Secure Distribution of Watermarked Images for a Digital Library of Ancient Papers

Christian Rauber, Joe Ó Ruanaidh, Thierry Pun[1]
Department of Computer Science, University of Geneva, 1211 Genève 4, Switzerland
Christian.Rauber@cui.unige.ch

http://cuiwww.unige.ch/~rauber

Abstract

The electronic publishing, storage and distribution of documents is growing increasingly important 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 MEDIA project (Mobile Electronic Documents with Interacting Agents) is the development of a system for the archival, retrieval, and distribution of electronic documents. For this purpose, a mobile agent platform is used to securely distribute these documents. Information is accessed by a search mechanism that allows the retrieval of text and images according to their content.

An important feature of the system is a *digital watermarking tool* which embeds hidden signatures in images. This provides copyright protection and helps to ensure that the image will not be copied and sold and without proper authorisation.

The management of the database of documents and images is accomplished by an extensible object relational database management system. In addition, documents and data can be accessed through the World Wide Web network.

1. This work is supported by the Swiss Priority Programme for "Information and Communications Structures", under grant nr. 5003-045334.

1. Introduction

Documents are becoming more and more available in digital form and moreover accessible through the World Wide Web. These documents are stored in databases and contain a mixture of text and pictures. Automatic retrieval of text is becoming increasingly common but commercial systems that contain text retrieval and image retrieval mechanisms are still uncommon.

With the recent development in multimedia technology, information systems are required to manage and distribute a growing mass of data. To access new information, an efficient query tool is a critical aspect of the entire system.

The MEDIA project aims at developing a complete system for the archival, retrieval, and distribution of electronic documents over the Internet. The major features of MEDIA will include:

- multimedia storage of text and images, with hyperlinks relating these documents;
- query mechanisms for document retrieval by means of textual and visual descriptors;
- secure mechanisms to distribute information, with in particular copyrighting protection methods based on digital watermarking;
- user-friendly interface through the Internet;
- digital cash service to pay for accessed information.

The MEDIA project is composed of three interdependent subprojects: HyperNews, Asap and Krypict. The first two are briefly described in this section, while Krypict is detailed in sections 4 to 6.

The HyperNews sub-project is directed towards the development of a system for the commercial distribution of an electronic magazine over the World Wide

Web [8]. HyperNews is conducted in collaboration with the Swiss weekly magazine "L'Hebdo", which will be made accessible electronically. The HyperNews system will help the reader to reduce the time spent searching for information items of personal interest. This will be accomplished by means of individual user profiles allowing one to build virtual personal newspapers tailored to one's preference. Furthermore, the electronic articles will be dynamically updated when new pieces of information become available.

The objective of the second subproject ASAP (Agent System Architectures and Platforms) is to realise an execution platform that will support the development of applications based on mobile software agents [19]. Each agent is a self-contained autonomous unit capable of moving around heterogeneous networks and interacting with both other agents and local services. In our environment, these agents are assigned to the distribution of documents around the world using the Internet.

The third of the MEDIA subprojects, KryPict, aims to provide an image database system that allows query by content [15] as well as a means of copyright enforcement through digital watermarking [11][12].

3. Global system architecture

Figure 1 shows the overall client-server architecture of the MEDIA project. The client side (top level of the figure) contains the interface between the user and the MEDIA server. It is based on a common browser, such as Netscape. The interface is built with Java Applets and HTML tags. This interface is in communication with a mobile agent module that controls the integrity of the data and communicates with the electronic payment module in order to obtain and control the incoming electronic cash of the users for the specific documents that have been accessed.

The flow of information is then propagated through the network and is captured by the server agent. This agent controls the data and transfers orders to the information retrieval part of the system.

Queries are then dispatched to the two main modules: the hypertext retrieval module (the HyperNews subproject) and the image query tools (the KryPict subproject). The hypertext retrieval module filters the information according to the customer's profile and therefore provides a drastic reduction of time spent on retrieving interesting specific information. The HyperNews retrieval system also establishes HyperLinks between intra- and cross-references of the newspaper information topics. The customer can build his/her own personalised electronic hypermedia newspaper by selecting his/her topics of personal interest.

The image query tools and the encryption mechanism of the KryPict subproject are detailed in the following sections.

