ISOLIERUNG UND IDENTIFIZIERUNG IDENTISCHER WASSERZEICHEN IN GROSSEN DATENBANKEN

ISOLATION AND IDENTIFICATION OF IDENTICAL WATERMARKS WITHIN LARGE DATABASES

H. Moreu Otal and J.C.A van der Lubbe
Delft University of Technology,
ICT Group, P.O. Box 5031, 2600 GA Delft, The Netherlands
email: {H.MoreuOtal, J.C.A.vanderLubbe}@tudelft.nl

Zusammenfassung

Wasserzeichen in Papier sind eine wichtige Informationsquelle zur Zeitbestimmung und Echtheitsprüfung von Kunstwerken. Sie werden in Sammlungen vorgehalten, die Experten bei der Suche identischer Wasserzeichen unterstützen. Seit einigen Jahren steigt die Nachfrage nach automatisierten Recherchemöglichkeiten. Die Printed Piccard Sammlung und die Piccard Online Datenbank bilden zusammen den weltweit größten Bestand an Wasserzeichen. Als Voraussetzung für eine digitale Weiterverarbeitung dieser Bestände (Merkmalserkennung, Abfragen) ist eine Methode zur Isolierung der Wasserzeichen erforderlich. In diesem Vortrag wird eine halbautomatische Methode vorgeschlagen, die auf beide Sammlungen anwendbar ist. Weiterhin wird ein Content-Based Image Retrieval System vorgestellt, das Recherchen nach ähnlichen Wasserzeichen in beiden Sammlungen erlaubt.

Abstract

Watermarks in paper are an important source of information to determine the dating and authenticity of artworks. They are gathered into watermark collections, which are consulted by experts searching for identical watermarks. Since a few years ago experts want to have at their disposal automatic techniques to retrieve these watermarks. The Printed Piccard collection and the Piccard Online database together form the largest watermark database in the world. In order to carry out further digital processing of these collections (e.g. watermark feature extraction, watermark retrieval), a method that first isolates every watermark is needed. In this paper a semiautomatic method with this purpose is proposed for both collections. Furthermore, we present a content-based image retrieval system which enables retrieval of similar watermarks from Piccard Online and Printed Piccard collection.

1. INTRODUCTION

Paper research provides dating information of artworks by means of exploiting paper features. It is possible to determine the artist and creation date of an artwork by discovering identical pieces of paper. Assume e.g. two etchings of Rembrandt. One was printed in 1648 and the other has an unknown date. However, since both were made on the same paper, it is reasonable to suppose that also the undated artwork was made in 1648. The reason is that in the 16^{th} -18^{th} century paper was sold in stocks of paper sheets coming from the same paper mill and mould, which implies an identical paper structure. There are many paper features which can be used to get information about the artwork, like chain and laid lines exploited by Van Staalduinen in [1]. In this work, we focus on watermarks which are line structures introduced into the paper during the paper manufacturing process. They are marked in the paper by copper wires which have been curved in a specific shape and which are fixed on the mould. According to the art experts watermarks are the most important paper feature in order to discover identical pieces of paper.

Due to the importance of watermarks in the field of paper research many watermark collections were created in the 20th century. These collections are used by art experts in order to search for identical watermarks by visual inspection. The largest watermark collection in the world is the Piccard database. It was created by Gerhard Piccard (1909-1989) who gathered and edited the collection of 95.000 watermarks. He visited archives and libraries and traced the watermarks on index cards from the original documents. Moreover, he added additional information such as chain and laid lines, the place where the document was printed, the date and type of document and the quality of the paper. Between 1961 and 1997, a part of Piccard's work was published in 17 inventories with 25 volumes. Approximately two thirds of the index cards were used. This means that this so-called Printed Piccard (*PP*) collection includes around 55.000 watermarks. At present, all the index cards of the original Piccard database have been digitized and are accessible via the worldwide web in the so-called Piccard Online (*PO*) database [2]. As may be expected, most of the watermarks of the printed collection are also stored in the Piccard Online database. However, there are watermarks in the printed collection which are not on the index cards and therefore they are not in the Piccard Online database. For watermark experts it is important to detect these watermarks in order to find out their origin and to complete Piccard Online.

