Patent Image Retrieval

Stefanos Vrochidis
IRF Symposium 2008
Vienna, November 6, 2008

Aristotle University of Thessaloniki
Overview

1. Introduction
2. Related Work in Patent Image Retrieval
4. Figure Extraction
5. Binary Content Based Image Retrieval
6. Exploitation of Linguistic Data
7. Image Retrieval Results
8. Performance and Evaluation
9. Future Work and Research
10. Conclusions
Introduction

- Nowadays vast numbers of patent documents are submitted to patent offices worldwide.
- Figures, drawings, and diagrams are contained in patents as a means to further specify the objects to be patented.
- The patent expert can get a clear idea of what is claimed just by looking at the images, and subsequently take the decision to read further or discard this patent in the search process.

- For this reason, a patent image retrieval system would be of great help to the patent experts.
- The retrieval functionalities of such a system should extend beyond figure browsing and metadata-based retrieval to include content-based search according to the paradigm of retrieval by example.
Related Work in Patent Image Retrieval

- **PATSEEK**
  - Image based retrieval system for the US patent database that can complement a text based search system.
  - Consists of two subsystems, one for the creation of feature vector and image database and another one for the retrieval of images similar to the query image.

- **LTU Technologies**
  - Image analysis system that indexes, recognizes, and describes images according to their visual content.
  - LTU applications are not implemented specifically for patents, so a more general approach is followed.

- **WISPER project**
  - Claims to cover image-based patent search.
  - The whole project remains under strict confidentiality.
Patent Image Module Framework

- Integrated solution
- Proposed framework
  - Preprocessing (figure extraction)
  - Image Retrieval
- 3 Layers
  - User
  - Query
  - Content
Figures Extraction

- A prerequisite process for other tasks within the image retrieval framework is the extraction of patent images from the raw patent documents.
- The figure extraction step of the framework comprises the following specific tasks:
  - Drawing page selection:
    - The drawing page selection from the raw PDF documents is based on section information encoded within the document files.
  - Page orientation detection.
  - Page segmentation to single figures.
Page Segmentation to Single Figures

- Frequently, a patent drawing page contains more than one figure.
- An image segmentation step that splits a multi-figure page to single images has to be applied.
  - Each separate figure in patents is accompanied by a label of the form “Figure x”.
  - OCR for Figure Label Detection.

- This task is particularly challenging in certain cases:
  - a single figure consists of spatially disjoint elements or when multiple figures are adjacent.
  - Figure Labels are hand written.
Content based Image Retrieval (CBIR)

- **Binary CBIR**
  - A special case of Content Based Image Retrieval
  - Patent images are mostly binary (black and white) images
    - they do not contain any color or texture information
  - Use shape-based feature vector extraction methods
    - describe the image geometric information accurately
  - Patent databases contain a vast amount of patents (and so patent drawings)
    - Feature vector extraction and retrieval techniques need to be computational inexpensive

- **Existing Techniques:**
  - Object-segmentation and weak-segmentation
    - both un-scalable and computational expensive
  - Non-segmentation
    - edge-direction orientated and consequently exposed to noise degradations and diverse creator drawing style.
    - Up to now Edge Orientation Autocorrelogram (EOAC) has proved superiority over other methods and was applied by PATSEEK
Feature Extraction

- Adaptive Hierarchical Geometric Centroids
  - Non segmentation technique.
  - This algorithm was selected because the majority of the figures are binary so the useful information could be extracted by the geometry and the shape of the depicted objects.
  - Low dimension feature vector (~ 100).
  - Black-and-white image is visualized as a distribution of black spots in a white 2-dimensional plane.
  - combines enhanced accuracy, low computational cost and scalability.
Exploitation of Linguistic Analysis

• Use Figure labels to match image with text description.
• Extract associated text.
• Further processing to extract “figure type” information.

Example: Patent EP 0545532 A1 / Figure 5

Figure 5 is a cross-sectional view of a portion of a hub of a removable cartridge and an alternate spindle mechanism for use in a computer system according to the present invention.

Label: 5
Type: Technical Drawing
Main Related Concept: Part
Related Concept: Removable Cartridge
Image Retrieval

- Image Retrieval combining all the available techniques is supported.
- Retrieval is performed in two steps
  - Initial query
    - Browse all figures
    - Visually similar figures to a given example
      - An existing image
      - Upload a new image
  - Further filtering
    - Category
    - Patent
    - Text
Results – Browse Figures
Results - Image similarity
Results – Retrieval based on Category

- Retrieval based on category is supported
- Search for “flowcharts”
Performance and Evaluation

• **Performance**
  - The experiments were conducted on a PC, with a P4 3.0GHz Intel CPU and 1GB RAM.
  - PostgreSQL was used to store the actual non-multimedia content and the links to the multimedia files.
  - The involved dataset included 2000 patent images that have been picked up from EPO.
  - In terms of time response and scalability
    - the vector of AHGC was tested for more than 10000 images (~1000 patents) and the real-life time-response did not exceed 10 sec, while in EOAC, the total time for a query is 90 sec.

• **Evaluation**
  - Precision and Recall graphs were produced.
  - For ground truth annotated images were used.
Performance and Evaluation

- Comparison with PATSEEK

![Graph comparing PATEXPERT vs PATSEEK]
Future Work and Research

• Further improvement of the segmentation algorithm
• Automatic Image Extraction is a complicated task due to handwriting, scanned documents etc.
• Further optimization of the retrieval algorithm.
• Further test scalability of the system by employing larger image database.
• Introduce indexing structures to further improve performance in large databases.
• Content-based classification and clustering of images.
• Map semantic information from associated text onto parts of the image performing segmentation.
• Extract more information with text-graphics separation and OCR.
Conclusions

• The patent image retrieval module is an innovative search engine for patent visual content.
  • The design and implementation of this module is tailored to the special nature of patents.
  • it builds upon advanced techniques from image analysis and content-based retrieval to enhance the performance of patent image retrieval.
  • It is capable of combining the content-based search with the annotation-based search and of providing relevant results.
• It can provide quickly results especially in multilingual patents with no much preprocessing.
• Such an image retrieval system could be an integral part of a more general patent management system but due to its independent architecture it could also serve as a standalone search engine.
• Comparison between the proposed framework and other similar systems (i.e PATSEEK) support the efficiency of the proposed patent image retrieval system.
• This work was supported by PATExpert project (http://www.patexpert.org).
• Demo of the proposed patent image module is publicly available at: http://mklab-services.iti.gr/ patexpert/
Thank you!

Questions / Discussion