Multi-level Semantic Characterization and Refinement for Web Image Search

Yuanxi Li\textsuperscript{a}, C.H.C. Leung\textsuperscript{b}

Department of Computer Science Hong Kong Baptist University Hong Kong
\textsuperscript{a}yxli@comp.hkbu.edu.hk, \textsuperscript{b}element@comp.hkbu.edu.hk

Abstract

With the increasing number of Web images and social photograph sharing sites, effective search of real-world images becomes a formidable challenge and an important necessity. Different indexing and annotation algorithms are required for different types of Web images. To meet the challenge and provide satisfactory search results to users, we present a multi-level method with four levels differentially applied to different types of Web images. The performance of our method is evaluated by experiments on thousands of Web images and tags in different subsets, and our approach is able to yield highly promising results compared with applying a single method to all types of Web images.

© 2011 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
Selection and/or peer-review under responsibility of the Intelligent Information Technology Application Research Association.

Keywords: Image retrieval; scene analysis; automatic annotation; CYC inference; MPEG-7

1. Introduction and related work

With the number of different types of Web images uploaded at rapid rates, effective approaches for searching diverse images are becoming critical.

Web-based images search differ from non Web-based search in several important ways. These include (i) The collection is potentially infinite so that some standard measures such as recall are not directly applicable; (ii) The lack of uniformity in metadata set, with some images containing a significant amount, while others may have none, and those providing metadata may also show significant variations in the nature of such data; (iii) The indexing and annotation algorithm must be efficient, since cumbersome or slow algorithm may not be able to adequately keep up with the rate of increase in the size of collection.

For the indexing and retrieval system, the previous research studies proposed many different approaches and framework [1, 2], which mainly fall into two categories. Concept-based method is approach that images are retrieved by high-level concept and objects [3]. While with content-based method, images are retrieved according to low-level content by extracting low level features and capturing image similarities and characteristics [4, 5, 6]. However, the existing works could only give promising result
within some specific image sets, such as the images with annotation, but not all image types. For variety
different types of images, such as raw images without any tags or captions, and images with full MPEG-7
annotation, in order to get high retrieval accuracy, different levels of image characterization schemes are
needed.

In this paper, in order to raise and measure the semantic power of searching and discovering of web
images, we propose a multi-level semantic characterization and refinement method which contains
different retrieval algorithms differentially applied to different types of Web images. With this method, a
higher level of semantic richness is endowed while higher accuracy of retrieval is attained.

2. Four level semantic characterisation

Web images could be mainly categorized into the following four different types, as shown in Table 1.

<table>
<thead>
<tr>
<th>Image Type</th>
<th>Caption</th>
<th>Annotation, Tag, Keyword</th>
<th>Full MPEG-7 Annotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>unavailable</td>
<td>unavailable</td>
<td>unavailable</td>
</tr>
<tr>
<td>2</td>
<td>available</td>
<td>unavailable</td>
<td>unavailable</td>
</tr>
<tr>
<td>3</td>
<td>available</td>
<td>available</td>
<td>unavailable</td>
</tr>
<tr>
<td>4</td>
<td>available</td>
<td>available</td>
<td>available</td>
</tr>
</tbody>
</table>

According to the information provided by the images, different characterization schemes are needed to
deal with different types of images. Our four levels of semantic characterization model are designed as
indicated in Figure 1. The corresponding tasks are: (i) automatic semantic class refinement by rule
induction and concept validation enhancement; (ii) object association and CYC enabled inference; (iii)
structural casting to MPEG-7 representation and index building.
3. Automatic Semantic Classification refinement by metadata and concept validation enhancement