The present study has two objectives. The first objective is to isolate and store the individual watermarks from both collections in image files. This is achieved by a software tool that automatically isolates every watermark. Manual isolation is a tedious and labour intensive task. The user should only check the outputs, which can be corrected for the wrong isolations. The second objective of our research is the retrieval problem. The software developed enables the experts to retrieve watermarks which are present in the Printed Piccard but not in the Piccard Online database.

In literature some solutions to detect and isolate watermarks in X-ray images are given; compare e.g. Moreu [3] and Wenger [4]. However, until now there is no published technique to isolate watermarks in watermark databases or printed collections as in the Piccard case. Regarding the retrieval tool, matching with watermarks is presented by Riley [5] and Rauber [6, 7]. Both works conclude that in order to have an accurate retrieval system the watermark should be properly detected and isolated.

2. ISOLATION OF WATERMARKS FROM THE PICCARD ONLINE DATABASE

The Piccard Online database, at the Hauptstaatsarchiv Stuttgart, contains 95.000 different records. They are classified according to 38 categories, representing e.g. crown, bull's head, tower, cross and horn. Furthermore, most categories are divided in several subcategories which represent specific characteristics of each motif.

Every single watermark within the Piccard Online database is stored in a separate image file. These images show the watermark but also extra information which hinders easy isolation. Surrounding the watermark there are different elements (see Figure 1, left) as chain lines, laid lines and meta-data. All these elements need to be identified in order to isolate the watermark properly.

The digital images are defined as follows. A gray scale image is given by I, and the intensity of pixel $\vec{x} = (x_1, x_2)$ is given by $I(\vec{x}) \in \Re \{0,1,\ldots,255\}$. The dimensions of an image are given by (w_1, w_2) for the width and the height, respectively. Our method works with binary images defined as

$$B_{PO}(\vec{x}) = \begin{cases} 1 & if \;\; I(\vec{x}) \le 100, \\ 0 & otherwise. \end{cases} \tag{1}$$

The isolations consist of finding where the watermark is located and then make a bounding box which only surrounds the watermark (see Figure 1, left). All the non-watermark information is left outside the bounding box.

Our method locates and removes the non-watermark elements by recognizing noisy patterns in the *BLOBs* within the image. A *BLOB* (binary large object) is defined as a group of connected pixels. For each *BLOB* some image features such as position, size and distance to the closest *BLOB* are computed in order to distinguish between watermark and noisy *BLOBs*. Thereafter, the bounding box of the watermark can be determined by considering the new image without the noisy elements. The isolation method depends on several parameters which are learned by maximizing the number of correct isolations for a training set of 300 watermarks coming from the 38 different categories.

The isolation process starts by detecting and removing the laid lines. In this collection, laid lines appear in two formats: short laid lines (very short lines marked on the chain lines) and long laid lines (they are drawn between two chain lines). Both kinds of laid lines are similarly detected; the only difference is the values given to the parameters. Our method exploits the fact that laid lines always have horizontal orientation, and similar line length and line density. First of all, a binary erosion, explained in Gonzalez [8], using a horizontal line as a structuring element is computed. This structuring element is defined as *HL* where the length of the horizontal line is equal to 15 or 5 pixels to detect long or short laid lines respectively. Thereafter, we study the size of the *BLOBs* coming from the eroded image after applying a dilation operation, described in Gonzalez [8], in order to gather all the laid lines in the same *BLOB*. We distinguish between a laid line *BLOB* and other structures by considering the *BLOB* dimensions (*BD*). The dimensions of a laid line *BLOB* are larger than 60*60 for long laid lines and 9*60 in the case of short laid lines. B_{LL} is the image which stores the laid lines within B_{PO} (defined in Equation 1).

