World Patent Information 32 (2010) 94–106

Contents lists available at ScienceDirect

World Patent Information

journal homepage: www.elsevier.com/locate/worpatin

Towards content-based patent image retrieval: A framework perspective

Stefanos Vrochidis [a,*], Symeon Papadopoulos [a], Anastasia Moumtzidou [a], Panagiotis Sidiropoulos [a], Emanuelle Pianta [b], Ioannis Kompatsiaris [a]

[a] Informatics and Telematics Institute, Centre for Research and Technology Hellas, Thermi-Thessaloniki, Greece
[b] Fondazione Bruno Kessler, Trento, Italy

ARTICLE INFO

Keywords:
Content-based search
Patent
Retrieval
Images
Drawings
Figures
Search engine
Hybrid

ABSTRACT

In this article, we discuss the potential benefits, the requirements and the challenges involved in patent image retrieval and subsequently, we propose a framework that encompasses advanced image analysis and indexing techniques to address the need for content-based patent image search and retrieval. The proposed framework involves the application of document image pre-processing, image feature and textual metadata extraction in order to support effectively content-based image retrieval in the patent domain. To evaluate the capabilities of our proposal, we implemented a patent image search engine. Results based on a series of interaction modes, comparison with existing systems and a quantitative evaluation of our engine provide evidence that image processing and indexing technologies are currently sufficiently mature to be integrated in real-world patent retrieval applications.

1. Introduction

Nowadays, vast numbers of patent documents are submitted to patent offices worldwide, in order to describe and protect innovative artifacts, processes, algorithms and other inventions. Figures, drawings and diagrams are almost always contained in patents, as a means to further specify the objects and ideas to be patented. Obviously, image examination is important to patent experts in their attempt to deeply understand the patent contents and find relevant inventions and for that reason, a tool that supports efficient patent image retrieval would be of great help to 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 query-by-example paradigm. To this end, techniques from document image analysis (DIA) need to be employed for patent pre-processing in order to extract the patent images, while an advanced image search engine needs to be implemented to support efficient retrieval in a way that combines the two main image retrieval approaches to date: annotation- and content-based.

Annotation-based retrieval relies on image metadata or keywords that refer to the visual content or the properties of the image file. More often than not, annotation-based search is insufficient when dealing with visual content. When searching for visual content, it is very common to look for images that are visually similar but may be annotated in a different way. In addition, in many cases, manual annotation is not available or is incomplete. To tackle this problem, a second complementary approach has been devised: content-based search [1]. The core idea is to apply image processing and feature extraction algorithms to visual content and extract low-level visual features such as colour layout and edge histogram [1]. The retrieval is performed based on similarity metrics between such features, attempting to imitate the way humans perceive visual similarity [2].

More recent works attempt to combine these two main approaches by supporting hybrid queries. Such search engines combine text and content-based search, in order to offer more options to the users and produce more meaningful results [3]. More sophisticated approaches include also ontology-based search, which complements the content-based mode and vice-versa [4].

In this article, we attempt to exhibit the potential of exploiting image retrieval technologies for patent retrieval applications. First, we present the related work in Section 2 and in Section 3 we proceed with a general discussion of the potential benefits, the challenges and the requirements involved in the development and deployment of a patent image retrieval (PIR) framework. The design and the architecture of the proposed retrieval framework are analysed in Section 4, while Section 5 contains a discussion on the implemented search engine and its evaluation. Finally, Section 6 concludes the article.

2. Related work

Given that most patent search/classification systems focus on text, metadata and Boolean-based search, there is very little published work in the field of image-based patent search. In fact, it

* Corresponding author. Tel.: +30 2311 257754; fax: +30 2311 257707.
E-mail address: stefanos@iti.gr (S. Vrochidis).

0172-2190/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.wpi.2009.05.010

seems that most of image retrieval work in the area of intellectual property is dedicated to the field of trademark search [5–10]; however, as discussed in [11], these efforts had limited success in satisfying the user requirements.

Patent image search is only rarely treated in the respective literature. A prototype system that attempts to tackle the patent image retrieval problem is PATSEEK [12], while another related effort comes from a French company, LTU Technologies [13]. In addition, some related research works are published in [14,15].

PATSEEK is an image-based retrieval system for the US patent database [12]. It consists of two subsystems: one for the image feature extraction and one for query-by-example image retrieval. The image feature representation is achieved through a shape-based image retrieval method called the Edge Orientation Autocorrelogram (EOAC). The PATSEEK search system interacts with the user through a simple interface that, given a certain query image, returns a set of visually similar images.

ImageSeeker is a tool in the field of patent image-based search coming from LTU Technologies [13]. Their technology has been used by the French patent office (INPI) to build an image-based patent retrieval system and was also applied in a European project called eMARKS [16] that aims at the development of services for access to trademark and image databases. The performance of the system is claimed to be better compared to existing image classification systems. However, since the LTU Technology is proprietary, its retrieval performance, as well as its scalability cannot be evaluated.

Apart from the aforementioned systems, retrieval algorithms that focus on PIR, have been published in [14,15], however they were never tested extensively in large scale databases and were not applied in an integrated patent retrieval framework, where patent images have to be extracted from documents and other metadata need to be taken into account.

3. Patent image retrieval: motivation and considerations

In this section, we discuss the potential benefits and the open challenges resulting from the introduction of patent image retrieval approaches as well as the requirements expressed by the usage scenarios of the patent searchers.

3.1. Usage scenarios, potential benefits and challenges

According to a recent survey on patent search tools [17], no systematic efforts have been conducted with the aim of developing a PIR system. The lack of such enabling technologies is more pronounced, when considering that the non-textual elements of patents may play a crucial role in patent search [18]. More specifically, in technological domains, where technical drawings and depictions comprise a fundamental means of specifying the object of protection, patent search constitutes an overwhelming task, since it involves the collection of a significant number of patent documents (by means of conventional patent retrieval techniques such as Boolean keyword search) and inspection of their visual content, in order to establish the novelty or potential infringement involved in a newly submitted application.

To this end, we envisage the integration of a PIR framework to the standard patent search infrastructure that is available to patent searchers today. The patent searchers should be enabled to exploit such technologies in a series of use cases employing visual and textual image search. More specifically patent specialists should be provided with capabilities to search for patent images according to the query-by-example paradigm, whereby an input image is provided to the PIR system as an example in order to retrieve visually similar images (from a geometric/structural point of view). Furthermore, patent searchers should be able to retrieve images based on their textual description, which is referenced in the

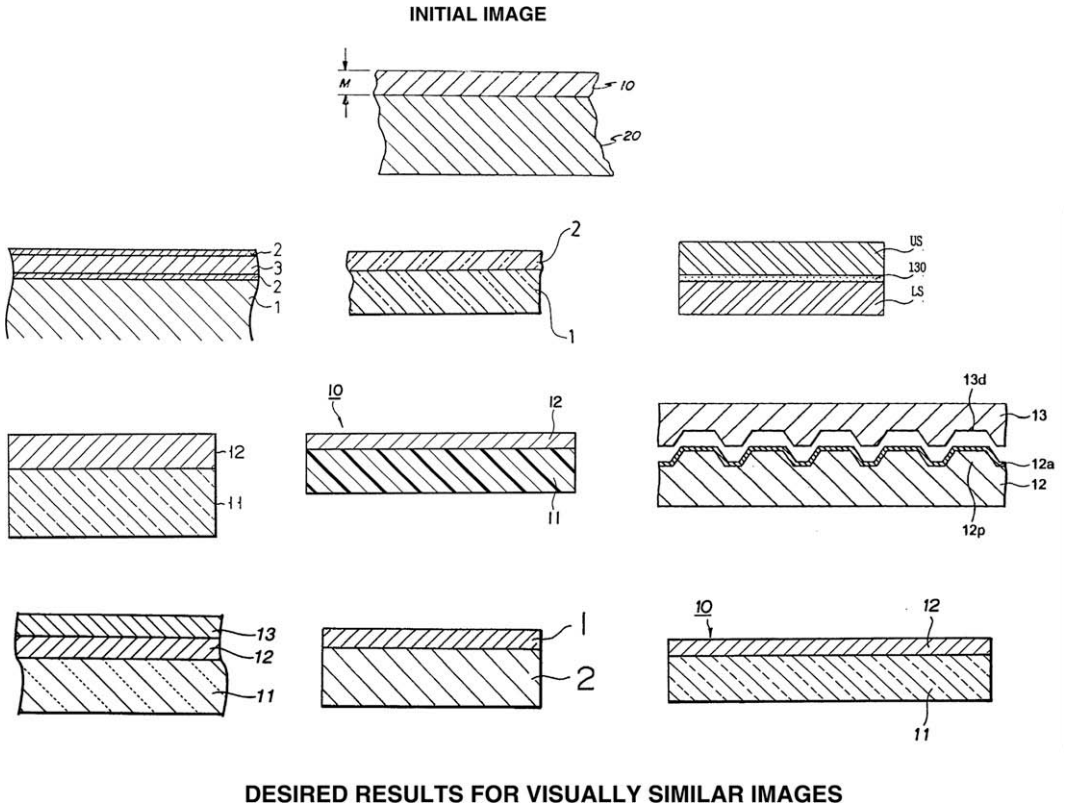


Fig. 1. The example image and visually similar figures.

document, as well as on their category (e.g. retrieve only images that are flowcharts). More complicated usage scenarios should include hybrid queries, where the aforementioned retrieval modes have to be combined efficiently and complement each other.

