Sport Photo Semantic Representation Using Image Captions to Aid the Image Retrieval System

Kraisak Kesorn, Stefan Poslad

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

Abstract—Image captions usually include semantic descriptions of the images generated by humans. Therefore, they are essential cues to represent the semantic of an image. This paper presents steps in a long process of representing, discovering, storing the knowledge base, and retrieving for visual information. We exploit a Natural Language Processing (NLP) framework in order to extract the knowledge from image captions and to transform those unstructured data to a hierarchical model (semantic model). The novelty of the proposed framework is to use the semantic model in order to find implicit relationships among concepts of photographs which are not mentioned directly in text captions. The Latent Semantic Indexing (LSI) is deployed to solve the ontology imperfections. The experiments showed that the major hypothesis of this work was clearly validated by the experimental results.

I. INTRODUCTION

The increasing popularity of sport leading to sport images and videos are produced every day. This media requires efficient methods for indexing and searching. Automatic discovery of the meaning of the image is one of the most fastest-growing research areas in the field of computer vision. A huge research effort focuses on the image analysis and automatic extraction of image descriptions at semantic level. The ultimate goal is to bridge the semantic gap between the low-level visual features and the high-level concepts capturing the conveyed meaning.

One promising approach to enhance visual image retrieval has been to supplement image content with textual information associated with the image. Text and image are two distinct types of information from different modalities as they represent the ‘thing’ in a quite different way. In the image retrieval research area, they can be used to enhance image retrieval by supplementing image content with text captions. Therefore, we try to exploit those image captions to create the knowledge base for semantic representing for images. As a result, the retrieval performance of the image retrieval system (IMR) could be improved.

The rest of this paper is organized as follows: Section II analyses state of the art frameworks; Section III describes the proposed framework infrastructure; Section IV describes the implementation, experimental results and discussion; and Section V concludes the strengths, weaknesses, and significance of the presented approach.

II. STATE OF THE ART FRAMEWORKS ANALYSIS

The use of textual information in order to improve image search results has been suggested by several researchers. Schreiber et al. [1] focused on establishing a photo-based ontology system. However, the author generated metadata for photographs manually which is not scalable in the real practice. The Image thesaurus [2] proposed a data-driven approach for image retrieval that uses Web images and their surrounding textual annotations as the source of training data to bridge the cognitive gap. However, it seems that natural language queries are not supported, that means, in the case of a purely text query no semantic search happens; only a traditional full-text search is executed. The novel methods for automatic classifying images based on knowledge discovered from annotated images in the form of media network have been presented in MediaNet [3]. Nevertheless, it does not support for the ontology imperfection.

Due to the fact that it is difficult to build a perfect ontology covering the whole domain in one step, this is so called the ‘ontology imperfection’ problem. Thus, the IMR should tolerate the imperfection of ontologies and semantic metadata. In additional, the retrieval algorithms often fail from a search operation and obtain irrelevant results because the metadata is incomplete. The main reason is the system could not generate complete metadata to describe image content. Thus the ontology-based IMR system should be able to handle the uncertainty of metadata efficiently. In addition, the IMR should fulfill the missing information (information which is not addressed in text captions) automatically. In other words, the IMR should find the implicit concepts of photographs automatically base upon associated textual information. From the analysis of the existing solutions, we can identify the main features and missing functions as shown in Table I.

<table>
<thead>
<tr>
<th>Surveyed system</th>
<th>Automatic metadata generation</th>
<th>Uncertainty metadata handling</th>
<th>Knowledge-based search</th>
<th>Ontologies imperfection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schreiber [1]</td>
<td>✗</td>
<td>✗</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Image thesaurus [2]</td>
<td>✓</td>
<td>✗</td>
<td>✗</td>
<td>✗</td>
</tr>
<tr>
<td>MediaNet [3]</td>
<td>✓</td>
<td>✗</td>
<td>✓</td>
<td>✗</td>
</tr>
</tbody>
</table>

From the Table I, two main features, uncertainty metadata handling and ontology imperfection handling, are...
III. PROPOSED FRAMEWORK

In this section, the novel methods for automatic metadata generation, uncertainty metadata handling, knowledge-based search for images, and ontology imperfection handling are described in subsequent sections respectively. Fig. 1 illustrates high-level architecture of the proposed framework.