Secondly, the text above the watermark is localized and then removed by taking into account the position of the *BLOBs* with respect to the largest *BLOB* which corresponds to the watermark. After that, the part of the noise located on the bottom of the watermark is filtered out by considering the size of the *BLOBs* that form the image. Every *BLOB* smaller than the noisy threshold (*NT)* is removed. From our training set we found out that the optimal *NT* is 50. The text and other noisy structures are stored in the image B_{RT}.

Finally, we eliminate the chain lines by removing those vertical lines which occupy more than 40% of the image height (w_2). This parameter is called vertical size threshold (*VST*) and it is defined as $VST=0.4*w_2$. Vertical lines are detected by using a binary erosion with a vertical line *VL* as structuring element. The length of *VL* is equal to 90 pixels. The chain lines of B_{PO} are saved in the image B_{CL}.

After we have identified and removed all the noisy elements, the result is a cleaned image that is used to compute the watermark bounding box. This final output is the isolated watermark B_{IW} given by

$$B_{IW}(\vec{x}) = bounding\_box\ \ [B_{PO} - (B_{LL} + B_{RT} + B_{CL})]\ \ (\mathrm{x}). \tag{2}$$

Once this automatic process has been carried out for all the images the user needs to check the outputs and label them as correct or erroneous isolations. This is easily done by using a graphical user interface which shows the original image and the automatic isolation given by our tool. Manual isolation is carried out by constraining the watermark within 4 line-markers which are easily moved by the image by clicking and dragging them with the mouse (see Figure 1, right). For every watermark our method automatically predicts where the 4 line-markers should be placed, in such a way that the user only needs to change the position of the wrong ones (one of the markers in most cases).

