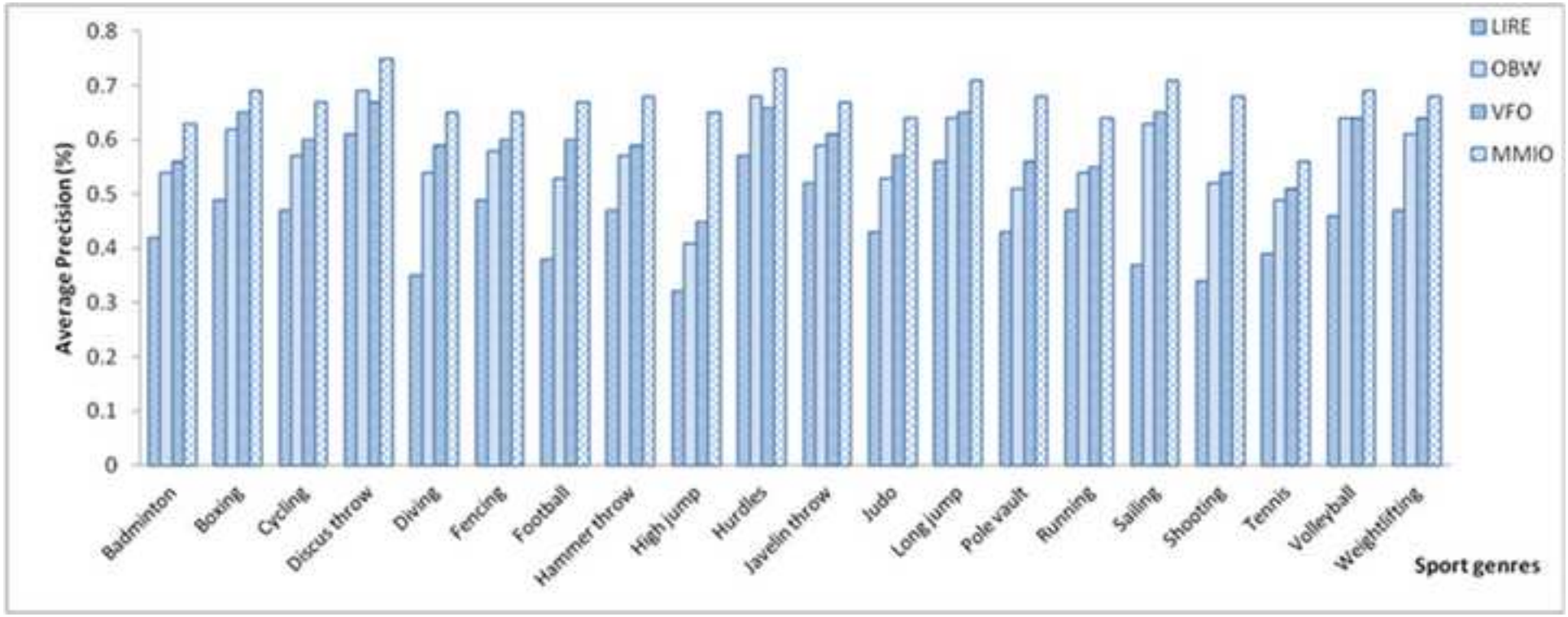


Fig.10 Precision graph comparison for textual queries across sp

Click here to download high resolution image

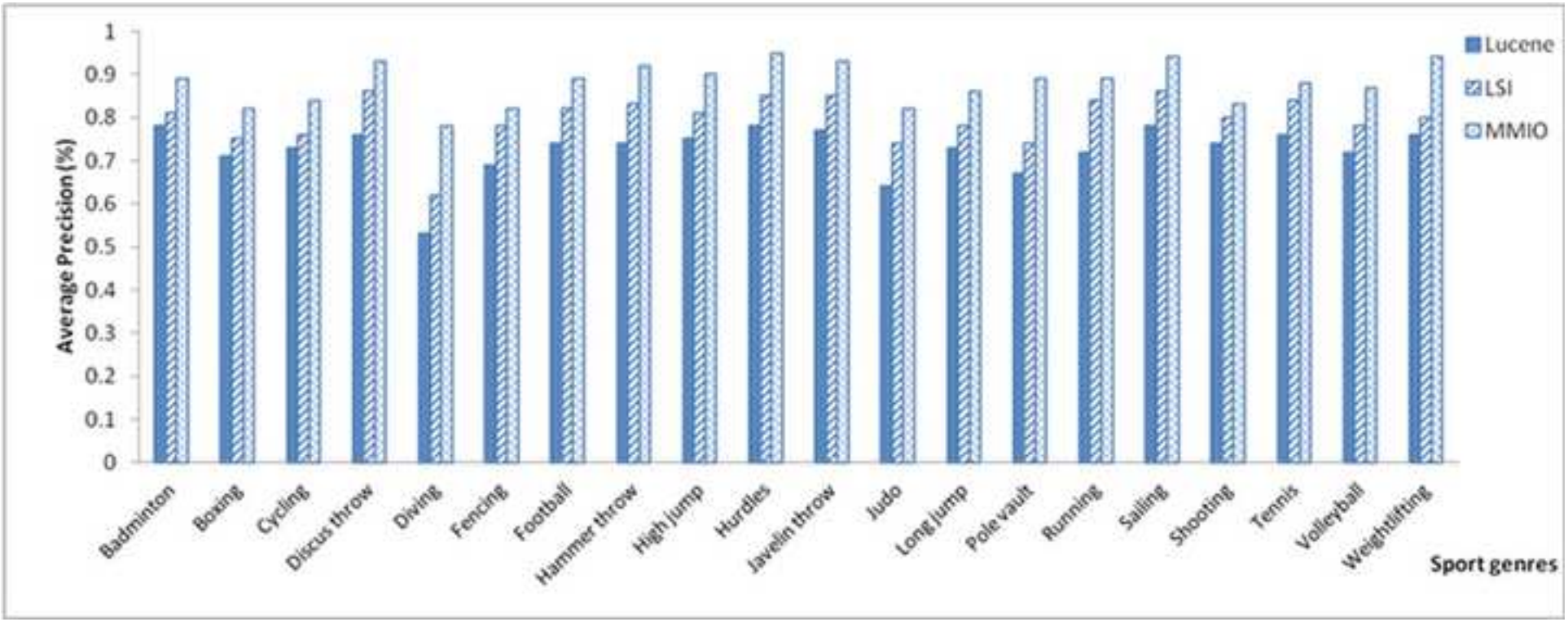


Fig. 11 Retrieval results of the query for MMIO, LSI and Lucene.
Click here to download high resolution image

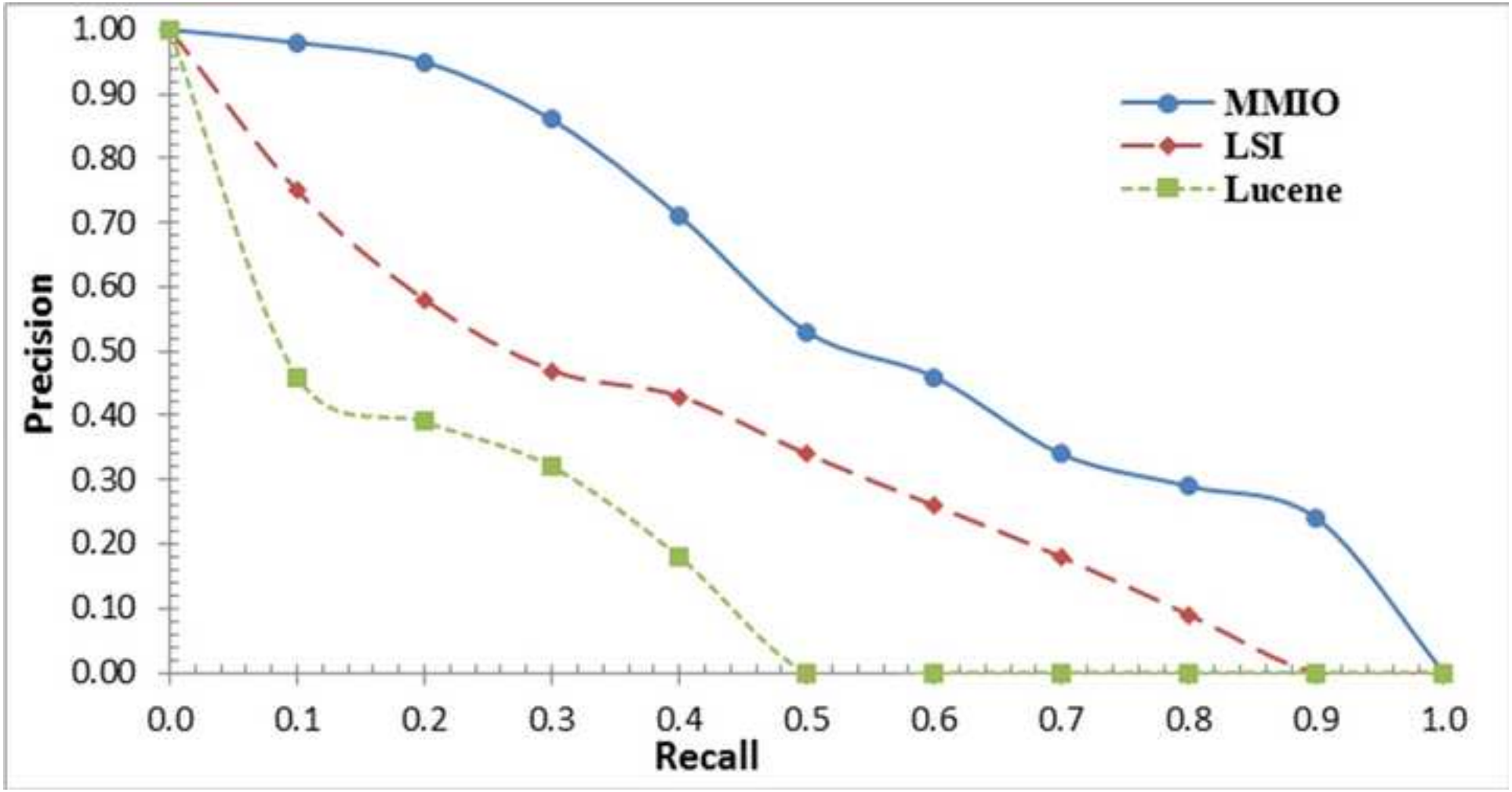


Fig. 13 Top five results for the example query "Vijender Singh..

Click here to download high resolution image

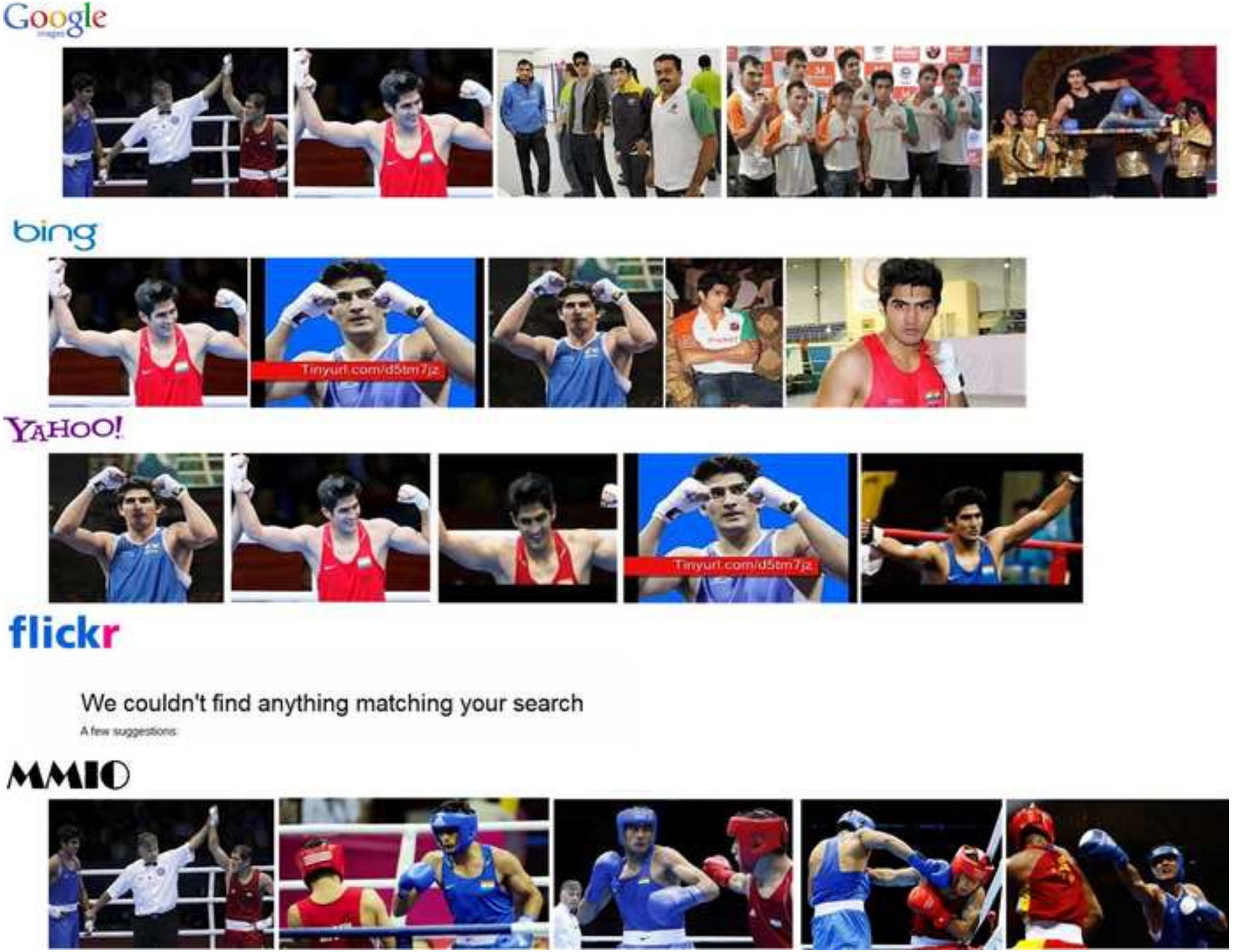


Fig. 14 Comparison of the top ten ranked results using a visual
Click here to download high resolution image

An Enhanced Bag-of-Visual Word Vector Space Model to Represent Visual Content in Athletics Images

Kraisak Kesorn and Stefan Poslad

***Abstract*—Images that have a different visual appearance may be semantically related using a higher level conceptualization. However, image classification and retrieval systems tend to rely only on the low-level visual structure within images. This paper presents a framework to deal with this semantic gap limitation by exploiting the well-known bag-of-visual words (BVW) to represent visual content. The novelty of this paper is threefold. First, the quality of visual words is improved by constructing visual words from representative keypoints. Second, domain specific "non-informative visual words" are detected which are useless to represent the content of visual data but which can degrade the categorization capability. Distinct from existing frameworks, two main characteristics for non-informative visual words are defined: a high document frequency (*DF*) and a small statistical association with all the concepts in the collection. The third contribution in this paper is that a novel method is used to restructure the vector space model of visual words with respect to a structural ontology model in order to resolve visual synonym and polysemy problems. The experimental results show that our method can disambiguate visual word senses effectively and can significantly improve classification, interpretation, and retrieval performance for the athletics images.**

***Index Terms*—Bag-of-visual words, non-informative visual words discovery, ontology model, visual content representation, visual words disambiguation.**

I. Introduction

THE continual rapid growth in digital content acquisition and visualization makes it increasingly challenging to find, organize, and access visual information. Research to better represent and understand visual content, content-based image retrieval (CBIR), has progressed over many decades. Typically, CBIR is based on two types of visual features: global and local features [1]. Global feature based algorithms aim at recognizing concepts in visual content as a whole. The main drawback is that they are often not directly related to any high-level semantics. Local features are an alternative choice and have several advantages over global features. Local feature algorithms focus mainly on *keypoints*, the salient image patches that contain the rich local information in an image. These can be automatically detected using various detectors, e.g., Harris corner [2] and difference of Gaussian (DoG) [3]. The scale invariant feature transform (SIFT) [4] is a promising low-level visual descriptor, which is invariant to scaling, translation, and rotation, and as well as partially invariant to illumination changes and affine projections. SIFT has also been used as the basis of a bag-of-visual words (BVW) model. The BVW model is proposed as a promising method for visual content classification [5], annotation [6], and retrieval [7]. However, the BVW model usually describes visual data at a non-semantic level. In contrast, humans often understand physical things more easily if they are represented semantically, i.e., the content of an image is represented in terms of relationships between concepts or instances as in an ontology model. Hence, an ontology-based model is used in this paper in order to bridge between low-level visual features and high-level semantic concepts and to support reasoning about data in order to promote semantic retrieval.

