Int J Digit Libr (2007) 6:351–361
DOI 10.1007/s00799-007-0008-7

REGULAR PAPER

A system for segmenting and extracting paper-based watermark designs

Hazem Hiary · Kia Ng

Published online: 14 February 2007

Abstract This paper presents a framework for the digitisation, extraction, and graphical representation of paper-based watermark designs embedded in paper texture: there is a growing need for this among librarians and antiquarians to aid with classification and preservation. The system is designed to handle manuscripts with foreground interference and defects; it uses a back-lighting scanning technique combined with image processing operations rather than radioactive techniques. Hence, it is faster, cheaper, safer, and easy to use. The system prototype includes a set of image processing operations which enhance, filter, and extract the watermark shape, and automatically convert it into a graphical representation. The paper focuses also on automated processes which determine the configuration of parameters in order to allow optimal content processing, in addition to the detection of watermark chainlines. With a machine readable graphical representation of the watermark, cataloguing and indexing of these heritage resources can be enhanced with the ease of digital content retrieval functionalities exploiting the advantages of digital technologies such as distribution and preservation.

Keywords Paper watermark · Digitisation · Image processing · Graphical representation · Preservation

H. Hiary · K. Ng (✉)
School of Computing, ICSRiM,
University of Leeds, Leeds LS2 9JT, UK
e-mail: hazemh@comp.leeds.ac.uk

K. Ng
e-mail: kia@kcng.org
Web: http://www.icsrim.org.uk/watermark/

1 Introduction

The production of paper watermarks was initiated over 700 years ago by papermakers in Italy; the oldest known watermarked paper was produced in 1292, originating in Fabriano [9]. Watermarks quickly spread through Italy and then over Europe. The original purpose for adding watermarks into the papers was to recognise the papers produced by specific workmen so they got paid [1], then it was used to specify the paper brand. Later, it was used to indicate paper formats and quality.

Papers were often made with two different moulds with two watermark designs that were very similar (but not necessarily identical), and hence watermarks are generally twins [17].

Nowadays, paper-based watermarks are considered important artefacts, because they identify paper owners to protect their intellectual property rights. It is also used for authentication to protect important documents such as bank notes, passports, entry tickets, etc. from forgery and theft. It could also be used to help tracing and studying old documents and artefacts to provide plausible historical relationships and background information. Watermarks in paper have attracted a wide range of interests from researchers for centuries. Watermarks designs not only available in several different forms but also dynamically change over time. This has introduced some complications that have hindered more systematic study of the artefacts which are decaying over time because of natural processes. Furthermore, there exist foreground interference (e.g., writing on front and back) and background interference (e.g., paper defects such as folding) which may overlap and obstruct the watermark pattern.

Due to the natural decay of paper materials, digitisation has been widely applied as one of the preservation approaches to keep a visual record of the artefacts, by creating a digital copy of the paper materials. However, current processes do not normally capture the complete representation—typically, are only concerned with the paper surface. There is hidden 'visual' information, particularly the embedded watermark in the paper, which is not being digitised and may be lost forever. Digitisation guidelines and best -practises are available from many important projects, such as the Pulman (http://www.pulmanweb.org), Minerva (http://www.minervaeurope.org), and MUSICNETWORK (http://www.interactivemusicnetwork.org).

In order to classify different paper materials, including watermarks, the International Association of Paper Historians (http://www.paperhistory.org) created taxonomy of terms for describing the components of laid paper, including the watermark.

This paper illustrates a methodology to extract paper watermarks. The main focus in this project will be on the digital acquisition, and automatic processing and analysis of watermark patterns, exploring beyond the paper surface data to extract and trace the watermark design, and to create its digital representation for long-term preservation and wider accessibility.

This paper is structured in five sections. Section 2 provides a brief background and related work, Sect. 3 discusses the stages of our proposed system with results taken from various system stages, Sect. 4 illustrates and discusses more results, and Sect. 5 gives conclusions drawn from the research, and discusses a number of future plans and work in hand.

2 Related Work

There are various existing approaches to extract paper watermarks: manual tracing; photosensitive paper 'Dylux 503' [1]; ilkley; phosphorescence watermark imaging [11]; transmitted light photographs or 'back-lighting'; and beta-radiography [2].

Manual tracing is a time consuming and highly subjective task.

