V14

Retrieval of Images from a Library of Watermarks for Ancient Paper Identification

Christian Rauber[1], Peter Tschudin[2], Thierry Pun[1]

1. Department of Computer Science, University of Geneva, 1211 Geneva 4, Switzerland
e-mail: Christian.Rauber@cui.unige.ch
http://cuiwww.unige.ch/~rauber

Telefon: (+41 22) 705 76 33, Telefax: (+41 22) 705 77 80

2. Schweizerisches Papiermuseum & Museum für Schrift und Druck,
St. Alban-Tal 37, CH-4052 Basel
e-mail: chbpm@datacomm.ch

1. Introduction

The importance of electronic publishing, storage and distribution of documents is increasing and will have profound implications for our economy, culture and society. The multimedia digitalisation of libraries and the distribution of the contents of museums is revolutionising these organisations and will make these resources available to a much wider audience than was previously possible.

The main goal of our project is to develop a system for the archival, retrieval, and distribution of electronic documents [1]. Information is accessed via the World Wide Web with a search mechanism that allows the retrieval of text and images according to their content [2].

This system was applied to an existing library. The internationally known Swiss Paper Museum in Basel houses thousands of images of historical papers as well as ancient watermarks [3]. The difficulty for historians is to determine the origin and date of creation of an unknown paper. For this purpose, an efficient method consists of comparing the watermark present inside the paper with another similar known watermark. It is then possible to determine whether this unknown paper comes from the same region and approximately same period as the reference watermark.

The objective of this project is to create an electronic database of known watermarks. This database contains an image of each watermark and a short description of each. The watermark's description consists of the textual characterisation of the paper containing the watermark, the origin, the date of creation, etc. More generally, a paper is described by approximately 150 different parameters [4].

The database is built from images of watermarks directly digitalised from the ancient original documents or from an encyclopedia. A specific digital image processing algorithm is used in order to extract the binary image of the watermark from the original scan.

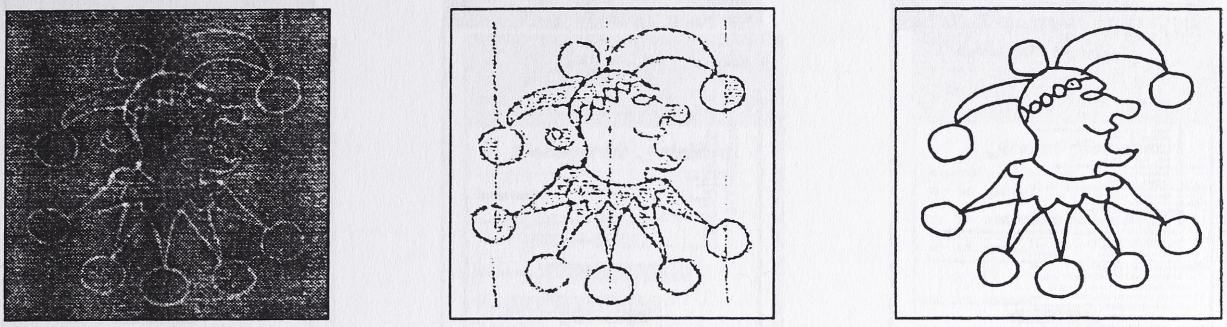
In order to be accessed by a large number of people around the world, this database is accessible via the Internet [5] by using a common browser (such as Netscape or IE from Microsoft). There exists the possibility of adding, removing or correcting a watermark (the image or the attached textual description) and of retrieving watermarks by different means. A watermark can be retrieved in six different ways: retrieval by using the class, the IPH code, specifying global features, comparing similar images, drawing a sketch or using a small pattern.

2. Images acquisition

To create the digital library of ancient watermarked papers, we need to digitise old documents. Each of the documents is scanned at 150 dpi by transparency. Different techniques have previously been used and tested to extract in an efficient way the watermarks from the ancient documents [6][7]. In our case, by using a scanner with a specific light and by using specific imaging processing algorithms, the watermark is extracted from the document in a better manner, see Figure 1.b. Some useless information is still present: the chain lines, the laid lines or some small spots. Some parts of the watermark can be missing too. For these various reasons, preprocessing algorithms are applied in order to remove useless information and to add small missing parts of the watermark.

