Biomedical Figure Search using Combination of Bag of Keypoints and Bag of Words

Asako Koike\textsuperscript{1,2} \quad Natsu Ishii\textsuperscript{2}\\ akoike@hgc.jp \quad natsui@cb.k.u-tokyo.ac.jp \\
Toshihisa Takagi\textsuperscript{2} \quad tt@k.u-tokyo.ac.jp

\textsuperscript{1} Cent. Res. Lab., Hitachi Ltd., 1-280 Higashi-koigakubo, Kokubunji-shi, 185-8601, Japan
\textsuperscript{2} Dept. of Comp. Biol. Grad. Sch. of Front. Sci., Univ. of Tokyo, 5-1-5 Kashiwanoha, 277-8561, Japan

Keywords: image search, text mining, full papers

1 Introduction

In recent years, many of scientific journals started to open their full papers for free after a certain period of publication through PubMed central or on their web site. Although there remains copyright problems in some journals, text mining targeting full papers are getting active. Since especially figures contain condensed information not explicitly described in text, the figure retrieval and information extraction from figures are crucial tasks. Under these situations, various researches for figure classifications and figure retrievals have been reported. Mulfy \textit{et al.} constructed a subcellular location classification system based on figure legends and image descriptors [1]. Koike and Takagi classified figures into 22 categories using figure features and figure legends [2]. Xu \textit{et al.} constructed Yale Image Finder to search figures using text information [3]. In this study, we have developed a full-text retrieval system which enabled us to treat figure information and text information in the same way utilizing bag of keypoints method based on the vector quantization of features (descriptors) for detected interest points in figures. In this system, cross-search between text and figure information and figure search with specifying both text context and figure content are available.

2 System and Method

2.1. Graphical feature generation from figures

Various kinds of graphical features have been studied in the field of computer vision to search similar figures. In this study, bag of keypoints are used. In this method, 1) interest points are detected in the first step, 2) features of the detected interest points are calculated in the second step, 3) the calculated features are quantized (features of interest points are converted into pseudo words), and 4) figures are represented by a bag of these quantized features (bag of keypoints) so that figures are searchable similarly with bag of key words in text retrieval. In this study, the Harris-affine detector, Hessian-affine detector, and Maximum Stable Extremal Regions (MSER) detector are used as interest point detection methods. The Harris-affine detector is based on multiple-scale interest point detections of Harris corner measures on the second moment matrix, while the Hessian-affine detector uses multiple-scale interest point detections on the Hessian matrix. As feature description methods of these detected interest points, scale-invariant feature transform (SIFT), GLOH (extended SIFT), and shape content (SC) are used. The quantization of features of interest points was done by k-means clustering. First k-means clustering was done using Dataset type A, and these cluster centers were used for the assignment of each feature of interest points of Data type B into clusters.

2.2. Development of Text and Figure Search System

For text and figure search, Apache Lucene library is used. Each figure is treated as a pseudo document consisting of pseudo words calculated based on k-means clustering. All full paper text files are separated into “title”, “abstract”, “introduction”, “method”, “results & discussion”, and “references” by analyzing html or pdf structures. These are indexed in different fields to specify a search field. Each field is searched...
dependently and when multiple fields or figure descriptions are searched simultaneously, the sum of tf*idf score in each field is used.

2.3. Test data set
We used two kinds of data set test data set A and B. Test data set A: Figures were extracted from “Proceedings of the National Academy of Sciences (PNAS)” and “Radiology” and were classified into 22 classes by a biologist [2]. In total, 6642 figures from 1305 full PNAS papers and 318 figures from 35 full Radiology papers were used as a data set. Figure legends and main text were also extracted from their html format data. Test data set B: Similar figures were manually extracted from “PNAS”, “Radiology”, and “American Journal of Neuroradiology (AJNR)” 2000-2008 years. In the test data set B, similar figures are experimental results of the same kind of modality or measurement equipment, and have common topics such as cellular function and disease. Totally 122 similar figure pairs were provided.

3 Results and Discussions

The similar figure extraction accuracies of each figure extraction method with/without text information are summarized in Table 1. The accuracy is the percentage of figures whose retrieved classes agree with query classes. Queries and the retrieval figures are the test data set A. Overall Harris-affine shows the best performance as an interest point detection method and this is expected to be dependent on the detected number of interest points. Since the accuracies with figure legend and figure legend+main text describing the corresponding figure are 50.6% and 52.4%, respectively, the figure feature information and text information compensate each other in some methods. For example, in the case of “chemical compound”, features of figures are useful, since in the main text or legend, chemical structure is omitted as structure and can not be distinguished from protein structure. Concerning “fluorescent microscope”, trivial term variations such as “immunofluorescent microscopy” and “confocal microscopy” lead to confusion with other microscopy data, while the features of figures can distinguish them. By contraries, in the case of “MRI” and “CT”, they are difficult to be distinguished by features of figures but are easily distinguished by text information.

In the data set B, addition to figure legends and main texts referring the corresponding figures, abstracts were also tested. The best combination of data (i.e., which kinds of information are effective for figure retrieval) depended on the query figure type. In the case of disease related figures, the disease full spells are sometimes not appeared in the main text referring the words and the figure legends. By addition of the abstracts containing full spells, term ambiguities are resolved. On the other hand, concerning cellular function related figures, addition of abstracts causes the decrease of retrieval performance. This is probably because cellular functions related figures are partial results of abstracts and are not always major results. The abstract contents are less related to the cellular function figures than disease related figures. To increase the retrieval performance, the effects of relevance feedback and usage of dictionaries are also discussed.

Table 1 The accuracies of top 10 search results in each feature extraction method with/without text information*.

<table>
<thead>
<tr>
<th>Feature description method; Interest point detection method;</th>
<th>SC</th>
<th>SIFT</th>
<th>GLOH</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harris-affine</td>
<td>57.4/63.2%**</td>
<td>57.2/61.9%</td>
<td>55.7/60.8%</td>
</tr>
<tr>
<td>Hessian-affine</td>
<td>56.5/60.0%</td>
<td>49.0/51.9%</td>
<td>49.1/52.0%</td>
</tr>
<tr>
<td>Maximum Stable Extremal Regions</td>
<td>51.2/56.5%</td>
<td>42.0/46.0%</td>
<td>34.7/38.1%</td>
</tr>
</tbody>
</table>

* Subclasses are ignored and k-means clustering center is set to 1000. Text information contains figure legend+main text describing the corresponding figure. **Accuracies without and with text information. The random value is 19.3%.

References