A. Automatic metadata generation from text captions

To define semantic annotations for sport photographs, two main ontologies are defined: Domain and Photo Annotation ontology. The proposed ontologies are inspired by ontologies in [1]. Domain ontology provides the vocabulary and background knowledge describing the content of the photo. It is subclassed to the various types of domain concepts that need to describe a sport photograph e.g., Athlete, Sport, Event, and Place. The Photo Annotation Ontology specifies an annotation’s structure, independent of the particular subject matter domain e.g., What does the photo depict? How, when, and where was the photo made? This ontology provides the description template for annotation construction.

In order to acquire knowledge from text captions, HTML documents will be parsed in order to extract text captions. Later, text captions which are described in natural language will be processed to detect the important name entities using NLP. We deploy the established NLP framework named ESpotter\(^1\). ESpotter provides a function for Named Entity Recognition and generates annotated documents in XML files. These annotated documents will then be extracted to form initial metadata and will be subsequently stored in a relational database. However, some metadata cannot match with any particular ontology entities but it might still be important to identify the image. To handle this uncertainty, the system does not discard these non-ontological named entities. They are mapped to the ‘otherDetails’ attributes of ontology.

B. Uncertainty metadata handling

The knowledge discovery process performs to find implicit relationship among ontology entities. To do this, semantic rules are applied to this task. However, some ontology entities might miss the required information because it is not available in text captions. The semantic rules also try to fulfill the missing information by interlinking to previous relevant semantic metadata and calculating based on the semantic rules. The example of a simple semantic rule is shown in following:

\[
\text{Add } x \text{ to } M \text{ (metadata) if all of the following conditions hold:}
\]

- \( \forall x \exists y \mid \text{Photo} (x) \land \text{PhotoDate} (x) \land \text{happensDuring} (x, y) \Rightarrow \text{Event} (y) \)

Where \(M\) denotes the current photo metadata; \(Event(x)\) denotes that \(x\) is instance of the concept \(Event\);

**Meaning:** add \(Event(x)\) to metadata of a given photo if a photo contains ‘PhotoDate’ that occurred during the given event, and then this event is considered relevant for a photo. In summary, the knowledge discovery step is able to fulfill the metadata uncertainty handling requirement. This is because the framework tries to find other relevant concepts of images and generate metadata for those relevant concepts automatically. Consequently, the system can retrieve images even a specific concept in a query is not mentioned directly in the text captions. Please note that in this paper we apply our framework to the Olympics Games domain only. This means we do not take other sport events into account at the current of implementation.

C. Knowledge-based search

Ontology-based information retrieval can be seen as an evolution of classic keyword-based retrieval techniques, where the keyword-based index is replaced by a semantic knowledge base. To query information from a semantic model, SPARQL\(^2\) is deployed in our framework. SPARQL query is executed against the knowledge base, which returns a list of instance triples that satisfy the query. When user input the query into the system, the query will be transform to SPARQL query format automatically. For example, “Find all photographs about field event”. Field event refers to all sports that athletes usually perform in field e.g., hammer throw. This user’s query would be formalized as:

\[
\text{SELECT ?photograph WHERE \{ ?photo sport:Sport ?subType.}
\]

---

\(^1\) http://kmi.open.ac.uk/people/jianhan/ESpotter

\(^2\) http://www.w3.org/TR/rdf-sparql-query


Finally, the results are ranked by using the cosine similarity technique to compute the similarity between query and image. Equation (2) shows cosine similarity formula. Let \( \{P_i\}_{i=1}^{n} \) be the set of all photos in the collection, \( d_i \) is the document vector of \( P_i \), query \( Q \) and photo \( P_j \) have \( t \) terms and their associated weights are \( QW_{ik} \) and \( PW_{ijk} \) respectively, for \( k=1 \) to \( t \). Similarity between the query and document (photograph) is measured using the following inner product:

\[
\text{sim}(Q, P_j) = \frac{\sum_{k=1}^{t} QW_{ik} \times PW_{ijk}}{\sqrt{\sum_{k=1}^{t}(QW_{ik})^2 \times \sum_{k=1}^{m}(PW_{ijk})^2}}
\]

