PATSEEK: Content Based Image Retrieval System for Patent Database

Avinash Tiwari[1], Veena Bansal[2]

[1,2] Industrial and Management Engineering Department, IIT Kanpur 208016 India
[2]veena@iitk.ac.in

ABSTRACT

A patent always contains some images along with the text. Many text based systems have been developed to search the patent database. In this paper, we describe PATSEEK that is an image based search system for US patent database. The objective is to let the user check the similarity of his query image with the images that exist in US patents. The user can specify a set of key words that must exist in the text of the patents whose images will be searched for similarity. PATSEEK automatically grabs images from the US patent database on the request of the user and represents them through an edge orientation autocorrelogram. L1 and L2 distance measures are used to compute the distance between the images. A recall rate of 100% for 61% of query images and an average 32% recall rate for rest of the images has been observed.

Keywords: Patent Search, Content Based Image Retrieval, Recall rate, Precision

1. INTRODUCTION

The sheer size of the information available about the patents has led many researchers to develop efficient and effective retrieval techniques for patent databases. The Web Patent Full-Text Database (PatFT) of United States Patent Office (USPTO) contains the full-text of over 3,000,000 patents from 1976 to the present and it provides links to the Web Patent Full-Page Images Database (PatImg), which contains over 70,000,000 images. The volume of this repository makes it prohibitive for humans to find similar images in them. This has motivated us to develop a content-based image retrieval system for the patent databases.

Initial image retrieval techniques were text-based that associated textual information, like filename, captions and keywords with every image in the repository. For image retrieval, keyword based matching was employed for finding the relevant images. The manual annotation required prohibitive amount of labor. Moreover, it was difficult to capture the rich content of images using a small number of key words apart from being an unnatural way of describing images.

Soon it was realized that image retrieval based on the contents is a natural and an effective way of retrieving images. This led to the development of Content-Based Image Retrieval (CBIR). The Content-Based Image Retrieval (CBIR) [8] is aimed at efficient retrieval of relevant images from large image databases based on automatically derived image features. The original Query by Image Content (QBIC) system [3] allowed the user to select the relative importance of color, texture, and shape. The virage system [1] allows queries to be built by combining color, composition (color layout), texture, and structure (object boundary information).

Moments [6] fourier descriptors [5], [11] and chain codes [4], [12] have also been used as features. Jain and Vailya [7] introduced edge direction histogram (EDH) using the Canny edge operator [2].

The edge orientation autocorrelogram (EOAC) classifies edges based on their orientations and correlation between neighboring edges in a window around the kernel edge [9].

2. SYSTEM ARCHITECTURE

PATSEEK consists of two subsystems- one for Creation of Feature Vector and Image Database and another one for Retrieval of Images similar to the query Image. The components of both subsystems and their interaction are shown in figure 1 (next page).

In order to add the feature vectors and images to the database, the user interacts with the system through its graphical user interface and provides a set of keywords. A snapshot of the user interface for specifying the search criteria for the patents to be grabbed is shown in figure 2 (next page). The patents are grabbed from the USPTO website. The image grabber searches the patent database and grabs the image pages from the patents that satisfy the search criteria.

A page image may contain more than one individual image. To extract the individual images from these page images, we need to identify the connected components or blocks. In these pages, the connectivity is present at a very gross level and an individual image is well separated in both directions from other images. The occasional captions are insignificant compared to the images and can be treated as noise. Our experiment shows that these captions do not change the feature vector of an image in any significant way. To identify the connected components, we start scanning a page image from the first pixel row. If a row has no black pixel, the row is discarded from further consideration. If we find a row that contains at least one black pixel, the position of the row is recorded as the potential start

of a connected component. We continue the horizontal scan as long as we continue to find rows that contain at least one black pixel.