For images uploaded with no explicit annotation, caption nor tag, this task will carry out a classification of its semantic content. In the case of JPEG images, metadata such as aperture, exposure time, shutter speed, resolution, date and time, or in some image capture devices, the GPS coordinates, are automatically embedded. With a given image

\[ I_i = (d_{i1}, ..., d_{id}) \]  

characterized by a number of dimensions \( d_i \) which correspond to the image acquisition parameters, each image may be represented by a point in multi-dimensional space. Figure 3 shows the image points of a set of images in which particular types of image scenes (categorized as \( S_i \)’s) tend to naturally cluster together. Our earlier paper [7] has developed a rule-base induction method to separate images into elementary semantic categories \( c_{mn} \). The induction rule is in the form of

\[ R_{mn} : U_1 \land U_2 \land ... \land U_k \rightarrow c_{mn} \]  

where the antecedent will consist of a range of dimensional values of the form \( U_i = \{ u_i | u_i \in D_i \} \), with \( u_i \) representing a particular parametric dimension, and \( D_i \) representing the associated restricted domain of values. This will produce a prior probability \( P[c_{mn}] \) for the particular semantic content category. Such categorization will provide a large scale pruning of the search tree whereby highly selective procedures may be applied differentially to different refined categories. Figure 2 categorizes different scenes of images, obtained from Flickr.

<table>
<thead>
<tr>
<th>Micro Scenes</th>
<th>Night Activities</th>
<th>Indoor Activities</th>
<th>Day Scenes</th>
<th>Portraits</th>
<th>Outdoor Activities</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1" alt="Image" /></td>
<td><img src="image2" alt="Image" /></td>
<td><img src="image3" alt="Image" /></td>
<td><img src="image4" alt="Image" /></td>
<td><img src="image5" alt="Image" /></td>
<td><img src="image6" alt="Image" /></td>
</tr>
<tr>
<td><img src="image7" alt="Image" /></td>
<td><img src="image8" alt="Image" /></td>
<td><img src="image9" alt="Image" /></td>
<td><img src="image10" alt="Image" /></td>
<td><img src="image11" alt="Image" /></td>
<td><img src="image12" alt="Image" /></td>
</tr>
<tr>
<td><img src="image13" alt="Image" /></td>
<td><img src="image14" alt="Image" /></td>
<td><img src="image15" alt="Image" /></td>
<td><img src="image16" alt="Image" /></td>
<td><img src="image17" alt="Image" /></td>
<td><img src="image18" alt="Image" /></td>
</tr>
<tr>
<td><img src="image19" alt="Image" /></td>
<td><img src="image20" alt="Image" /></td>
<td><img src="image21" alt="Image" /></td>
<td><img src="image22" alt="Image" /></td>
<td><img src="image23" alt="Image" /></td>
<td><img src="image24" alt="Image" /></td>
</tr>
</tbody>
</table>

Figure 2 Each column shows the different categories of “Micro”, “Night Scenes”, “Indoor Activities”, “Day Scenes”, “Portraits” and “Outdoor Activities”
Figure 3 Image distribution in 3D space of some Flickr images

A sample rule set of Automatic Semantic Annotation is shown in Figure 4.

- \( \text{Exposure Value} \leq 8 \text{ AND Aperture} < 3.0 \text{ AND } 1/160 < \text{Exposure time} < 1/40 \text{ AND Focal length} < 6.0 \rightarrow \text{Micro} \)
- \( \text{Timeslot} \in \text{Night} \text{ AND } 3.5 < \text{Aperture} < 5.6 \text{ AND } 1/50 < \text{Exposure time} < 1/2 \text{ AND 18 < Focal length < 88} \rightarrow \text{Night Scenes} \)
- \( 3.5 < \text{Aperture} < 5.6 \text{ AND } 1/50 < \text{Exposure time} < 1/6 \text{ AND 18 < Focal length < 33} \rightarrow \text{Indoor Activities} \)
- \( \text{Subject Distance} > 30 \text{ AND } 3.5 < \text{Aperture} < 4.5 \text{ AND } 1/4000 < \text{Exposure time} < 1/1600 \text{ AND 18 < Focal length < 39} \rightarrow \text{Day Scenes} \)
- \( 1.4 < \text{Aperture} < 5 \text{ AND } 1/4000 < \text{Exposure time} < 1/500 \text{ AND 20 < Focal length < 50} \rightarrow \text{Portraits} \)
- \( 4 < \text{Aperture} < 10 \text{ AND } 1/2000 < \text{Exposure time} < 1/80 \text{ AND 18 < Focal length < 62} \rightarrow \text{Outdoor Activities} \)