Photosensitive paper, back-lighting, ilkley, and the phosphorescence watermark imaging are time saving techniques. However, these techniques do not remove any interference (e.g., the ink of the writing, defects, and others), together with chainlines (lines that appear as part of the watermark background). Hence, the output combines the watermark design with any foreground. It works for clean paper with only the watermark, but not manuscripts that have been written or printed onto.

The chemicals used in Dylux proofing papers release 'offgas', so users should be cautious and use a well-ventilated environment [1].

The beta-radiography method gives the best result, with minimum foreground interference. Unfortunately, this method is time consuming and the resource required is expensive [2] and not easy to handle due to the radioactive source needed, together with darkroom conditions [3,11].

Combining the back-lighting digitisation technique with various image processing operations offers an effective and simple to use the technique for extracting the watermark design from paper. The system has the ability to isolate and remove most of noise and different foreground interferences since most of the watermarked manuscripts have writings and other defects which obstruct the watermark design.

Other related work that also used this combination approach includes Whelan et al. [20]. In this case, they used the back-lighting technique for digitisation and apply morphological and Fourier approaches for segmentation. However, this project focused on the localisation of the watermark from new paper without any foreground interference.

SHREW [15] 'Shape Retrieval System for Watermarking Images' worked on watermark-related issues including a comparison of watermark digitisation techniques and focuses on a watermark image retrieval system based on the watermark shape.

Rauber et al. [12,14] proposed a system for the management of a database of historical papers which contains watermarks. The system supports several different types of queries to locate and retrieve an existing watermark design from the database, including textual search and shape-based queries [13]. For the database input, the watermark paper used is free of foreground interference; their system is not fully automatic, and they apply semi-automatic image processing to enhance the resulted watermark.

Karnaukhov et al. [8] used the beta-radiographic method combined with Fourier transform techniques and image filters. However, they concentrated on paper-based watermarks without any foreground interference.

Wenger et al. [19] proposed a project for developing a distributed watermark library, which allows access to local watermark databases with different structures in different locations; another objective of this project was to improve watermark recording by new radiographic methods.

Gants [5] proposed a watermark reproduction method by first using the Dylux digitising technique [1], followed by shifting the contrast and brightness as image enhancements to strengthen the watermark signal.

3 A watermark extraction system

The developed system can be divided into four main stages. The first stage is digitisation in which the data are acquired (an image of the paper). This is followed by the pre-processing steps where image processing modules are applied to highlight the watermark and remove foreground- and background- interferences. This is an important stage that provides the key advantage to this system since it handles typical noise and writings/markings on both surfaces of the paper. After this, the segmentation stage offers the localisation and extraction of watermark pattern and chainlines. The final processing stage is optional. It provides a function to vectorise the watermark designs segmented in order to create a graphical representation in vector description for better storage and visualisation.

An overall flow chart of the system is illustrated in Fig. 1.

3.1 Digitisation

The extraction system starts with the digitisation process using the back-lighting technique, with even lighting behind the paper in order to improve the visualisation of the watermark pattern. The image is digitised using a high-resolution digital back CCD, PhaseOne PowerPhase FX+ (http://www.phaseone.com), which can capture an image with a resolution up to 10,500 × 12,600 pixels. The back-lighting technique reveals watermark designs that are usually hidden. See Fig. 2 for two digitised images, one with back-lighting and the other with normal lighting.