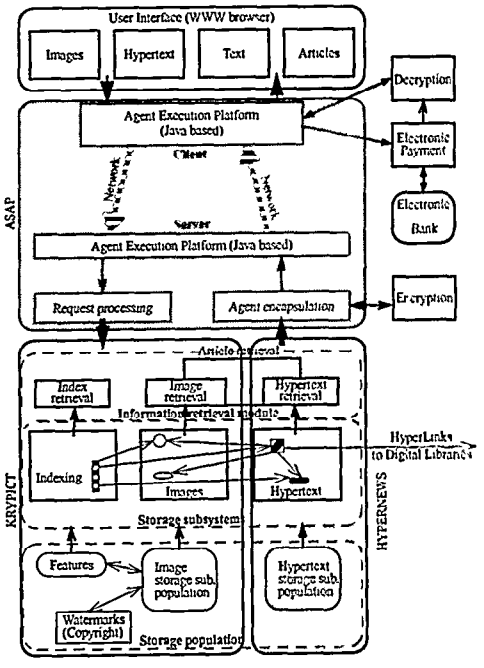


Figure 1: Overall architecture of the MEDIA project.

4. KryPict

Nowadays, a wide range of techniques has been developed to retrieve images depending on their content [20]. Sometimes these systems are tuned to specific application domains (f. ex. in geography [16] or in medical domains [5]) but some systems provide basic content based image retrieval tools of wider applicability [14]. Notable examples include the Query By Image Content system developed by IBM [4] or the Iris system that can deal also with videos [1]. For image retrieval, these systems use combinations of colour, shape and texture for specifying queries. Typically, the aim of these different systems is to retrieve an image from the database if it contains characteristics specified by the user. Features can be standard and simple characteristics such as the size of the image, the colours of different regions or more complex such as a sketch drawn manually by the user. Another common way to retrieve a picture is to query using an example image. An important point to note in this last method is that one is not

necessarily seeking an exact match. Instead, the goal is to retrieve images that are similar to a given query image. There are a few commercial systems that support retrieval of images by pictorial specification and comparison [21]. It has been shown that if one desires quality and efficiency in image retrieval techniques based on image understanding, then these systems must be application dependent.

The main goal of the KryPict subproject is to develop a system for handling pictorial documents and for copyright protection. This database pictorial module will be capable of containing images and their associated textual descriptions. Krypict will include mechanisms allowing one to retrieve images depending on their content. Copyright enforcement will be implemented using digital watermarking methods that are integrated with the pictorial database management system, allowing the user to encrypt his/her own images with a specific signature. This system will therefore make it possible to distribute documents and to make images accessible over the network without having to fear that these documents be copied.

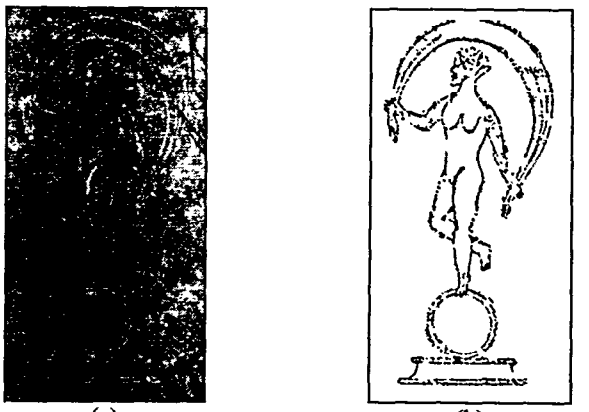
One of the databases we use to test and develop our system comes from the internationally known Swiss Paper Museum in Basel that houses images of historical papers as well as ancient watermarks. This database grows by the day and the number of known watermarks is currently approximately of the order of 600,000.

4.1 The original papers

It is very difficult for historians to work with original ancient documents. The first reason is that they are fragile and delicate. The second reason is that because of the rarity of these papers, they are confined to libraries, townhalls or museums. Often, these onerous and exceptionally valuable documents cannot leave these establishments. The advantage of working with digital documents becomes evident: it is not necessary to go to a specific museum to inspect and study a particular document, on the contrary, an historian can stay in his office and examine a digital paper on his computer screen by accessing the digital library server through the network. In addition, different specialists can study the same virtual digitized document at the same moment in different places.