To obtain the high-level semantics for an image, several challenges need to be overcome. First, visual data need to be analyzed and transformed into a format that represents the visual content effectively. For the BVW technique, keypoints are detected and extracted. Unfortunately, some keypoints are just noise and are not useful to represent visual content. Therefore, a method to *enhance the quality of visual words* is required. Second, some visual words generated from keypoints might not represent any useful visual content. Those visual words should be eliminated as well. Third, low-level features of visual information can often be ambiguous. In comparison to words in text documents, multiple text concepts may share similar features, use synonyms, or one word may have several meanings (polysemy). Therefore, image retrieval systems should be able to handle these ambiguities properly in order to achieve high image classification accuracy. To solve these problems, this paper proposes novel techniques to resolve these three problems.

The remainder of this paper is organized as follows. Section II presents a survey of "state-of-the-art" frameworks and their limitations. Section III describes our proposed technique. Section IV discusses our experimental results. Finally, Section V summarizes our key contributions, and it discusses limitations and further work.

Manuscript received October 31, 2010; revised March 18, 2011, June 15, 2011, and September 06, 2011; accepted September 06, 2011. Date of publication October 06, 2011; date of current version January 18, 2012. This work was supported by Naresuan University and A New Researcher Scholarship of CSTS, Coordinating Center for Thai Government Science and Technology Scholarship Students, National Science and Technology Development Agency, Thailand. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Ming-Ting Sun.

K. Kesorn is with the Computer Science and Information Technology Department, Science Faculty, Naresuan University, Phitsanulok 65000, Thailand (e-mail: kraisakk@nu.ac.th).

S. Poslad is with the School of Electrical and Electronic Engineering and Computer Science, Queen Mary University of London E1 4NS, U.K. (e-mail: stefan@eecs.qmul.ac.uk).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2011.2170665

II. State of the Art

The advantages of local features versus global features for object recognition and visual content categorization have been summarized by Lee [8]. Lowe [3] proposed a technique for visual features extraction which is invariant to image scaling, translation, rotation, and partially invariant to illumination changes and affine projections. This technique is called scale invariant feature transform (SIFT). Ke *et al.* [9] improved the SIFT technique by applying principal components analysis (PCA) to reduce the dimensions of SIFT descriptors. As a result, local descriptors are more distinctive, more robust to image deformations, and more compact, compared to the standard SIFT representation, increasing image retrieval accuracy and matching speed. The BVW technique is motivated by an analogy to the "*bag of words*" representation for text categorization [13]–[15]. The BVW model has limitations when representing visual content. Just as words in traditional text documents may be ambiguous, so may visual words be in an image. The ambiguity lies in two areas: synonymy and polysemy. One possible way to disambiguate multiple word senses is to combine visual words into a larger unit, a so-called visual sentence [5]. Yuan *et al.* [10] proposed a solution to tackle this issue using the "*visual phrase*" technique based upon a likelihood ratio test method and an improved frequency itemset mining (FIM) algorithm [11]. FIM is a data mining technique using association rules. It has been applied to discover meaningful patterns of visual words because it is closely related to finding frequent itemsets in a transaction database. As a result, meaningful item sets are discovered. The major weakness of the visual phrase approach is that it merely considers the co-occurrence information among visual words but *neglects spatial information* amongst the visual words. In this paper, spatial information refers to the location of keypoints or visual words in an image.

Chen *et al.* [12] proposed a novel method that combines SIFT descriptors with the real spatial constitution of image content called a "Gaussian mixture model" (GMM) [13]. GMM provides a spatial weighting for visual words to facilitate content-based image retrieval. Quack [14] applied the FIM algorithm to incorporate spatial information for visual words when mining representative objects, actors, and scenes in videos. However, the test set was too small (only two videos) and may be not general enough for a larger image collection. Tirilly *et al.* [5] proposed a method that exploits the spatial information of visual words as a *visual sentence* using PCA. In addition, a language modeling (LM) method is applied in order to classify keypoints of images. However, the problem with this method is that the use of PCA over the coordinates of visual words requires effective background elimination of visual words. Furthermore, applications of PCA tend to focus on images that contain only one object.

Besides the spatial information issue, another limitation of the BVW model is "*noise visual word removal*". In text-based information retrieval, noise words (also called *stop words*) represent frequently occurring, insignificant words that appear in a text document, e.g., a, an, the, in, of, on, are, be, if, into, which, etc. In contrast, in visual content processing, it is difficult to define what the noise words are. They are sometimes called meaningless visual words or "*non-informative visual words*" because they are often domain specific; thus, it is difficult to create a standard list for visual content. Sivic *et al.* [15] considered useless visual words in relation to the most frequent visual words that occur in almost all images. However, considering only the frequency of visual words occurring in images is not adequate. Correlations between visual words and image concepts in the collection are also important since these indicate how strongly a visual word and image concepts in the collection are related. Thus, this information should not be ignored. Tirilly *et al.* [5] proposed a method to eliminate *useless visual words* based upon the geometric properties of the keypoints and the use of probabilistic latent semantic analysis (pLSA). Yuan *et al.* [16] tried to discover unimportant information at a larger unit of visual words, *the meaningless phrases*, through measuring the likelihood ratio (the statistical significance measure) of those visual phrases. However, this method ignores the coherency (the ordering of visual words) of component visual words in a visual phrase. Fulkerson [17] proposed a technique to reduce a number of visual words using the information bottleneck principle. Nonetheless, a major limitation of existing BVW models [5], [16], [17] is that they ignore the correlations between a specified word and other concepts in the collection. Some words might appear less in conjunction with one concept but appear more in conjunction with other concepts—these words could be the featured words. In such a case, deleting low-probability words will decrease the accuracy of categorization.

Another challenge for image classification using BVW is *visual heterogeneity*. This challenge is perhaps the greatest obstacle for image and video categorization and for retrieval systems which rely solely on visual appearance. It is common for different visual appearances to be semantically similar at a higher semantic conceptualization. Recently, the use of probability distributions of visual word classes has been proposed [7] based upon the hypothesis that semantically similar visual content will share a similar class probability distribution. However, this method is less useful when unrelated visual words exist that coincidently have a similar probability distribution. Gemert *et al.* [18] proposed uncertainty modeling to handle the ambiguity of visual words in order to improve the image features representation. However, the proposed method ignores the elimination of uninformative visual words. Consequently, the image feature dimensionality is high and this could reduce the visual content representation power. To solve visual heterogeneity problem, a hierarchical model has been exploited by Jiang *et al.* [19] to tackle the issue using a novel technique called a soft-weighting scheme. A hierarchical model is constructed using an Agglomerate clustering algorithm to capture the "*is-a*" relationship of visual words. However, the model of Jiang [19] is not practical in real situations because the hierarchical model is only a binary model and has no multiple-parent relationships. However, using a hierarchical model to index image features can dramatically improve retrieval performance of the image retrieval system. Similar ideas of using a hierarchical model for indexing image features are also proposed in [20] and [21].

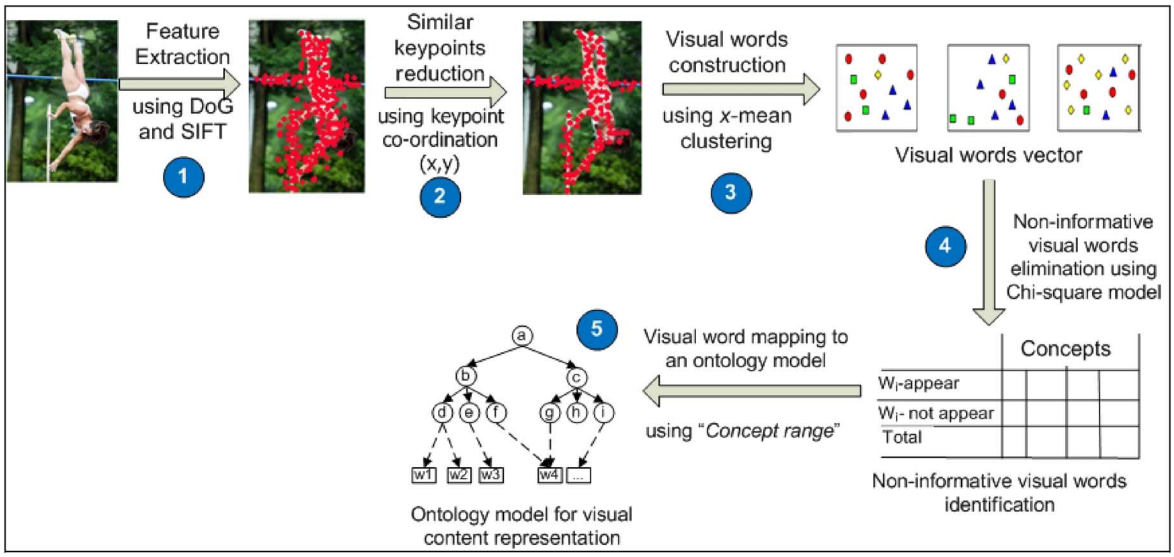


Fig. 1. Proposed framework to discover semantically similar visual words to facilitate visual invariance.

From the survey, the various limitations of existing BVW models can be summarized as follows. First, the existing models construct visual words using all detected keypoints. Some of these keypoints are not useful and decrease the quality of visual words. If visual words contain many noisy keypoints, visual word quality is decreased and, as a result, they cannot represent visual content effectively. Second, uninformative visual word detection based solely on document frequency (DF) does not account for visual words that appear in only a few concepts, leading to them having a low DF and being considered as noisy visual words. Sometime these visual words could be important visual features of images even if they only appear in a few concepts. Third, the construction of visual phrases or visual sentences helps disambiguate multiple word senses using statistical computation. However this method does not represent the actual semantic relationship between visual words.

A more effective model, to disambiguate visual word senses and to represent the semantics of visual content, is a hierarchical ontology model. Nevertheless, existing hierarchical clustering algorithms, e.g., the Agglomerate clustering algorithm, are often impractical to capture semantic relationships between concepts of visual information as they do not represent the semantics of visual content efficiently. To this end, this paper proposes a framework to generate a new representation model which addresses these three challenges.

III. Visual Content Analysis and Representation Framework

Visual features alone might not be sufficient to allow computer systems to analyze and interpret the meaning of visual data [22]. It needs other useful cues to guide the decision to attribute a higher-level meaning such as using any accompanying textual information or external knowledge base, e.g., WordNet [23]. However, there are situations when no accompanying textual information for the visual data may be present. Therefore, a method is proposed in this paper to deal with the situation when no textual information for the visual content may be supplied and to try to predict the content of visual data based on a structural analysis of visual features. This section presents a framework to represent a higher level conceptualization for visual data that can be related to lower level features. A brief overview of the proposed framework is shown in Fig. 1.

The deployment of the proposed framework comprises five major steps which are described as follows:

1) *Feature extraction*: several local patches are extracted which are considered as candidates for the basic elements—the "words". Interest point detectors aim to detect "keypoints" in an image or the salient image patches using an existing detector.
2) *Similar keypoints reduction*: each image is abstracted from several local patches. This step applies the SIFT technique to convert each patch to a multi-dimensional (128) vector. As a result, each visual content is a collection of vectors of the same dimension (128 for SIFT) where the order of different vectors is of no importance. All detected keypoints from step 1 are processed in order to construct visual words. In contrast to existing techniques, the detected keypoints are processed to find representative keypoints in order to improve quality of visual words (see Section III-A).
3) *Visual words construction*: vectors representing patches of "*visual words*" produce a "bag of visual words" represented in the form of a vector (histogram) using x-mean clustering algorithm.
4) *Non-informative visual words identification*: some of the generated visual words may not be useful to represent visual content. Hence, this kind of visual word needs to be detected and removed in order to reduce the size of visual word feature space, computation cost, and represent visual content effectively.
5) *Visual word mapping to an ontology model*: semantically related visual words are mapped to a hierarchical model which describes the visual content more explicitly and efficiently than a feature space model using conceptual

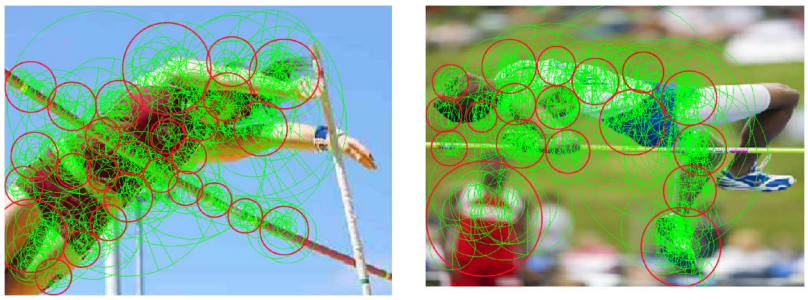


Fig. 2. Example of the noisy keypoint reduction. The red circles indicate the nearby keypoints that are grouped together.

structures and relationships. This structured model is used to disambiguate visual word senses effectively. Hence, it can aid visual information retrieval systems in interpreting or understanding the meaning of visual content more accurately.

The first two steps will not be discussed further in this paper since our work does not contribute to those areas. Instead, we focus on quality of visual word improvement, non-informative visual words elimination, and visual word mapping to an ontology model.

A. Visual Word Quality Enhancement

Typically, thousands of keypoints may be detected in an image. Some extracted keypoints may also pertain to a noisy image background and to redundant nearby keypoints. Such keypoints directly affect the dimensionality of the feature space. If visual words contain many noisy keypoints, the quality of visual words is reduced. This consequently affects the classification performance and the quality of visual words. In addition, a large number of keypoints leads to a large size for the visual word feature space and a high computation cost. Therefore, if the numbers of similar keypoints are reduced and the informative keypoints are preserved, the constructed visual words can represent visual content more effectively while reducing computational cost. Rudinac [24] proposed a method to detect duplicate keypoints by combining the nearby keypoints that have a distance of less than 1 pixel between the keypoints and centroid. The main weakness of this technique is its computational complexity because it needs to consider every pixel in an image. Jamshy [25] reduces the size of the keypoint database through learning which keypoints are beneficial for a specific application and using this knowledge to filter out a large portion of the keypoints. However, the classification performance is reduced using this technique.

Here, we propose a technique to detect redundant keypoints. Similar keypoints are identified through considering the physical location (spatial criteria) in visual content. Basically, a keypoint contains four useful properties, physical coordination (x, y), angle, scale, and SIFT descriptor. Our hypothesis is that keypoints located in nearby positions in the visual content could potentially be similar and can be grouped together. Hence, a new centroid of the group can be used as *a representative keypoint*. Similar keypoints are grouped together based on coordination, angle, the scale of a keypoint, and a SIFT descriptor using the x-mean algorithm [26]. Let k be a group of similar keypoints, $k_1, k_2, k_3, \ldots, k_n \in K, n \geq 1$. In each k, the average value of a SIFT descriptor will be used as a representative keypoint to generate visual words. Consequently, the numbers of keypoints in the feature space is reduced and only representative keypoints are used to generate visual words using the x-mean algorithm (Fig. 2). The main benefit of the x-mean algorithm over the k-mean algorithm is its speed as it does not need to specify the cluster numbers (k value). At this stage, we obtain a set of visual words which are improved compared to the traditional model because they are generated from the informative keypoints which represent visual content better and semantic information is preserved via a linkage between physically similar keypoints and the high level semantics of the objects.

However, not all of them are useful to enhance the classification algorithm, and some of generated visual words can even degrade the categorization power because they are noisy visual words, so-called *non-informative visual words*. Therefore, these visual words should be eliminated.