There exists another way to add watermarks to our digital library. Some authors have traced by hand thousands images of watermarks and published in encyclopedia [8][9]. For example, Briquet has printed in four volumes more than 16'000 im-

ages of watermarks [8]. By scanning these images and by applying image pre-processing algorithms, watermarks can be easily added to the database. See [10] for details on techniques used to extract the laid lines, chain lines and other information from the images.



(a) (b) (c)

Figure 1 : Original input image of an ancient paper. (b) After the digitalisation by transparency. (c) After the pre-processing steps, the image is enhanced.

3. Retrieval by using the class

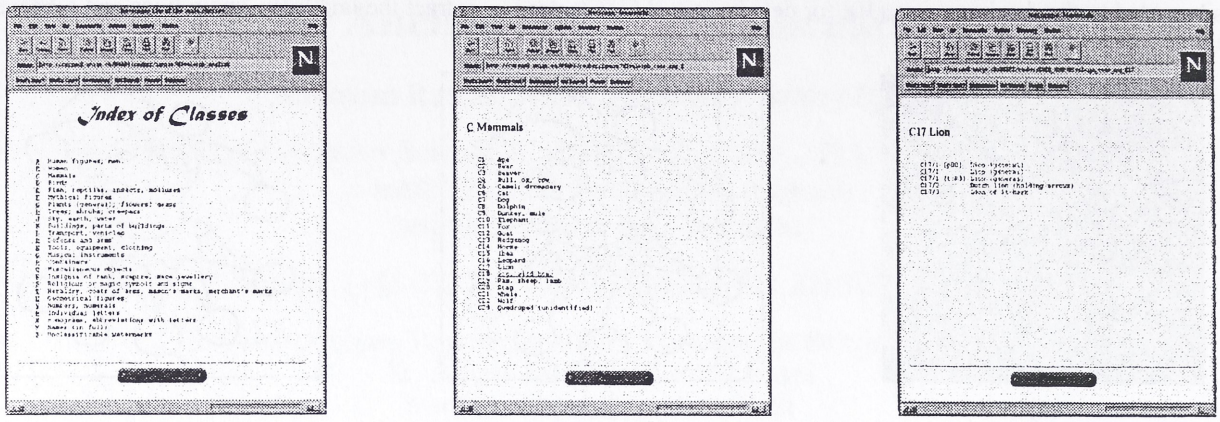
The first method of accessing a watermark is by taking the original classification as presented by Briquet in his encyclopedia [8]. He has classified his 16'000 watermarks into more than 200 different classes. For example, the first two classes are *Agnus Dei* and *Eagle of St-John*. Our digital library of watermarks is composed of approximately 4'000 images of watermarks and is classified into 107 different classes. They can be accessed by using two different pages: the first one consists of selecting textual signification of the desired class (Figure 2.a) and the second way is by selecting the iconic representation of the class (Figure 2.b).

(a) (b) (c)

Figure 2 : Internet main page to access watermarks by using the class as the retrieval mechanism. (a) Textual list of classes. (b) Iconic representation of the classes. (c) Main page for the class Eagle (in French).

4. Retrieval by using the IPH Code

The second way to access watermarks consists of using the IPH code [4]. This code, which was proposed by the International Association of Paper Historians, defines each watermark by a unique textual description and index (such as E8 for the *snake*). One can retrieve watermarks by browsing through the HyperLink Web pages. These different codes are arranged into a tree structure (e.g. Birds→Eagle→double-headed). The retrieval module accepts the supposed description (or code) of the watermark (for example M6 for "crossbow") and the system responds automatically by listing all watermarks corresponding to the IPH code and the verbal description mentioned. The advantage of this second method is the fact that this textual classification is accepted by the international community of historians and there should be no more confusion between different terms representing the same watermark (or the same term used for two different watermarks); for example, by describing precisely the difference between a watermark representing a *dragon* and an other representing a *griffin*. The different pages can be accessed in three different languages: French, German and English. The Spanish language will be added in the next revision of the IPH-code revue by the end of 1997, while Italian is planned for future releases.