In order to obtain a digitized document from an ancient document, the paper has to be scanned three times: from the front (Figure 2.a), from the back and by transparency (Figure 2.b). The last scan allows one to extract the information which is contained in the paper [17][18]. For example the laid lines that mark the paper with fine horizontal lines (due to the wire mesh that was used to hold the paper during its fabrication), and the chain lines that mark the paper with fine vertical white lines (due to the thicker wires that were used to support the laid lines). In fact, the most important information that is needed to determine the origin of an ancient paper is the watermark hidden in the paper.

This specific signature was already present in papers as early as 1282 in medieval Europe and is of important historical value. The pattern that composes a watermark reflects the evolution of the commercial and cultural exchanges that took place between cities in the Middle Ages. With this old copyright signature it is possible for historians to determine the origin and date of creation of an ancient unknown paper by comparison with known documents bearing a similar watermark. For these reasons, a digital database is being constructed with old documents associated with a textual description that contains detailed information about the paper (color, texture, format, composition, etc.), the place of origin, the date of creation and the watermark (shape, size, position, etc.). Figure 2.a is an example of a part of an ancient document (only the front of the paper is presented) and Figure 2.b is the representation of the watermark extracted from the transparency digitalisation of Document 2.a.



(a) (b)

Figure 2: (a) Original scanned image of an ancient paper. (b) After the semi-automatic extraction process.

4.2 Preprocessing step

Prior to archiving images in the database, a series of operations is necessary to transform the coarse images of watermarks resulting from the scanning process into clear images that can then be used for archival and retrieval. The following sequence of semi-automated

steps allow one to extract more efficiently the features required to compute the different indexes that define each image:

- global contrast enhancement;
- contour enhancement;
- grey-level inversion;
- removal of the written annotations;
- removal of the chain lines and laid lines;
- completion of the watermark shape if some small parts are missing;
- reduction of the size of the image by a factor of two.

Figure 3 shows three examples where this procedure has been applied to various images. After this main preprocessing step, it is now possible to extract global primitives and local features in order to define automatically the particularities of each watermark. With these disparate simple primitives, it will be possible to retrieve rapidly one or more watermarks that correspond to or most resemble a known watermark. By comparing watermarks, it should then be trivial to determine the origin and date of creation of an unknown document.

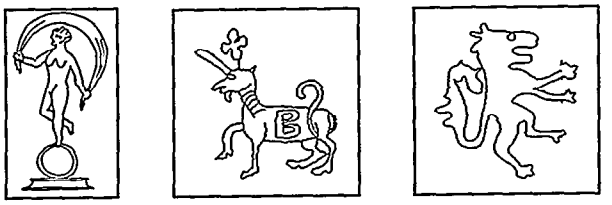


Figure 3: Images of a watermark after the preprocessing steps.

4.3 Classification

In the past, historians have classified their watermarks by identifying some simple characteristics. For example, Briquet in [2] has classified about 16,000 images of watermarks in about 80 textual classes (f. ex. the *Eagle* or the *Bottle* class). In [3], Del Marmol has sorted watermarks according to their date of creation.

However, in many cases, it is very difficult to classify a watermark. It can be incomplete, unidentifiable or of unknown date. In addition, a given watermark can be classified into more than one class, such as the drawing representing a *Cross on a Bear* which could be stored in the *Bear* class or in the *Cross* class.

By using a textual query on the description that is stored with each watermark it is possible to retrieve a given watermark in less than one second on an ordinary personal computer. There are however two major limitations to such queries: the textual descriptions are subjective (despite the existence of a standard set by the community of historians [7]) and are incomplete (some patterns that compose a watermark are not always described in the text).

For these two main reasons, queries based on the content of the images, such as local or global morphological characteristics, are essential [10][9].

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 images. There actually exists about twelve different features, such as the size of the watermarks, their position on the paper, the distance between two chain lines, the position of the regions that define the watermarks, etc. Figure 4 shows some examples of the characteristics that were used.