B. Non-Information Visual Words Identification

Non-informative visual words are the local visual content patterns which are not needed for retrieval and classification tasks. They are relatively "safe" to remove [27] in the sense that their removal does not cause a significant loss of accuracy but rather significantly improves the classification accuracy and computation efficiency of categorization. In a text-based document, *stop words* need to be removed before further processing, e.g., text categorization. Likewise, in visual data processing, there exist uninformative visual words that are insignificant local visual content patterns and useless for retrieval and classification. These visual words need to be eliminated in order to improve the accuracy of the results and to reduce the size of visual word feature space and computation cost. In this paper, we utilize a statistical model to automatically discover non-informative visual words, to eliminate them strengthening the discrimination power. Here, non-informative visual words are identified based on a document frequency method using the Chi-square model. In addition, the visual words are normalized in order to compensate for discrepancies in the size of the images.

Definition 1: Non-Informative Visual Words

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

1) It usually appears in many visual content in the collection; thus, it has a high *DF*. Since it occurs in several images, it

cannot be used to represent any particular image or object and

2) It has a small statistical association with all classification categories.

From the given definitions, non-informative visual words can be extracted from the visual word feature space using a statistical method, e.g., the Chi-square model. Having created visual words in the previous step, visual words can be quantized into a Boolean vector space model to express each visual content vector. Assuming that the appearance of the visual word i (ϖ_i) is independent of any concepts $C, C \in Z, Z = \{C_1, C_2, \ldots, C_n\}$ where $n \geq 1$, the correlation between ϖ_I and concepts could be expressed in the form of a contingency table. The contingency table (Table I) $T = \{n_{ij}\}_{j=1}^{k}, 1 \leq i \leq 2$ contains the number of images containing visual words in each category. A matrix A_{NxM} represents T, where n_{1j} is the number of images containing a visual word ϖ_i for the concept C_j; n_{2j} is the number of images which do not contain visual word ϖ_i in the concept C_j; n_{+j} is the total number of images in the concept C_j; n_{i+} is the number of images in the collection containing the visual word ϖ_i; N is the total number of images in the training set, where

$$x^2 = \delta = \sum_{i=1}^{2} \sum_{j=1}^{k} \frac{(Nn_{ij} - n_{i+}n_{+j})^2}{Nn_{i+}n_{+j}}. \quad (1)$$

To measure the independence of each visual word from all the concepts, the Chi-square statistic is deployed [(1)]. Having calculated the degree of independence, the Chi-square values are sorted in descending order. The Chi-square value indicates the association between visual words and their concepts. However, there exists a problem concerning terms which appear in a small number of documents, leading to them to have a small Chi-square value. These terms sometimes could be the feature words. In such a case, we need to weight the Chi-square value using (2) [28]

$$x^2_{weighted}(\beta) = \frac{x^2_{2*p}}{DF_r} \quad (2)$$

where DF_r denotes the document frequency of the word r. This model balances the strength of the dependent relationship between a word, all concepts, and document frequency of a word. As a result, those visual words which have β less than a threshold (chosen experimentally) are designated as non-informative visual words and are removed. Obviously, non-informative visual words identified in this manner are collection specific. This means by changing the training collection, one can obtain a different ordered list. The remaining visual words are informative and are useful for the categorization task.

C. Visual Words Disambiguation and Image Annotation Based Upon an Ontology Model

Several researchers have tried to restructure visual words as a hierarchical model in order to disambiguate word senses more explicitly and effectively. Deng *et al.* [29] proposed a method using visual attributes on ImageNet [30], a large scale ontology of images built upon WordNet. The main differences from the presented framework and the method in [29] are the presented framework exploits local features (SIFT descriptors) whereas [29] uses global features, e.g., shape, texture, and color. In addition, an ontology in this proposed framework is not built based upon WordNet. These methods convert an unstructured visual words vector space model into a hierarchical structure model using well-known clustering algorithms, e.g., the Agglomerate clustering algorithm [19], hierarchical spatial Markov model [31], and the hierarchical latent Dirichlet allocation algorithm [32]. Nevertheless, the hierarchical models generated from these algorithms have some drawbacks. Firstly, they are a binary hierarchical model which is not always efficient in representing visual content data. In practice, types of relationships among concepts are more diverse. Secondly, the generated hierarchical model is such that there are an equivalent number of child nodes in every parent node but this is not always the case. Thirdly, multiple-relationships between a parent and child node do not exist which means that a child node cannot have more than one parent with similar or with different relationships.

TABLE I
CONTINGENCY TABLE OF ϖ_i

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

Rather than combining multiple visual words to disambiguate word senses, the visual word vector space model is transformed into a structural ontology model in order to resolve the three limitations (as described in Section II). In addition, the proposed method can enhance the annotation, interpretation, and retrieval performance of the system. Since the ontology model is usually domain specific, e.g., natural scene or sports, the structure of concepts and relationships among concepts for each application differs for each knowledge domain. Furthermore, it is impossible to exploit standard clustering algorithms or expect human beings to generate a general ontology model for every application. Therefore, a pre-designed ontology model for the sports domain is needed to enable the system to retrieve information semantically and precisely.

Typically in text documents, word sense disambiguation can be performed using external knowledge, e.g., WordNet [23]. However, WordNet cannot be used in this way because visual words do not provide any linguistic information. Therefore, an alternative method is to use mathematic calculations. In the training phase, the semantic concept of each individual object is defined. Each concept will be used to disambiguate the informative visual words and to assign the concept(s) for each visual word under the *pre-designed ontology* model. To transform visual words from the vector space model to an ontology model, the algorithm is shown in Fig. 3 which enhances the method of Wu [33] and described as following. First, the visual objects of interest are manually separated from the background in order to reduce noise. Second, 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. Third, the keypoints of objects

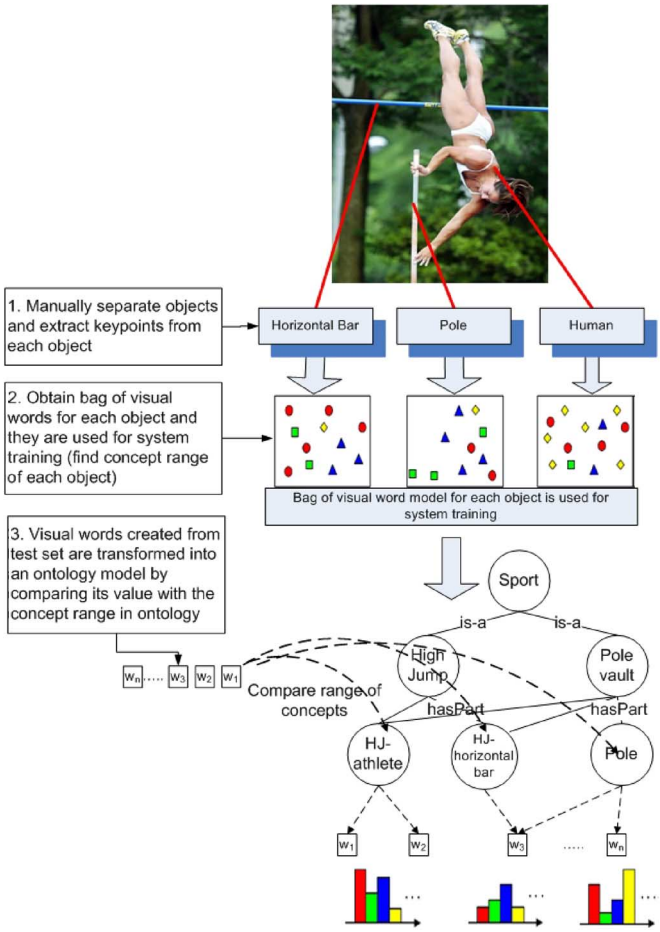


Fig. 3. Example of a structural ontology model and the different kinds of relationships between concepts for athletic sports. To disambiguate word senses, each visual word is compared to the concept range and assigned to concept(s) in the ontology model.

will be further processed to generate visual words. The links between the visual words and high level semantics for an object category can be obtained [31]. This serves to connect low level features to high level semantic objects. It is noted that when performed manually, such object separation is not an efficient method for a large-scale multimedia system. However, this method is applied for training only in order to allow the system to learn the proper sets of visual words. In the testing phase, images are processed automatically applying the processes shown in Fig. 1. In each object category, all the related objects are clustered using the x-mean algorithm to generate visual words. As a consequence, a set of visual words 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.

Definition 2: The concept range [31] of the key object

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

$$r_i = \max|\nu - c_i|, \quad \nu \in \varpi. \tag{3}$$

The concept range is useful for the visual word sense disambiguation and image classification. If a visual word is inside the range of any concept, the concept is assigned to the visual word; otherwise, the visual word does not respond to any concept and is discarded. 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 [19], [31] which cannot represent multiple-parent relationships (Fig. 3). Multiple assignments of visual words using a range of concepts can handle polysemy problem more effectively. Since the concepts in an ontology model are generated from different visual appearances of different parts of an object in the training phase, the visual diversity of objects leads to the semantics of visual content being represented using different visual words. Consequently, the range of concepts is invariant to the *visual heterogeneity* of an object. The mechanism proposed here is similar to a soft-assignment technique. Nonetheless, the proposed technique does not assign visual words fractionally. In other words, it does not calculate the degrees of membership of a visual word to each cluster because the framework does not use this information. Instead, the proposed mechanism checks whether a visual word belongs to a concept. A visual word can be a member of multiple concepts with a similar degree. This mechanism allows the framework to handle the polysemy or visual heterogeneity problem (as described in Section II) effectively.

Furthermore, this model can be used to annotate and interpret visual content at a higher level conceptualization. If the frequency of related visual words, $f(v_i)$, in each object (concept in the ontology model) is higher than a threshold (chosen experimentally), the visual content will be annotated with that concept label. However, direct use of $f(v_i)$ may be unfair to every visual content in the collection due to their different scales. Hence, we need to normalize $f(v_i)$ in order to compensate for discrepancies in the frequency of the visual words. Equation (4) shows the normalization formula

$$\eta_i = f(v_i) / \sum_{j=1}^{N} f(v_j). \tag{4}$$

In this paper, the ontology model is represented in the Resource Description Framework (RDF) format. RDF is selected as the knowledge representation rather than Resource Description Framework Schema (RDFS) or Web Ontology Language (OWL) because of several reasons, e.g., cardinality, transitivity, and inverse constraints, which are available in OWL, but have not been exploited by the presented system; RDF is much simpler to parse and query reducing the computation complexity. To interpret the high-level semantics of visual content, the simplest way is using the detected object information. Nevertheless, there are some uncertainties in image interpretation. Basically, there are three main causes of uncertainty for visual content interpretation [34]. First, uncertainty can arise from using an *incomplete image* as an input. The image may not contain enough information to make an interpretation. Second, uncertainty may be caused by the *ambiguity of an object*, e.g., an object can belong to several sport types. Finally, the uncertainty can occur from *object recognition errors* because an object recognition algorithm might not be able to detect some key objects in an image due to noise or the quality of an image. For a more robust and

Fig. 4. Example of a pole vault image in which a pole is missing; it is difficult for the system to interpret that this image is relevant to a pole vault event or a high jump event.

reliable system, a method is required that can handle the uncertainty and ambiguity of image interpretation. The basic idea of uncertainty management for visual content interpretation is to model how likely the scene in an image should be if some objects cannot be detected due to the object recognition uncertainties or object ambiguity. To this end, probability theory seems to be the prevailing method for dealing with uncertainty.

D. Uncertainty and Ambiguity of Visual Interpretations

Sometimes, the system cannot interpret the type of sport in an image properly using the detected objects because some key objects are absent; for instance, a pole object in a pole vault image (Fig. 4) is missing. The system could classify the content assigning a "high jump" event for the image. In this case, it can be difficult for a computer system to recognize that this image concerns a pole vault event because a pole object is missing.

Usually, humans use their past experiences to interpret visual content when it is ambiguous. Likewise, a computer system can be designed to interpret the meaning of an image based upon previous data. To handle visual uncertainty, a Bayesian network is applied to minimize the uncertainty of visual interpretation. A Bayesian network complements the ontology model to aid the interpretation of visual content. A Bayesian network [35] is a directed acyclic graph (DAG). When used in conjunction with statistical techniques, the graphical model has several advantages for data analysis. One, because the model encodes the dependencies among all variables, it readily handles situations where some data entries are missing. Two, because the model has both causal and probabilistic semantics, it is an ideal representation for combining prior knowledge and data. The conditional probability of a node B having value b was caused by a node A having value a can be described in the following expression:

$$P(B=b|A=a)=\frac{P(B=b)P(A=a|B=b)}{P(A=a)}. \quad (5)$$

A Bayesian network integrates the detected key object information to better determine how likely the image represents a specific sports event. To do this, the frequencies of occurrence of the objects in images have been counted and a Bayesian network has been used to model the frequency of occurrence. The system interprets images based on the probability of objects from previous data. This can enhance the categorization ability of the proposed system and can handle uncertainty.

TABLE II
WEIGHTING SCHEMES FOR VISUAL-WORD FEATURE (t_i)

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

IV. EXPERIMENTAL RESULTS AND DISCUSSION

There exist standard test collections that provide a "golden standard" to evaluate the retrieval performance of image retrieval systems. One of the standard collections of sport images is the "*Event dataset*" provided by Stanford University.[1] It contains only 1579 relevant images, which we decided was too few for testing the system. Furthermore, we mainly focus on eight athletics sports (high jump, long jump, pole vault, running, hurdles, javelin throw, discus throw, and hammer throw) because they are visually similar and, as a result, they are very challenging to classify. Thus, a new test collection was created to focus on this domain. Images are collected from the Olympic organization website[2] and the Google image search engine[3] as the basis for the test collection. The resulting image collection contains 16 000 images from the athletic sports. We divide the image collection into two sets: 4000 images (500 images from each category) are selected for training which produces total 2 483 765 keypoints; and the rest are used for testing which generates 6 843 945 keypoints by the DoG algorithm. The vector quantization technique is utilized for clustering the visual words based on the extracted keypoints using the SLAC clustering algorithm.

A. Hypotheses to Evaluate the Proposed Framework

To evaluate the retrieval performance of our framework, three hypotheses were established based upon the limitations of the existing systems (as discussed in Section II).

Hypothesis 1 (H1): The use of representative keypoints can improve the quality of classification performance.

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

Hypothesis 3 (H3): Restructuring the visual words space model using an ontology model can resolve the visual heterogeneity problem more effectively than the traditional BVW model and the content-based image retrieval (CBIR).

B. Evaluation and Results

The evaluation consists of two main parts, visual content representation (classification performance) and information retrieval (precision and recall [36]).

[1]Event dataset (http://vision.stanford.edu/lijiali/event_dataset/).

[2]Olympic Organization (http://www.olympic.org).