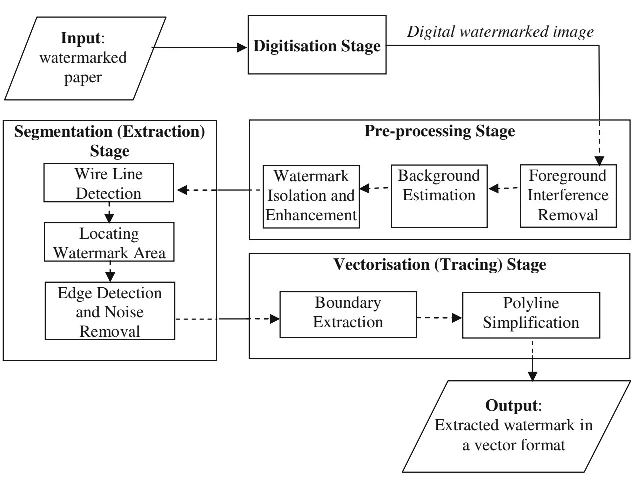

Fig. 1 Flow chart of the overall watermark extraction system

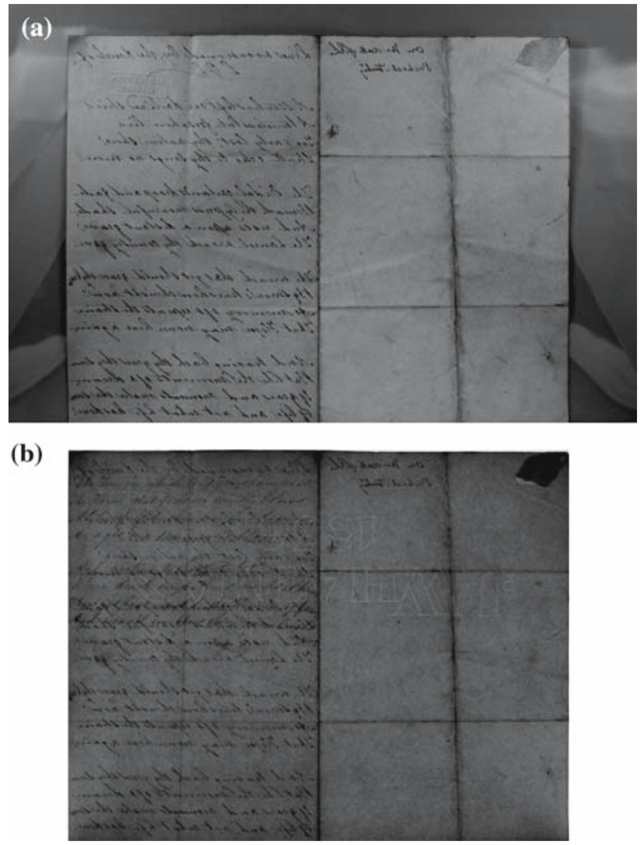

Fig. 2 Image acquisition: **a** Normal, and **b** with back-light, the watermark is becoming visible

3.2 Pre-processing

The digitised image normally consists of the paper (in the centre) with some background. For better estimations of the dynamic threshold setting, the preprocessing stage starts with the localisation of region of the paper in the image by analysing its grey-level distribution. See Fig. 3a. A series of steps are then applied in order to extract the watermark design, by separating the image into a number of layers as presented in the following sub-sections. Firstly, foreground interference such as writing is removed by producing an intermediate image (I_a) with the background and the watermark. Next, the background of the image (e.g., paper texture, noise, folding marks) is estimated (I_b). After that, the difference image of I_a and I_b is produced (I_w) which contains the watermark (and some noise).

3.2.1 Foreground interference removal

The watermark design is extracted from the foreground interference by applying mathematical morphological operations. The size of structuring element to use in

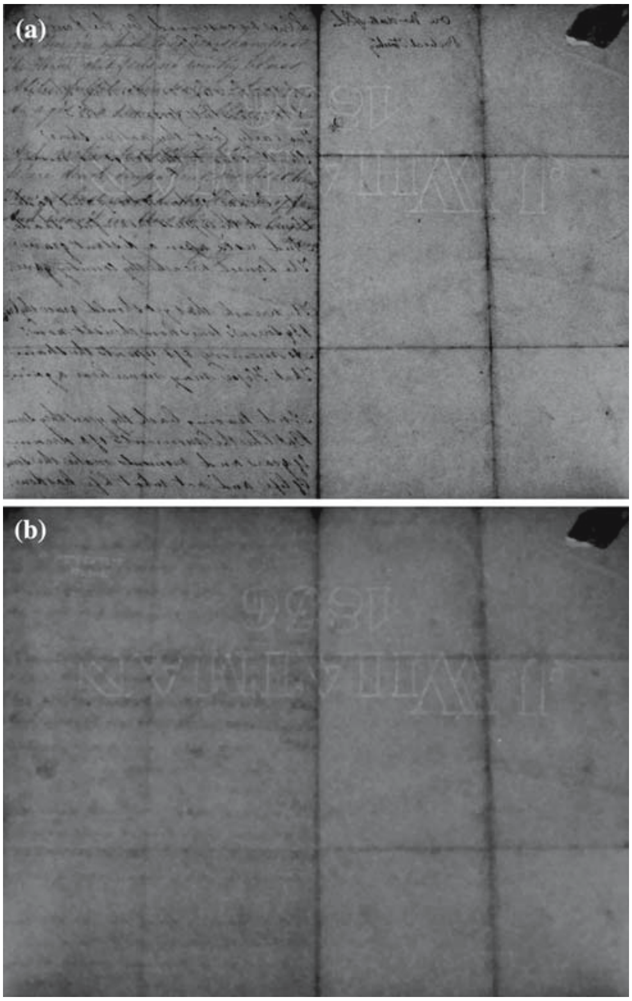

Fig. 3 **a** Input image after border removal, **b** after foreground removal: watermark is visible, and most foreground interference is removed

dilation [6] to remove foreground interference is critical, and we seek to determine this automatically:

1. Applying a contrast stretching process [6], so the darkest pixels will take 0 intensity value.
2. Determining the percentage of such pixels: $x\%$.
3. Within the original image, determine the grey level g such that $x\%$ of pixels are at intensity g or less.
4. Dilate the input image, starting with structuring size 1, and increasing the structuring size until all pixels values are above g.
5. The final structuring size is taken as the optimal value to remove foreground interference.

Figure 3b illustrates the intermediate result after this stage.

3.2.2 Background estimation

The next step is focused on the removal of background. If the image does not have uniform illumination (i.e., some areas are brighter than others), it can be corrected by first estimating the background illumination.

Background estimation is done by using morphological opening which is an erosion operation to be followed by a dilation operation. Both of these operations use the same structuring element [6]. The resulted output is the removal of image objects that are completely contained in the structuring element used in the opening operation. Opening operator is useful for separating touching objects, and removing small regions and sharp peaks. To remove the watermark pattern, it is necessary to choose a structuring element that is large enough to cover a single feature of that pattern.

The automatic selection of this optimal size is an interesting challenge for this step. One of the successful approaches is to estimate the width of the watermark pattern, and choose the structuring element size that is larger than this value. Granulometry [16] is used to determine the size distributions of objects in an image without segmenting each object. This is achieved by applying a series of morphological openings with structuring elements (SE) of increasing size. The sum of pixel intensity values in the output image after each opening is stored. See Fig. 4.

Taking the difference of total intensities (sum of pixel intensity values) between two sequential openings will give the distribution of objects sizes at that scale. This definition is also referred to as the pattern spectrum of the image. Figure 5 illustrates the granulometry or the pattern spectrum of image objects, which can be viewed as the first derivative of the intensity surface area distribution.

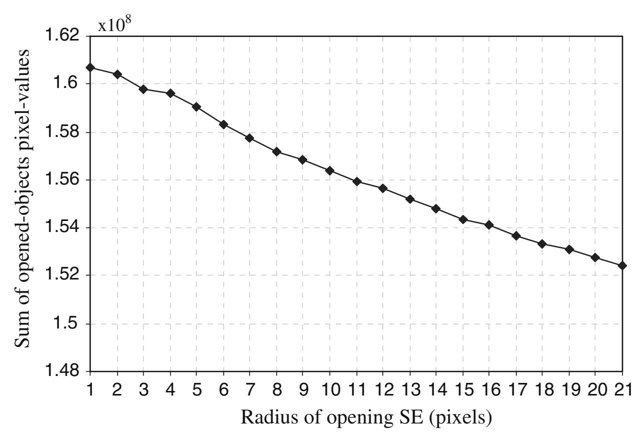

Fig. 4 Cumulative intensities plotted against SE radius

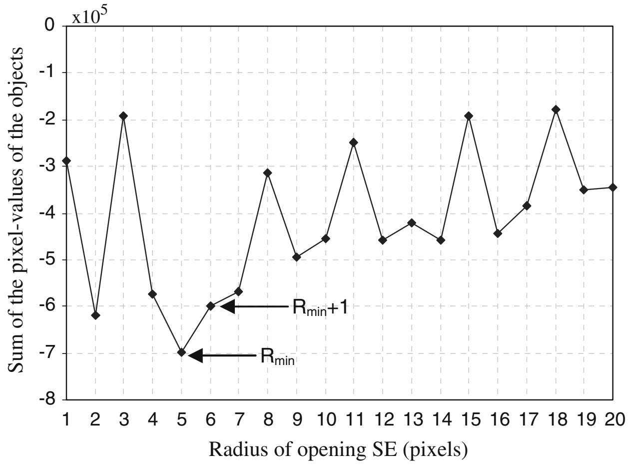

Fig. 5 Granulometry (size distribution) of image objects: first differences of the plot in Fig. 4