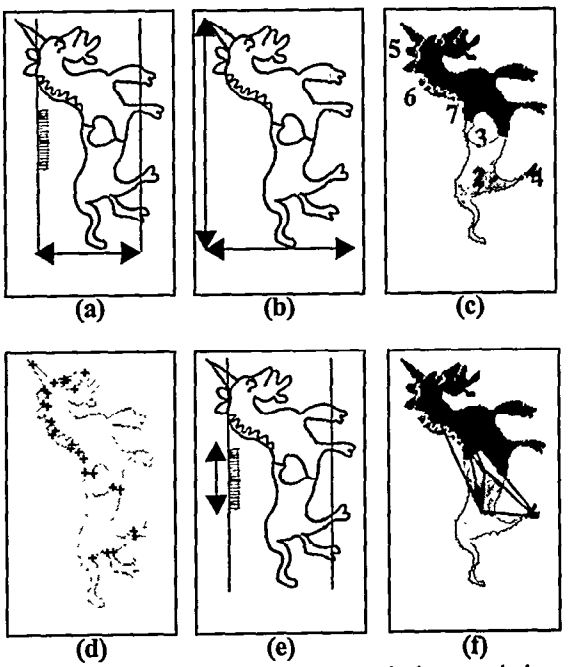


Figure 4: Automatically computed characteristics of watermarks. (a) Space between chain lines. (b) Height and width. (c) Number of regions. (d) Cross junctions. (e) Density of laid lines per unit length. (f) The respective locations of the regions.

4.4 Results

The system is capable of retrieving images depending on different features:

- textual description associated with each image;
- precomputed characteristics based on the global documents (f.ex. space between laid and chain lines);

- characteristic of the ancient papers (size, texture, origin, etc.);
- features computed from the watermarks (number of regions, complexity of the watermarks, etc.);
- morphological shape of the watermark (patterns corresponding to part of or to entire watermarks).

At retrieval times, queries can be made according to one or more of these disparate characteristics. Textual queries and simple characteristic queries are processed very rapidly (less than one second) and therefore can be made interactively by the users. On the other hand, retrieving images on the basis of on the shape of the watermarks is a complex operation. In order to retrieve similar watermarks, a convolution-like operation is applied between the search pattern (of size w x h) and a bidirectional hash table (of size H x W). The size of the bidimensional table typically is of order of W=1000 and H=1000. The complexity of a request is then of the order of O(WHwh).

The computation time of this retrieval method is of the order of four seconds per watermark inspected on a Sun Sparcstation 10. For this reason, this matching algorithm has been implemented on a parallel computer (IBM SP2 system with 14 processors). On this computer the matching algorithm retrieves similar pattern with a computation time of 13 ms per watermark. Finally, in order to use this parallel algorithm on a common computer, PVM (Parallel Virtual Machine) libraries and mechanisms were employed to allow communication of data and results between this specific machine and other common workstations.

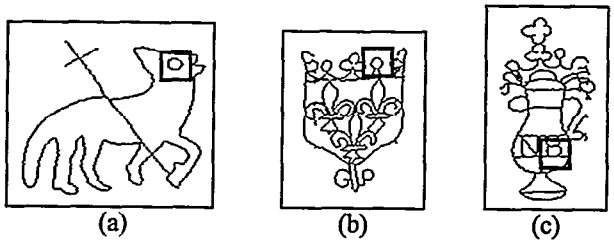


Figure 5: Results of shape-based retrieval: The pattern used for this query is given by the surrounded circle. Twenty images have been retrieved with this small circle pattern. Only the first three are displayed.

5. Storage and distribution

Various libraries are now accessible through the Internet by using a common browser. This allows one to retrieve information from anywhere around the world. The interface to such system is an important component for the users: it should be easy to use, attractive and provide fast responses to queries. On the other side of the system, the provider side, it an efficient database management system is essential to provide access to information in a reasonable amount of time.

5.1 The interface

We have developed an interface that is accessible through the network using a Netscape browser [13]. For visitors, permission to add, edit or remove a watermark is disabled. The database currently contains about 3,000 images of watermarks and in most cases a textual description is attached.

A retrieval page has been built for searching a list of watermarks based on similar features. There are currently six different access primitives: the height and the width of the watermark, the date of origin, the space between two chain lines, the class name and, finally, a textual primitive based on the description. These various primitives can be combined in order to obtain a list of similar watermarks. For example, one can request the database to retrieve watermarks with a width of approximately 170 pixels (2.5 cm) and with originate from "Geneva". The result of this query is displayed in Figure 6.b, where five watermarks have been found satisfying this description.