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

TABLE III
COMPARISON OF THE PERFORMANCE IN CLASSIFYING USING THREE CLASSIFIERS FOR EIGHT SPORT GENRES USING VISUAL WORDS (INCLUDING NON-INFORMATIVE VISUAL WORDS)

Classifiers	Weighting schemes	Average Precision (AP)							
		High jump	Long jump	Pole vault	Running	Hurdles	Javelin throw	Discus throw	Hammer throw
Naïve Bayes	BIN	0.451	0.583	0.554	0.571	0.452	0.547	0.412	0.528
	TFY	0.463	0.594	0.596	0.604	0.514	0.638	0.517	0.539
	TFN	0.463	0.614	0.671	0.616	0.551	0.641	0.524	0.592
	TFI	0.544	0.615	0.592	0.518	0.537	0.574	0.527	0.494
	TIN	0.589	0.668	0.572	0.553	0.513	0.617	0.531	0.591
SVM-Linear	BIN	0.534	0.586	0.594	0.584	0.563	0.582	0.492	0.498
	TFY	0.521	0.667	0.667	0.613	0.587	0.507	0.524	0.376
	TFN	0.564	0.656	0.666	0.627	0.611	0.581	0.537	***0.596***
	TFI	0.585	0.638	0.575	0.528	0.551	0.597	0.558	0.512
	TIN	0.624	0.664	0.573	0.569	0.532	0.574	0.560	0.486
SVM-RBF	BIN	0.567	0.631	0.601	0.578	0.570	0.572	0.511	0.486
	TFY	0.592	0.635	0.672	0.622	0.592	0.588	0.537	0.503
	TFN	***0.634***	***0.679***	***0.695***	***0.635***	***0.610***	***0.643***	***0.569***	0.586
	TFI	0.547	0.647	0.634	0.542	0.563	0.597	0.551	0.552
	TIN	0.612	0.677	0.685	0.579	0.545	0.628	0.559	0.564

1) Evaluation of the Categorization Algorithms: To evaluate the classification methods, we calculated the average precision value for three different classifier algorithms:[4] Naïve Bayes, SVM-Linear, and SVM-RBF were used to cluster the same set visual words produced by SLAC using different term weighting schemes. Since term weighting is a key technique in information retrieval, we explored its use in visual-word feature representation. Two major factors in term weighting are tf (term frequency) and idf (inverse document frequency). A third factor is *normalization*, which converts the feature into unit-length vector to eliminate the difference between short and long documents (small and big images' size). Some popular term weighting schemes in information retrieval area are then applied to the visual-word feature vectors. These schemes are summarized in Table II. The classification results for the five types of classifiers and the precision values that are maximal for each sport category are printed in bold-italic face to help identify the best algorithm. Based on the classification results in Table III, the SVM-RBF classifier with the TFN weighting scheme is a good choice as it always produces the best performance in our experiments (except for the Hammer throw event in which the SVM-Linear with TFN weight scheme provides a slightly better classification performance than SVM-RBF). Thus, these techniques will be applied to further experiments in subsequent sections.

2) Evaluation of the Quality of Visual Words Constructed From the Representative Keypoints: To evaluate H1, we compared the results of image classification between visual words created from all detected keypoints (μ), selected keypoints (Φ) using the state-of-the-art framework [24], and the representative keypoints (Ω) using the proposed method in this paper. All methods are compared using three different categorization algorithms in Table III with the TFN weighting scheme. The results of this comparison are shown in Table IV. The classification results using visual words created by the proposed framework (Ω) and the method (Φ) (proposed in [24]) appear to be superior to μ. This figure indicates that the visual words produced by the representative keypoints have a better quality than another because some redundant keypoints are reduced. Therefore, only informative keypoints are used to generate visual words. As a result, these visual words represent the visual content more effectively and enhance the categorization power. In contrast, the redundant keypoints affect the quality of visual words because they could be noise in the vector space. Consequently, the generated visual words cannot represent visual content properly. This uncertainty leads to decreasing classification performance. Although a large number of keypoints makes a feature become more discriminative, keypoints also make the feature vector less generalizable and might contain more noise. These are major factors in the classical degradation of the performance for visual words. A large number of keypoints also increases the cost in computing visual-word features and in generating supervised classifiers. The H1 hypothesis is validated under the domain of eight sport genres. However, the classification performance of Ω and Φ are slightly different since both methods reduce the number of keypoints before visual word construction.

3) Evaluation of Classification Performance After Non-Informative Visual Word Reduction: Next, we illustrate the effect of non-informative visual words $\{\psi\}$ removal on the classification performance. This elimination may affect the classification performance (Table III). Thereafter, the effect of non-informative visual words removal on the classification performance needs to be investigated. We compared the classification results between visual words with uninformative visual words $\{\varpi + \psi\}$, state-of-the-art framework (pLSA) and the proposed technique (Chi-square). All techniques apply the SVM-RBF classification algorithm with TFN weighting scheme.

Fig. 5 shows that the classification accuracy is heavily influenced by the proposed method (eliminating non-informative visual words). For example, in the pole vault event, the classification accuracy is increased from 69% to 71% when non-informative visual words is removed and using the Chi-square technique to detect non-informative visual words based upon two main characteristics defined in definition 1. The improvement

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

TABLE IV
COMPARISON OF CLASSIFICATION RESULTS BETWEEN VISUAL WORDS CREATED FROM SELECTED KEYPOINTS (Φ) USING METHOD IN [24] AND FROM THE REPRESENTATIVE KEYPOINTS (Ω) USING THE PROPOSED METHOD WITH THREE DIFFERENT CLASSIFIERS

	High jump			Long jump			Pole vault			Running		
Classifiers	μ	Φ	Ω	μ	Φ	Ω	μ	Φ	Ω	μ	Φ	Ω
Naïve Bayes	0.519	0.539	0.535	0.531	0.628	0.634	0.505	0.691	0.685	0.589	0.632	0.635
SVM-Linear	0.582	0.581	0.587	0.548	0.683	0.681	0.604	0.675	0.681	0.614	0.638	0.642
SVM-RBF	0.601	0.657	0.649	0.551	0.685	0.692	0.623	0.705	0.711	0.627	0.657	0.651

	Hurdles			Javelin throw			Discus throw			Hammer throw		
Classifiers	μ	Φ	Ω	μ	Φ	Ω	μ	Φ	Ω	μ	Φ	Ω
Naïve Bayes	0.554	0.586	0.594	0.571	0.653	0.658	0.495	0.548	0.547	0.467	0.604	0.608
SVM-Linear	0.583	0.62	0.618	0.605	0.625	0.624	0.563	0.61	0.613	0.538	0.625	0.627
SVM-RBF	0.621	0.66	0.665	0.582	0.63	0.628	0.517	0.589	0.592	0.565	0.617	0.615

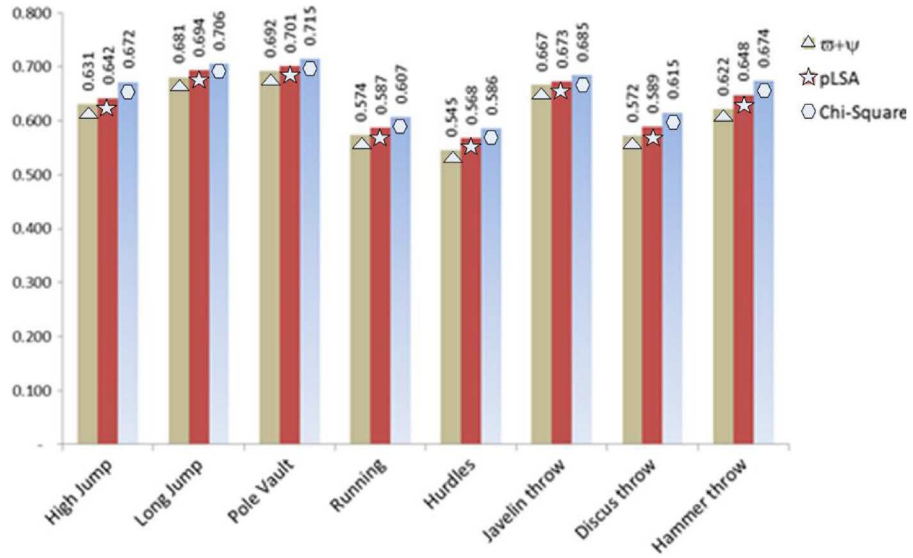


Fig. 5. Classification performance comparison of the visual words + uninformative visual words ($\varpi + \psi$), the state-of-the-art framework (pLSA), and Chi-square (the proposed framework).

of classification accuracy trends similarly in other categories. This consistent improvement suggests that non-informative visual words removal does not cause a loss of classification accuracy but instead it is able to improve the discrimination power.

Compared to the state-of-the-art framework, the proposed technique obtains a better classification performance because pLSA [5] ignores the correlations between a specified word and other concepts in the collection. However, some words could be the feature visual words. They might appear less often in conjunction with one concept but appear more often in conjunction with other concepts. pLSA relies only on visual word probability and it deletes low-probability visual words which are considered as the non-informative visual words. Therefore, only the probability value is insufficient to consider the useless visual words. This decreases the accuracy of categorization.

The classification performance shown in Fig. 5 does not show a significant improvement because there is a very small difference (2%–3% on average) between pLSA and the Chi-square. In several cases, differences between two methods are small but statistically significant due to the large sample size. To test the statistical significance, a T-test is calculated by comparing the average value of some variables obtained for two groups. This experiment contained 12 000 images for testing which are used to construct visual words (ϖ) and before the non-informative visual words (ψ) are removed. Then, the image classification is performed based upon two groups of data (containing eight sport types), pLSA and Chi-square. The T-test can be calculated using (6):

TABLE V
T-TEST CALCULATION FOR THE EXPERIMENTAL RESULTS

	pLSA	Chi-square
Number of observations	8	8
Mean ($\bar{X}$)	0.645	0.673
Variance (var)	0.00262	0.002338

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{var_1}{n} + \frac{var_2}{n}}} \tag{6}$$

where $\overline{X}$ is the mean value of precision value, var is the variance value of precision value, and n is the number of observation data (sport types). To perform a T-test, the results are transformed into a table as shown in Table V.

Here, the null hypothesis is "*removing non-informative visual words* (ψ) *will provide the same classification performance as non-removing* ψ" and a probability of error level $(\text{alpha level}) = 0.05$. From the distribution table[5] with $df = 8$

[5]Distribution table (http://www.statsoft.com/textbook/distribution-tables/).

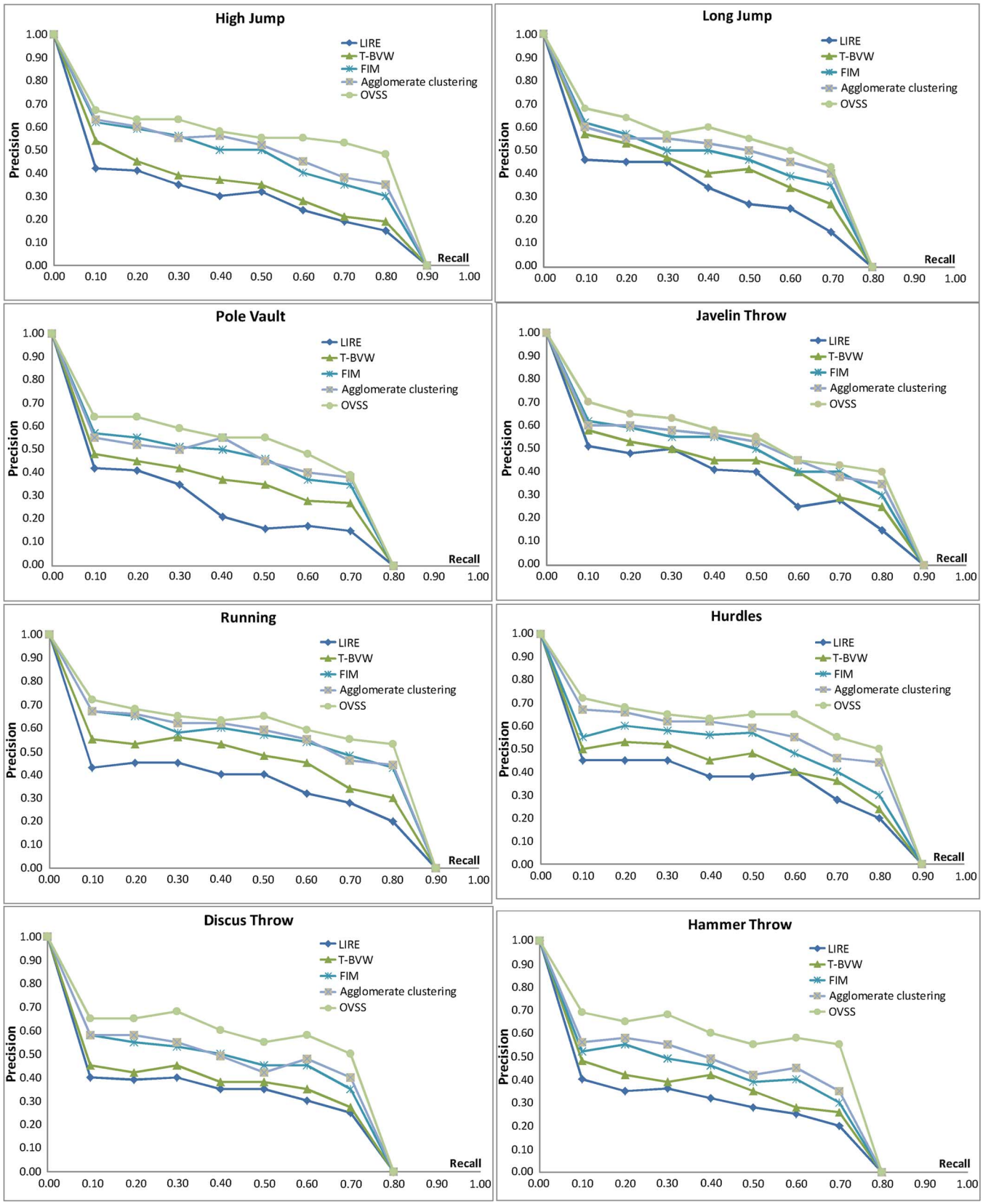


Fig. 6. Retrieval performance results comparison of LIRE, T- BVW, FIM-based, Agglomerative clustering, and OVSS.

$(n-2)$ and $p = 0.05$, t must equal or greater than 1.645. Here, we obtain the $t = 1.733$ and, therefore, this can conclude that there is a statistically significant probability that a relationship between the two variables exists and that this is not due to chance. The null hypothesis is rejected. This lends support to the research hypothesis and indicates that 3% is a significant

improvement of the image categorization performance when non-informative visual words are removed.

4) Evaluation of Image Retrieval Performance Using Query-by-Example: To evaluate H3, we compared the retrieval performance (precision-recall graph) of the traditional CBIR technique (LIRE framework),[6] the traditional bag-of-visual-words model (T-BVW), the FIM-based BVW technique, the Agglomerative clustering-based technique, and the proposed method (ontology-based for visual semantic search, OVSS). A T-BVW model is constructed from visual words generated by the k-mean algorithm. The FIM-based BVW technique exploits the data mining library from Weka to discover correlated visual words. The Agglomerative clustering-based technique [37] restructures visual words into a hierarchical model (Binary tree) in order to represent visual content. However, these state-of-the-art methods ignore how to perform noisy visual word removal. Fig. 6 shows that the retrieval performance is affected by the proposed technique. Since the OVSS technique analyzes an image query and translates it into an ontology model, the search engine is able to perform conceptual searching by using these annotations (concepts' name in the ontology) as searching keywords rather than just simple low-level feature matching. As a consequence, more relevant images can be recognized and retrieved since all images in the same concept will be considered as relevant images. The OVSS outperforms the Agglomerative clustering-based technique because the algorithm clusters visual words into a hierarchical model in which a parent node can have only two child nodes and the members of each cluster cannot overlap. Thus, it results in lower precision than OVSS due to the binary tree structure being inadequate to represent visual content.

FIM discovers the related visual words using mathematic calculation and ignores the high-level semantic relationship among those visual words. In contrast, OVSS obtains a high level of semantic information of visual words during the training phase and uses this information to cluster the generated visual words. This leads to an OVSS technique obtaining a higher precision and recall compared to FIM-based technique. LIRE retrieves all images which have similar low-level features. Unfortunately, some of them are not semantically relevant to the image query. As a result, LIRE obtains a lower precision and recall compared to other techniques. T-BVW attains a better retrieval performance compared to LIRE because the use of the associative visual words efficiently distinguishes visual content more than the color and texture features used in LIRE. However, it obtains a lower precision and recall than the OVSS method because the T-BVW technique represents visual content based on the feature space model, whereas OVSS deploys a hierarchical model which expresses visual content more explicitly and efficiently than the feature space model using conceptual structures and relationships. This structured model can disambiguate visual word senses more effectively with the aid of a Bayesian network. The Bayesian network complements the ontology model to aid the interpretation of visual content. It exploits the conditional probability to cope with the uncertainty that may occur during the classification process. This probability aids the classifier in making a decision about which category (sport genre) an image should be in when an uncertainty occurs, e.g., when the underlying objects are missing. As such, this mechanism leads to a higher classification performance of the system and enhances the visual content interpretation and retrieval performance. This allows systems to recognize all semantically relevant images (all images under the same node), although their visual appearances may differ according to the query. In other words, the proposed technique is highly invariant to visual appearance. Therefore, the H3 hypothesis is successfully evaluated under the domain of eight sport genres.