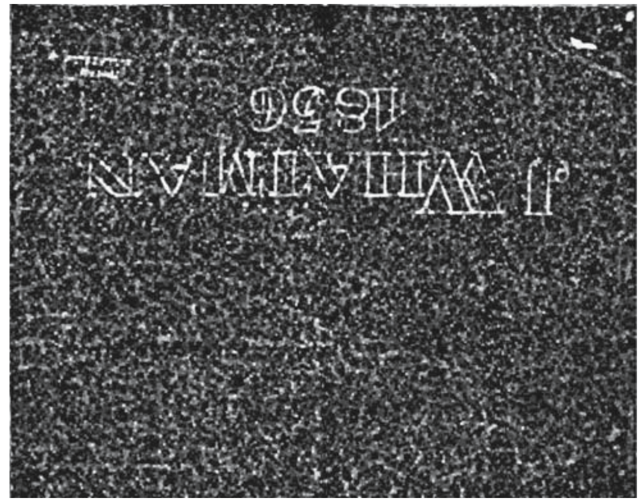

Fig. 6 The result after pre-processing stages

By investigating this distribution, a local minimum at a specific radius will indicate the existence of many image objects of that radius. The global minimum, $R_{\min}$, will indicate the highest cumulative intensity of objects at that radius. The most suitable structuring element size for background estimation will have the value $R_{\min}$+1; choosing a smaller size will not isolate the watermark pattern from the background.

3.2.3 Watermark isolation and enhancement

The pre-processing stage is then finalised by subtracting the estimated background from the image after foreground removal. The result will have a uniform background, noisy regions such as folding will be eliminated in this process. The signal for the watermark has less interference from foreground noise. However, the intermediate output after the differencing operation is low in contrast due to the numerical subtraction. To correct this, the process of contrast stretching is applied for better visualisation and to enhance the contrast of the image (numerically). See Fig. 6.

3.3 Segmentation

The extraction of watermark patterns works by applying edge detection followed by the identification of noise image features and interior segments in preparation for automatic tracing. This stage also includes the detection of chainlines (lines that appear as part of the watermark background), if they exist.

3.3.1 Chainlines detection

Chainlines are the traces left in the paper where the parallel wires forming the surface of the mould are attached to the wooden ribs below. These lines are parallel to each other and are spaced roughly 25–30 mm apart [11]. The distance between individual chainlines varies across the mould and hence the sequence of varying spaces (between two chainlines) is unique to the specific mould. This can be analogous to the fingerprint identification of the mould since such sequences can be used to identify papers made from the same paper mould. A specific function of this watermark analysis system has been developed to output the sequence of spaces between chainlines which can be very useful for the research and scholarship in paper studies [18].

The process of detecting chainlines in the image uses either the Hough or Radon transforms [10,16]. This process redraws the detected lines in the case some of them do not appear due to the digitisation process, or because of paper folding and cutting. Further, image skew can be adjusted depending on detected chainlines, in the case the paper was moved during digitisation.

The Radon transform computes projections of an image matrix along specified directions; it computes the line integrals from multiple sources along parallel paths by rotating the source around the centre of the image.

The Radon transform of the image in Fig. 7a is illustrated in Fig. 7b; detected lines (white spots) were located when applying a projection of angle 1 which is the same direction as angle (181). Detected lines are in Fig. 7c.

The process takes the detected line coordinates, finds spacing and thickness, and then redraws the lines. The