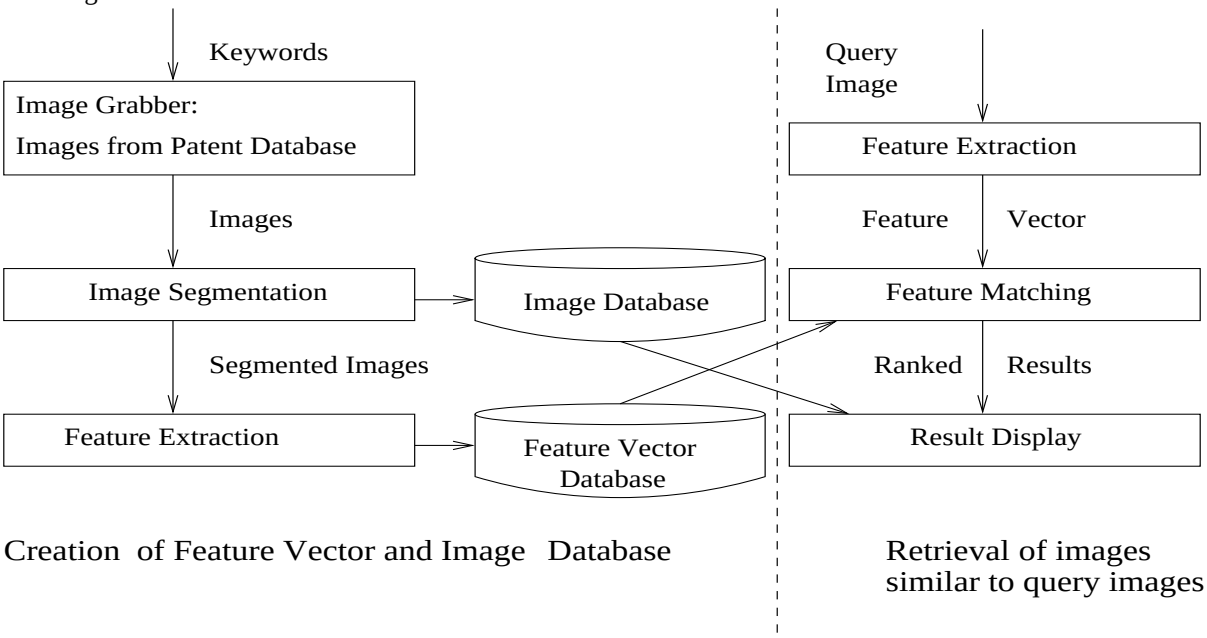


Figure 1. System Architecture of PATSEEK

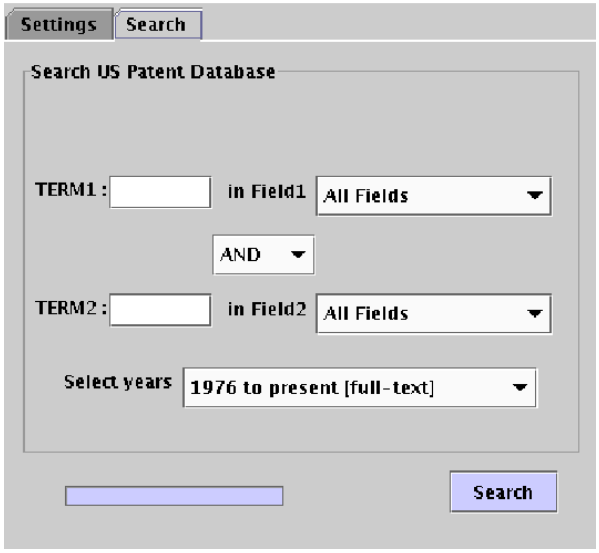


Figure 2. Patent Grabber

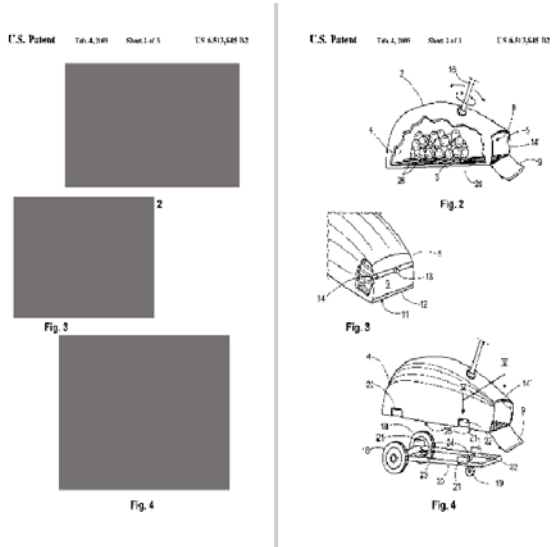


Figure 3. Blocks Identified for Individual Images

If we find enough (design parameter) contiguous horizontal rows that contain no black pixels, we have reached the end of the previous block if any. We continue horizontal scan till the end of the page image. For each horizontal block identified, we scan it vertically to segment it vertically using vertical threshold. These thresholds are set to 1 mm. A block that is less than 5 square cm is discarded as noise.