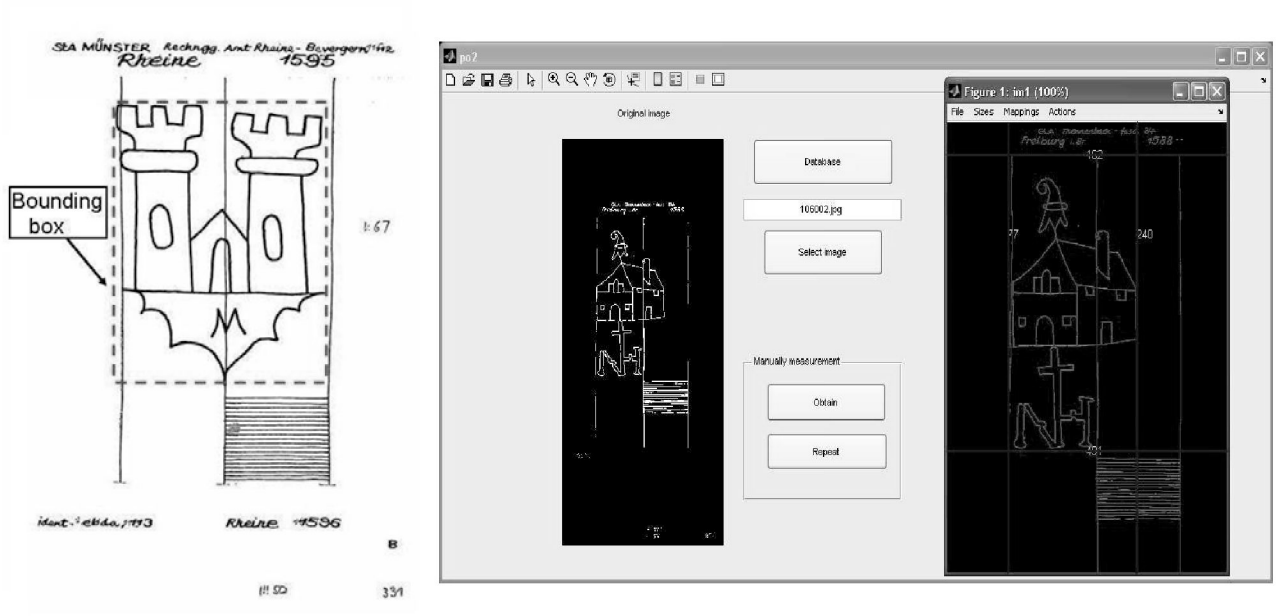

Figure 1. Left: Watermark from the Piccard Online database. The three vertical lines are the chain lines. The set of horizontal lines represents the laid lines existing in the paper. The upper text says where the document is stored and where and when it was certified. The lower text gives information about the kind of document and condition of the paper where the watermark is. The numbers on the bottom of the image are the distance between chain lines and a classification number. The broken line represents the bounding box. Right: Matlab GUI developed to correct the wrong isolations carried out by the automatic isolation tool.

3. ISOLATION OF WATERMARKS FROM THE PRINTED PICCARD COLLECTION

The isolation problem with respect to the Printed Piccard collection is completely different to that of the Piccard Online database. As mentioned, the 55.000 watermarks of the printed collection are stored in 25 volumes ordered according to the watermark motifs. Every page of the collection includes several watermarks as it is shown in Figure 2. The inputs for the isolation method are scanned images of the pages of the volumes. This means that each image file contains several watermarks as well as the page number, volume number and watermark numbers. As in section 2 the original images are defined as $I(\vec{x}) \in \Re\ \{0,1,\ldots,255\}$ and the binary images we work with are defined as follows

$$B_{PP}(\vec{x}) = \begin{cases} 1 & if \quad I(\vec{x}) \leq 200, \\ 0 & otherwise. \end{cases} \tag{3}$$

In this case our goal is not only to isolate the watermarks inside their bounding box (what we called retrieval watermarks), but also we want to offer another output whereby each watermark is stored together with its corresponding watermark number, chain lines and a white frame around the watermark (these outputs are called catalogue watermarks). Both outputs are shown on the right side of Figure 2.

The first step is to compute the number of watermarks (N_{WM}) stored in each page. This is done by locating and counting the watermark numbers in each page. These numbers can be seen on the left side of Figure 2 below each watermark. It was found out that the positions of the watermark numbers are always within fixed boundaries inside the page as well as that their dimensions are almost constant. Furthermore, the watermark numbers are never within other *BLOBs*.

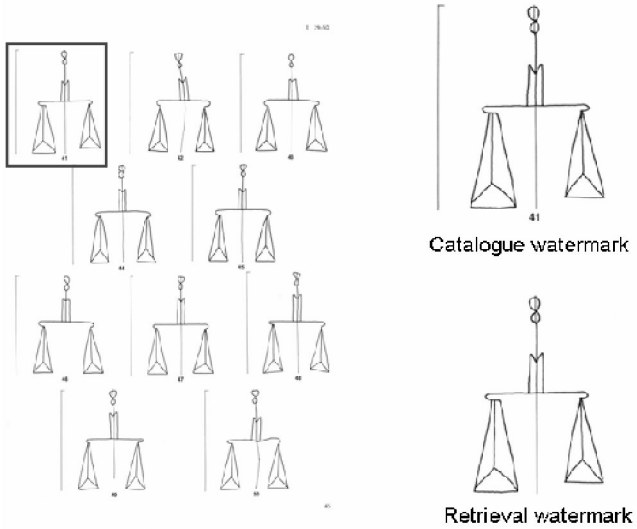

Figure 2: The left image represents a piece of paper from the Printed Piccard collection. Inside the box is the watermark selected to be isolated. The right image shows the two isolations provided by our method.

After studying the training set for the Printed Piccard collection, we can ensure that only *BLOBs* with watermark information can be inside other *BLOBs*. This is very important since it allows us not to confuse small watermark details with noisy *BLOBs*. In this case the training set is formed by 150 scanned pages from 14 different volumes of the collection and all the parameters are trained by maximizing the number of correct isolations. From our training set, we also know that all the watermark parts are larger than a fixed value called size threshold (*ST*) equal to 170 pixels. These watermark *BLOBs* are saved in the image B_{SB}.