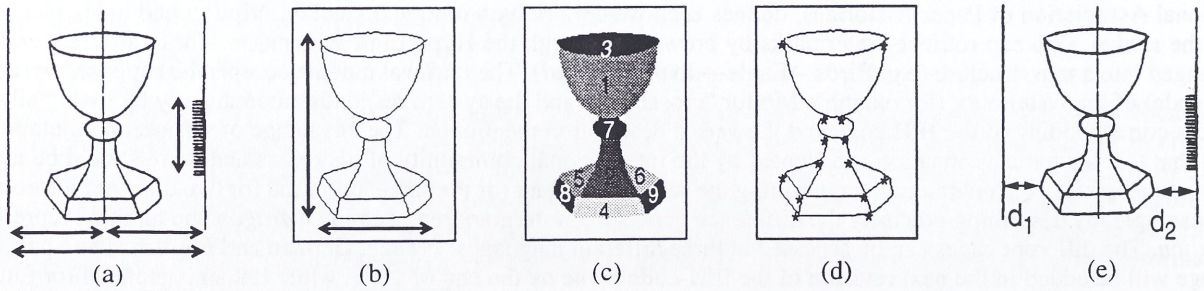
(a) (b) (c)

Figure 3 : Main pages for accessing watermarks by using the IPH code. (a) Main index in French. (b) Sub-class Mammals. (c) Sub-Sub class Lion.

5. Retrieval by specifying global features

A list of morphological characteristics has been set and for each of them an automatic algorithm has been developed to extract these features from the input images. There presently exist about twelve different features [10]:

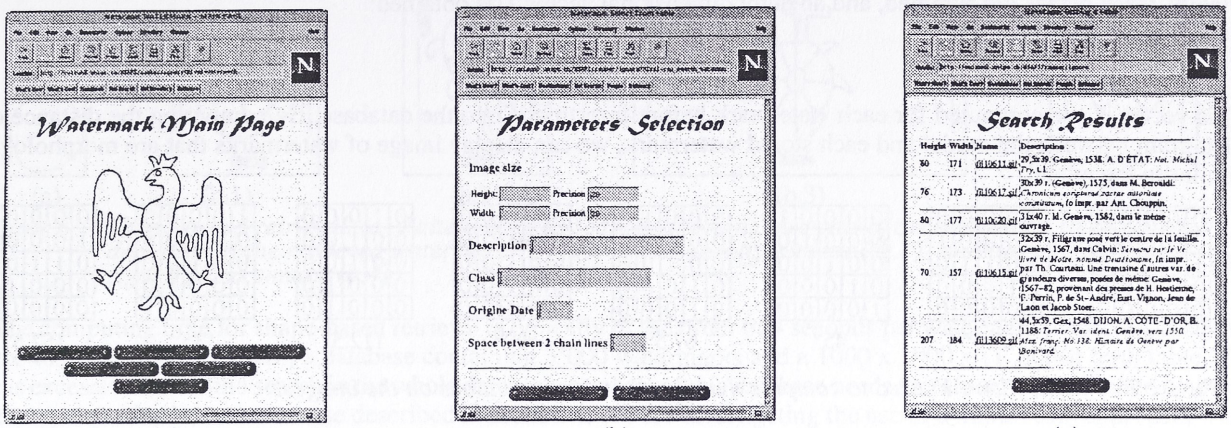
- the size of the watermarks. The height and the width is computed (Figure 4.b).
- the position on the paper. The distance between the left and right side of the watermark and the nearest chain line is given as two measures (Figure 4.e).
- the distance between two chain lines. Each distance is determined automatically (Figure 4.a).
- the number of regions. This feature is very sensitive of the fact that the contours should be closed. For this reason, a mechanism that completes some small missing part of the contour of the watermark is applied before to extract the regions (Figure 4.c).
- the respective location of the regions. The centres of gravity of the three main regions are selected to build a new normalised coordinate frame. The largest region is the origin of the new coordinate frame and the two other regions form the two independent coordinate axes. The remaining regions from the watermark are then described in this new frame by the means of two cardinals. This description provides a morphological representation of the watermark and is invariant to rotation, translation and scaling.
- the number and the position of the junctions. Cross-junctions and T-junctions are extracted from the input image (Figure 4.d).
- the density of laid lines per unit length. This information is related to the quality of the paper (Figure 4.a).
- the percentage of black pixels present in the bounding box defined by the watermark extent. This value allows to assess the complexity of the watermark: older (more ancient) watermarks have less black pixels because they contain less details than recents ones.
- the shape of individual regions, described by the invariant moments as proposed in [11];



(a) (b) (c) (d) (e)

Figure 4 : Global features on watermarks. (a) Space between chain lines and density of laid lines per unit length. (b) Height and width of the watermark. (c) Number of regions. (d) Number and position of junctions. (e) Position of the watermark depending on the chain lines.