There are two major potential benefits that patent examiners could reap from the adoption of such a framework. First, since it will be possible to carry out searches based on the visual similarity between patent drawings or the concepts depicted by the drawings, patent image searches will become more targeted and thus will consume less effort and time. Furthermore, endowing patent search systems with visual retrieval capabilities could be beneficial for increasing the breadth of the search, i.e. improve the recall performance of the system. That is especially important when considering that due to the swiftly changing and often inconsistent terminology in emerging technical domains (e.g. electronic devices), keyword-based and Boolean search may frequently return only a subset of the documents that are related to an input document. Visual similarity can overcome the limitation of language inconsistency and novelty, since technical drawings of the same domains typically share style and semantics. Improving the recall (i.e. the number of returned results that are relevant to the input query) of patent searches in emerging domains is crucial for avoiding cumbersome infringement-related litigations caused by inefficiencies in the patent search process.

Apart from the potential benefits that we envisage from the deployment of a PIR system, it is also important to identify the main research challenges that the development of such a framework involves. The primary challenge in dealing with patent images is the inherent difficulty in extracting and indexing them in a reliable way. By nature, patent images are bi-level (black and white) since they depict technical information in diagrammatic form. Existing content-based image retrieval systems rely heavily on colour and texture image features. Such features are completely absent from patent figures; to make up for this, one would need to extract features that quantify the figure geometry, e.g. the shapes depicted within the figure and the spatial relations between them.

Though geometric information is invaluable for characterizing bi-level images, the bulk of patent images are the result of either vector graphics rasterisation or digitisation of hand-made sketches by means of scanning. Through this process significant geometric information contained in the original image, such as textual elements, shape and segment information is lost. In addition, the majority of binary images are sketches created with a variable drawing style, which depends on the preferences of the designer and the idiosyncrasies of the particular technical domain. Consequently, a generic binary image indexing algorithm should demonstrate invariance to drawing style and robustness to noise.

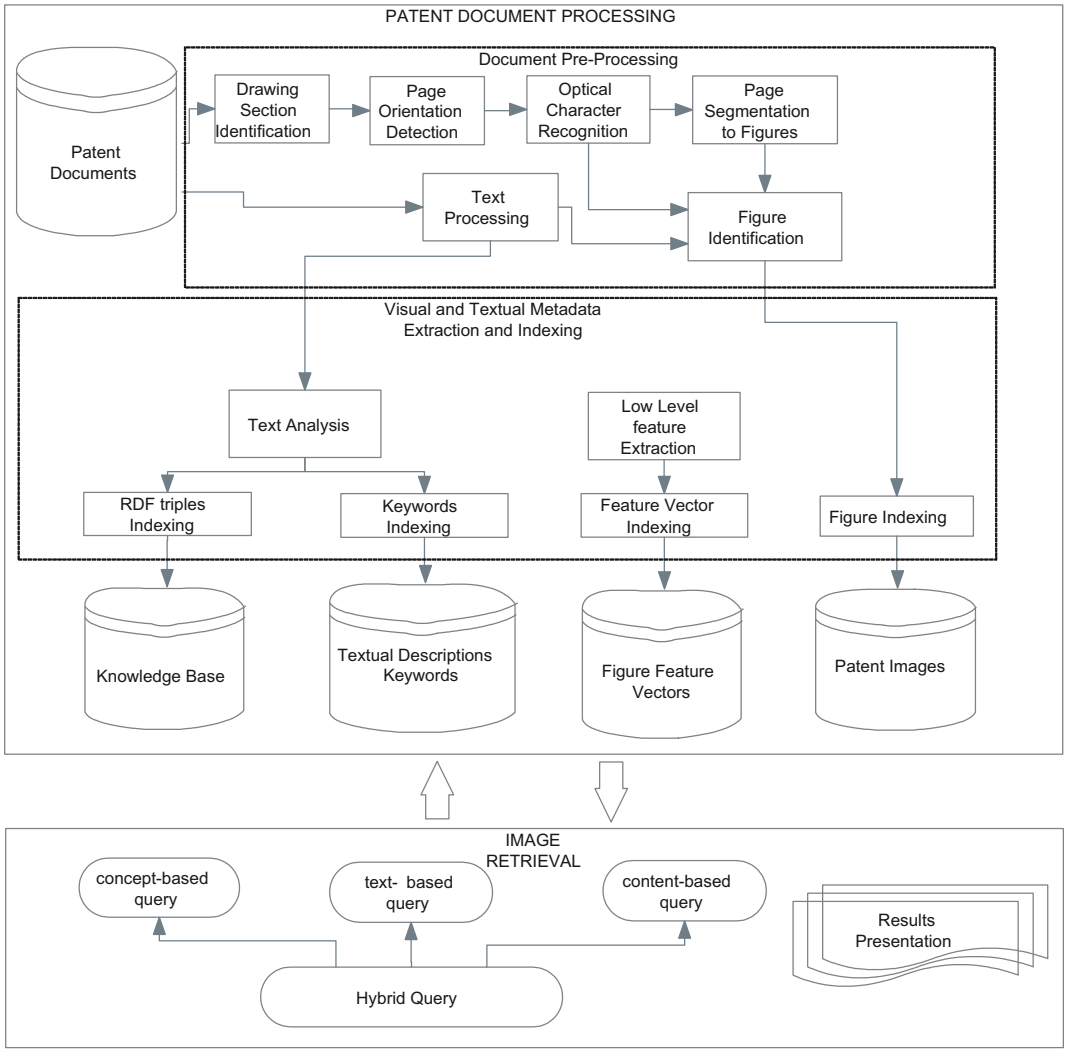


Fig. 2. The patent image retrieval framework.

Finally, the application of image retrieval technologies in real-world patent search scenarios brings to surface the problem of scalability. Millions of patent applications are processed by patent offices on a yearly basis. This leads to the need for indexing and searching in patent image collections comprising several millions of images. Thus, only low computational complexity image indexing and search algorithms will be able to cope with a problem of this magnitude.

3.2. Requirements

Based on the aforementioned usage scenarios, a set of requirements is defined for the architectural design of a PIR framework. These requirements deal mostly with functionality and performance issues. Below, the most important high level requirements are presented:

- *Automated figure processing and database population:* The framework should support automatic figure identification and extraction from patent documents. Furthermore, it should support the extraction of content-related metadata from patent documents to enable advanced retrieval functionalities. As such, specific image features, which describe efficiently the shape and the geometry of the depicted objects, are considered, as well as textual descriptions of the figures. Further processing of the metadata should lead to concept detection and association.
- *Retrieval functionalities:* The system should support content-based binary image search to output images that are visually similar to an input example as it is illustrated in Fig. 1. Additionally, text-based image retrieval should be possible. Furthermore, the system should enable concept-based image retrieval, i.e. the retrieval of images that are associated (or depict) given concepts of the technological domain of interest. Finally, the system should support the submission of hybrid queries pertaining to the different aspects of patent images, i.e. textual, conceptual and visual.
- *Performance:* The system should be responsive, e.g. its response time should be below 10 s. In addition, the system should be scalable, i.e. it should cope with vast amounts of content (in the order of millions of patent images).
- *Adaptability:* The search engine needs to build upon open technologies and standards in order to be integrated into the established patent search platforms used by patent searchers.

4. Patent image retrieval framework

In order to meet the requirements described above, we introduce a PIR framework that combines advanced techniques from text and document image analysis, as well as image retrieval methodologies which are catered for patent visual content. The proposed architecture is illustrated in Fig. 2.