Figure 4 Sample rules of Automatic Semantic Annotation

4. Object association and cyc enabled inference

The existing image retrieval techniques seldom consider the context of the keywords present in the user's queries. When the keywords are typed into the searching engine, some information may be lost. Therefore, the refinement and expansion of keywords should be considered to assist semantic searching. The concepts in the user queries and the relationships among concepts should be understood in semantic based information retrieval techniques.

From the sub-categories derived, additional semantics will be incorporated and enriched to support higher precision retrieval. Apart from images without explicit annotation, there are those with some basic caption. Such primitive level of information may be exploited to carry out inferential reasoning based on...
domain content. It has been found in our study in [1, 7] that captions may sometimes harm annotation correctness, and QBE techniques will be additionally deployed to attempt to filter out the misleading captions and provided keywords. Common sense and domain knowledge from the extensive CYC knowledge base [8, 9] will be used for such enrichment; in particularly, we make use of the specific CYC collections and individuals (#Sisa and #Sgenls) and CYC Microtheories to carry out the expansion. Unlike the study [10], the correlation of concepts will not only be limited to hypernymy/hyponymy and holonomy/meronomy relationships, but also includes related concepts and common sense non-hierarchical associations. In addition, while [10, 11] is using WordNet, here the application of CYC commonsense knowledge is expected to be far richer, and more comprehensive than WordNet. Using CYC, certain objects in an image may be linked to related objects. Such inferences will entail examination of the conditional probabilities $P(J_i | J_j)$, where $J_i$, $J_j$ are objects and $J_j$ is given to be present in an image. Common sense association and ontology in CYC are used to construct an inference tree, which allows the index elements $X_i$’s of an image to be automatically expanded according to the probability linked to the underlying ontology of the domain, see Figure 5.

![Figure 5 CYC assisted concept expansion tree](image)

Figure 5 CYC assisted concept expansion tree

An example of demonstrating concept expansion is shown as follows in Figure 6 [12].

![Figure 6 Example of concept expansion](image)

Figure 6 Example of concept expansion
Table 2 illustrates the experimental results of query expansion using the tagged Flickr images. The second column gives the precision of the query “Springfield”; the third column gives the precision of the expanded query “Springfield+Missouri”; the third column gives the precision of the further expanded query “Springfield+Missouri+USA”.

Table 2 example of query concept expansion

<table>
<thead>
<tr>
<th>Plain query words</th>
<th>+1 holonymy</th>
<th>+2 holonymy</th>
<th>…</th>
<th>+k holonymy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Springfield</td>
<td>Missouri</td>
<td>USA</td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>Precision</td>
<td>5.6%</td>
<td>61.1%</td>
<td>22.2%</td>
<td>-</td>
</tr>
</tbody>
</table>

5. Multiple characterisation and structural casting into MPEG-7

As demonstrated in our MPEG-7 Core Experiment [13], the indiscriminate use of keyword will result in an inordinate number of returns and weaken the retrieval precision. The object, attribute, relationship in accordance with the Ternary Fact Model from the previous stages will be analyzed and the first four elements in the MPEG-7 Structured Annotation Description Scheme will be filled. In the case of images without any captions or tags, the second to fourth elements in the MPEG-7 Structured Annotation Description Scheme will be filled, which gives definite semantic properties of an image. MPEG-7 represents the final most semantically rich level of characterization from which the search index is constructed for image identification.