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

V. Conclusions and Future Work

This paper proposes to replace the visual word vector space model with a structural ontology model for visual content representation, which provides a semantic-based classification, interpretation, and retrieval solution for images of athletic sports. The novelty of this research is as follows. First, a technique to enhance the quality of visual words using the representative keypoints is researched and developed. Second, a technique to detect domain specific *non-informative visual words* that provide no value to represent the content of visual data using a Chi-square statistical model is given. The main difference of our technique to existing techniques is that not only *DF* is considered but also a statistical association with all the concepts in the collection. In addition, visual words that appear in a small number of concepts are normalized since they could be featured visual words. Thus, all visual words are normalized before performing elimination. Finally, a method is outlined to restructure the visual word vector space model into an ontology-based model in order to disambiguate visual word senses. Unlike the hierarchical model in other state-of-the-art frameworks, the ontology model in the presented framework can capture the knowledge of athletic sport domain in an improved way, e.g., by avoiding a binary tree and by sharing concepts through the use of an ontology. The ontology incorporates with Bayesian network is very useful for image classification, interpretation, and retrieval tasks that do not only rely on visual similarity but are also based on concept similarity. In other words, the technique can resolve the visual heterogeneity problem.

Although the proposed framework has been tested for only eight athletic sports, these sports are very challenging to classify because they produce several similar visual words. This work could be used as a prototype for other types of images or other types of events and it could be applied to other sports equipment (e.g., to differentiate racket or ball). These objects can be detected using the keypoints detection algorithm (DOG) and SIFT descriptors followed by the main steps shown in Figs. 1 and 3. Nonetheless, applying this technique to other domains would require the ontology structure to be modified to make the concepts in the ontology more relevant to the domain in order to enhance the reasoning mechanism.

Another limitation of the framework is *ontology incompleteness*[38]. It has already been recognized that developing ontologies is a laborious, expensive, and time-consuming task. It is also technically difficult to build in advance a perfect ontology covering the whole domain of knowledge [39]. This is because some important aspects cannot be modeled in present-day standard ontology languages, e.g., uncertainty and gradual truth values. These cannot directly be represented in a strong ontology language representation such as OWL that hardwires a specific logic, i.e., a description logic, into the ontology representation. Therefore, we plan to design ontologies having some

degree of openness rather than being fixed (*closed ontology* [40]) at development. The most important benefit of this approach is that it is not limited to the scope of topics provided by a training set. Therefore, the system will not rely solely on the information in the ontology model and the system will be better equipped to find any relevant information.

REFERENCES

[1] F. Alhwarin, C. Wang, D. Ristic-Durrant, and A. Graser, "Improved SIFT-features matching for object recognition," in *Proc. Int. Academic Conf. Vision of Computer Science-BSC*, 2008, pp. 179–190.

[2] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. 4th Alvey Vision Conf.*, 1988, pp. 147–151.

[3] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proc. Int. Conf. Computer Vision*, 1999, vol. 2, pp. 1150–1157.

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

[5] P. Tirilly, V. Claveau, and P. Gros, "Language modeling for bag-of-visual words image categorization," in *Proc. 2008 Int. Conf. Content-Based Image and Video Retrieval*, 2008, pp. 249–258.

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

[7] Y.-T. Zheng, S.-Y. Neo, T.-S. Chua, and Q. Tian, "Toward a higher-level visual representation for object-based image retrieval," *Vis. Comput.*, vol. 25, no. 1, pp. 13–23, 2008.

[8] Y. Lee, K. Lee, and S. Pan, "Local and global feature extraction for face recognition," in *Proc. 5th Int. Conf. Audio- and Video-Based Biometric Person Authentication*, 2005, pp. 219–228.

[9] Y. Ke and R. Sukthankar, "PCA-SIFT: A more distinctive representation for local image descriptors," in *Proc. 2004 IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2004, vol. 2, pp. 513–506.

[10] J. Yuan, Y. Wu, and M. Yang, "From frequent itemsets to semantically meaningful visual patterns," in *Proc. 13th ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining*, 2007, pp. 864–873.

[11] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *ACM SIGMOD Record*, 1993, vol. 22, pp. 207–216.

[12] X. Chen, X. Hu, and X. Shen, "Spatial weighting for bag-of-visual-words and its application in content-based image retrieval," in *Proc. 13th Pacific-Asia Conf. Advances in Knowledge Discovery and Data Mining*, 2009, pp. 867–874.

[13] G. J. McLachlan and K. E. Basford, *Mixture Models: Inference and Applications to Clustering*. New York: Marcel Dekker, 1988.

[14] T. Quack, V. Ferrari, and L. Van Gool, "Video mining with frequent itemset configurations," in *Proc. 5th Int. Conf. Image and Video Retrieval*, 2006, vol. 2006, pp. 360–369.

[15] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *Proc. 9th IEEE Int. Conf. Computer Vision*, 2003, vol. 2, pp. 1470–1477.

[16] J. Yuan, Y. Wu, and M. Yang, "Discovery of collocation patterns: From visual words to visual phrases," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[17] B. Fulkerson, A. Vedaldi, and S. Soatto, "Localizing objects with smart dictionaries," in *Proc. 10th European Conf. Computer Vision: Part I*, Marseille, France, 2008, pp. 179–192.

[18] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, Jul. 2010.

[19] Y.-G. Jiang and C.-W. Ngo, "Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval," *Comput. Vis. Image Understand.*, vol. 113, no. 3, pp. 405–414, 2009.

[20] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, 2006, vol. 2, pp. 2161–2168.

[21] Š. Obdrzálek and J. Matas, "Sub-linear indexing for large scale object recognition," in *Proc. British Machine Vision Conf.*, 2005, pp. 1–10.

[22] R. Moller and B. Neumann, "Ontology-based reasoning techniques for multimedia interpretation and retrieval: Theory and applications," in *Semantic Multimedia and Ontologies*, Y. Kompatsiaris and P. Hobson, Eds. London, U.K.: Springer, 2008, pp. 55–98.

[23] G. A. Miller, "WordNet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[24] M. Rudinac, B. Lenseigne, and P. Jonker, "Keypoint extraction and selection for object recognition," in *Proc. IAPR Conf. Machine Vision Applications*, 2009, pp. 191–194.

[25] S. Jamshy, E. Krupka, and Y. Yeshurun, "Reducing keypoint database size," in *Proc. 15th Int. Conf. Image Analysis and Processing*, Vietrisul Mare, Italy, 2009, pp. 113–122.

[26] D. Pelleg and A. W. Moore, "X-means: Extending K-means with efficient estimation of the number of clusters," in *Proc. 17th Int. Conf. Machine Learning*, 2000, pp. 727–734.

[27] Y. Yang and J. Wilbur, "Using corpus statistics to remove redundant words in text categorization," *J. Amer. Soc. Inf. Sci.*, vol. 47, no. 5, pp. 357–369, 1996.

[28] L. Hao and L. Hao, "Automatic identification of stop words in Chinese text classification," in *Proc. Int. Conf. Computer Science and Software Engineering*, 2008, pp. 718–722.

[29] O. Russakovsky and L. Fei-fei, "Attribute learning in large-scale datasets," in *Proc. ECCV 2010 Workshop Parts and Attributes*, 2010, pp. 1–14.

[30] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-fei, "Imagenet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

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

[32] J. Sivic, B. C. Russell, A. Zisserman, W. T. Freeman, and A. A. Efros, "Unsupervised discovery of visual object class hierarchies," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008, pp. 1–8.

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

[34] P. B. Cullen, J. H. Hull, and S. N. Srihari, "A constraint satisfaction approach to the resolution of uncertainty in image interpretation," in *Proc. 8th Conf. Artificial Intelligence for Applications*, 1992, pp. 127–133.

[35] F. Ruggeri, R. Kenett, and F. W. Faltin, *Encyclopedia of Statistics in Quality and Reliability*. London, U.K.: Wiley, 2008, vol. 1.

[36] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2008.

[37] B. Walter, K. Bala, M. Kulkarni, and K. Pingali, "Fast agglomerative clustering for rendering," in *Proc. IEEE Symp. Interactive Ray Tracing*, 2008, pp. 81–86.

[38] K. Kesorn and S. Poslad, "Semantic restructuring of natural language image captions to enhance image retrieval," *J. Multimedia*, vol. 4, no. 5, pp. 284–297, 2009.

[39] G. Nagypál, *Possibly Imperfect Ontologies for Effective Information Retrieval*. Karlsruhe, Germany: Univ. Karlsruhe (TH), 2007.

[40] S. Poslad, *Ubiquitous Computing Smart Devices, Environments and Interactions*. Chichester, U.K.: Wiley, 2009.

Kraisak Kesorn received the B.Sc. degree in computer science from Chiang Mai University, the M.Sc. degree in information technology (IT) from King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand, and the Ph.D. degree in electronic engineering and computer science from Queen Mary College, University of London, London, U.K.

He is currently teaching and researching at the Computer Science and IT Department, Faculty of Science, Naresuan University, Phitsanulok, Thailand, e.g., courses in information retrieval for undergraduates and post-graduates. His current research interests include semantic multimedia retrieval, knowledge-based modeling for information retrieval, and cross-language (Thai-English) information retrieval, as well as semantic business intelligence (BI 2.0).

Stefan Poslad received the Ph.D. degree from the University of Newcastle upon Tyne, U.K.

He is Lecturer at the School of Electronic Engineering and Computer Science, Queen Mary, University of London, London, U.K. His research interests are ubiquitous computing [he is an author of the book *Ubiquitous Computing: Smart Devices, Environments and Interaction* (New York: Wiley, 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.

Expert Systems with Applications 38 (2011) 11472–11481

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Visual content representation using semantically similar visual words

Kraisak Kesorn [a,*], Sutasinee Chimlek [b], Stefan Poslad [c], Punpiti Piamsa-nga [b]

[a] Department of Computer Science and Information Technology, Naresuan University, Phitsanulok 65000, Thailand
[b] Department of Computer Engineering, Kasetsart University, 50 Phahon Yothin Rd., Chatuchak, Bangkok 10900, Thailand
[c] School of Electronic Engineering and Computer Science, Queen Mary, University of London, Mile End Rd., London E1 4NS, United Kingdom

ARTICLE INFO

Keywords:
Bag-of-visual words
SIFT descriptor
Visual content representation
Semantic visual word

ABSTRACT

Local feature analysis of visual content, namely using Scale Invariant Feature Transform (SIFT) descriptors, have been deployed in the 'bag-of-visual words' model (BVW) as an effective method to represent visual content information and to enhance its classification and retrieval. The key contributions of this paper are first, a novel approach for visual words construction which takes physically spatial information, angle, and scale of keypoints into account in order to preserve semantic information of objects in visual content and to enhance the traditional bag-of-visual words, is presented. Second, a method to identify and eliminate similar key points, to form semantic visual words of high quality and to strengthen the discrimination power for visual content classification, is given. Third, an approach to discover a set of semantically similar visual words and to form visual phrases representing visual content more distinctively and leading to narrowing the semantic gap is specified.

1. Introduction

Research in multimedia retrieval has been actively conducted since many years. Several main schemes are used by visual content (e.g. image and video) retrieval systems to retrieve visual data from collections such as content-based image retrieval (CBIR) (Bach et al., 1996; Rui, Huang, & Chang, 1999; Smeulders, Worring, Santini, Gupta, & Jain, 2000; Smith & Chang, 1996), automatic classification of objects and scenes (Forsyth & Fleck, 1997; Naphade & Smith, 2003; Tseng, Lin, Naphade, Natsev, & Smith, 2003), and image and region labeling (Barnard et al., 2003; Duygulu, Barnard, Freitas, & Forsyth, 2002; Hironobu, Takahashi, & Oka, 1999). The research on CBIR started using low-level features, such as color, texture, shape, structure, space relationship to represent the visual content. Typically, the research on CBIR is based on two types of visual features: global and local features. Global feature-based algorithms aim at recognizing objects in visual content as a whole. First, global features (i.e. color, texture, shape) are extracted and then statistic feature classification techniques (i.e. Naïve Bayes, SVM, etc.) are applied. Global feature-based algorithms are simple and fast. However, there are limitations in the reliability of object recognition under changes in image scaling viewpoints, illuminations, and rotation. Thus, local features are also being used.

Several advantages of using local rather than global features for object recognition and visual content categorization have been addressed by Lee (2005). Local feature-based algorithms focus mainly on *keypoints*. Keypoints are salient patches that contain rich local information about visual content. Moravec (1977) defined the concept of "point of interest" as distinct regions in images that can be used to match other regions in consecutive image frames. The use of the Harris corner detector (Harris & Stephens, 1988) to identify interest points and to create a local image descriptor at each interest point from a rotationally invariant descriptor in order to handle arbitrary orientation variations has been proposed in Schmid and Mohr (1997). Although this method is rotation invariant, the Harris corner detector is sensitive to changes in image scale (Alhwarin, Wang, Ristic-Durrant, & Graser, 2008) and, therefore, it does not provide a good basis for matching images of different sizes. Lowe (1999) has overcome such problems by detecting the key locations over the image and its scales through the use of local extrema in a Difference-of-Gaussians (DoG). Lowe's descriptor is called the Scale Invariant Feature Transform (SIFT). SIFT is an algorithm for visual feature extraction invariant to image scaling, translation, rotation, and partially invariant to illumination changes and affine projection. Further improvements for object recognition technique based on SIFT descriptors have been presented recently by. Ke & Sukthankar (2004) who have improved the SIFT technique by applying Principal Components Analysis (PCA) to make local descriptors more distinctive, more robust to image deformations, and more compact than the standard SIFT representation. Consequently, this technique increases image retrieval accuracy and matching speed. Recently, keypoints represented by the SIFT descriptors also have been used in a special technique namely "bag-of-visual words" (BVW). The BVW visual content representation has drawn much attention by computer vision communities, as it tends to code

* Corresponding author. Tel.: +66 81555 7499.
E-mail addresses: kraisakk@nu.ac.th, kraisak@msn.com (K. Kesorn).

0957-4174/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2011.03.021

the local visual characteristics towards the object level (Zheng, Neo, Chua, & Tian, 2008). The main advantages of the BVW technique are its simplicity and its invariance to transformations as well as occlusion, and lighting (Csurka, Dance, Fan, Willamowski, & Bray, 2004). There are hundreds of publications about visual content representation using the BVW model as it is a promising method for visual content classification (Tirilly, Claveau, & Gros, 2008), annotation (Wu, Hoi, & Yu, 2009), and retrieval (Zheng et al., 2008).

The BVW technique is motivated by an analogy with the '*bag-of-words*' representation for text categorization. This leads to some critical problems, e.g. the lack of semantic information during visual words construction, the ambiguity of visual words, and is computationally expensive. Therefore, this paper proposes a framework to generate a new representation model which preserves semantic information throughout the BVW construction process that can resolve three difficulties; (i) loss of semantics during visual word generation; (ii) similar keypoints and non-informative visual words discovery; and (iii) semantically similar visual words identification. In the remaining sections of this paper, we propose and analyze our solution.

2. Key contributions

A framework for semantic content-based visual content retrieval is proposed. It reduces similar keypoints. It preserves semantic relations between keypoints and objects in the visual content. This generates visual words for visual content representation through reducing the dimensions of keypoints in feature space and enhancing the clustering results. It also reduces the less computation cost and memory needed. Visual heterogeneity is a critical challenge for content-based retrieval. This paper proposes an approach to find semantically (associatively) similar visual words using a similarity matrix based on the semantic local adaptive clustering (SLAC) algorithm to resolve this problem. In other words, semantically related visual content can be recognized even though they have different visual appearances using a set of semantically similar visual words. Experimental results illustrate the effectiveness of the proposed technique which can capture the semantics of visual content efficiently, resulting in higher classification and retrieval accuracy.

The rest of this paper is organized as follows. In Section 3, the proposed technique is described. In Section 4, experimental results are discussed. Section 5 concludes the paper by presenting an analysis of the strengths and weaknesses of our method and describes further work.