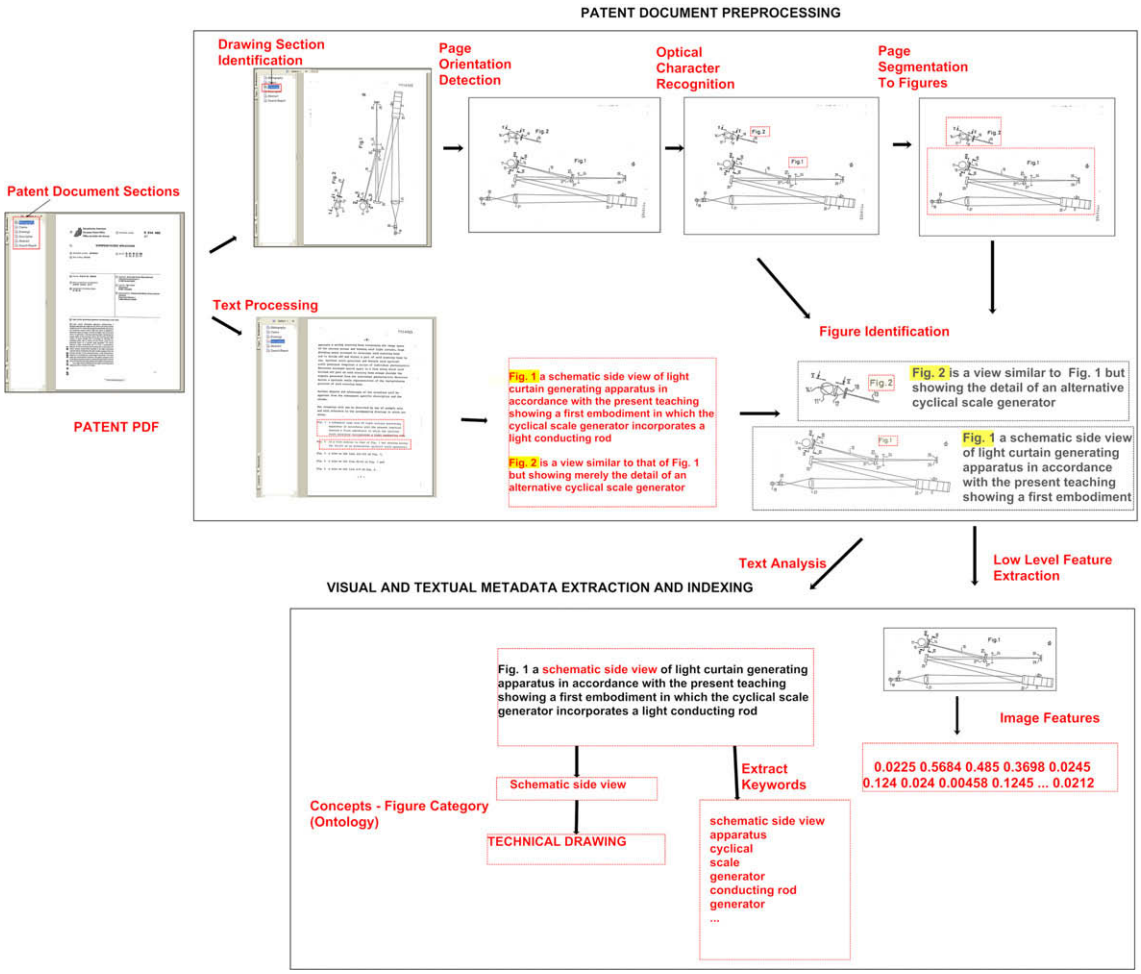


Fig. 3. A schematic example of the patent document processing phase.

The main part of the framework deals with the document processing in order to extract all the required metadata, while the image retrieval part retrieves and presents the images after a query (or combination of queries) submission by the user. Subsequently, these two main parts will be described in detail.

4.1. Patent document processing

This part comprises two basic components: (a) document preprocessing and (b) figure and metadata extraction. A patent document is considered as the sole input to the system. This is processed by the first component, in order to identify its figures and the related textual descriptions. Subsequently this information is fed to the second component, which takes care of extracting and indexing visual and textual metadata. A detailed example of the whole process is described in Fig. 3.

4.1.1. Document pre-processing

This component consists of several modules: (a) drawings section identification, (b) page orientation detection, (c) optical character recognition (OCR), (e) figure segmentation, (d) text processing and (f) figure identification.

The first module is necessary due to the fact that the input patent documents that include figures, are usually available in PDF format. The drawing page selection from the raw PDF documents is based on section information encoded within the document files. In most cases, a section denoted "Drawings" contains all the drawing-holding pages of a given patent document. Once this section is identified, all pages contained in it are extracted.

Once the patent drawing pages are extracted, it is necessary to detect their orientation and compensate for the cases where the true orientation of the image is not the correct one. The orientation detection process is carried out in three steps. First, connected-components regions are extracted. In the second step, regions are classified as "text" and "non-text" based on their spatial alignment, their spatial relations and their size. The algorithm used to carry out this step is a variant of the one described in [19]. Finally, the page orientation is estimated by the use of the "text" regions. If the majority of the text regions are aligned along the horizontal orientation, then the page orientation is classified as "horizontal", otherwise it is classified as "vertical".

Frequently, a patent drawing page contains more than one figure. Thus, it is necessary to employ techniques to identify the number and the position of the figures on the page in order to isolate them. The employed technique separate figure in patents is accompanied by a label of the form "Figure x". Therefore, if we first count the number of occurrences of such a label on a page, we directly know how many figures are contained in that page. The figure label detection can be based on existing OCR tools. The output of this subcomponent includes the figure label (e.g. "FIG. 7"), as well as the relative location of the label on the page.

Subsequently a segmentation step is required to isolate the images. The segmentation is based on the connected-components technique, which identifies the parts in the page that can be considered as separate objects. Overlapping objects are merged in a repetitive process until the main concrete objects that can be considered as separate drawings are identified. Performance of this component can be significantly improved by the introduction of heuristic axioms derived from the observation of a large patent document set (e.g. relatively small objects should be merged with neighbouring objects even in the case they do not overlap).

At the same time, a text processing step is applied to take advantage of the references to the image throughout the patent text. In most patent documents, there is a separate paragraph under the title "BRIEF DESCRIPTION OF DRAWINGS", which contains descriptive text for each of the figures. References to the patent figures and their components are also made in other parts of the patent text. The description of the invention refers to different drawings by specifying the figure number and to the different parts of a drawing by reference letters or numerals. The description of the figures is an important source of information, since the image category (i.e. waveform, diagram, etc.) can be extracted together with other useful information (the type of view, the depicted artifact, etc.).

Although the aforementioned modules employ sophisticated analysis techniques, it is possible that errors are introduced into the results for a variety of reasons. More specifically, handwritten or low quality scanned labels will possibly lead to failure of the OCR tools, while complicated textual descriptions or lack of figure references in the text could introduce errors in the output of the text extraction module. In addition, the segmentation process could fail in certain cases, especially when a single figure consists of spatially disjoint elements or multiple figures are adjacent.

In order to minimise these errors, another step is employed, where the results of the above procedures are combined in order to produce more reliable output. This is performed in the figure identification module. First, the two sets of labels, extracted by the OCR and text processing subcomponents, are merged and a new updated set of labels is defined. Subsequently, a correction procedure takes place, assuming that the figures labels are appearing, in most cases, sequentially. For instance, if we come up with a sequence of labels: "Fig 1, Fig 2, Fig Unknown, Fig 4...", it is very likely, that the unknown label is "Fig 3". Subsequently this information, as well as the coordinates of the labels extracted by the OCR are compared with the extracted figures from the segmentation component. In that way it is possible to correct also cases where the segmentation has failed. For example, when label "Fig 1" is recognised by the OCR in page 1, label "Fig 2" in page 2 and segmentation process has outputted two figures in page 1, we are able to merge these images and associate them with the label "Fig 1".

4.1.2. Visual and textual metadata extraction and indexing

This part comprises the following modules: (a) low-level feature extraction, (b) text analysis and (c) concept, text and visual features indexing.

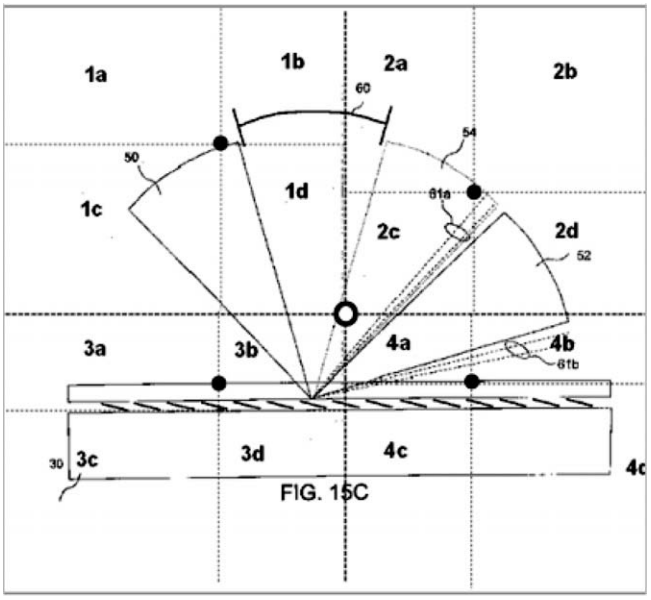


Fig. 4. Processing for the generation of the Adaptive Hierarchical Density Histogram; ○, the first geometric centroid and ●, the geometric centroids during the second iteration.