(2)

D. Ontologies imperfection handling

As mentioned previously, ontologies are usually incomplete and this fails IMR from finding relevant images. To solve this problem, we need second indexing algorithm working in parallel with ontology indexing. Thus, LSI is selected because its performance is better than traditional text-based indexing and can find implicit relationship between keyterms and images. Text captions are analysed and stop words are eliminated. The remaining keywords are used to form a term-image matrix. Each term will be assigned weight to show the importance of that term to the image. Term weight (\( TW \)) is a product of local weight, global weight, and normalization factor. Equation (1) shows the term weight formula [4].

\[
TW = \sqrt{f_j-0.5} \times \left(\frac{N}{n_i}\right)^{1/2} \times \left(\frac{1}{\sqrt{\sum_{i=1}^{m} \left(\frac{N}{n_i}\right)^{1/2} \times (f_j-0.5)+1}}\right)
\]

(1)

where \( f_j \) is the frequency of term \( i \) in document \( j \), \( N \) is the number of documents in the collection and \( n_i \) is the number of documents in which term \( i \) appears. The normalization factor compensates for discrepancies in the lengths of the documents and has to be done after the local and global weighting. Fig. 2 shows an example of LSI matrix with term weight (\( TW \)).

<table>
<thead>
<tr>
<th>Keywords</th>
<th>img1</th>
<th>img2</th>
<th>img3</th>
<th>img4</th>
<th>img5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Olympics</td>
<td>0.57735</td>
<td>0.707107</td>
<td>0.707107</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Tennis</td>
<td>0.57735</td>
<td>0</td>
<td>0.707107</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>AGASSI</td>
<td>0.57735</td>
<td>0</td>
<td>0</td>
<td>0.57735</td>
<td>0.57735</td>
</tr>
<tr>
<td>USA</td>
<td>0</td>
<td>0.57735</td>
<td>0.57735</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Celebrates</td>
<td>0</td>
<td>0.57735</td>
<td>0.57735</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Figure 2. LSI matrix after assigning weight to each term

Term weight value is used to determine the degree of important of that term to the image. Term weight in each document is represented in the matrix format which row heading is keyword and column heading is image (image ID). When the IMR fails to find relevant images using the semantic model, LSI is activated to compute the similarity between query and indexing term. The results of LSI will be used instead of the knowledge-based search results. In other words, the performance of the framework will degrade nicely when ontology is imperfect by compensating from LSI results. This technique can fulfill the ontology imperfection problem.

IV. EMPIRICAL RESULTS AND EVALUATION

For the purpose of experimental evaluation, a collection of sport photographs from the Olympic organization was assembled. Since we only deal with associated textual information to build the knowledge base in the current implementation step, our framework (knowledge-based search) was tested by some sample queries, and compared to the Lucene (keyword-based search), full-featured text search engine.

To evaluate the retrieval performance of the proposed solution, some hypotheses were established against the missing features of the IMR addressed in section II. The main hypotheses are:

**Hypothesis 1 (H1):** The missing metadata of ontologies should be generated automatically even it is not supplied in the text captions. This leading to a presented framework is able to find the implicit concepts of images and improve the retrieval performance. This hypothesis aims to evaluate the uncertainty metadata handling.

**Hypothesis 2 (H2):** The ontology-based search provides good results even metadata is incomplete. Good IMR system should tolerate the incompleteness problem by not return blank results to user even it fails from search operation.

To validate these hypotheses, two sample queries were selected with their specific purposes. The experimental results have been reported by the 11-point interpolated average precision graph (Fig. 5a and b).

Query 1 (Q1): “Find all photographs using keywords which are not explicitly appeared in text captions” e.g., ‘Aquatics’. This query aims to test the H1 hypothesis. Aquatics refers to all kinds of sports that athletes usually perform in water e.g., swimming and driving but it is not usually presented in the image captions. As shown in Fig. 5a, the knowledge-based search outperforms the keyword-based search because semantic rules expand initial metadata to the relevant concepts automatically. This is because semantic rule links swimming concept to field event concept. Therefore, the proposed system is able to recognize images which text captions addressed all kinds of Aquatics sports even the word ‘Aquatics’ is not presented in the text captions. This leads to the knowledge-based search obtaining better precision and recall than text-based approach. Then H1 is verified.