Another approach for images retrieval is based on similarity between images of watermarks. The user is only required to present a normalised watermark (model) and the system retrieves similar watermarks by means of the watermarks' global similarities in shape. The similarity task processing algorithm is based on the global outline of the watermarks. Furthermore, the historians has the possibility of manually specifying the model by drawing its approximate shape. A Java-drawing interface is provided to allow the user to sketch this model directly on the browser.

There is another way to access watermarks which is to use the IPH code. This code, which has been proposed by the International Association of Paper Historians [7], defines each watermark by a unique textual description and index. One can retrieve watermarks by browsing through the HyperLink Web pages. These different codes are arranged in a tree structure (e.g. Transport→Ship→ motor ship). The retrieval module accepts the supposed description (or code) of the watermark (for example "bottle") and the system responds automatically by listing all IPH pages that contain this word.

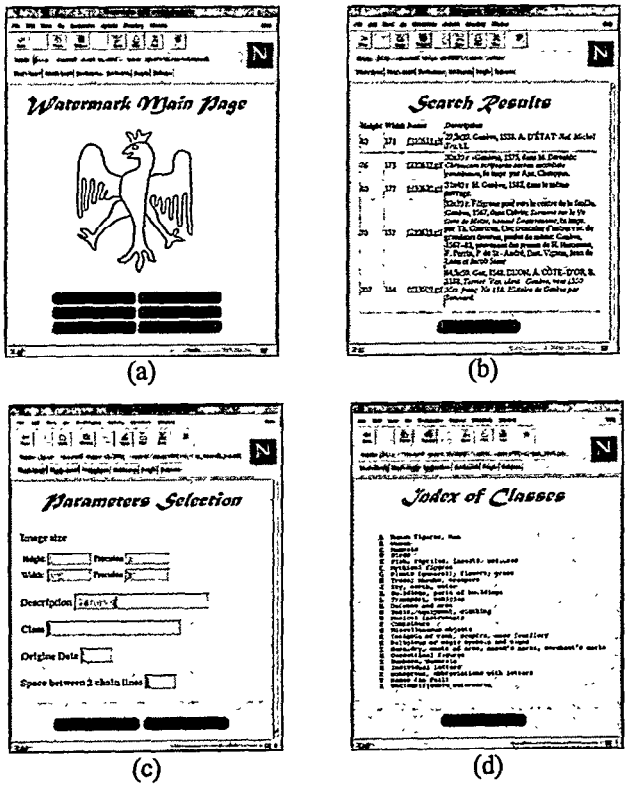


Figure 6: (a) Main page for accessing the digital library of watermarks images. (b) Result of a retrieval query. (c) Retrieval main page. (d) Index of classes for the IPH code.

5.2 The database

Images and textual descriptions are stored in the object relational database management system (ORDBMS) Illustra from Informix. This database allows one to store images, text files (such as HTML pages or Word documents) as well as video or audio data. Illustra also allows the creation of new access methods as well as of new data structures (named "DataBlades"). Consequently, we have developed a specific DataBlade for our historical watermark application (Figure 7). In addition, another DataBlade has been developed to access this database directly from the WWW through a Netscape interface.

The database for the Swiss Paper Museum currently contains two main data structures: the watermark structure and the IPH code structure. For each watermark, the black and white watermark image (which in average requires about 30 kilobytes), the front and the back scan of the original paper where the watermark was found, the different features that describe the watermark and the textual description of the ancient paper and of the watermark are stored. Even without the two original scans of the paper about 100 kilobytes of storage is required for each new watermark. Currently, more than 3,000 watermarks are stored in our database.

The second structure used in the database is the IPH code. This code make it possible to obtain a unique index key for each of the different watermark shapes. For example, the Eagle of St. John has the code D5/3 and is categorised under the main class Bird (D) and the subclass Eagle (5). In order to be used by international historians, these codes are defined in three different languages: in French, German and English. The IPH codes are stored in the database in HTML page format. There are more than 350 pages for the description of these codes. It is important to note that all this information is available using the Netscape browser.