As far as the stand-alone figures are provided by the document pre-processing analysis, it is possible to extract representative features for each image, in order to be able to compare them.

To cope with the requirements of Section 3.2, the feature extraction was implemented as a variation of the technique described in [20]. In most of the weak-segmentation or non-segmentation based techniques a binary image is considered as a complex geometrical shape or a set of simple geometrical primitives, whose orientations and relative positions would produce a fair description of its topological structure. Instead, we introduce a novel technique, where a black and white image is considered as a distribution of black points on a two-dimensional background, and we propose that an estimation of both the global and local density of this distribution will provide a consistent quantification of the image structure. The employed feature is called Adaptive Hierarchical Density Histogram (AHDH). First, the algorithm involves a pre-processing phase for noise reduction, coordinate calculation and normalisation. After the pre-processing has taken place, the first geometric centroid of the image plane is calculated and the image area is split into four regions based on the position of this centroid. Subsequently, the feature vector is initialised by estimating the distribution of the black points in each region. This procedure is repeated in a recursive way (Fig. 4) for a manually specified number of iterations and the feature vector is updated.

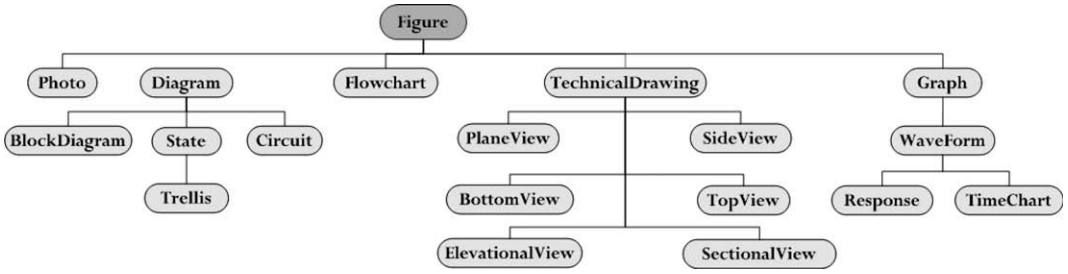


Fig. 5. Patent Drawings Ontology.

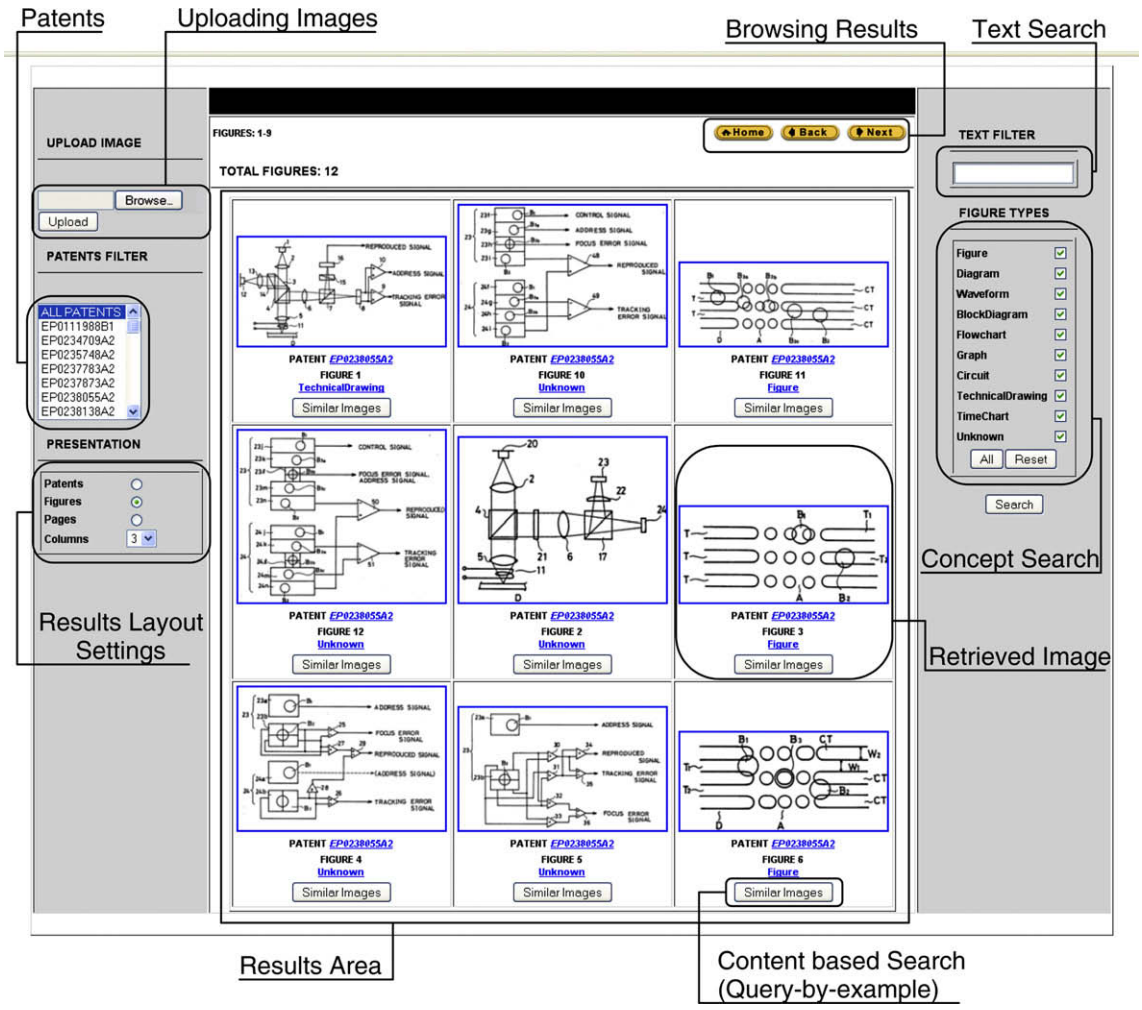


Fig. 6. PatMedia graphical user interface.

This non-segmentation point-density orientated technique seems to combine high accuracy at low computational cost.

At the same time, the textual descriptions are fed into the text analysis module, which carries out a semantic analysis of the description. This step involves parsing the description, with the aim of finding a concept which represents the image category. Detected concepts are taken from the patent figures ontology (Fig. 5), which provides a formal structure of the figure categories [21]. The main parent concept is *Figure*, which is considered as a special kind of image which is found in patent documents. The *Figure* concept has several subclasses, namely *Photo*, *Diagram*, *Flowchart*, *TechnicalDrawing* and *Graph*. The semantic analysis of the image description tries to annotate the image with one of the concepts in the *Figure* hierarchy. For instance, from an image description like: "Fig 2 is a diagrammatic, radial cross-sectional view of an optical recording element according to the invention", the concept *sectionalView* is extracted representing the figure category. Conceptual annotations are stored in the Knowledge Base in the form of Resource Description Framework (RDF) triples [22], to enable concept-based image retrieval.

Finally, the extracted data and metadata (i.e. figures, feature vectors, keywords and concepts) are indexed and stored into the image, visual feature and text databases and in the Knowledge Base respectively (Fig. 2).

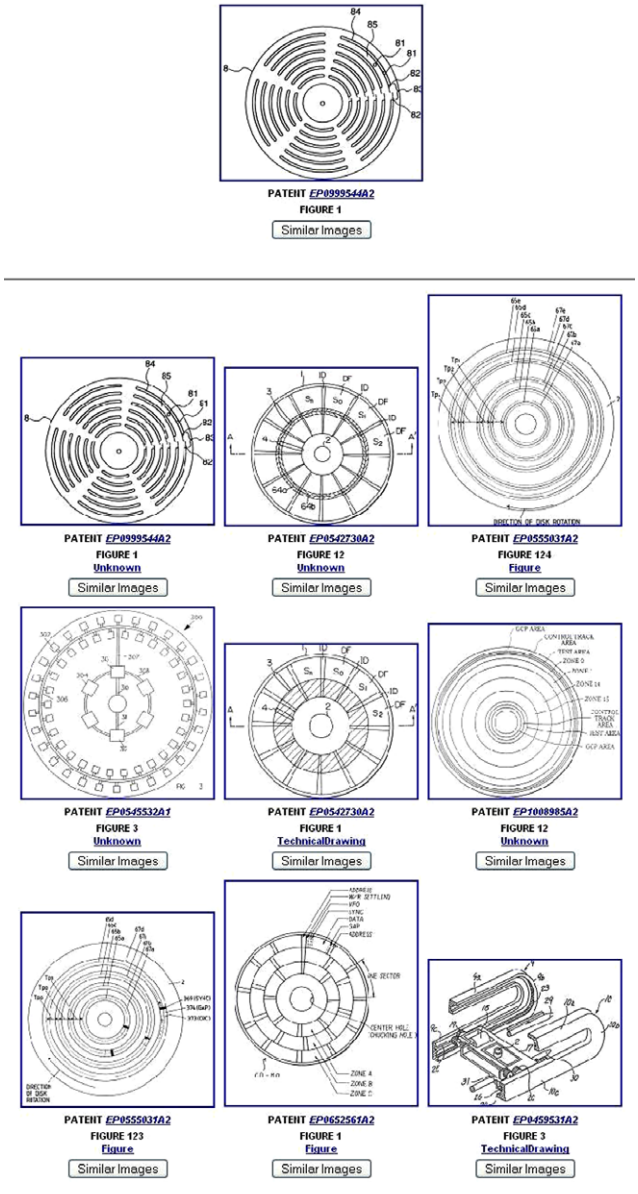


Fig. 7a. Content based search; the user searches for images that looks like optical discs.