6. Controlled experiments

To measure the effectiveness of the approach, controlled experiments are performed. 1,000+ Images are collected using an unbiased randomized mechanism from Flickr.com to form the basis of the experimentation. These represent a cross-section of the different types of Web images, and they consist of three main subsets:

- images where text information is completely absent – subset 1
- images with basic caption – subset 2
- images annotated with keywords and tags – subset 3

Measures of performance are taken between the unaided approach similar to that in search engines and the present approach for each individual subset as well as collectively for their union. A set of representative semantic queries is designed which is used for all the experiments. The following are used to measure system performance of both approaches on the same image collection:

- precision
  \[
  \text{Precision} = \frac{|\{\text{relevant \_ images}\} \cap \{\text{retrieved \_ images}\}|}{|\{\text{retrieved \_ images}\}|} \tag{3}
  \]

- recall
  \[
  \text{Recall} = \frac{|\{\text{relevant \_ images}\} \cap \{\text{retrieved \_ images}\}|}{|\{\text{relevant \_ images}\}|} \tag{4}
  \]

- average precision
- fallout
  \[
  \text{Fallout} = \frac{|\{\text{non\_relevant \_ images}\} \cap \{\text{retrieved \_ images}\}|}{|\{\text{non\_relevant \_ images}\}|} \tag{5}
  \]
- F\alpha-score for \( \alpha \ll 1 \)
Recall is not included as a direct measure, since for potentially infinite collections, the total number of relevant images cannot be directly determined. However, the $F_{\alpha}$-score gives some indication of recall, which may be ascertained for the finite collections in these experiments. Substantially higher weight is assigned to precision ($\alpha=0.01$), which is much more important for Internet search.

The experimental results are shown in Table 3. We see that the precision, $F$-score and average precision are all satisfactorily higher compared with the unaided approach, while the fallout or false alarm (which is the proportion of non-relevant images retrieved) is kept reasonably low. These results indicate that significant improvement in performance may be attained from using the proposed approach.

Table 3 Experimental result

<table>
<thead>
<tr>
<th>Subset</th>
<th>Unaided approach</th>
<th>Proposed approach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>57.09%</td>
</tr>
<tr>
<td></td>
<td>Fallout</td>
<td>38.60%</td>
</tr>
<tr>
<td></td>
<td>F$_{\alpha}$</td>
<td>58.20%</td>
</tr>
<tr>
<td>Subset 1</td>
<td>Precision</td>
<td>66.20%</td>
</tr>
<tr>
<td></td>
<td>Fallout</td>
<td>35.00%</td>
</tr>
<tr>
<td></td>
<td>F$_{\alpha}$</td>
<td>66.20%</td>
</tr>
<tr>
<td>Subset 2</td>
<td>Precision</td>
<td>82.29%</td>
</tr>
<tr>
<td></td>
<td>Fallout</td>
<td>16.74%</td>
</tr>
<tr>
<td></td>
<td>F$_{\alpha}$</td>
<td>79.92%</td>
</tr>
<tr>
<td>Subset 3</td>
<td>Precision</td>
<td>70.26%</td>
</tr>
<tr>
<td></td>
<td>Fallout</td>
<td>21.13%</td>
</tr>
<tr>
<td></td>
<td>F$_{\alpha}$</td>
<td>72.68%</td>
</tr>
<tr>
<td>All Subsets</td>
<td>Average Precision</td>
<td>51.02%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>74.76%</td>
</tr>
</tbody>
</table>

7. Conclusion

As shown from the results of our experiments, with our proposed approach of the four levels of semantic characterization, the accuracy of Web image searching has seen significant improvements. By the systematic analysis of embedded image metadata and parametric dimensions, the query refinement with object association and CYC enabled inference, as well as the image metadata enhancement with MPEG-7 Structured Annotation Description Scheme, different types of Web images could be retrieved with higher precision. The semantic meanings of Web images, including raw images, images with captions or tags, and the images with MPEG-7 full annotation, are enriched and standardized. Additional refinement is no doubt possible and desirable in future to further increase user satisfaction. Our proposed approach is an important first step towards this.

References