The users are prompted to enter values corresponding to the watermark they wish to retrieve (Figure 5.a). This method is very useful for simple primitives such as the width between two chain lines or the size of the watermark. But for more complex features, the parameters to set are nor simple nor intuitive to choose. For these reasons, other approaches have been retained and developed.



(a) (b) (c)

Figure 5 : (a) Main page. (b) Page that allows the setting of the different values. (c) Result page returned by a query.

6. Retrieval by comparing similar images

For historians, the greatest utility of this database is to retrieve watermarks that are similar to a model. For this purpose, we have developed a similarity task processing algorithm. The user is only required to present a watermark normalised with respect to size and the system retrieves similar watermarks by comparing the watermarks' similarities in shape. There are two different algorithms and they can be constrainted by size or by IPH class.

Circular histogram algorithm:

The first algorithm used to retrieve a similar image of a watermark model consists of computing a circular histogram around the centre of gravity of the watermark (Figure 6.a). In Figure 6.b, each bar corresponds to the number of pixels present inside a quadrant. We have chosen to split the circle into 16 equal parts. The histograms are computed once for each existing watermark and stored inside the database. Two watermarks are globally similar if their respective histograms are similar. The resemblance $d(H_1,H_2)$ between two histograms is computed as following:

$$d(H_1, H_2) = \sum_{i=1}^{16} |H_1[i] - H_2[j]|$$

The result of the similarity retrieval is the first n watermarks where the distance $d(H_1,H_2)$ with the model H_1 is the smallest.

(a) (b)

Figure 6 : Graphical representation of the circular histogram. (a) The centre of gravity of the watermark is used as the middle point to compute the circular histogram. (b) Representation of the histogram computed from the figure (a).

Directional algorithm:

The second algorithm consists of filtering the input image by height directional filters:

$$G_j(x, y) = I(x, y) \cdot F_j \qquad j = 1..8$$

Figure 7 displays the eight filters F_j used to computed G_j. After this first operation, eight new planes are obtained. We compute the eight $K(x,y)_j$ planes by taking the highest value from these last planes.

$$K_j(x, y) = \begin{cases} G_j(x, y) & \text{if} \quad G_j(x, y) = max_j(G_j(x, y)) > 1 \\ 0 & otherwise \end{cases}$$

Finally, these planes are summed, and an eight dimensional vector T_j is obtained:

$$T(j) = \sum_{x, y} K_j(x, y)$$

This vector is pre-computed for each watermark before inserting it into the database. By computing the distance of the histogram between the model and each stored watermark, we can display image of watermarks that are morphologically similar to the model.