A document containing three images and their corresponding blocks that have been identified are shown in figure 3.

The separated images are stored in the image database along with patent number and the page number within the patent where these images were found.

The graphic content are then used to calculate the image feature vector. We selected EOAC for our work because it is computationally inexpensive and independent of translation and scaling. Also, the size of the feature vector is small, 144 real numbers. We computed magnitude and gradient of the edges using Canny edge operator.

The edges that have less than 10 percent of the maximum possible intensity are ignored from further consideration.

The gradient of edges is then used to quantize edges into 36 bins of 5 degrees each. The edge orientation autocorrelogram is then formed which is a matrix, consisting of 36 rows and 4 columns. Each element of

this matrix indicates the number of edges with similar orientation. Columns 1, 2, 3 and 4 give the number of edges that are 1, 3, 5 and 7 pixels apart. Each row corresponds to 5 degrees bin.

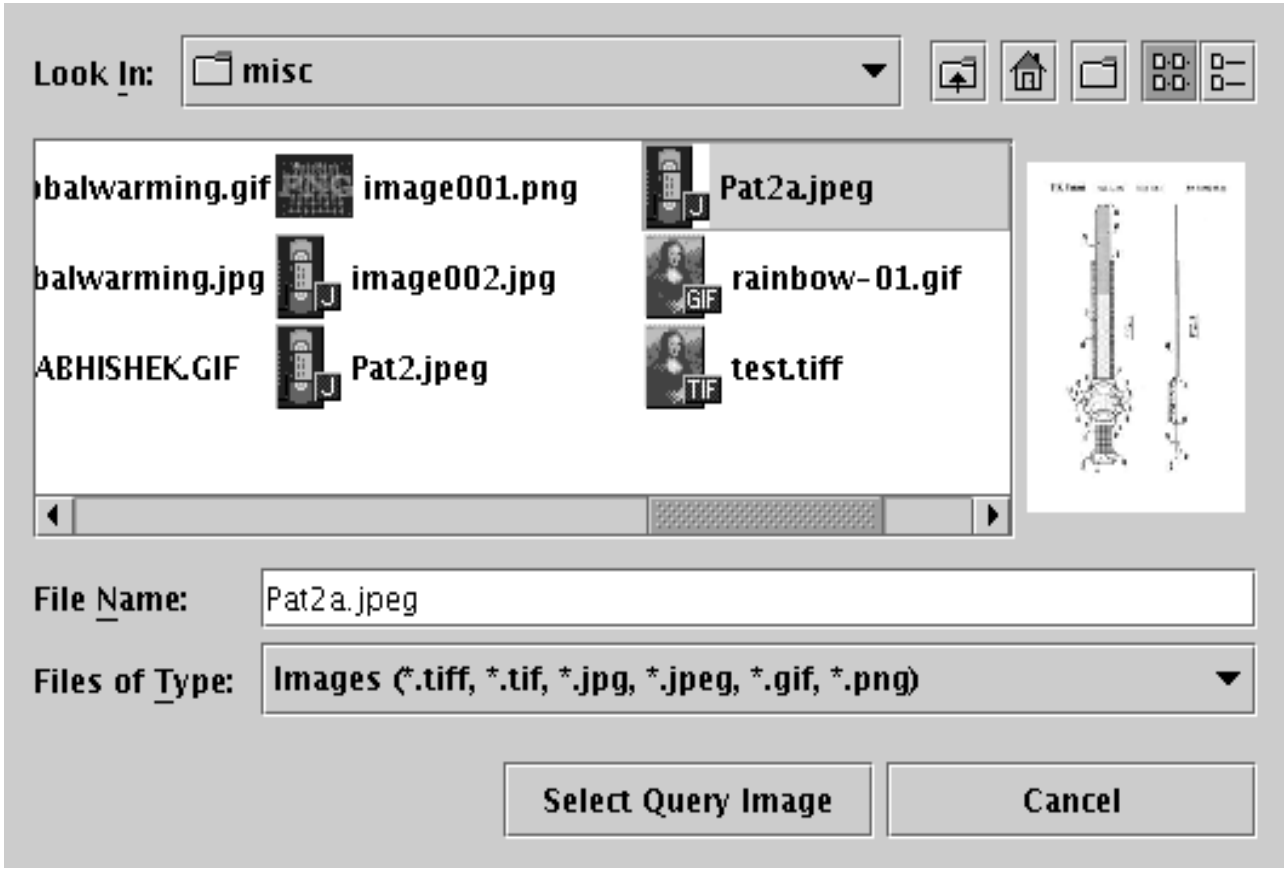


Figure 4. Query Image Selection

Two edges with k pixel distance apart are said to be similar if the absolute values of their orientations and amplitude differences are less than an angle and an amplitude threshold value, respectively [5]. These are user defined thresholds. The edge orientation autocorrelogram is stored in the database along with patent number and the page number within the patent where this image was found.

Feature vectors for the images are stored in an RDMBS table. We have created the feature vector database on Oracle as well as on Mysql.

3. THE DATABASE RETRIEVAL SYSTEM

For querying the database for images similar to a query image, the user interacts with the system through a graphical user interface (shown in figure 4). The user can specify the query image by providing its name and path. The image can be in any of the popular format such as tiff, gif, jpeg etc.

PATSEEK gives user an opportunity to specify the rotation angle for the query image. The angle may range between 0 degrees and 180 degrees.
We have used L1 and L2 distance measures to select 12-nearest neighbors and both have given almost identical results. The top twelve images, ranked on the basis of the distance are displayed as thumbnails along with the respective distance.

The graphical user interface displays the query image and the results for browsing to the user. A snapshot of the user interface is shown in the figure 5.

4. EXPERIMENTS

In this section, we report experimental results. All experiments were performed on an Intel Pentium IV Processor 2.4 GHz with 512 MBytes of RAM. The system was implemented in Java (Sun JDK 1.4.1). The feature vector database was initially created on Oracle and later on moved to MySQL Ver 12.21. We did not use the client of Oracle. We implemented the front-end using Java and its utilities.