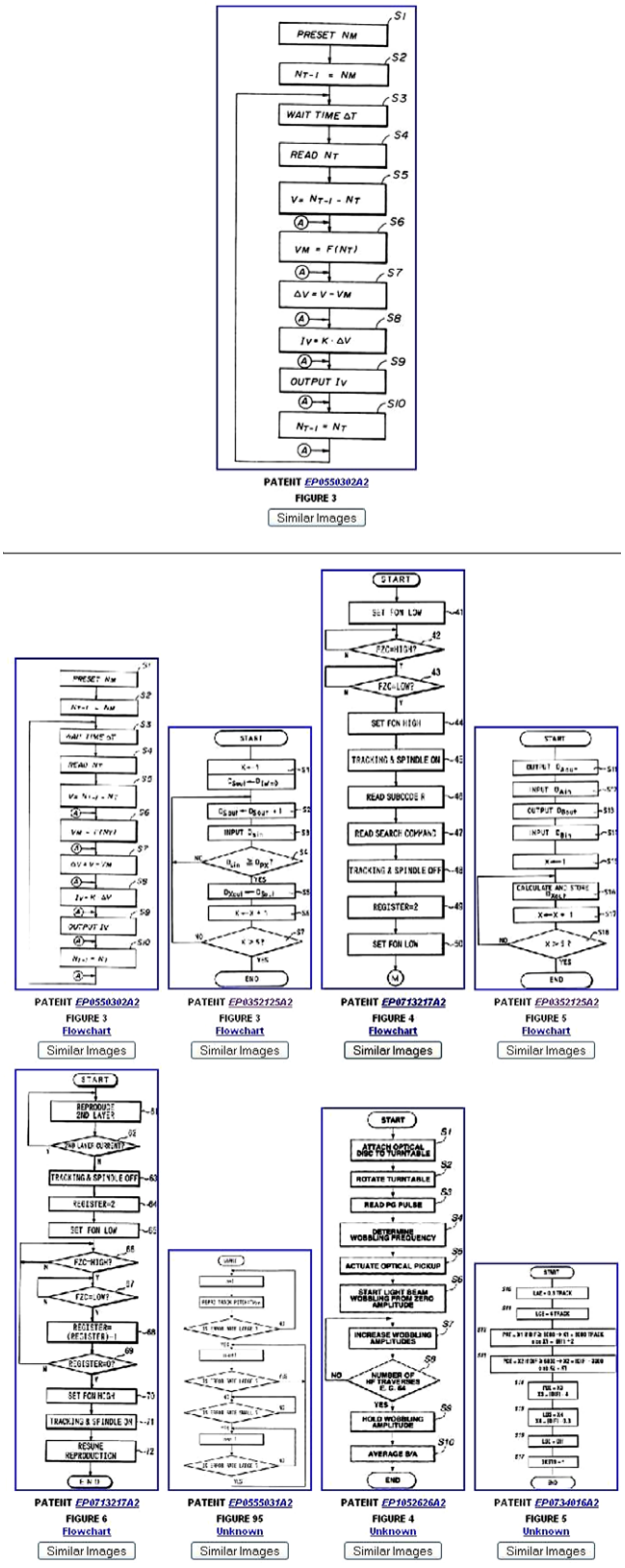


Fig. 7b. Content based search; the user searches for flowcharts.

4.2. Image retrieval

During the retrieval process, a hybrid query (i.e. a combination of queries) is submitted by a user, and the results are retrieved from the databases. Based on the metadata extraction, the user has the capability of retrieving images based on keywords, concepts referring to category and visual similarity. This means that the system is capable of retrieving content in three different modes: (a) text-based, (b) concept-based and (c) content-based. These modes are complementary to each other and they can serve different user needs.

Full text search is realised with inverted index technologies exploiting the associated textual descriptions and keywords for each image. Concept-based search is performed by submitting queries to the Knowledge Base, where images of a specific concept (i.e. category) are retrieved using standard query languages (e.g. SPARQL).

The search based on image visual similarity supports retrieval according to the query-by-example paradigm [23]. This involves direct comparison of the generated feature vectors. Such comparison is usually implemented by means of a certain type of distance (e.g. L1 distance) between the vector of the example image and the corresponding vectors for the rest of the figures. Then, the images are sorted according to this distance and presented to the user.

The system also involves a User Interface (UI) component that enables the user to combine all the aforementioned search modes in order to submit a single query that combines all the available functionalities.

5. A patent image search engine and evaluation corpus

The evaluation procedure took place by means of a search engine, called PatMedia, which realises the aforementioned PIR framework. This section aims at demonstrating the advanced functionality of the PIR module, through usage modes. Furthermore, it provides insights into the performance of the search engine and presents a quantitative evaluation of the attained results.

5.1. The PatMedia retrieval engine

The PatMedia engine offers a user-friendly interface for hybrid query submission and for results presentation. A snapshot of the Graphical User Interface (GUI) of the search engine is illustrated in Fig. 6, along with the descriptions of the supported functionalities.

Several advanced software solutions were employed to support the implementation of individual components. For instance, for the OCR, a widely known commercial solution was used [24], while text analysis was performed with the MiniPar parser [25].

5.2. Patent document set collection

A primary task for testing the search engine and evaluating the results was the collection of a sufficiently large patent document reference set consisting of EPO documents [26] from the optical recording domain (IPC class G02B). A number of criteria were considered for selecting this reference data set. First, the documents

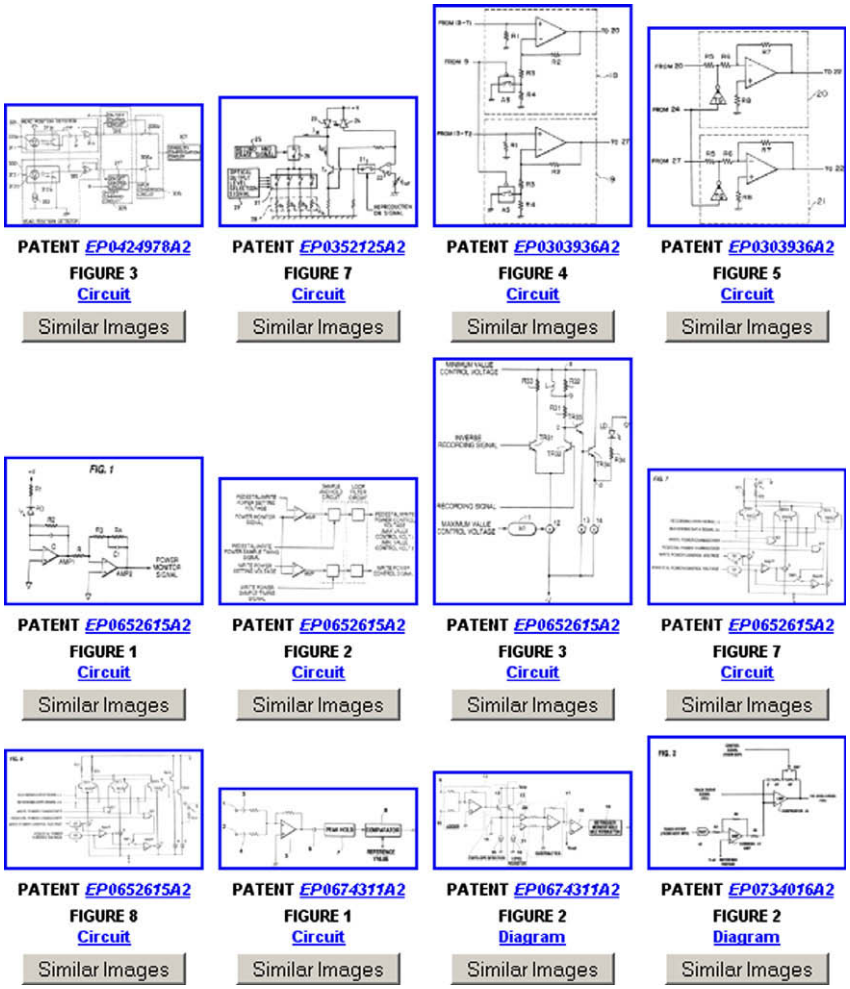


Fig. 8a. Concept-based search to retrieve images of type "circuits".

were selected to contain many representative types of images. In addition, we made sure that groups of similar figures exist in the dataset, so that the classification/matching retrieval algorithms can be evaluated.

