Athletics images interpretation using structural ontology model
Kraisak Kesorn*
Computer Science and Information Technology Department, Science Faculty,
Naresuan University, Phitsanulok, Thailand 65000

Received June 2012
Accepted November 2012

Abstract
The continual rapid growth in digital content acquisition and visualization makes it increasingly challenging to find, organize, and access visual information. Typically, image classification and retrieval systems tend to rely only on the lowlevel visual structure within images. Image classification methods usually perform based upon a vector space model. This paper presents a framework 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.

Keywords: Visual content representation, Image interpretation, Visual words disambiguation, Ontology model.

*Corresponding author. Tel.: +66 (0) 8 1555 7499
Email address: kraiskk@nu.ac.th
1. Introduction

Image retrieval aims at aiding users to find desired images more easily, speedily and efficiently. However, human users usefully retrieve images at higher levels of semantics but this is still far from being achieved in practice. 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. The Scale Invariant Feature Transform (SIFT) [2] 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 [3], annotation [4], and retrieval [5]. 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 deployed 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. A framework to generate a new representation model which preserves semantic information throughout the BVW construction process that can resolve the ambiguity of the generated visual words is proposed.

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 interpretation accuracy.

The remainder of this paper is organized as follows. Section 2 presents a survey of ‘state-of-the-art’ frameworks and their limitations. Section 3 describes our proposed technique. Section 4 discusses our experimental results. Finally, section 5 summarizes our key contributions, discusses limitations, and further work.

2. State of the art

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 [5] which is based upon the hypothesis that semantically similar visual content will share a similar class probability distribution. Yuan et al. [6] 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. To solve visual heterogeneity problem, a hierarchical model has been exploited by Jiang et al. [7] 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 [7] is not practical in real
situations because the hierarchical model is only a binary tree 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. Thus, 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, is often impractical to capture semantic relationships between concepts of visual information as theydo not represent the semantics of visual content efficiently. Typically, the object detection/recognition task using low-level features still have some limitations. Several researchers have tried to restructure visual words as a hierarchical model in order to disambiguate word senses more explicitly and effectively. Several 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 [7], Hierarchical Spatial Markov model [8] and the Hierarchical Latent Dirichlet Allocation algorithm [9]. 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. To this end, this paper proposes a framework to generate a new representation model which addresses the mentioned limitation.

3. Proposed framework

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 limitation (as described in section 2). 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 [10]. 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, 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 Figure 1 which enhances the method of Wu [4] 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 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 [8]. 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 Figure 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 \( \{\mathcal{V} \in C_i\} \) are obtained for each object category \( C_i \). Different visual words represent different views of different parts of an object.

![Figure 1](image)

**Figure 1** 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.

3.1 Visual word disambiguation using the concept range

DEFINITION 1. The concept range [8] of the key object \( C_i \) is the maximum distance of a visual word’s centroid \( \mathcal{V} \) to the concept’s centroid \( C_i \) and can be calculated using the following formula:

\[
 r_i = \max \left| v - c_i \right|, \quad v \in \mathcal{V} \tag{1}
\]

The concept range is useful for the visual word sensedisambiguation 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 [7], [8] which cannot represent multiple-parent relationships. 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 2) 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 (2) shows the normalization formula,

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

In this paper, the ontology model is represented in the RDF (Resource Description Framework) format. RDF is selected as the knowledge representation rather than RDFS (Resource Description Framework Schema) or OWL (Web Ontology Language) 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 [12]. 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 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.

### 3.2 Handling the uncertainty 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 hammer object in a hammer throw image (Figure 2) is missing. The system might could not classify the content for the image or misclassify. In this case, it is difficult for the system to interpret that this image is relevant to a hammer throw event. 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.

![Figure 2](image.png) A hammer throw image which a hammer is missing

A Bayesian network (BN) complements the ontology model to aid the interpretation of visual content. A Bayesian Network 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.

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.

4. Results and discussions

In the prototype of this research, a new test collection was established to focus on only five athletics sports (high jump, long jump, pole vault, javelin throw, and hammer throw). This is because they are visually similar and, thus, they are very challenging to classify. Images are collected from the Olympic organization website and the Google image search engine as the basis for the test collection. There exist standard test collections that provide a “golden standard” to evaluate the retrieval performance of image retrieval systems. One of these standard collections of sport images is the “Event dataset” provided by Stanford University. It contains only 1,579 relevant images which was too few for testing the system. A main hypothesis is that 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).