By using the number of watermarks in each page we are able to determine the main part of each watermark considering the size of the *BLOBs*. This is possible since each page gathers very similar watermarks; therefore all of them have similar sizes. Main watermark parts are stored in the image B_{MP}.

Afterwards, we proceed to associate the non-noisy *BLOBs* to those ones that were previously selected as main watermark parts. This step is done by linking every unlinked *BLOB* from B_{SB} with one main watermark part from B_{MP} in such a way that the Euclidean distance between the gravity centres of the *BLOBs* is minimized. After this, we independently store each watermark of the page as images RW_i where $i \in \{1, N_{WM}\}$. They are the so-called retrieval watermarks.

In order to compute the catalogue watermarks we need to add the watermark numbers as well as the chain lines to the already isolated watermarks (RW_i). Chain lines are detected by recognizing the vertical lines which are not part of the isolated watermark as explained in section 2. Then, chain lines and watermark numbers are attributed to a main watermark part by minimizing the Euclidean distance between BLOBs. Finally, every catalogue watermark is stored as an image CW_i with $i \in \{1, N_{WM}\}$.

As for the Piccard Online case, the user needs to check whether the watermark isolation was made correctly. Now, the user has to check that the automatic step succeeded for every watermark in a page; otherwise that page should be processed manually. The wrong isolations are corrected by the user by using a semi-automatic tool. This consists of a graphical interface where the user needs to surround every watermark (including watermark number and chain lines) within the image; thereafter the software is able to directly compute the catalogue watermarks. Then, the retrieval watermarks are also automatically generated by our method since it is able to find and remove the watermark numbers and chain lines.

4. IDENTIFICATION OF IDENTICAL WATERMARKS

Due to the large number of watermarks to deal with and the similarity among the watermarks of the same motif, a method known as zoning is used for retrieval. It consists of subdividing the previously isolated watermark images in zones and then counting the number of watermark pixels in each of them. This information is then stored within a vector for each watermark in both Piccard databases. According to our experiments, the optimal number of zones to retrieve identical watermarks in the Piccard collections is equal to 1000. In this manner our method takes into account small details of the watermarks and is able to find identical watermarks among very similar images. Matchings are found by considering for each vector in one database that vector in the other database which leads to the smallest Euclidean distance.

Thanks to the watermark isolation carried out by using the above mentioned tools, this method provides an important help for searching for similar watermarks. However, it is impossible to automatically determine for all the watermarks whether they are identical or very similar. This problem holds even for watermark experts. This is due to the following two facts. First of all, we need to remember that the watermarks were traced manually, which makes them imprecise. Secondly, sometimes there are many watermarks of one type which makes it very difficult to decide whether we have two identical watermarks or just two very similar watermarks.

In these situations some additional information, like the chain lines position or laid lines density, needs to be used in order to decide whether two watermarks are identical. Following suggestions of watermark experts, we decided to develop a graphical interface where the seven most similar watermarks to the query are shown. After this the experts should make the final decision, whether they are identical or not. This is already a big improvement for art experts since the retrieval method is able to provide a set with the seven most similar watermarks instead of having to work with all the watermarks of the same motif.

5. RESULTS

The tool for the automatic isolation of the watermarks from the Piccard Online database has been tested in a representative test set of 500 watermarks. Its performance is 84%. After analyzing, the following causes of error were identified: chain lines are considered as watermark because they are in contact with the watermark (65%), watermark parts are wrongly removed since they are confused with noise (30%) and the incorrect recognition of the laid lines (5%).

Regarding the performance of the isolation tool for the printed collection the results are slightly better; 88% of the pages are correctly isolated. These results come from a test set of 150 pages with almost 1500 watermarks in total. For this collection the sources of error are completely different. Most of the errors are due to the fact that some watermark parts are assigned to a wrong watermark since this one is the closest to the gravity centre of the unlinked watermark part (70%). This mistake mostly happens when the catalogue watermarks are isolated, since chain lines can be linked to neighbouring watermarks. The other errors are caused by an incorrect estimation of the number of the watermarks in each page (30%).