5.3. Presentation of results

In this section, we will exemplify the usage of the proposed patent image retrieval module through some interaction modes. In the first usage scenario, we consider a user who uses an example image depicting an optical disc and searches for visually similar images in all patents (Fig. 7a). Clearly, the output images include cyclic objects that look very similar to the example image. Even in the case that the patent searcher is looking for discs with a particular characteristic that is not properly described by the image features, using this module it is possible at least to discard many dissimilar images, speeding in that way the search process. In the second usage mode, presented in Fig. 7b, the input image is a flowchart. The retrieved figures are also flowcharts that share similar structure with the query image. In the third example (Fig. 8a), the user asks for images that fall into the category of "circuits". Based on the concept information, such figures are retrieved. Furthermore, the user is capable of searching for a keyword in the description of the figure. In this example (Fig. 8b), images that include the word "lens" in their descriptions are retrieved. In the last interaction mode the user starts with a content-based search to retrieve visually similar images to a diagram. Subsequently the user fires a hybrid search including, besides the query-by-example, the concept filtering option. In that way, the content-based results, which provide visually similar figures with the image example, are filtered according to the selected category, which is "diagram" in this example. Observing the results for both cases in Fig. 9a and b, it is clear that in the second search, the quality of results has been improved, as several irrelevant images were discarded due to the concept filtering.

5.4. Evaluation of results

In this section, we present a more detailed investigation into the efficiency and the retrieval performance of the PatMedia search engine.

5.4.1. Performance insights

The experiments were conducted on a PC, with a P4 3.2 GHz Intel CPU and 1 GB RAM. PosgreSQL was used to store the actual

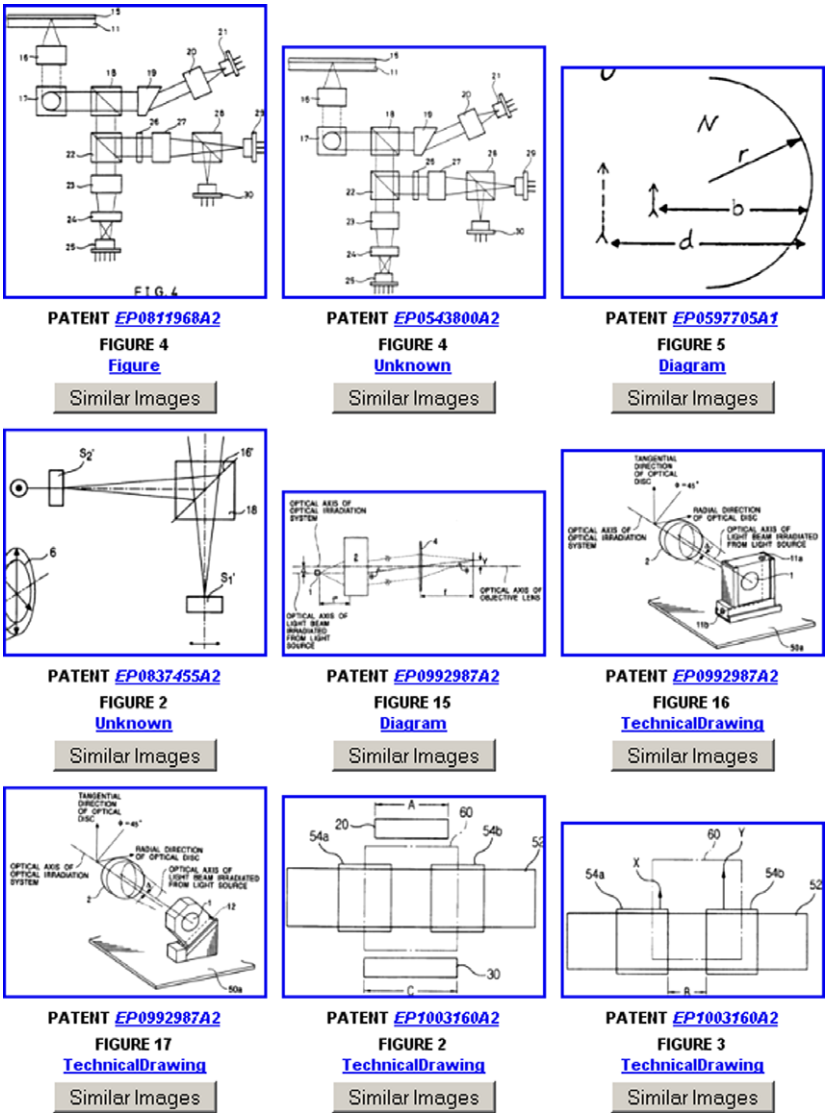


Fig. 8b. The user searches for figures that include the keyword "lens" in their description.

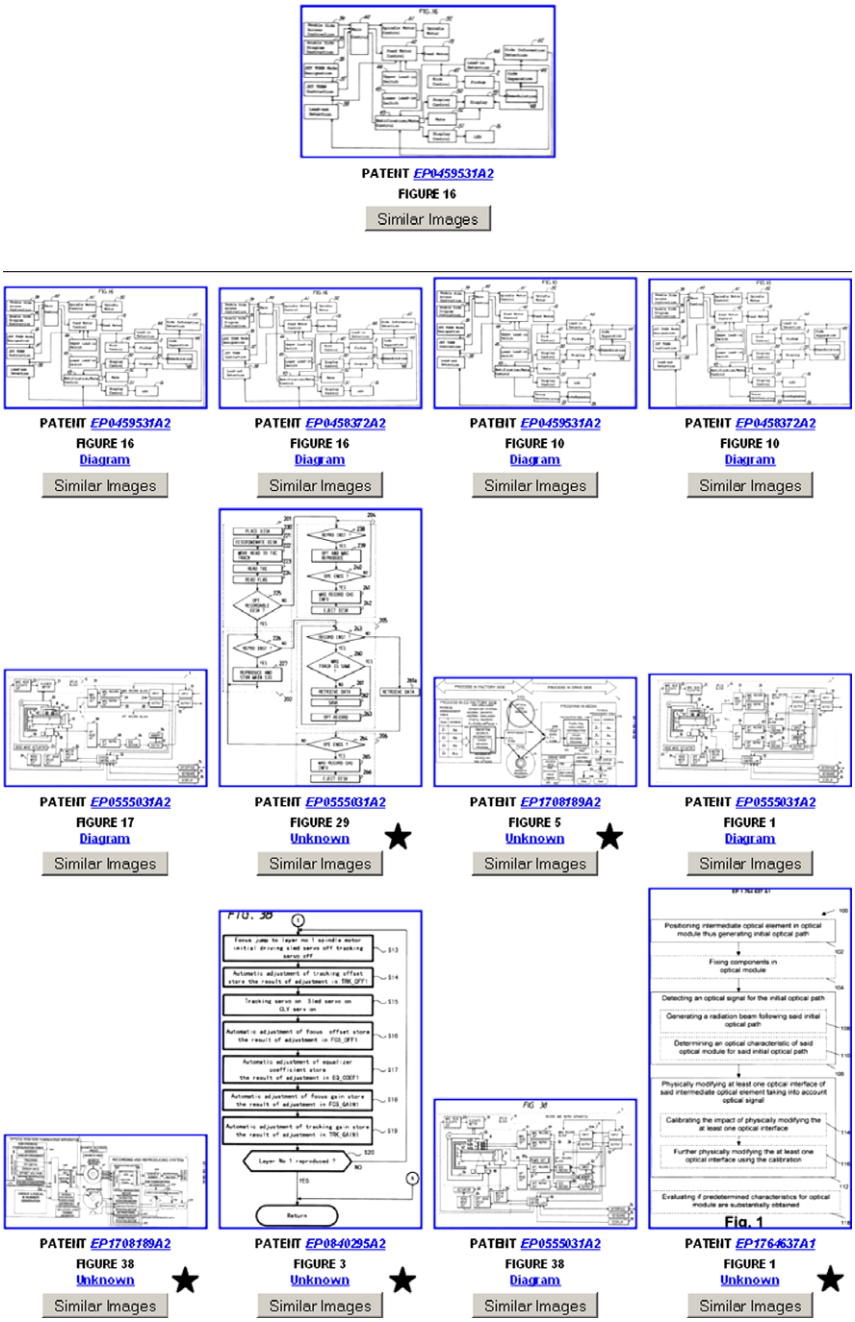


Fig. 9a. Content based search for a diagram; visually similar results, which are not thematically relevant are marked with a star (★).

non-multimedia content. More specifically the database contained the textual associations and the category information, which was extracted from RDF files. The involved dataset included 2000 patent images that have been extracted from about 200 EPO patents. For the evaluation of the results, the widely used precision and recall metrics [27] were employed.