4.1 Retrieval performance evaluation

In this research, the two classical measures used to evaluate the performance of the retrieval mechanism, precision and recall. Let A denote all relevant images (as specified in a user query) in the image collection. Let B denote the retrieved images which the system returns for the user query. Definition 2. The portion of relevant images in the retrieved image set is precision.

\[
\text{precision} = \frac{|A \cap B|}{|B|}
\]  

Definition 3. The portion of relevant images that were returned by the system and all relevant images in the collection is recall.

\[
\text{recall} = \frac{|A \cap B|}{|A|}
\]

The retrieval performance (precision-recall graph) between the proposed method, the so-called Ontology-based Visual Semantic Search (OVSS), a traditional bag-of-visual word model (TBVW) and a traditional CBIR (LIRE framework) are compared. The results (Figure 3) show that the retrieval performance is affected by the proposed technique. Since the OVSS technique analyses an image query and interprets it into a high-level conceptualisation, the search engine is able to perform conceptual searching rather than a simple low-level feature matching. As a consequence, more relevant images can be recognised and retrieved. This leads to the OVSS technique obtaining the highest precision and recall compared to other techniques. The content-based search (LIRE) retrieves all images which have similar low-level features. Unfortunately, some of them are not semantically relevant to the image query. As a result LIRE, obtains a lower precision and recall compared to other techniques. T-BVW attains a better retrieval performance compared to LIRE because the use of associative visual words efficiently distinguishes visual content more than the colour and texture features.
used in LIRE. However, it obtains a lower precision and recall than the OVSS method because the T-BVW model represents visual content based on the feature space model whereas OVSS deploys a hierarchical model which expresses visual content more explicitly than the feature space model. This structured model is able to disambiguate visual word senses more effectively. As a result, this structural model enhances the visual content interpretation and the retrieval performance. The use of the ontology model incorporated with the BVW model allows the systems to recognise all semantically similar images even when their visual appearance is different. In other words, the proposed technique is highly invariant to visual appearance.

4.2 Image interpretation evaluation

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. An Ontology-based Bayesian Network technique has been deployed in order to enhance the visual content classification and uncertainty management. In this section, the performance for visual classification using a hierarchical Bayesian network will be evaluated and compared with other categorization frameworks. In this paper, three major classification algorithms have been compared.

**Figure 3** Retrieval performance results comparison of OVSS (the presented method), BVW and LIRE (the CBIR framework)

**Figure 4** Object-based classification performance comparison between an Ontology-based Bayesian Network (OBN), SVM-Linear and SVM-RBF

BN usually has classification performance lower than the SVM. However, BN in this paper is modified by incorporating with the ontology model, so-called OBN, which contains additional information about sport concepts. This can enhance the classification power of BN. The classification results in Figure 4 indicate that OBN can improve the classification power compared to SVM-Linear and SVM-RBF. This is because OBN classifies data based on the hierarchical structure. The structural model addresses the relationship between the key objects and all possible concepts of sport explicitly. In addition, it exploits the conditional probability to cope...
with the uncertainty which may occur during the classification process. This probability aids the classifier in making a decision about which category (sport genre) an image should be in when an uncertainty occurs, e.g. when the underlying objects are missing. As such, this mechanism leads the proposed system to obtain a higher classification performance than the other two methods. Since SVM-Linear and SVM-RBF performs categorization based purely upon a statistical calculation, they do not have a mechanism to deal with the uncertainty. Therefore, they obtain a lower classification performance than OBN. Among all sports, the hammer throw event obtains the lowest classification accuracy, about 74% (for OBN). From analysing the classification data, the main cause is that the system cannot detect a hammer object in the hammer throw images. From the observation of images in the test collection, in several cases, a hammer object is very small and is merged into the image background. Hence, the keypoints of a hammer object are very difficult to detect and extract (Figure 5) because there are a lot of noises from the background. Consequently, a hammer object cannot be detected because the generated visual words are different from the visual words generated in the training phase. When the system cannot detect a hammer object in an image, it may misclassify a hammer throw image as other sports e.g. a long jump using the probability table.

5. Conclusions

This paper proposes to replace the visual word vector space model with a structural ontology model for visual content representation, which provides an ontology-based classification, interpretation solution for images of athletic sports. 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 othersports equipment (e.g. to differentiate racket or ball). 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.

A major limitation of the framework is ontology incompleteness [13]. 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 [14]. 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

Figure 5 Examples of hammer throw images which a hammer object is merged with the background.
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 [15]) 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.

The framework is also currently extended to a personalized image retrieval (PIMR) system which is identified as a key step in order to cope with the variety of users and the continuous growth in the number of multimedia documents in the future. The main challenges for PIMR include:

1) Automatic user preference acquisition-manual user profile creation is not possible for a large scale system; automatic user preference acquisition is more scalable;

2) Dynamic capture of users' interests-users profiles are usually not static but vary with time and depend on the situation and, thus, profiles should be automatically updated based on observations of users actions; and

3) Richness of the semantic representation-user preferences should be represented in a richer, more precise and less ambiguous way than a keyword/text-based model.

6. Acknowledgment

This research has been supported by the National Science and Technology Development Agency (NSTDA), Thailand. Project no: Sci_T55011.

7. References