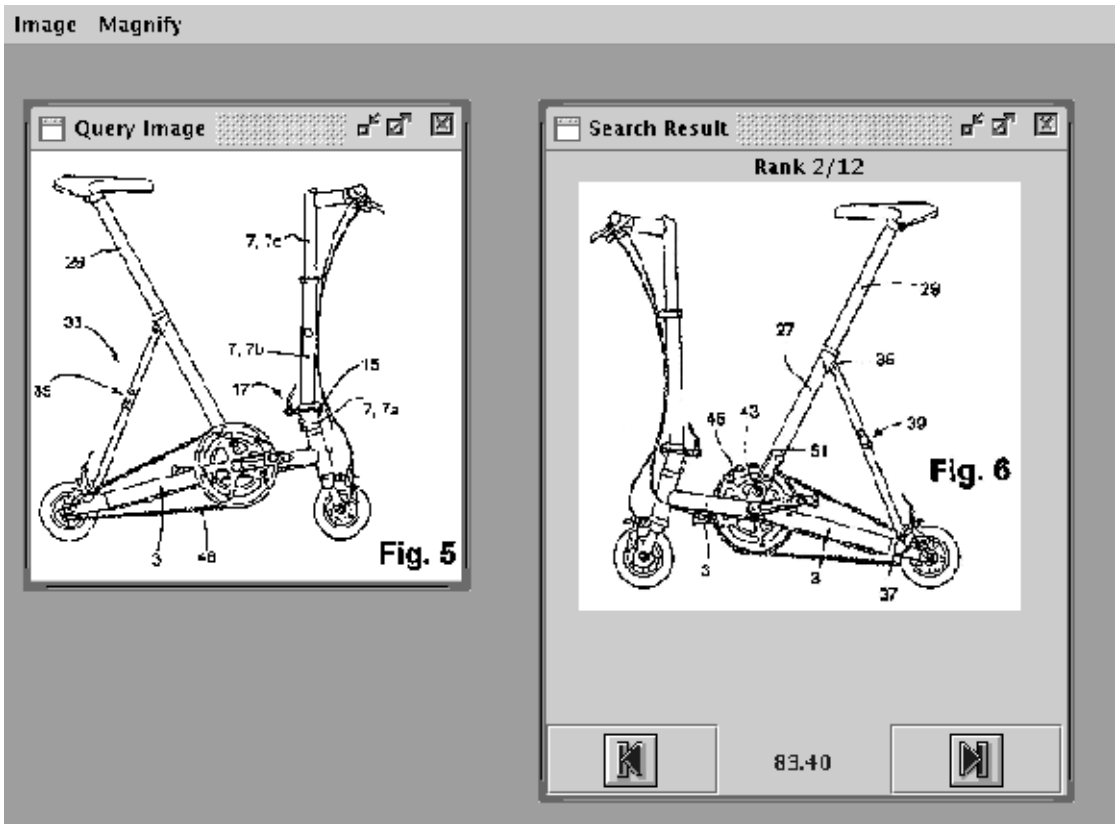


Figure 5. The User Interface for query result navigation

Our database contains approximately 200 images that have been picked up from the patent database of United States Patent Office. For the performance evaluation, we have arbitrarily chosen 15 images from our collection. For each query image, a set of relevant images in the database have been manually identified. An ideal image retrieval system is expected to retrieve all the relevant images. One of the popular measures is *recall rate* [10] that is the ratio of number of relevant images retrieved and total number of relevant images in the collection. The *precision rate* is the ratio of the number of relevant images retrieved and total number of images in the collection.

Minkowski-form distance is used assuming that each dimension of image feature vector is independent of each other and is of equal importance. L1 and L2 (also called Euclidean distance) are some of the widely used Minkowski-form distance measures.