3. Related work

Hundreds of papers have been published, about visual content representation using local features, over the last two decades. This survey focuses on visual content representation based upon Scale Invariant Feature Transform (SIFT) descriptors and the bag-of-visual words model because these methods are the major techniques and are the basis of the method in this paper. The survey focuses on three major limitations of visual content classification and retrieval: loss of semantics during visual word generation; similar keypoints and non-informative visual words reduction; semantically similar visual words discovery.

3.1. Loss of semantics during visual word generation

The main disadvantage of existing methods (Jiang & Ngo, 2009; Yuan, Wu, & Yang, 2007a; Zheng et al., 2008) is the lack of spatial information, i.e., the physical location, between keypoints during visual words construction when using only a simple *k*-mean clustering algorithm. To tackle this issue, Wu et al. (2009) tried to preserve the semantic information of visual content during visual word generation by manual separating objects in the visual content during a training phase. Therefore, all detected keypoints that are relevant, are put into the same visual word for each object category, so, that the linkage between the visual words and high level semantic of object category can be obtained. One possible way to preserve semantic information is to generate visual words based upon physical location of keypoints in the visual content. Our hypothesis is that the nearby keypoints in the visual content are more relevant and can represent the semantic information of the visual content more effectively. Therefore, rather than using a manual object separation scheme in order to obtain the relevant keypoints or using simple *k*-mean clustering algorithm, a technique is needed that clusters relevant keypoints together based on their physical locations that can preserve semantic information between keypoints and visual content to improve the quality of visual words.

3.2. Similar keypoints and non-informative visual words

Keypoints are salient patches that contain rich local information in an image. They can be automatically detected using various detectors, e.g. Harris corner (Harris & Stephens, 1988) and DoG (Lowe, 2004) and can be represented by many descriptors, e.g. SIFT descriptor. However, some of these keypoints may not be informative for the clustering algorithm as they are redundant and they even can degrade of the clustering performance. Therefore, similar keypoints should be detected in order to reduce noise and to enhance clustering results. To the best of our knowledge, the reduction of similar keypoints has never been addressed before. Instead researchers in this area focus more on elimination of unimportant visual words. Wu et al. (2009) consider noisy visual words using the range of visual word (maximum distance of a keypoint (feature) to the center of its cluster (visual word). If the keypoint is inside the range of any visual word, the keypoint is assigned to that visual word, otherwise the keypoint is discarded. The weakness of this technique is computationally expensive because every keypoint needs to be compared to the range between keypoint and the visual word's center. For example, if there is N keypoints and M visual words, the complexity of this algorithm will be $O(MN)$. This also leads to a scalability problem to the large-scale visual content system.

Besides the similar keypoints issue, some of the generated visual words may be uninformative when representing visual content and degrade the categorization capability. Non-informative visual words are insignificant local image patterns which are useless for retrieval and classification. These visual words need to be eliminated in order to improve the accuracy of the classification results and to reduce the size of visual word vector space model and computation cost. Rather than to identify unimportant visual words, Yuan et al. (2007a) have attempted to discover unimportant information for larger units of visual words, *meaningless phrases* (word-sets created from a frequency itemset data mining algorithm), by determining the likelihood ratio (the statistical significance measure) of those visual phrases. The top-*k* most meaningful word-sets with largest likelihood ratio will be selected and the rest are considered as meaningless visual phrases and will be discarded. However, this method ignores the coherency (the ordering of visual words) of component visual words in a visual phrase. Tirilly et al. (2008) deployed probabilistic Latent Semantic Analysis (pLSA) to eliminate the noisiest visual words. Every visual word *w,* whose probability between visual word w and concept z, $\Pr(w|z)$, is lower

than $|w|/|z|$, is considered to be an irrelevant visual word since they are not informative for any concept. The main disadvantage of this method is that it ignores the correlations between word and other concepts in the collection. Some words might appear less in one concept but appear more in other concepts and these words could be the featured words. Deleting low-probability words decreases the accuracy of categorization.

3.3. Semantically similar visual words

Visual heterogeneity is one of the greatest challenges when categorization and retrieval relies solely on visual appearance. For example, different visual appearances might be semantically similar at a higher semantic conceptualization. One of the challenges for the BVW method is to discover a relevant group of visual words which have semantic similarity. Recently, a number of efforts have been reported including, among others, the use of the probability distributions of visual word classes (Zheng et al., 2008) which is based upon the hypothesis that semantically similar visual content will share a similar class probability distribution. Yuan et al. (2007a), Yuan, Wu, and Yang (2007b) overcome this problem by proposing a pattern summarization technique that clusters the correlated visual phrases into phrase classes. Any phrases in the same class are considered as synonym phrases. A hierarchical model is exploited to tackle the semantically similar of visual content issue by Jiang and Ngo (2009). A soft-weighting scheme is proposed to measure the relatedness between visual words and the hierarchical model constructed by the agglomerate clustering algorithm and then to capture is-a type relationships for visual words. Although these methods can discover the semantically related visual word sets, there are some remaining issues that need to be overcome. Identifying the semantically related or synonym visual words based on probability distributions (Zheng et al., 2008) might be not always reliable because unrelated ones can accidentally have a similar probability distribution. Finding semantically similar visual words based only on the distance between visual words in the vector space model, e.g. (Jiang & Ngo, 2009; Yuan et al., 2007a) is not effective if those visual words are generated by a simple clustering algorithm because distances in the feature space do not represent the semantic information of visual words.

To this end, this paper proposes a framework to discover semantically similar visual words which preserves the semantic information throughout the BVW construction process and improves the resolution of these three major limitations of visual content classification and retrieval.

4. A semantic-based bag-of-visual words framework

Before going into the details of our proposed system, we will briefly give an overview of the framework (Fig. 1) as follows:

(1) Feature detection: extracts several local patches which are considered as candidates for basic elements, "visual words". Interest point detectors detect the 'keypoint' in an image or the salient image patches using a well-known algorithm. For example, the Difference of Gaussian (DoG) detector (Lowe, 2004) is used in our method to automatically detect keypoints in images. The DoG detector provides a close approximation to the scale-normalized Laplacian of Gaussian that produces very stable image features compared to a range of other possible image functions, such as the gradient, Hessian, and Harris corner detectors.
(2) Feature representation: each image is abstracted in terms of several local patches. Feature representation methods deal with how to represent the patches as numerical vectors. These methods are called feature descriptors. One of the most well-known descriptors is SIFT. SIFT converts each patch to 128 dimensional vector. After this step, each visual content is a collection of vectors of the same dimension (128 for SIFT); the order of different vectors is of no importance.
(3) Semantic visual words construction: this step converts a vector representing patches into "visual words" and produces a "bag-of-visual words" that is represented as a vector (histogram). A visual word can be considered as representative of several similar patches. One simple method performs clustering (i.e. k-means or x-mean clustering algorithm) 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. This representation is analogous to the bag-of-words document representation in terms of form and semantics because the bag-of-visual-word representation can be converted into a visual-word vector similar to the term "vector" for a document.
(4) Non-informative visual words identification: some of the generated visual words may not be useful to represent visual content. Hence, this kind of visual word is need to be detected and removed in order to reduce the size of visual word feature space and computation cost. This can be done using the Chi-square model.
(5) Semantically similar visual words discovery: only a visual word cannot represent the content of an image because it

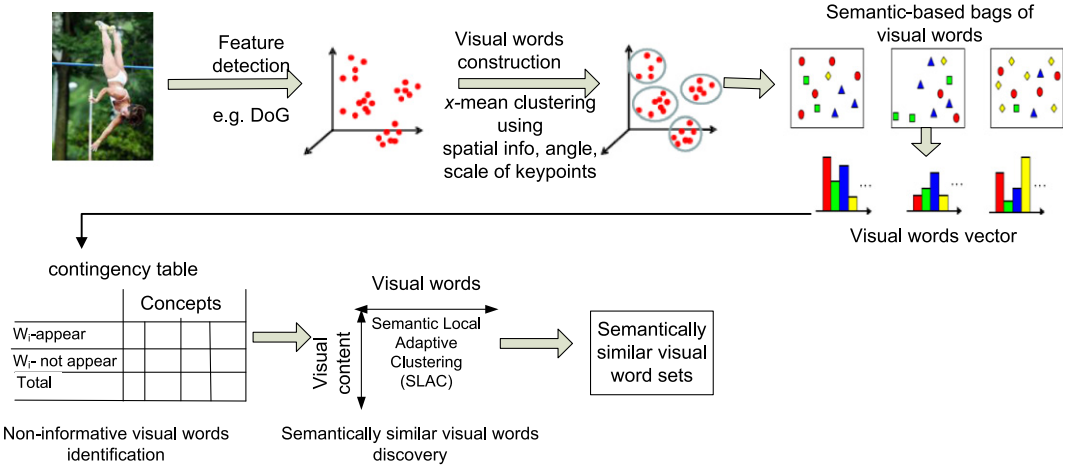


Fig. 1. The semantically similar visual word discovery framework architecture.

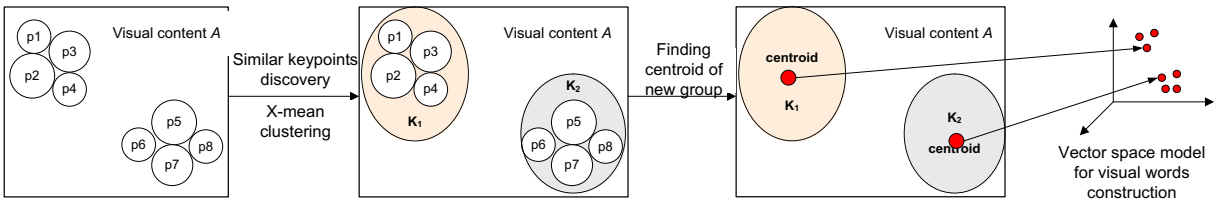


Fig. 2. Similar keypoints are merged together using a new centroid of cluster as a representative keypoint. Those representative keypoints are used to construct the semantic visual words.

may have multiple word senses. In other words, a visual word is ambiguous. To tackle this issue, the associatively similar visual words can help to disambiguate word senses and represent visual content more distinctively. To find these associatively visual words, a semantic local adaptive clustering technique (AlSumait & Domeniconi, 2008) has been proposed to complete this task.

In this paper, the first two steps will not be discussed in detail since our work makes no contributions to those areas. Instead we focus our framework on semantic visual word construction and semantically similar visual words discovery.

4.1. Semantic visual word construction

The high dimension of keypoints in feature space can lead to the feature space containing noise and redundancy. As a result, these kinds of keypoints directly affect the quality of visual words leading to the following serious drawbacks. First, noisy keypoints can confuse the clustering algorithm resulting in poor visual word quality. Second, large numbers of keypoints lead to a large size of the visual word feature space and a subsequent high computation cost. Therefore, we propose a novel method for visual word generation which aims to eliminate the noisy keypoints and reduce similar keypoints as well as preserve semantic information between those keypoints in order to generate semantic-preserved visual words, in short, *semantic visual words*.

4.1.1. Similar keypoints detection

Similar keypoints (κ) will be identified by considering their physical location in visual content. Basically, the keypoint generated from the DoG algorithm provides four useful properties, physical coordination (x, y), angle, scale, and SIFT descriptor. Our hypothesis is that any keypoints located in nearby positions in the visual content could potentially be similar so we can group them together and find a *representative keypoint* (ξ) for those similar keypoints. Therefore, we discover similar keypoints using coordination, angle, scale of the keypoint and SIFT descriptor. Similar keypoints will be grouped together using the x-mean algorithm (Fig. 2). Let κ be a set of the similar keypoints (k), $\kappa_i = \{k_1, k_2, \ldots, k_n\}$ where $n \geqslant 1$. Similar keypoints will be grouped together using the x-mean algorithm. This serves to connect the low-level feature to high level semantic objects and thus the semantic information is preserved. Having grouped the similar keypoints, we use the centre value of the group to represent the whole group of the similar keypoints and this representative value will be used to generate visual words. In each κ, the average value of the SIFT descriptor will be used as ξ to generate the semantic visual words ($\{\bar{\omega}\}$). Consequently, the number of keypoints in feature space is reduced and only ξ are used to generate $\{\bar{\omega}\}$ using the x-mean algorithm. The main benefit of the x-mean algorithm over the k-mean algorithm is its speed; it does not need to specify the cluster numbers (k-value). At this stage, we obtain $\{\bar{\omega}\}$ which is improved in contrast to traditional models because noisy and similar keypoints are reduced with respect to the original visual content using ξ. In addition, semantic information is preserved via the similar keypoints which are the building block of ξ and connected to a high level semantic visual content by their physical co-ordination.

4.1.2. Non-informative visual word identification and removal

Non-informative visual words are often referred to as local visual content patterns that are not useful for retrieval and classification tasks. They are relatively 'safe' to remove in the sense that their removal does not cause a significant loss of accuracy but significantly improves the classification accuracy and computation efficiency of the categorization (Yang & Wilbur, 1996). By analogy a text-based document, there usually are unimportant words, so-called *stop words* (a, an, the, before, after, without etc.), containing in a text document. These stop words need to be removed before further processing, e.g. text categorization, to reduce the noise and computation costs. Likewise, in visual data processing technique, there exist unimportant visual words, the so-called *non-informative visual words* $\{\psi\}$. The non-informative visual words are insignificant local visual content patterns that do not contribute to visual retrieval and classification. These visual words need to be eliminated in order to improve the accuracy of the results and to reduce the size of visual word feature space and computation cost. In this paper, we utilize a statistical model to automatically discover ψ and eliminate them to strengthen the discrimination power. Yang, Jiang, Hauptmann, and Ngo (2007) evaluate several techniques usually used in feature selection for machine learning and text retrieval, e.g. Document frequency, Chi-square statistics, and Mutual information. In contrast to (Yang et al., 2007), in our framework, non-informative visual words are identified based on the document frequency method and the statistical correlation of visual words.

Definition 1. Non-informative visual words $\{\psi\}$

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

1. Usually does not appear in much visual content in the collection (Yang et al., 2007; Yang & Pedersen, 1997). Thus, it has a low document frequency (DF).
2. Has a small statistical association with all the concepts in the collection (Hao & Hao, 2008).

Hence, ψ can be extracted from the visual word feature space using the Chi-squared model. Having created $\{\bar{\omega}\}$ in the previous step, $\{\bar{\omega}\}$ will be quantized into a Boolean vector space model to express

Table 1
The 2^*p contingency table of $\{\bar{\omega}\}_i$.

	C_1	C_2	...	C_k	Total
$\{\bar{\omega}\}_i$-appear	n_{11}	n_{12}	...	n_{1k}	n_{1+}
$\{\bar{\omega}\}_i$-not appear	n_{21}	n_{22}	...	n_{2k}	n_{2+}
Total	n_{+1}	n_{+2}	...	n_{+k}	N

each visual content vector. Assuming that the appearance of the visual word i ($\{\bar{\omega}\}_i$) is independent of any concepts C, $C \in Z$, $Z = \{C_1, C_2, \ldots, C_n\}$ where $n \geqslant 1$, the correlation between $\{\bar{\omega}\}_i$ and its concepts could be expressed in the form of a 2^*p contingency table as shown in Table 1.

Definition 2. The Boolean vector space model.

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

Definition 3. The 2^*p contingency table.