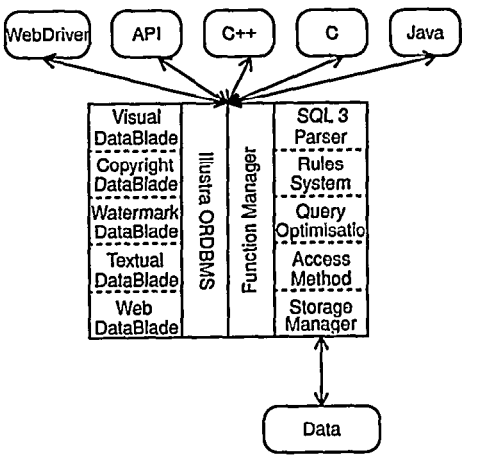


Figure 7: Overview of the storage catalogue sub-system.

6. Digital image watermarking

A major impediment to the use of electronic distribution and storage is the ease with which electronic images and documents can be intercepted, copied and redistributed in their exact original form. As a result, publishers and art houses are understandably reluctant to use this means of disseminating material. The commercial possibilities for the World Wide Web are steadily becoming more appreciated. However it is clear that in order for these possibilities to be realized that an integrated approach is required for the secure handling, issue and duplication of issued documents.

Public key encryption systems such as the RSA algorithm [6] do not completely solve the problem of unauthorised copying because of the ease with which images may be reproduced from previously published documents and enhanced using software such as "Corel

Draw" or "Adobe Photoshop". All encrypted documents and images need to be decrypted before they can be inspected or used. Once encryption is removed the document can be passed on in an electronic form. If there is more than one recipient of an image, there is no direct proof that any particular authorised recipient is responsible for passing it on to unauthorized users.

The idea of using an indelible digital watermark to identify uniquely both the source of an image and an intended recipient has therefore stimulated much interest in the electronic publishing and printing industries. The aim of digital image watermarking to hide robust invisible labels inside an image.

The task of detecting and decoding the mark could (and should!) be regarded as a problem in digital communications. For several decades a communication technology known as spread spectrum communications has existed to carry out secret military broadcasts using very low amplitude signals. These signals take the form of pseudo-random sequences to blend in with natural background noise and thus avoid detection. In digital watermarking the scenario is somewhat similar. The aim is to communicate a relatively small amount of information with limited power. Ideally the embedded mark should merge with the image features and be visually imperceptible. In addition the embedded mark must be secure and resistant to attacks (such as cropping, low pass filtering, translation or re-scaling).

We have developed various perceptually adaptive methods for watermarking images based on a frequency transform domain representation of a digital image. One approach which we briefly describe here has proven to be extremely robust to lossy image compression methods such as JPEG.

The novelty of this approach lies in the fact that the watermark is embedded in the most significant frequency component of the image. The algorithm is loosely based on the JPEG image compression algorithm itself. The image is divided into blocks and each block is mapped into the transform domain using a Discrete Cosine Transform [6] (other domain transforms have also been tested and also work well). The watermark is embedded in the form of a pseudo random sequence (a spread spectrum signal). The watermark can be expressed as a binary bit stream. The bits of which can be grouped together to form symbols. An error control code (such as Reed Solomon codes [6]) is then applied. This has the effect of increasing the overall length of the bit stream. The next layer of encoding is to express each symbol as a random signal vector. The elements of this signal vector are added directly to the Discrete Cosine Transform coefficients. The process can be enhanced by taking account of the sensitivity of the human eye to these spatial frequencies. Only the components that are most significant to image integrity are marked. Please consult [11][12] for more details of this algorithm.

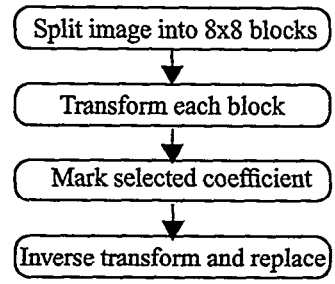


Figure 8: Diagram of the watermarking algorithm.

Currently, much improved faster image watermarking algorithms are being developed. These are more secure, are resistant to scanning, photocopying and common image processing operations and do not require an original image for comparison. Use of the algorithm will also involve formal registration of copyright and therefore will be legally binding. It is intended to make this service available for general use through the World Wide Web.