We have compared the performance of two similarity measures, L1 and L2 distance, in our experiments. For our experiment, we calculated precision and recall rate for each image. A recall rate of 100% for 61% of query images and an average 32% recall rate for rest of the images were observed. The precision rate varied between 10% and 35%. Precision rate greatly depend on the number of the images in the database. Graph 1 and graph 2 show the precision rates and recall rates for 15 query images. The image shown in figure 6 is one of the images selected for the query image shown on the left in figure 4 but it is not one of the top ranking images. For the same query image, when we specify the angle of rotation as 180 degrees, the ranking of this image improves considerably.

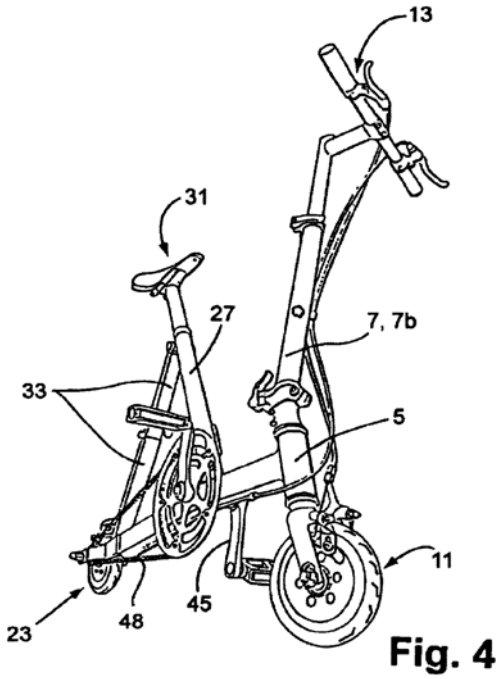
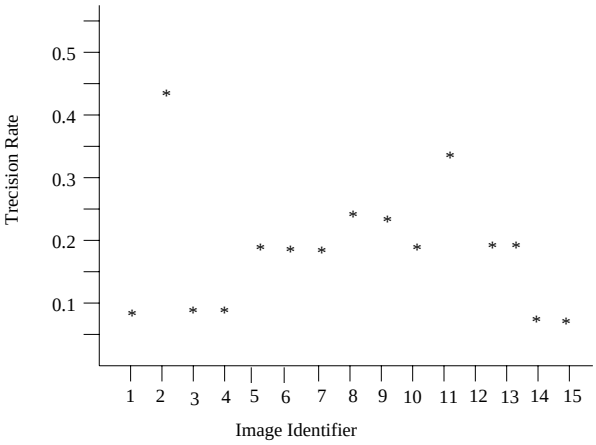
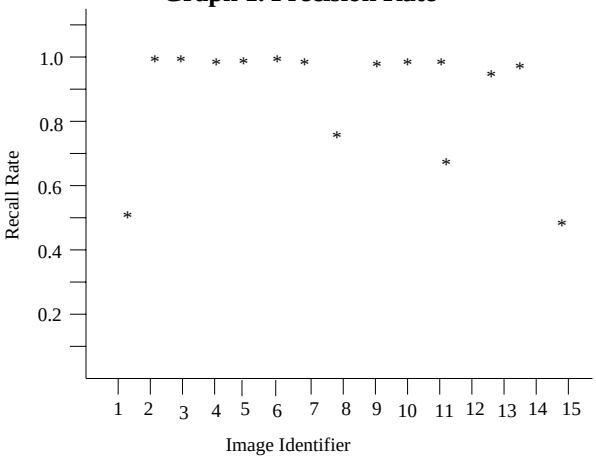