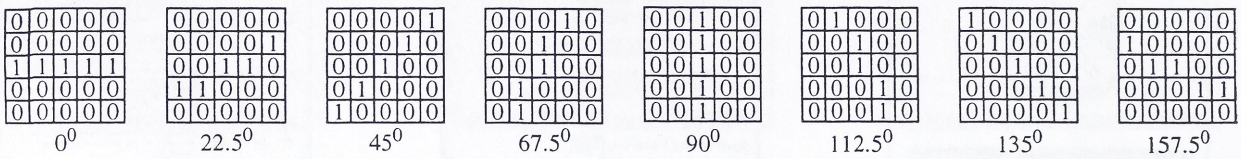
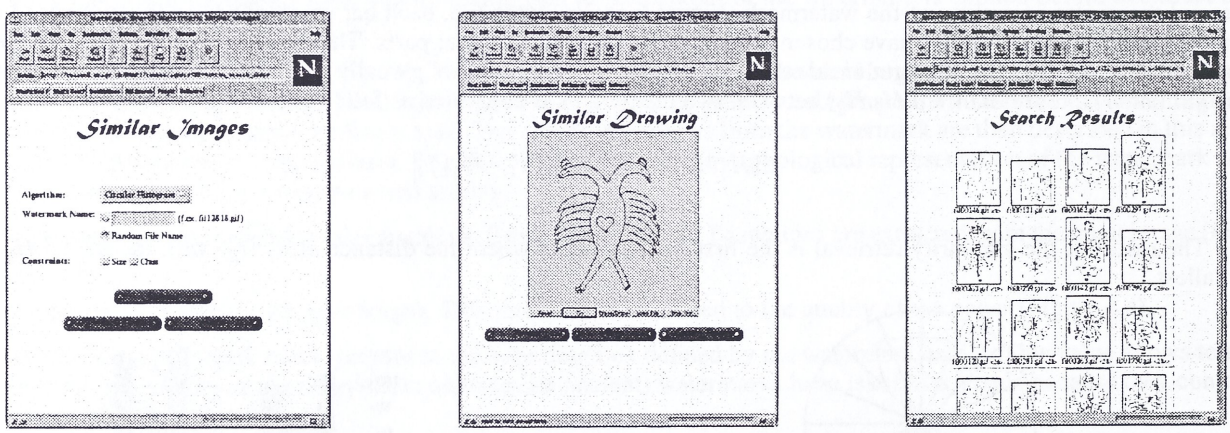


Figure 7 : Eight filters F_j used to compute the directional informations on the images.

7. Retrieval by drawing an approximate shape

If a historian does not have access to an image of the watermark to retrieve, the possibility exists of manually drawing an approximate shape of the watermark. This approximate shape is then used as the model for retrieval by similarity. A Java-drawing interface is provided to allow the sketching of the contour directly from the browser (Figure 8.b). With this interface, it is possible to load a local image of watermark and to apply some image pre-processing algorithms. Erosion, dilation with connectivity four or eight can be applied. A thinning algorithm is present and a module to add small missing parts of the contour of the watermark can be used. Finally, small spots or noise can be automatically removed.

After having drawn the model, the same distance algorithm is used in order to compute the shape representation of the hand-drawn sketch image. The histogram is calculated localy and the time complexity depends on the computer used. For a Pentium 200 computer, and an image of size 400 by 400 pixels, the algorithm takes about 30 seconds to extract features and 15 seconds to transfer data.



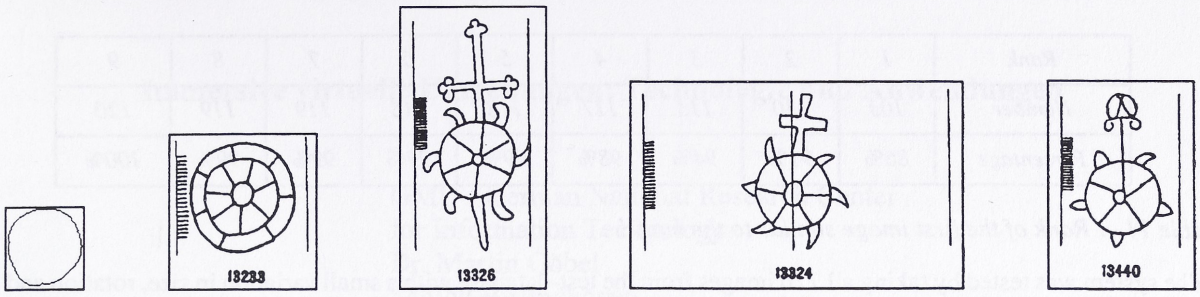
(a) (b) (c)

Figure 8 : (a) Main page of the similarity retrieval module. (b) Drawing applet interface to draw a similar shape. (c) Result of similar images.

8. Retrieval by using a small pattern

It was shown that the whole shape of a watermark is a very important information for accessing and retrieving a given watermark. In some cases, the watermark is not complete or only a small part of the watermark can be distinguished. For these reasons, our system offers the possibility of accessing watermarks by means of patterns corresponding to a small part of a watermark. Rather than using a correlation type approach, which would be prohibitive in terms of computing time, a hashing mechanism has been developed where all watermarks are indexed via a bi-directional table whose access points are the watermarks themselves. In order to retrieve a similar watermark, a sort of "convolution" is applied on the hash table with the search pattern.