The identification method has been tested by discovering identical watermarks within one motif in both collections. The chosen category was scales (balances) which gathers 798 watermarks in the Piccard Online database and 1.722 in the printed collection. In order to validate the results the experts provided us with a set of 170 matching watermarks between both collections. For this test set the identification method works out for the 91% of the cases. This means that for the 91% of the watermarks coming from the Piccard Online database the user can find their matching watermark within the first seven closest watermarks of the printed collection provided by our software. In the 62% of the cases the software provides the identical watermark within the printed collection in the first position. For 9% of the watermarks the identification

software fails. This is due to the fact that there are many watermarks very similar to the query and moreover the watermarks are not completely identical. In some cases two watermarks supposedly identical present some differences either because of errors when they were traced or because the isolation software is not able to suppress the chain lines in contact with the watermark.

6. CONCLUSION AND FUTURE WORK

The wide variety of watermarks and mostly noisy structures within the Piccard Online database makes the automatic isolation a complex task. Although the performance provided by the automatic isolation tool is high, it is still necessary to verify all the isolated watermarks and to manually correct the wrong outputs. The outputs can be easily checked and corrected by means of the graphical user interface designed in Matlab. This GUI stands out for the ease of use and because watermarks can be quickly isolated by displacing the line-markers to select the bounding box.

As a future work, it would be interesting to develop a method able to suppress the chain lines in the outcomes for the Piccard Online database. Sometimes chain lines are within the bounding box of the watermarks and therefore in the final result. This task is complicated due to the fact that chain lines and watermark parts are occasionally in contact, which makes it very difficult to automatically recognize the chain lines.

Our isolation tool for the Printed Piccard collection provides both the images used by the retrieval tool and also the watermarks to be published in catalogues or watermark collections. The errors committed by the automatic isolation tool are corrected by the user by means of the Matlab GUI for this purpose.

The results provided by the identification tool are a big help for watermark experts. The searching job is considerably reduced. Instead of dealing with a large number of watermarks, experts only need to check the seven most similar watermarks presented by our software. The performance would be improved, as mentioned before, by removing chain lines in contact with some watermarks within the Piccard Online database. In the retrieval watermarks coming from the printed collection there are no chain lines. This fact makes an important difference when there is a set of similar watermarks some with chain lines and some without.

REFERENCES

[1] M. van Staalduinen, J.C.A. van der Lubbe, E. Backer and P. Paclik, “Paper retrieval based on specific paper features: Chain and laid lines.” In MRCS pages 346-353, Istanbul, Turkey, 2006.

[2] *Piccard database. http://www.piccard-online.de/start.php*

[3] H. Moreu Otal, M. van Staalduinen, P. Paclik and J.C.A. van der Lubbe. “Watermark detection in X-ray images from paper for dating artworks.” In SPPRA, pages 66-71, Innsbruck, Austria, 2008.

[4] E. Wenger, V. Karnaukhov, A. Haidinger and M. Stieglecker. “A Digital Image Processing and Database System for Watermarks in Medieval Manuscripts.” In ICHIM, pages 259-265, 2001.

[5] K. J. Riley and J. P. Eakins. “Content-Based Retrieval of Historical Watermark Images: I-tracings.” In CIVR, pages 253-261, 2002.

[6] C. Rauber, P. Tschudin, S. Startchik and T. Pun. “Archival and retrieval of historical watermark images.” In ICIP, volume 2, pages 773-776, 1996.

[7] C. Rauber, P. Tschudin and T. Pun, “Retrieval of images from a library of watermarks for ancient paper identification”, In EVA'97, Vol.14.

[8] R. Gonzalez, R. Woods. “Digital Image Processing”, 2nd edition. Prentice-Hall, Upper Saddle River 2002.