For the performance evaluation of the content-based retrieval part, which is the most important functionality of the search engine, we have chosen 110 images of diverse content from this collection. For each query image, a relevant response set was manually identified, containing from 3 to 25 images. The ideal patent retrieval system would return all the relevant images. We performed two different sets of experiments. In the first experiment, we retrieved the 25 most similar images for each query image. The resulting image set contained an average of 86.9% of the figures, which had been manually identified as visually similar to the query images. In the second experiment, we employed a distance threshold, which discriminates relevant from irrelevant images and we computed the precision and recall metrics. A fair compromise between these complementary metrics leads to 77.4% recall for 49% precision. For both experiments, the total time required for a query did not exceed 1 s. As far as scalability is concerned, PatMedia was tested for over 10 000 images and the retrieval time was below 10 s.

5.4.2. Comparison with existing approaches

In this section, a comparison between the PatMedia and the PATSEEK search systems is presented, since we lack details regarding the implementation details and performance of the other systems described in Section 2.

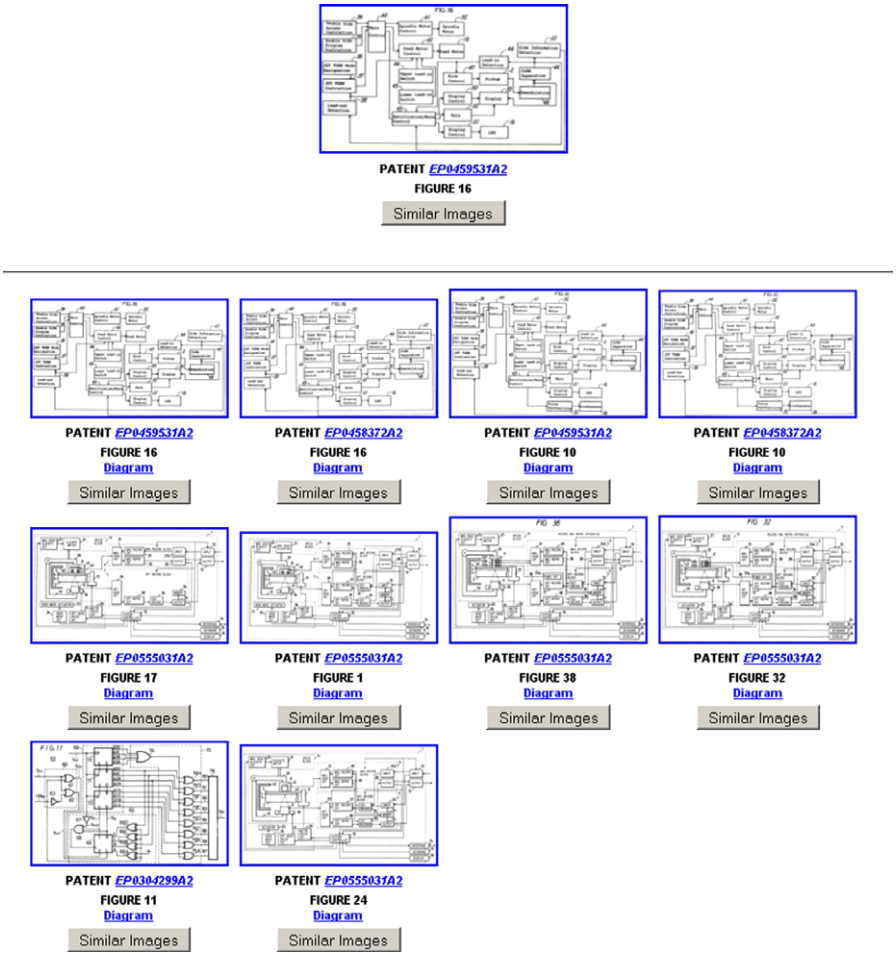


Fig. 9b. Hybrid search combining content and concept-based retrieval; the irrelevant figures are discarded by filtering the initial visual-based results with the concept "diagram".

Regarding the functionalities provided by the two systems, it is quite evident that PatMedia offers more query options by incorporating text, concept and visual search, whereas PATSEEK is strictly an image-based search system.

Although satisfactory results are presented in [12], the evaluation of PATSEEK was performed using a relatively small dataset of 200 patent images. With the intention of performing a realistic comparison, we simulated the algorithm that was introduced in PATSEEK and compared the content-based functionality of the two systems under identical conditions. The database contained 1317 patent images from the machine tools and the optical recording domains. The query set consisted of 120 images, while a relevant set of at most 73 figures was defined for each query image. In both methods, L1 distance was utilised as a similarity measure. The precision and recall curves that were produced by using a variety of similarity thresholds are depicted in Fig. 10.

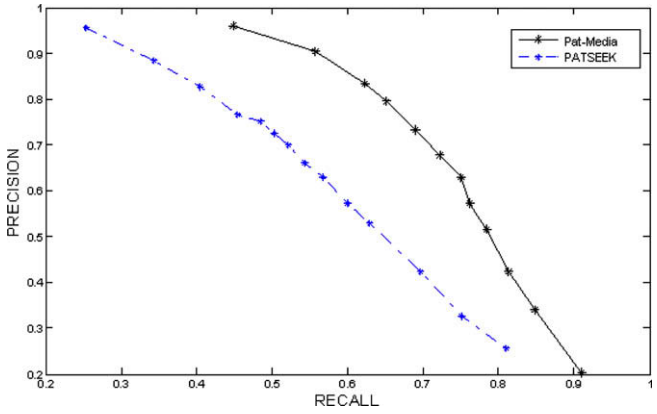


Fig. 10. Comparison between PatMedia and PATSEEK.

6. Conclusions

This article provided an overview of the benefits, the challenges and the requirements involved in the development and deployment of a PIR framework. The design and implementation of this framework is tailored to the special nature of patents as it builds upon advanced techniques from image analysis and content-based retrieval to enhance the performance of patent image search.

The modular architecture of the proposed framework was designed with extensibility, interoperability and adaptability in mind. The PIR framework could be an integral part of a more general patent management system (as described in [28]) but due to its independent architecture it could also serve as a stand-alone search engine. The framework is capable of combining the content-based search with the annotation-based search and of providing results of improved relevance to the patent searchers.

To evaluate this framework, the patent search engine PatMedia was developed. The results, of this search engine can be considered promising, as they are validated through specific interaction modes, and comparisons with existing systems (i.e. PATSEEK), and evaluated by means of precision and recall measurements.

Future work could deal with content-based classification and clustering, as well as with the introduction of indexing structures to further improve performance in large databases. Finally, additional pre-processing steps that could allow the user to search for specific parts of the figure (i.e. objects) could be investigated.

Acknowledgments

This work was supported by the projects PATExpert [29] and CHORUS [30], both funded by the European Commission. Parts of this work were presented in IRFS 2008 [31] and in IPI–ConfEx 2009 [32].

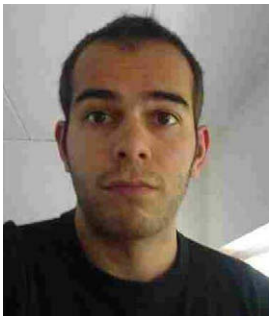
Appendix A. Glossary

- *Image analysis techniques:* Extraction of meaningful information (i.e. vectors describing image colour, texture, geometry, etc.) from images.
- *Indexing techniques:* Data structures that enable efficient and fast lookup.
- *Content-based (image) search*: It aims at retrieving visually similar images. It is based on extracting several features (low-level visual features) such as colour and shape.
- *Concept-based (image) retrieval:* The retrieval of images that are associated (or depict) given concepts of the technological domain of interest.
- *Query-by-example paradigm:* It is a query technique where the user provides an image example to the system and expects to retrieve visually similar figures.
- *Document image analysis:* It refers to algorithms and techniques that are applied to images of documents to obtain a computer-readable description from pixel data. A well-known document image analysis product is the optical character recognition (OCR).
- *Hybrid queries:* Queries formed by the user that combine multiple types of search such as content-, textual- and concept-based search.
- *Ontology:* It is a formal hierarchy of a set of concepts within a domain and the relationships between those concepts.
- *Ontology-based search:* Utilize semantics captured in the ontology to process the query. In case the search is limited only to ontology concepts, it also referred as *concept-based search.*
- *Patent image retrieval (PIR) framework:* A basic conceptual structure, which consists of different integrated technological modules, and is capable of performing patent image retrieval.
- *Use cases:* Use cases describe the interaction between a primary actor (e.g. the user) and the system itself, represented as a sequence of simple steps.
- *Optical character recognition (OCR):* It refers to the branch of computer science that involves reading text from paper and translating the images into a form that the computer can manipulate.
- *Figure segmentation:* The process of partitioning a digital image into multiple parts in order to identify objects and boundaries (lines, curves, etc.).
- *Connected-components regions:* They represent the different pieces of a graph.
- *Knowledge Base:* A collection of knowledge in the form of triplet (i.e. subject–predicate–object) information that pertains to a specific topic or subject of interest.
- *Resource Description Framework (RDF):* It is a general-purpose language for representing information.