This matching algorithm yields an ordered list of the most similar watermarks, and includes the possibility of retrieving watermarks having only a part of a pattern in common with the one specified at query time (Figure 9). Shape and size variations are therefore allowed with respect to the model.



(a) (b.1) (b.2) (b.3) (b.4)

Figure 9 : Shape-based retrieval. (a) Pattern used as search criterion, here part of the watermark (b.1). (b) Ordered list of the retrieved watermarks: on the left is the most similar ((b.1), similarity 100%), on the right the less similar (b.4).

The computing time for shape-based retrieval is typically of the order of 4 seconds per watermark inspected, assuming a 42 x 48 pixels search pattern, a database containing 3'000 watermarks and a 1000 x 1000 x [1..3000] hash table. This time is excessive for a blind search in the whole database; a subset of this database needs therefore to be specified by means of other query criteria, such as those described in section 4. We are investigating the use of a multi-scale approach, where the initial search would be performed at the largest scale only; only the most promising candidates would be inspected at the finest scale.

This matching algorithm has been implemented on a parallel computer (IBM 9076 Scalable POWERparallel System, SP 2, with 14 processors). Assuming the same size as previously (42 x 48 pixels) for the search pattern with a database containing 3'000 watermarks, a computing time is obtained of the order of 13 ms per watermark (speed-up factor of 300)[12]. This last algorithm is not yet accessible through Internet due to the fact that a local specific transputer is used.

9. Results

In order to test our system and algorithms, twelve different classes have been chosen from our data base and each of them contains ten similar watermarks. Before computing performances and robustness of the retrieval mechanism, we need to define five terms:

S :Total number of images in the database.
I :Total number of images that the system should return ($I \leq S$).
P :Number of images returned by the system ($P \leq S$).
R :Number of correct images returned ($R \leq I$ et $R \leq P$).

A retrieval result can then be characterised by these five values:

Correct match : the number of images returned that are similar to the model by the total number of images returned by the system: R/I.
False match : the number of image classified as similar to the model but that are in reality not similar: $P-R\,/\,(S-I)$.
Correct non match : the number of images that are not similar to the model and that are not returned: $S-I-(P-R)\,/\,(S-I)$.
Missed match : the number of similar image that are not returned by the system: $I-R/I$.

Figure 10 displays these four measures for the two algorithms used to retrieve similar images. Each image of this test-database is used as a model (i.e. 120 models) in order to compute the average number of images retrieved for a given model and a given P. With these measures, we can conclude that in average, if the system returns thirteen images, there are eight images that are really similar to the model (for the circular algorithm).

	Circular algorithm		*Directional algorithm*	
	P=10	***P=13***	***P=10***	***P=13***
Average R	*6.9*	*8.0*	*7.9*	*8.3*
Correct match	*57.5%*	*80%*	*65.8%*	*83%*
False match	*2.8%*	*4.5%*	*1.9%*	*4.3%*
Correct non match	*97.2%*	*95.5%*	*98%*	*95.7%*
Missed match	*31%*	*20%*	*21%*	*17%*

Table 10 : Estimation of the results for P=10 and P=13.

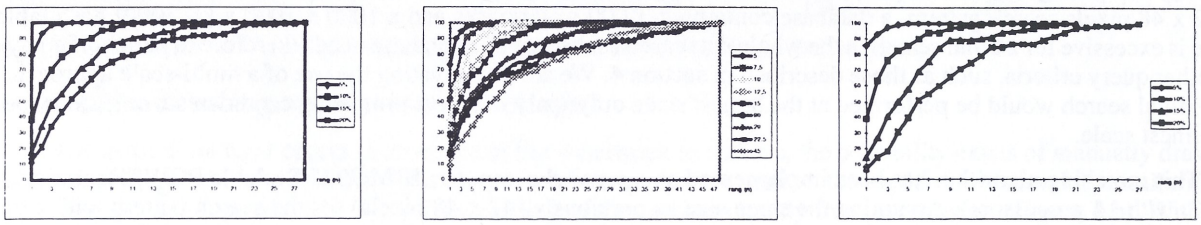
The system was tested by computing the rank of the first image that belongs to the same class as the model (see Figure 11). We can conclude that on an 86% basis, an image is retrieved that is similar to the model in the first position, i.e. that the distance between the two images is minimal. If the system returns nine images, then it is certain that, at the minimum,

there is one image similar to the model.