Query 2 (Q2): “Find all photos which the ontology does not contain the information about those photos”. To do this, I delete some information of images in ontology resulting in incompleteness of ontology e.g., swimming photographs in the Sydney. This query aims to test H2 hypothesis. Since information of swimming photographs in Sydney Games is deleted, the knowledge-based search then fails from finding

---

3 http://www.olympic.org
4 http://lucene.apache.org

relevant images in a semantic model. Then, the presented system will turn to turn to use LSI results.

![Figure 2. Experimental results comparison between two approaches](image)

**TABLE II**

<table>
<thead>
<tr>
<th>Types of Queries</th>
<th>Queries</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KB-search</td>
<td>KW-search</td>
<td>KB-search</td>
</tr>
<tr>
<td>Simple</td>
<td>Q3</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>Q4</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>Q5</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td>Complex</td>
<td>Q6</td>
<td>89%</td>
<td>56%</td>
</tr>
<tr>
<td></td>
<td>Q7</td>
<td>100%</td>
<td>75%</td>
</tr>
<tr>
<td></td>
<td>Q8</td>
<td>100%</td>
<td>8%</td>
</tr>
<tr>
<td></td>
<td>Q9</td>
<td>63%</td>
<td>100%</td>
</tr>
<tr>
<td></td>
<td>Q10</td>
<td>100%</td>
<td>63%</td>
</tr>
<tr>
<td></td>
<td>Average</td>
<td>94%</td>
<td>75%</td>
</tr>
</tbody>
</table>

Fig. 5b shows that a performance of the two algorithms is not significantly different. However, a presented framework obtains slightly better performance compared to the keyword-based search because LSI can perform semantic search resulting in it can find implicit relationships between keyterms and images. This clearly shows that the proposed system is able to tolerate the imperfection of ontologies and semantic metadata. To summarize the discussion, a proposed framework still provides a good result even the knowledge base is incomplete and H2 is clearly evaluated.

More sample queries were used for testing a proposed framework. The sample queries are divided into two categories. The first category is related to simple query formulation. The simple query contains keywords which directly appear in the text captions. In the second category, the query keywords do not appear explicitly in the image captions. Therefore, the searching system needs to find the relevant concepts from semantic relations among ontology instances. Corresponding numerical values are reported in Table II. Overall, the approaches of both searches obtain similar performance in simple category (Q3-Q5) but the knowledge-based search outperforms both precision and recall for the keyword-based search when the query is more complex. This is because the knowledge-based search could recognize the concepts of a photograph which have semantic relations with query keywords. However, the keyword-based approach obtains better recall in Q9, “Find all photographs about opening ceremony”. This is because it retrieves all photos containing “opening” or “ceremony” keywords. Thus, all relevant photos about opening ceremony are retrieved in addition to some other irrelevant photos e.g., photos about medal ceremony or closing ceremony. Therefore, it obtains higher recall but lower precision than the knowledge-based search. Based on the average from all queries, 94% of all relevant documents in a collection are recognized by the knowledge-based technique whereas the keyword-based system retrieves only 75% from all relevant documents in a repository.

V. CONCLUSIONS AND FURTHER WORK

In this paper we described the framework which is able to represent the semantic of images effectively by using text captions. The main innovation is the employed knowledge infrastructure uses ontologies for the extracted information from text captions and for expanding this information to relevant concepts in domain ontologies leading to narrow the semantic gap. The empirical results shown the knowledge-based technique improves the retrieval performance significantly compared to the keyword-based search technique and solve addressed problems in the existing IMR. In conclusion, the major hypotheses of this work were clearly validated by the experimental results.

In the future, we would like to exploit low-level features of images in order to deal with uncertainty problem e.g., when text captions are not supplied for images. These issues will be addressed in our future work.

REFERENCES


Retrieval, Computer Science and Mathematics Division, Oak Ridge National Laboratory, March, 1999.