Figure 6. An image obtained by rotating the query image



Graph 1. Precision Rate



Graph 2. Recall rate

5. CONCLUSION

In this paper, we have described PATSEEK that is an image retrieval system for the US patent database. The system has shown good performance and can be effectively utilized to locate similar patents before issuing a new patent by the patent office. A researcher or developer can locate all the images similar to the image in his document before filing for a patent. PATSEEK can compliment a text based search system.

The image feature vector database for PATSEEK is expected to grow and the retrieval system must give real-time performance even when database is large. At present, we have not made any effort to optimize the speed. The total time elapsed from the moment a query image was given and to the moment relevant images are retrieved was about 90 seconds. We plan to use multidimensional indexing to cut down the retrieval time.

REFERENCES

[1] Bach, J. R., et al., "The virage image search engine: an open framework for image management", *Proc. SPIE: Storage and retrieval for Still Image and Video Databases IV*, vol 2670, pp 76-87, 1996.

[2] Canny, J., "A computational approach to edge detection", *IEEE trans. Pattern Analysis and Machine Intelligence*, PAMI-8, pp 679-698, 1986.

[3] Flickener, M., et al., "Query by image and video content: the QBIC system", *IEEE Computer*, vol 28 , pp 23-32, 1995.

[4] Freeman, H., "On the encoding of arbitrary geometric configurations", *IRE Trans. on Electronic Computers*}, vol EC-10, pp 260-268, 1961.

[5] Gonzalez R.C. and Wints P., *Digital Image Processing*, Addison-Wesley Reading, MA, 1992.

[6] Hu, M. K., "Visual pattern recognition by moments invariants", *IRE Transactions on Information Theory*, IT-8, pp. 179-187, 1962.

[7] Jain, A.K. and Vailaya, A., "Image Retrieval using Color and Shape", *Pattern Recognition*, vol 29, pp 1233-1244, August 1996.

[8] Kato, T., "Database architecture for content-based image retrieval in Image Storage and Retrieval Systems" (Jambardino A and Niblack W eds), *Proc SPIE 2185*, pp 112-123, 1992.

[9] Mahmoudi, F., Shanbehzadeh J., Eftekhari A.M., and Soltanian-Zadeh H., "Image retrieval based on shape similarity by edge orientation autocorrelogram", *Pattern Recognition*, vol 36, pp 1725-1736, 2003.

[10] Muller, H., Mullerm, W., McGSquire, D., Marchand-Maillet, S., and Pun, T., "Performance evaluation in content-based image retrieval: overview and proposals", *Pattern Recognition Letters*. vol 22, pp 593-601, 2001.

[11] Persoon, E. and Fu, K. S., "Shape discrimination using Fourier descriptors", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol 8, pp 388-397, 1986.

[12] Zhang, D. and Lu, G., "Review of shape representation and description techniques", *Pattern Recognition*, 37(1), pp 1-19, 2004.