Rank	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*
Number	*103*	*110*	*113*	*117*	*118*	*118*	*119*	*119*	*120*
Percentage	*86%*	*91%*	*94%*	*98%*	*99%*	*99%*	*99%*	*99%*	*100%*

Table 11 : Rank of the first image similar to the model.

The system was tested by taking all 120 images from the test-database with a small variation in size, rotation and SNR. The goal was to retrieve the original model without transformation. The Figure 12 displays the number of images the system should return in order to find the original model depending on the variation. We can see (Figure 12.a), that if the image is scaled by a factor of 10%, then 40% of the original models are still found in the first position (all models are found in the first 12 images). If 20% of the pixels that compose the watermark are removed (Figure 12.c) then 85% of the models are found in the first 13 images.



(a) (b) (c)

Figure 12 Test of the robustness in (a) size, (b) rotation and (c) SNR for the directional algorithm.

10. Conclusion

A global system for distributing documents containing textual and pictorial data has been created. Textual and visual information is accessible using a complete and efficient system available via the global network Internet.

The current database, that contains approximately 4,000 images of historical watermarks allows the searching and retrieving of a specific document on the basis of global features, textual criteria and/or morphological measures. The next goal of this project is to integrate a digital watermarking algorithm and a security module of our system. A payment module should complete the system that will allow the Swiss Paper Museum to benefit from this work. Ultimately, the whole system associated with an european project should manage a database of approximately 600,000 different watermarks (images, descriptions and features).

11. References

[1] C. Rauber, J. O. Ruanaidh, T. Pun, "Secure Distribution of Watermarked Images for a Digital Library of Ancient Papers", ACM DL97, Second ACM International Conference on Digital Libraries, Philadelphia, july 23-27, 1997.

[2] C. Rauber, P. Tschudin, S. Startchik and T. Pun, "Archival and retrieval of historical watermark", IEEE Signal Processing Society, ICIP 1996 International Conference on Image Processing, Lausanne, Switzerland, Sept. 16-19, 1996.

[3] P. Tschudin, "Papiergeschischte als Hilfswissenschaft", Sonderdruck aus "Das Papier", 37, Jahrgang, Heft 7, pp. 285-295, 1983.

[4] International Association of Papers Historians - IPH, International Standard for the Registration of Watermarks, Provisional Ed., P.F. Tschudin, Ed., Riehen, Switzerland, 1992.

[5] Address for accessing the digital library of watermarks: `http://cuisun8.unige.ch/NSAPI/rauber/watermark`. Enter with the login: `guest_watermark` and with the password `guest_watermark`.

[6] D. Stewart, R. A. Scharf, J. S. Arney, "Techniques for Digital Image Capture of Watermarks", Journal of Imaging Science and Technology, N. 30, pp. 261-267, 1995.

[7] P. Zamperoni, "Wasserzeichenextraktion aus digitalisierten Bildern mit Methoden der digitalen Bildsignalverarbeitung", Das Papier, 43, Jahrgang, Heft 4, pp. 133-143, 1989.

[8] C. M. Briquet, "Les filigranes", Dictionnaire historique des marques de papier dès leur apparition vers 1282 jusqu'en 1600, Tome I à IV, Deuxième édition, Verlag Von Karl W. Hiersemann, Leipzig, 1923.

[9] F. Del Marmol, "Dictionnaire des filigranes classés en groupes alphabétique et chronologiques", Namur : J. Godenne, 1900. - XIV, 192 p., 1987.

[10] C. Rauber, P. Tschudin, S. Startchik and T. Pun, "Archivage et recherche d'images de filigranes", CNED'96, 4ème Colloque National sur l'Ecrit et le Document, Nantes, 3 juillet 1996.

[11] M. F. Zakaria, L. J. Vroomen, P. J. A. Zsombor-Murray, J. M. H. M. Van Kessel, "Fast algorithm for the computation of moment invariants", Pattern Recognition, Vol. 20, No 6, pp. 639-643, 1987.

[12] J. Raki, "Parallélisation de la recherche de filigranes", Mémoire de licence, Université de Genève, Faculté des Sciences, Département d'Informatique, 90 pages, Février 1997.