7. Conclusion

We have presented a global system for distributing secure documents containing textual and pictorial data. Textual information is accessible through a complete and efficient system which will allow the distribution of a weekly electronic newspaper. A robust secure digital watermarking algorithm has been developed to protect documents and digital images.

A real application was used to demonstrate the complete structure of the system. The current database which contains thousands of images of historical watermarks allows one to search and retrieve a specific document and the basis of global features, of textual criteria or of morphological measures. The next goal of this project is to integrate the digital watermarking algorithms with the database management and the security module of our system. The payment module should complete the system which will allow the Swiss Paper Museum to benefit from this work. In the end, the whole system should manage a database of approximately 600,000 different watermarks.

8. References

[1] P. Alshuth, Th. Hermes, Ch. Klauck, J. Kreyß, M. Röper, "IRIS - Image and Video Retrieval", Proc. of CASCON 96, Toronto, Ontario, Canada, 12-14 Nov. 1996.

[2] 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.

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

[4] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele, P. Yanker, "Query by image and video content: the QBIC system", IEEE Computer, Special Issue: Finding the right image: Content-based image retrieval systems, V.N. Gudivada and V.V. Raghavan, Eds., September 1995, 23-32.

[5] A. Gupta, S. Moezzi, A. Taylor, S. Chatterhjee, R. Jain, M. Goldbaum, S. Burgess, "Content-based retrieval of ophtalmological images", IEEE Signal Processing Society, ICIP 1996 International Conference on Image Processing, Lausanne, Switzerland, pp. 703-706, Sept. 16-19, 1996.

[6] S. Haykin, "Communications Systems", Wiley, 3rd edition, 1994.

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

[8] J. H. Morin, D. Konstantas, "Towards Hypermedia Electronic Publishing", Proceedings of the Second IASTED/ISMM International Conference in Distributed Multimedia Systems and Applications, Stanford, CA, August 7-9, 1995.

[9] 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.

[10]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.

[11]J.J.K. Ó Ruanaidh, W.J. Dowling and F.M. Boland, "Phase Watermarking of Digital Images", IEEE Signal Processing Society, ICIP 1996 International Conference on Image Processing, Lausanne, Switzerland, Vol III, pp 239-241, Sept. 16-19, 1996.

[12]J.J.K. Ó Ruanaidh, W.J. Dowling and F.M. Boland, "Watermarking Digital Images for Copyright Protection", IEE Proceedings on Vision, Signal and Image Processing, Vol 143, No. 4, August 1996.

[13]Adress for the main page: http://cuisun8.unige.ch/NSAPI/rauber/watermark, enter with the login "guest_watermark" and with the password "guest_watermark".

[14]E. Saber, A. Murat Tekalp, "Integration of color, shape, and texture for image annotation and retrieval", Proc. Int. Conf. on Image Proc., ICIP'96, Lausanne, Vol III, pp. 851-854, Sept. 16-19, 1996.

[15]A. W. M. Smeulders, R. Jain, "Image Databases and Multi-Media Search", Proceedings of the First International Workshop, IDB-MMS'96, Amsterdam, The Netherlands, Aug. 22-23, 1996.

[16]T. R. Smith, "A digital library for geographically referenced materials", in Computer, pp. 54-60, May 1996.

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

[18]P. F. Tschudin, "DOCSCAN und KRYPICT - zwei Basisprojekte der historischen Disziplinen", Vortrag anlässlich der Tagung des Fachausschusses für Papiergeschichte und Wasserzeichenkunde des Vereins ZELLCHEMING und des 23.Kongresses der IPH in Leipzig, August/September 1996.

[19]J. Vitek, "Compact Dispatch Tables for Dynamically Types Programming Languages", Object Application, Edited by D. Tsichritzis, Centre Universitaire d'Informatique, University of Geneva, August 1996, pp. 81-138

[20]Special Issue: Finding the right image: Content-based image, IEEE Computer, retrieval systems, V.N. Gudivada and V.V. Raghavan, Eds., September 1995.

[21]J. R. Bach, C. Fuller, A. Gupta, A. Hampapur, B. Horowitz, R. Humphrey, R. Jain, C. Shu, "The Virage Image Search Engine: An open framework for image management", SPIE conference, Storage and Retrieval for Still Image and Video Databases IV, Feb. 1, 1996.

DL 97 Philadelphia PA, USA