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

$$n_{+j} = \sum_{i=1}^{2} n_{ij}, \quad n_{i+} = \sum_{j=1}^{k} n_{ij}, \tag{1}$$

$$N = \sum_{i=1}^{2}\sum_{j=1}^{k} n_{ij} = \sum_{i=1}^{2} n_{i+} = \sum_{j=1}^{k} n_{+j}, \tag{2}$$

$$x^2_{2^*p} = \delta = \sum_{i=1}^{2}\sum_{j=1}^{k} \frac{(Nn_{ij} - n_{i+}n_{+j})^2}{Nn_{i+}n_{+j}}, \tag{3}$$

To measure the independence of each $\{\bar{\omega}\}$ from all the concepts, the Chi-square statistic (δ) is deployed (Eq. (3)). Having calculated this independence, the δ values are sorted in descending order. The δ value indicates the degree of correlation relationships between $\{\bar{\omega}\}$ and concepts; the smaller the δ value, the weaker the correlation relationship. However, there exists a problem for terms which appear in a small number of documents leading them to have a small δ value. These terms sometimes could be the feature words. In such a case, we need to weight the δ value using Eq. (4) (Hao & Hao, 2008) where DF_r denotes the document frequency of the word r

$$x^2_{weighted} = \frac{x^2_{2^*p}}{DF_r}. \tag{4}$$

This model balances the strength of the dependent relationship between a word, all concepts, and the document frequency of a word. As a result, those $\{\bar{\omega}\}$ which have δ values less than a threshold (chosen experimentally) can be considered to be non-informative visual words and removed. The remaining $\{\bar{\omega}\}$ are informative and useful for the categorization task. Obviously, ψ identified in this manner is collection specific which means by changing the training collection, one can obtain a different ordered list.

Nevertheless, each individual visual word could be ambiguous when it is used for classifying visual content alone. For example, a visual word might represent different semantic meanings in different visual context (polysemy issue) or multiple visual words (different visual appearances) and can refer to the same semantic class (synonymy issue) (Zheng et al., 2008; Zheng, Zhao, Neo, Chua, & Tian, 2008). These problems are crucial and will be addressed in the next section.

4.1.3. Associatively similar visual words discovery

One possible way to disambiguate multiple word senses is to combine visual words into a larger unit, a so-called 'visual phrase' (Yuan et al., 2007b; Zheng et al., 2008) or 'visual sentence' (Tirilly et al., 2008). However, visual phrases are usually constructed using the FIM algorithm which is purely based on frequent word collocation patterns without taking into account the term weighting and spatial information. The latter information is crucial in order to discover the semantic similarity between words. Typically in text documents, there are two ways to form phrases. The first is syntactical: linguistic information is used to form the phrases. The second is statistical, i.e. a vector space model, where co-occurrence information is used to group together words that co-occur more than usual. In this paper, the use of a statistical model to find the similarity of visual words and to construct visual phrases seems suitable because such visual content does not provide any linguistic information. The vector space model is the most commonly used geometric model for the similarity of words with the degree of semantic similarity computed as a cosine of the angle formed by their vectors. However, there are two different kinds of similarity between words, *taxonomic similarity* and *associative similarity*. Taxonomic similarity, or categorical similarity, is the semantic similarity between words at the same level of categories, so-called synonyms. Associative similarity is a similarity between words that are associated with each other by virtue of semantic relations other than taxonomic ones such as a collocation relation and proximity. In this paper, we mainly focus on discovering the associatively similar visual words using the semantic local adaptive clustering algorithm (SLAC) (AlSumait & Domeniconi, 2008) which takes the term weighting and spatial information (distance between $\{\bar{\omega}\}$) into account.

The elimination of the noise visual words (ψ) and the co-occurrence information makes visual content representation more efficient to process. This can be added using semantic similarity techniques but these are still challenging to use in practice. To tackle this challenge of determining semantic similarity, we exploit the SLAC clustering algorithm to cluster the semantically similar words into the same visual phrase. Nevertheless, it is difficult to define the semantic meaning of a visual word, as it is only a set of quantized vectors of sampled regions of visual content (Zheng et al., 2008). Hence, rather than defining the semantics of a visual word in a conceptual manner, we define the semantically similar visual words using their position in a feature space.

Definition 4. Associatively similar visual words.

Associatively similar visual words (φ) are a set of semantic visual words ($\{\bar{\omega}\}$) which have a different visual appearance but which are associated with each other by virtue of semantic relations other than taxonomic ones such as collocation relations and a proximity. A set of associatively similar visual words is called a *visual phrase*.

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

$$K_j(d_1, d_2) = \phi(d_1) Sem_j Sem_j^T \phi(d_2)^T, \quad (5)$$

$$Sem_j = R_j P, \quad (6)$$

where d is a document in the collection, and $\varphi(d)$ is document vector, Sem_j is a semantic matrix which provides additional refinements to the semantics of the representation. P is the proximity matrix defining the semantic similarities between the different terms and R_j is a local term-weighting diagonal matrix (Fig. 3) corresponding to cluster j, where w_{ij} represents the weight of term i for cluster j, for $i = 1, \ldots, D$. One simple way to compute weights to w_{ij} is to use the inverse document frequency (*idf*) scheme. However, the *idf* weighting scheme concerns only document frequency without taking the distance between terms into account. In other words, the *idf* weighting scheme does not concern the inter-semantic relationships among terms. In this paper, therefore, a new weighting measure based on the local adaptive clustering (LAC) algorithm is utilized to construct matrix R.

LAC gives less weight to data which are loosely correlated and this has the effect of elongating distances along that dimension (Domeniconi et al., 2007). In contrast, features along which data are strongly correlated receive a larger weight which has the effect of constricting distances along that dimension. Different from the traditional clustering algorithm, LAC clustering of concepts is not only based on points among features, but also involves weighted distance information. Eq. (7) shows the formula of the LAC term weight calculation

$$w_{ij} = \frac{\exp\left(-\frac{1}{|S_j|}\sum_{x \in S_j}(c_{ji} - x_i)^2/h\right)}{\sum_{i=1}^{D}\exp\left(-\frac{1}{|S_j|}\sum_{x \in S_j}(c_{ji} - x_i)^2/h\right)}, \quad (7)$$

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

P has nonzero off-diagonal entries, $P_{ij} > 0$, when the term i is semantically related to the term j. To compute P, the Generalized Vector Space Model (GVSM) (Wong, Ziarko, & Wong, 1985) is deployed to capture the correlations of terms by investigating their co-occurrences across the corpus based on the assumption that two terms are semantically related if they frequently co-occur in the same documents. Since P holds a similarity figure between terms in the form of co-occurrence information, it is necessary to transform it to a distance measure before utilizing it. Eq. (8) shows the transformation formula:

$$P_{ij}^{dist} = 1 - (P_{ij}/\max(P)), \quad (8)$$

where P_{ij}^{dist} is a distance information, $max(P)$ is the maximum entry value in the proximity matrix. Consequently, a *semantic dissimilarity matrix* for cluster j is a $D \times D$ matrix given by Eq. (9)

$$Sem_j^{dissim} = R_j P^{dist}, \quad (9)$$

which represents semantic dissimilarities between the terms with respect to the local term weightings. The algorithm of SLAC (Algorithm 1) starts with k initial centroids and equal weights. It partitions the data points, re-computes the weights and data partitions

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

Fig. 3. The local term-weighting diagonal matrix (R_j).

accordingly, and then re-computes the new centroids. The algorithm iterates until convergence or a maximum number of iterations are exceeded. The SLAC uses a semantic distance. A point x is assigned to the cluster j that minimizes the semantic distance of the point from its centroid. The semantic distance is derived from the kernel in Eq. (5) as follows:

$$L_w(c_l, x) = (x - c_l) Sem_l^{dissim} Sem_l^{dissim^T} (x - c_l)^T. \quad (10)$$

Hence, every time the algorithm computes S_j, the semantic matrix must be computed by means of the new weights.

Algorithm 1. The associatively similar visual words discovery algorithm

Input: N points $x \in R^D$, k, and h
Output: semantically similar visual word sets (visual phrases)
1. Initialize k centroids $c_1, c_2, \ldots, c_k$;
2. Initialize weights: $w_{ij} = \frac{1}{D}$, for each centroid c_j, $j = 1, \ldots, k$ and for each term $i = 1, \ldots, D$;
3. Compute P; then compute P^{dits};
4. Compute Sem^{dissim} for each cluster j (Eq. (9));
5. For each centroid c_j, and for each point x, set: $S_j = \{x | j = \arg\min_l L_w(c_l, x)\}$,
where $L_w(c_l, x) = (x - c_l) Sem_l^{dissim} Sem_l^{dissim^T} (x - c_l)^T$
6. Compute new weights:
for each centroid c_j, and for each term i:
Compute Eq. (7)
7. For each centroid c_j:
Recompute Sem_l^{dissim} matrix using new weights w_{ij};
8. For each point x:
Recompute $S_j = \{x | j = \arg\min_l L_w(c_l, x)\}$;
9. Compute new centroids: $c_j = \frac{\sum_x x 1_{S_j}(x)}{\sum_x 1_{S_j}(x)}$ for each $j = 1, \ldots, k$
where $1_S(\cdot)$ is the indicator function of set S
10. Iterate 5–9 until convergence, or maximum number of iterations is exceeded

SLAC clusters visual words according to the degree of relevance and thus generates visual phrases which are semantically related. In addition, the randomly generated visual phrases will not occur as in the FIM-based method (Yuan et al., 2007a; Zheng et al., 2008). However, the disadvantage of the SLAC algorithm is computation complexity. The running time of one iteration is $O(kND^2)$, where k is the number of clusters, N is the number of visual contents, and D is the number of visual words. Since use of the Chi-square model reduces the thousands of visual words, D is reduced. Fig. 4 illustrates the advantage of the SLAC algorithm.

5. Visual content indexing, similarity measure, and retrieval

Semantic visual words (φ) representation for visual contents are indexed using the inverted file scheme due to its simplicity, efficiency, and practical effectiveness (Witten, Moffat, & Bell, 1999; Zheng et al., 2008). The *cosine similarity* (Eq. (11)) is deployed to measure the similarity between the queried and the stored visual content in a collection. Let $\{P_i\}_{i=1}^N$ be the set of all visual contents in the collection. The similarity between the query (q) and the weighted φ associated with the visual content (p) in the collection is measured using the following inner product:

$$sim(p, q) = \frac{p \cdot q}{\|p\|\|q\|}. \quad (11)$$

The retrieval visual content is ranked by descending order, according to its similarity value with the query image. In order to evaluate the retrieval performance of φ, the two classical

Fig. 4. The SLAC algorithm classifies visual words A, B, C, D and E into two groups of φ, {A, B} and {C, D, E}. In this example, visual words A and B alone cannot distinguish image (a) and (b) because they share a visual similarity with words C and D: A is similar to C and B is similar to D. However, the combination of the visual word E with {C, D} can effectively distinguish the high jump event (a) and the pole vault event (b).

measures used to evaluate the performance of information retrieval systems are *precision* and *recall*. Let A denote all relevant documents (visual contents) in the collection. Let B denote the retrieved documents which the system returns for the user query.

- Precision is defined as the portion of relevant documents in the retrieved document set.

$$Precision = \frac{|A \cap B|}{|B|}. \quad (12)$$

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

$$Recall = \frac{|A \cap B|}{|A|}. \quad (13)$$

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

$$P_{interp}(r) = max(r'), \quad r' \geq r. \quad (14)$$

6. Experimental results and discussions

We evaluate the performance of our proposed technique introduced in Section 2 against a sport image collection. The image collection contains 2000 images of four sport genres (high jump, long jump, pole vault, and swimming). We divide the image collection into two sets: 800 images (200 from each category) are selected for training which have total 259,790 keypoints and the rest are used for testing which are generated 649,475 keypoints by the DoG algorithm and these keypoints are further processed to find the representative keypoint (ξ). The vector quantization technique is utilized for clustering the visual words based on ξ using the x-means clustering algorithm which later produces the semantic visual words ($\{\bar{\omega}\}$).

6.1. Evaluation of the noisy keypoint reduction and the representative keypoints

First, we compare the proportion of number of keypoints (γ) and ξ in order to investigate how much the technique presented in Section 4.1.1 can reduce the similar keypoints. This comparison is performed on the original number of γ generated from the DoG algorithm and the number of ξ.

Table 2
A number of keypoints and representative keypoints for four different sport categories.

Sports	Keypoints {γ}	Representative keypoints {ξ}	Decrease (%)
High jump	105,410	69,244	34
Long jump	186,348	126,232	32
Pole vault	245,328	168,467	31
Swimming	112,389	73,592	35
Average			33

Table 3
The computation cost comparison of two clustering algorithms to generate the 6987 visual words (including non-informative visual words) for the pole vault event.

Clustering algorithm	{γ} (s)	{ξ} (s)	Ratio
x-Mean	3624	1139	3.2
k-Mean	4356	1613	2.7

Second, we evaluate the computation cost for visual words construction between two algorithms, k-mean and x-mean algorithms using Weka.[1] The two algorithms are used to cluster γ and ξ from all categories. To illustrate this, we present the computation time for the pole vault data set which has γ and ξ shown in Table 2 respectively. We first cluster data with the x-mean algorithm and then we specify the k value in k-mean that is equal to the number of clusters from the x-mean. Therefore, both algorithms generate equal number of visual words. From the results shown in Table 3, the x-mean clustering obtains the lowest computation cost which is 3.2 times faster than γ whereas the k-mean algorithm obtains a cost which is 2.7 times faster than γ. This shows that the use of ξ instead of the original keypoints (γ) significantly reduces the computation time.

However, the reduction of keypoints may degrade the categorization performance as well as the retrieval power. We study these aspects in subsequent sections.

6.2. Evaluation of the non-informative visual word reduction

Before going to compare the classification and retrieval performance, we would like to illustrate the proportion of visual words which are generated from both γ and ξ and are further processed to terminate the non-informative visual words. The number of

[1] Weka data mining tool, See www.cs.waikato.ac.nz/ml/weka/.

Table 4
The proportion of normal to representative semantic visual words using *x*-mean clustering before and after the non-informative visual word $\{\psi\}$ removal.

Keypoints	$\{\bar{\omega}\}$	$\{\psi\}$	$\{\varepsilon\} = \{\bar{\omega}\} - \{\psi\}$	Proportion $(\psi/\{\bar{\omega}\})$
Normal keypoints $\{\gamma\}$	6978	2622	4356	38%
Representative keypoints $\{\xi\}$	4893	971	3922	20%
Difference	30%	63%	10%	–

$\{\bar{\omega}\}$ = visual words; $\{\psi\}$ = non-informative visual words. $\{\{\bar{\omega}\} - \{\psi\}$ = informative visual words $\{\varepsilon\}$.

keywords shown in Table 4 is from all categories and generated using the *x*-mean clustering algorithm.

From Table 4, the number of $\{\bar{\omega}\}$ generated from γ is much larger, by 30%, compared to the number of $\{\bar{\omega}\}$ generated from ξ. This indicates that using γ without similar keypoint reduction in the clustering algorithm generates a greater number of noisy visual words, about 38%, and takes a longer time (Table 3) to construct $\{\bar{\omega}\}$. In contrast, the use of ξ to construct the $\{\bar{\omega}\}$ obtains less computational cost and generates a smaller number of ψ. This illustrates that the similar keypoint detection enhances the clustering algorithm by reducing the clustering numbers leading to generating less numbers of ψ, about 20%. The use of ξ can reduce the number of ψ by 63%.

However, the difference in informative visual words (ε) in both techniques is only 10%. This demonstrates that representative keypoints (ξ) can produce the semantic visual words $(\{\bar{\omega}\})$ which contain less number of non-informative data (ψ) but still obtain high numbers of informative visual words similar to the method using normal keypoints while being computationally less expensive. Next, we study about the effect of ψ removal to the classification performance. From this point, we will call a set of visual words created from ξ as a semantic visual words set $\{\bar{\omega}\}$ and a set of visual words created from γ which includes ψ as a normal visual words set $\{\alpha\}$. We compare the classification results between visual words with uninformative visual words $\{\alpha + \psi\}$, visual words without uninformative visual words $\{\alpha - \psi\}$, semantic visual words with uninformative visual words $\{\bar{\omega}\} + \Psi$ and semantic visual word without uninformative visual words $\{\bar{\omega}\} - \Psi$ using the SVM-RDF algorithm. The classification results are shown in Fig. 5.

Fig. 5 shows that the classification accuracy is heavily influenced by eliminating ψ. For example, in the pole vault event, the classification accuracy is increased from 45% to 56% when ψ is removed from normal visual words set (α) whereas the semantic visual words set $(\{\bar{\omega}\})$ obtained is 2% higher in the same category and the improvement of classification accuracy trends similarly to all categories. This consistent improvement suggests that ψ plays an important role in strengthening the discriminative power. Although the large number of normal visual words (α) makes a feature become more discriminative, α also makes the feature vector less generalizable and might contain more noise. A large number of α also increases the cost associated with clustering keypoints in computing visual-word features and in generating supervised classifiers.