References

[1] Smeulders A, Worring M, Santini S, Gupta A, Jain R. Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 2000;22:1349–80.
[2] Brunelli R, Mich O, Modena C. A survey on video indexing. J Vis Commun Image Represent 1999;10:78–112.
[3] Luo B, Wang X, Tang X. World wide web based image search engine using text and image content features. In: Proc SPIE 2003, Santa Clara, USA, vol. 5018; 2003. p. 123–30.
[4] Vrochidis S, Doulaverakis C, Gounaris A, Nidelkou E, Makris L, Kompatsiaris I. A hybrid ontology and visual-based retrieval model for cultural heritage multimedia collections. Int J Meta Semant Ontol 2008;3:167–82.
[5] Eakins JP. Trademark image retrieval. In: Lew M, editor. Principles of visual information retrieval. Berlin: Springer-Verlag; 2001.
[6] Jain AK, Vailaya A. Shape-based retrieval: a case study with trademark image databases. Pattern Recog 1998;31:1369–90.
[7] Kim YS, Kim WY. Content-based trademark retrieval system using a visually salient feature. Image Vis Comput 1998;16:931–9.
[8] Wu JK et al. Content-based retrieval for trademark registration. Multimedia Tool Appl 1996;3:245–67. Available from: <http://www.springerlink.com/content/u5t11g636m21t103/>.
[9] Eakins JP, Boardman JM, Graham ME. Similarity retrieval of trademark images. IEEE Multimedia 1998;5:53–63.
[10] Alwis S, Austin J. Trademark image retrieval using multiple features. In: Presented at CIR-99: the challenge of image retrieval, Newcastle-upon-Tyne, UK; 1999.
[11] Schietse J, Eakins JP, Veltkamp RC. Practice and challenges in trademark image retrieval. In: Proc 6th ACM int conf on image and video retrieval (CIVR); 2007. p. 518–24.
[12] Tiwari A, Bansal V. PATSEEK: content based image retrieval system for patent database. In: Proc int conf on electronic business-04, Tsinghua University, Beijing, China; 2004.
[13] LTU Technologies. <http://www.ltutech.com/en/>.
[14] Huet B, Kern NJ, Guarascio G, Merialdo B. Relational skeletons for retrieval in patent drawings. ICIP 2001;2:737–40.
[15] Zeng Z, Zhao J, Xu B. An outward-appearance patent-image retrieval approach based on the contour-description matrix. In: Proc 2007 Japan–China joint workshop on frontier of computer science technology; 2007. p. 86–9.
[16] eMARKS Project. <http://emarks.iisa-innov.com/>.
[17] List J. How drawings could enhance retrieval in mechanical and device patent searching. World Patent Informt 2007;29(3):210–8.
[18] Adams S. Electronic non-text material in patent applications – some questions for patent offices, applicants and searchers. World Patent Informat 2005;27:99–103.
[19] Hoenes F, Lichter J. Layout extraction of mixed mode documents. Mach Vision Appl 1994;7:237–46.
[20] Yang M, Qiu G, Huang J, Elliman D. Near-duplicate image recognition and content-based image retrieval using adaptive hierarchical geometric centroids. In: Proc 18th Int Conf Pattern Recog, vol. 2; 2006. p. 958–61.
[21] Wanner L, Baeza-Yates R, Brugmann S, Codina J, Diallo B, Escorsa E, et al. Towards content-oriented patent document processing. World Patent Informat 2007;30(1):21–33.
[22] Resource Description Framework (RDF). <http://www.w3.org/RDF/>.
[23] Izquierdo E, Casas J, Leonardi R, Migliorati P, O'Connor N, Kompatsiaris I, et al. Advanced content-based semantic scene analysis and information retrieval: the schema project. In: Proc workshop on image analysis for multimedia interactive services, London, UK; 2003. p. 519–28.
[24] ABBYY. <http://www.abbyy.com/>.
[25] MiniPar. <http://www.cs.ualberta.ca/lindek/minipar.htm>.
[26] EPO. <http://www.epo.org/patents/patent-information/european-patent-documents/publication-server.html>.
[27] Baeza-Yates R, Ribeiro-Neto B. Modern information retrieval. Addison-Wesley; 1999.
[28] Codina J, Pianta E, Vrochidis S, Papadopoulos S. Integration of semantic, metadata and image search engines with a text search engine for patent retrieval. In: Semantic search 2008 workshop, Tenerife, Spain, 2nd June 2008.
[29] PATExpert Project. <http://www.patexpert.org/>.
[30] Coordinated approacH to the EurOpean e®oRt on aUdiovisual Search engines (CHORUS) Project. <http://www.ist-chorus.org/>.
[31] Vrochidis S. In: Patent image retrieval, information retrieval facility symposium (IRFS 2008), Vienna, Austria, 5–8 November 2008.
[32] Vrochidis S. In: Towards patent image retrieval, international patent information conference and exposition (IPI–ConfEx 2009), Venice, Italy, 1–5 March 2009.

Stefanos Vrochidis received the Diploma degree in Electrical and Computer Engineering from Aristotle University of Thessaloniki, Greece and the MSc degree in radio frequency communication systems from University of Southampton, UK, in 2000 and 2001, respectively. He is a Research Associate with the Multimedia Knowledge Lab at the Informatics and Telematics Institute, Centre for Research and Technology Hellas. His research interests include semantic multimedia analysis and retrieval, as well as patent search. He has successfully participated in many European and National projects related to image and video analysis, patent retrieval and digital TV technologies and he has been involved as a co-author in more than 15 related journal and conference publications.



Symeon Papadopoulos received the Diploma degree in Electrical and Computer Engineering from Aristotle University of Thessaloniki (AUTH), Greece in 2004. In 2006, he received the Professional Doctorate in Engineering (P.D. Eng.) from the Technical University of Eindhoven, The Netherlands. Since September 2006, he has been working as a researcher in the Multimedia Knowledge Laboratory on a wide range of research areas such as information search and retrieval. His current research interests pertain to data mining on the Social Web. He is currently a Ph.D. candidate in the Informatics Department of AUTH and works towards the completion of an MBA degree in the Blekinge Institute of Technology, Sweden.

Anastasia Moumtzidou received the Diploma degree in Electrical and Computer Engineering and the MSc degree Advanced Computer and Communication Systems both from Aristotle University of Thessaloniki, Greece in 2006 and 2009, respectively. Since January 2007, she has been working as Research Assistant in the Multimedia Knowledge Lab at the Informatics and Telematics Institute, Centre for Research and Technology Hellas. Her research interests include semantic multimedia analysis and retrieval.

Panagiotis Sidiropoulos received the diploma degree in Electrical and Computer Engineering and the MSc degree in Computer Science both from Aristotle University of Thessaloniki (AUTH), Greece in 2003 and 2007, respectively. Since January 2007, he is working as Research Assistant in the Multimedia Knowledge Lab at the Informatics and Telematics Institute, Centre for Research and Technology Hellas. His research interests include multimedia content processing and retrieval.

Emanuele Pianta graduated in Linguistics at the University of Padova in 1991, with a thesis on the Relevance Theory applied to the automatic generation of texts. He is currently tenured researcher at FBK-irst in Trento. His research interests include development of multilingual resources (e.g. MultiWordNet), basic linguistic processors for Italian and English, parsing and information extraction from texts. He has participated in a number of national and European Projects, including PatExpert, dealing with content distillery from patents.

Ioannis Kompatsiaris received the Diploma degree in Electrical Engineering and the Ph.D. degree in 3-D model based image sequence coding from Aristotle University of Thessaloniki in 1996 and 2001, respectively. He is a Senior Researcher with CERTH–ITI and Director of its Multimedia Knowledge Laboratory. His research interests include multimedia content processing, multimodal techniques, multimedia and Semantic Web, multimedia ontologies, knowledge-based analysis, and context aware inference for semantic multimedia analysis, personalization and retrieval. He is the co-author of 10 book chapters, 30 papers in refereed journals and more than 100 papers in international conferences.