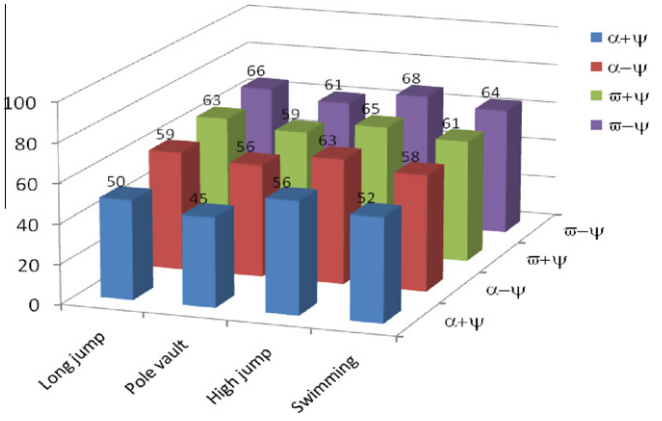


Fig. 5. The classification performance comparison of the normal visual words (α) and the semantic visual words $(\{\bar{\omega}\})$ before and after the ψ removing using SVM–RDF.

6.3. Evaluation of visual content classification using the semantically similar visual words

Next, we study the classification performance using the associatively (semantically) similar visual words φ. This classification has been tested after performing non-informative visual words removal, term weighting (LAC) and associatively similar visual word discovery (SLAC). Fig. 6 shows the performance of sport genre classification using three different classification algorithms: Naïve Bayes, SVM (Linear) and SVM (RBF). Among these classification methods, two types of BVW models are deployed for comparing the classification performance, the semantic-preserved BVW model (SBVW) and the Traditional BVW model (TBVW). SBVW refers to the bag-of-visual words model containing φ which preserves the semantic information among keypoints and visual words whereas TBVW constructs visual words using the simple *k*-mean clustering algorithm without preserving semantic information among keypoints and visual words. The evaluation criterion here is the mean average precision (MAP) collected from the Weka classification results. The experimental result (Fig. 6) shows that the proposed model (SBVW) which classifies sport types based on φ is superior to the TBVW method in all cases (classifiers and sport genres). This suggests that the physical spatial information between keypoints, term weighting and semantic distance of the SLAC scheme in vector space are significant to interlink the semantic conceptualization of visual content and visual features, allowing the system to efficiently categorize the visual data. In other words, semantically similar visual words (φ) which are computed based on term weight and semantic distance in the vector space can represent visual content more distinctively than traditional visual words. Specifically, SVM-linear outperforms other categorization methods by producing the highest performance up to 78% MAP in the high jump event. Among all sports, the high jump event obtains the highest classification accuracy. This could be because there are fewer objects in this sport event, e.g. athlete, bar and foam matt. As a consequent, there is less noise among objects in the visual content for the high jump event compared to other sports which contain more objects. The categorization algorithm seems to classify the high jump data more effectively compared to the other sport disciplines.

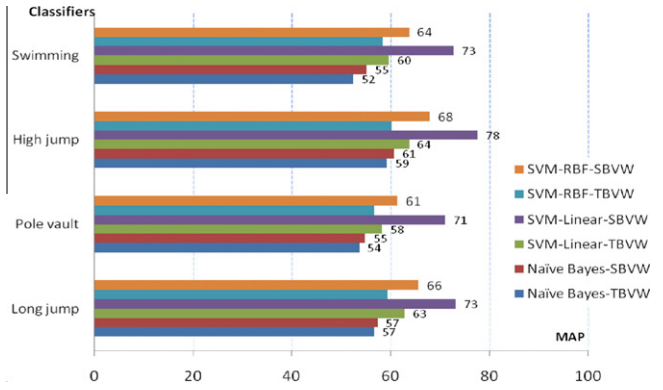


Fig. 6. The comparison of classification performance on four sport types between the preserve semantic information (SBVW) method and the normal BVW method (NBVW).

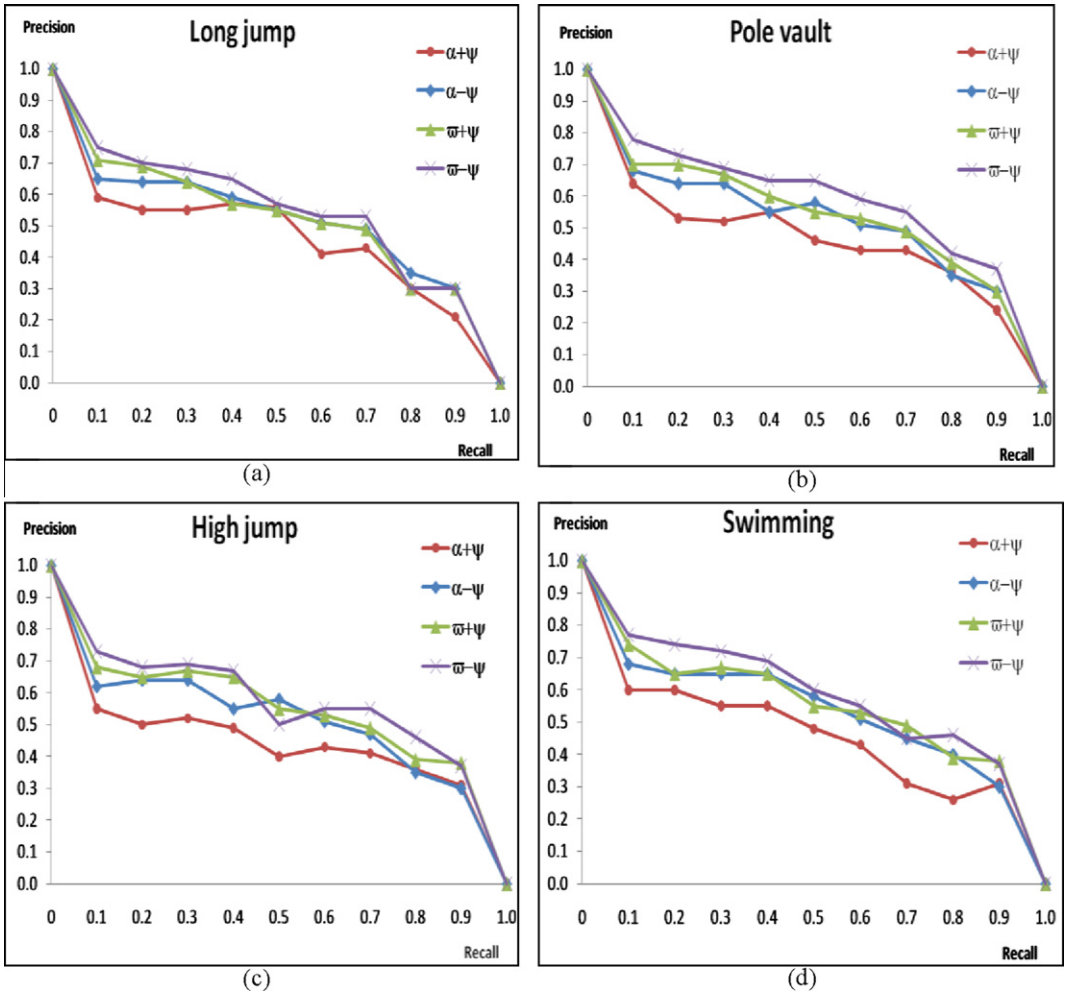


Fig. 7. Retrieval efficiency comparison of the normal visual words (α) and the semantic visual words ($\{\bar{\omega}\}$) before and after removing ψ.

6.4. Evaluation of content-based image retrieval based on the semantically similar visual words

In this section, we evaluate the retrieval performance based on the representation of φ for visual contents using a different number of visual words respect to the information given in Table 4. Fig. 7 illustrates the retrieval performance of the four different sport genres. Overall, the retrieval performance using the semantic visual words without non-informative visual words $\{\bar{\omega}\} - \Psi$ outperforms the others in all sport categories. In other words, the semantically similar visual words (φ) can deliver superior results over other methods using a more compact representation (small number of visual words). Although retrieval performance seems to not directly depend on the number of visual words, too many visual words could nevertheless bring poor results. This is because a large number of visual words might contain meaningless visual words and, thus, this could cause the retrieval mechanism to retrieve visual content that is not very semantically similar to the query. As a result, precision is reduced. In summary, we attribute the retrieval performance to two factors. First by preserving the semantic information throughout the processes of the visual word construction, this leads to visual words representing visual content more effectively. This semantic information makes the visual content distribution in the feature space more coherent and produces smaller variations within the same class. Second the use of the associative visual words efficiently distinguishes visual content and reduces the dimension of visual words in vector space using the Chi-square statistical model. It is able to resolve the curse of the dimensionality problem. As a result, these two factors enhance the retrieval performance.

7. Conclusions and future work

We presented a framework for visual content representation which has three main advantages. First the use of representative keypoints reduces their dimensions in feature space, consequently, improving the clustering results (quality of visual words) and reducing computation costs. Second a method is proposed for generating semantic visual words based on the spatial information of keypoints which are the linkage between visual words and high level semantic of objects in visual content. Third an approach to find semantically similar visual word to solve the visual heterogeneity problem is presented. Based on the experimental results, the proposed technique allows the bag-of-visual words (BVW) to be more distinctive and more compact than existing models. Furthermore, the proposed BVW model can capture the semantics of visual content efficiently, resulting in a higher classification accuracy. However, the major disadvantage of the proposed technique is the computation complexity of the SLAC algorithm. We proposed to resolve this problem by reducing the number of visual words (as dimensions in vector space) using the Chi-square model.

We are currently extending our technique to restructure bag-of-visual words model as a hierarchical, ontology model to

disambiguate visual word senses. The unstructured data of visual content is transformed into a hierarchical model which describes visual content more explicitly than the vector space model using conceptual structures and relationships. Hence, this could aid information systems to interpret or understand the meaning of visual content more accurately, i.e., this helps in part to narrow the semantic gap.

Acknowledgement

This work has been supported by Science Ministry Research Funding, the Royal Thai Government, Thailand.

References

Alhwarin, F., Wang, C., Ristic-Durrant, D., & Graser, A. (2008). Improved SIFT-features matching for object recognition. In *International academic conference on vision of computer science-BSC* (pp. 179–190).

AlSumait, L., & Domeniconi, C. (2008). Text clustering with local semantic kernels. In M. W. Berry & M. Castellanos (Eds.), *Survey of text mining II: Clustering, classification, and retrieval* (pp. 87–105). London, United Kingdom: Springer-Verlag London Limited.

Bach, J., Fuller, C., Gupta, A., Hampapur, A., Horowitz, B., Humphrey, R., et al. (1996). Virage image search engine: An open framework for image management. In *Storage and retrieval for still image and video databases IV* (pp. 76–87).

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

Csurka, G., Dance, C. R., Fan, L., Willamowski, J., & Bray, C. (2004). Visual categorization with bags of keypoints. In *International workshop on statistical learning in computer vision* (pp. 1–22).

Domeniconi, C., Gunopulos, D., Ma, S., Yan, B., Al-Razgan, M., & Papadopoulos, D. (2007). Locally adaptive metrics for clustering high dimensional data. *Data Mining and Knowledge Discovery, 14*(1), 63–97.

Duygulu, P., Barnard, K., Freitas, J. F. G., & Forsyth, D. A. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Proceedings of the 7th European conference on computer vision-part IV* (pp. 97–112).

Forsyth, D., & Fleck, M. (1997). Body plans. In *Proceedings of IEEE computer society conference on computer vision and pattern recognition* (pp. 678–683).

Hao, L., & Hao, L. (2008). Automatic identification of stop words in Chinese text classification. In *2008 International conference on computer science and software engineering* (Vol. 1, pp. 718–722).

Harris, C., & Stephens, M. (1988). A combined corner and edge detector. In *Proceedings of the 4th Alvey vision conference* (pp. 147–151).

Hironobu, Y. M., Takahashi, H., & Oka, R. (1999). Image-to-word transformation based on dividing and vector quantizing images with words. In *Proceedings of international workshop on multimedia intelligent storage and retrieval management* (Vol. 4, pp. 405–409).

Jiang, Y., & Ngo, C. (2009). Visual word proximity and linguistics for semantic video indexing and near-duplicate retrieval. *Computer Vision and Image Understanding, 113*(3), 405–414.

Ke, Y., & Sukthankar, R. (2004). PCA-SIFT: A more distinctive representation for local image descriptors. In *Proceedings of the 2004 IEEE computer society conference on computer vision and pattern recognition, CVPR 2004* (Vol. 2, pp. 513–506).

Lee, Y., Lee, K., & Pan, S. (2005). Local and global feature extraction for face recognition. In *Proceedings of the 5th international conference on audio- and video-based biometric person authentication* (pp. 219–228).

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *Proceedings of the international conference on computer vision* (Vol. 2, pp. 1150–1157).

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60*(2), 91–110.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. London, United Kingdom: Cambridge University Press.

Moravec, H. (1977). Towards automatic visual obstacle avoidance. In *Proceedings of the 5th international joint conference on artificial intelligence* (p. 584).

Naphade, M., & Smith, J. (2003). Learning regional semantic concepts from incomplete annotation. In *Proceedings of the international conference on image processing* (Vol. 2, pp. 603–606).

Rui, Y., Huang, T. S., & Chang, S. (1999). Image retrieval: Current techniques, promising directions, and open issues. *Journal of Visual Communication and Image Representation, 10*(1), 39–62.

Schmid, C., & Mohr, R. (1997). Local gray value invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19*(5), 530–535.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transaction on Pattern Analysis and Matching Intelligent, 22*(12), 1349–1380.

Smith, J. R., & Chang, S. (1996). VisualSEEk: A fully automated content-based image query system. In *Proceedings of the 4th ACM international conference on multimedia* (pp. 87–98).

Tirilly, P., Claveau, V., & Gros, P. (2008). Language modeling for bag-of-visual words image categorization. In *Proceedings of the 2008 international conference on content-based image and video retrieval* (pp. 249–258).

Tseng, B., Lin, C., Naphade, M., Natsev, A., & Smith, J. (2003). Normalized classifier fusion for semantic visual concept detection. In *Proceedings of the international conference on image processing* (Vol. 2, pp. 535–538).

Witten, I. H., Moffat, A., & Bell, T. C. (1999). *Managing gigabytes: Compressing and indexing documents and images* (2nd ed.). London, United Kingdom: Academic Press.

Wong, S. K. M., Ziarko, W., & Wong, P. C. N. (1985). Generalized vector spaces model in information retrieval. In *Proceedings of the 8th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 18–25).

Wu, L., Hoi, S. C., & Yu, N. (2009). Semantics-preserving bag-of-words models for efficient image annotation. In *Proceedings of the 1st ACM workshop on large-scale multimedia retrieval and mining* (pp. 19–26).

Yang, J., Jiang, Y., Hauptmann, A. G., & Ngo, C. (2007). Evaluating bag-of-visual-words representations in scene classification. In *Proceedings of the international workshop on multimedia information retrieval* (pp. 197–206).

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Proceedings of the 14th international conference on machine learning* (pp. 412–420).

Yang, Y., & Wilbur, J. (1996). Using corpus statistics to remove redundant words in text categorization. *Journal of the American Society for Information Science, 47*(5), 357–369.

Yuan, J., Wu, Y., & Yang, M. (2007a). Discovery of collocation patterns: From visual words to visual phrases. In *IEEE conference on computer vision and pattern recognition* (pp. 1–8).

Yuan, J., Wu, Y., & Yang, M. (2007b). From frequent itemsets to semantically meaningful visual patterns. In *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 864–873).

Zheng, Y., Neo, S., Chua, T., & Tian, Q. (2008). Toward a higher-level visual representation for object-based image retrieval. *The Visual Computer, 25*(1), 13–23.

Zheng, Y., Zhao, M., Neo, S., Chua, T., & Tian, Q. (2008). Visual synset: Towards a higher-level visual representation. In *IEEE conference on computer vision and pattern recognition* (pp. 1–8).