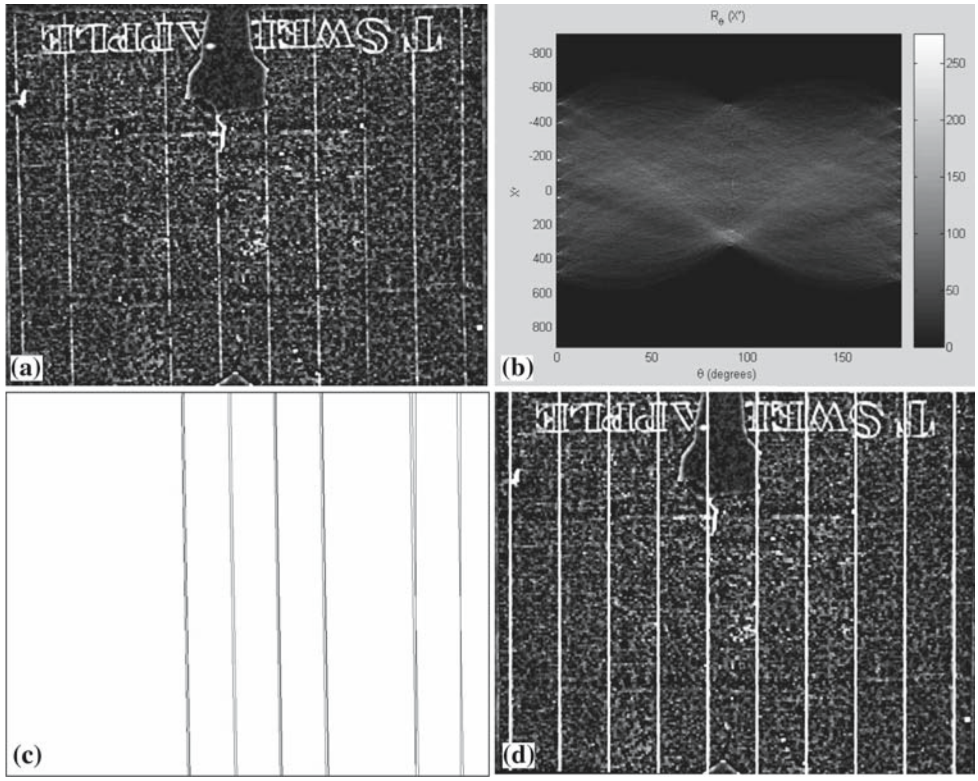

Fig. 7 **a** Image before chainlines detection, **b** Radon transform, **c** Detected lines, **d** Image after chainlines detection

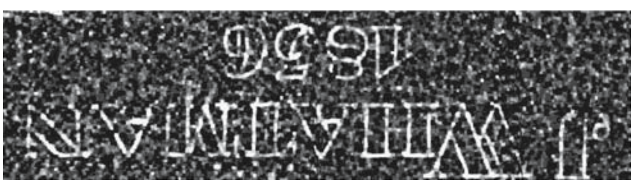

Fig. 8 Image after locating watermark area

Fig. 9 Results after segmentation

direction of the resulted image is then adjusted depending on the direction of chainlines; see Fig. 7d.

3.3.2 Locating watermark area

The location of the watermark area can be approximated by first applying data projection in both x and y directions, and picking the areas which have the highest variance and intensity values. The watermark pattern can be in one or both of these areas; the located watermark area is illustrated in Fig. 8.

3.3.3 Edge detection and noise removal

A Canny detector [4] is then used to locate edges. This method gave the best watermark design detection; others provided less shape detail, with more unnecessary image features. Both automated threshold settings and user-controllable settings for the edge detection processing are provided.

Next, a noise removal process is applied. Small gaps between image features are eliminated by applying a morphological closing operation; this operation reduces the number of image features (and hence reduces processing time needed). Image noise is then located and removed. To do this, three assumptions were made: (i) noisy image features are small-sized; (ii) noisy image features are isolated; and (iii) isolated, small groups of neighbouring image features are noise.

A process to output improved results is performed by interior filling of small unwanted holes. The result after these stages is shown in Fig. 9a; another result with chainlines is in Fig. 9b.

Fig. 10 Output after vectorisation stage

3.4 Vectorisation

The bitmapped watermark design output from the segmentation stage is then automatically traced and converted to a simplified vector graphical representation [7]; a vectorised watermark pattern is in Fig. 10.

4 Results

This section presents and discusses several set of watermark images to demonstrate the effectiveness and the results of the processing. The proposed approach, as reported in this paper, is generally robust and is capable of resolving a wide range of foreground and background interferences. However, it is limited to typical thickness of writing (e.g., radius of the nip). Clearly, any large region of dark interference cannot be supported. For example, a black circle that is larger than the watermark design that overlapped onto the watermark will prevent its complete segmentation (on the part that has been overlapped) since the texture of the watermark is not available on the input image. The system has been prototyped in MATLAB application with specially designed graphical user interface to provide easy operations, with default settings and abilities to handle any manual interventions. The system can also be run in standalone mode, without MATLAB environment. Further details of the developed project, updates, sample test data and results are available online at the project website: http://www.icsrim.org.uk/watermark/.

Figure 11 presents the main interface of the prototype with a window for the rendering of the input image and

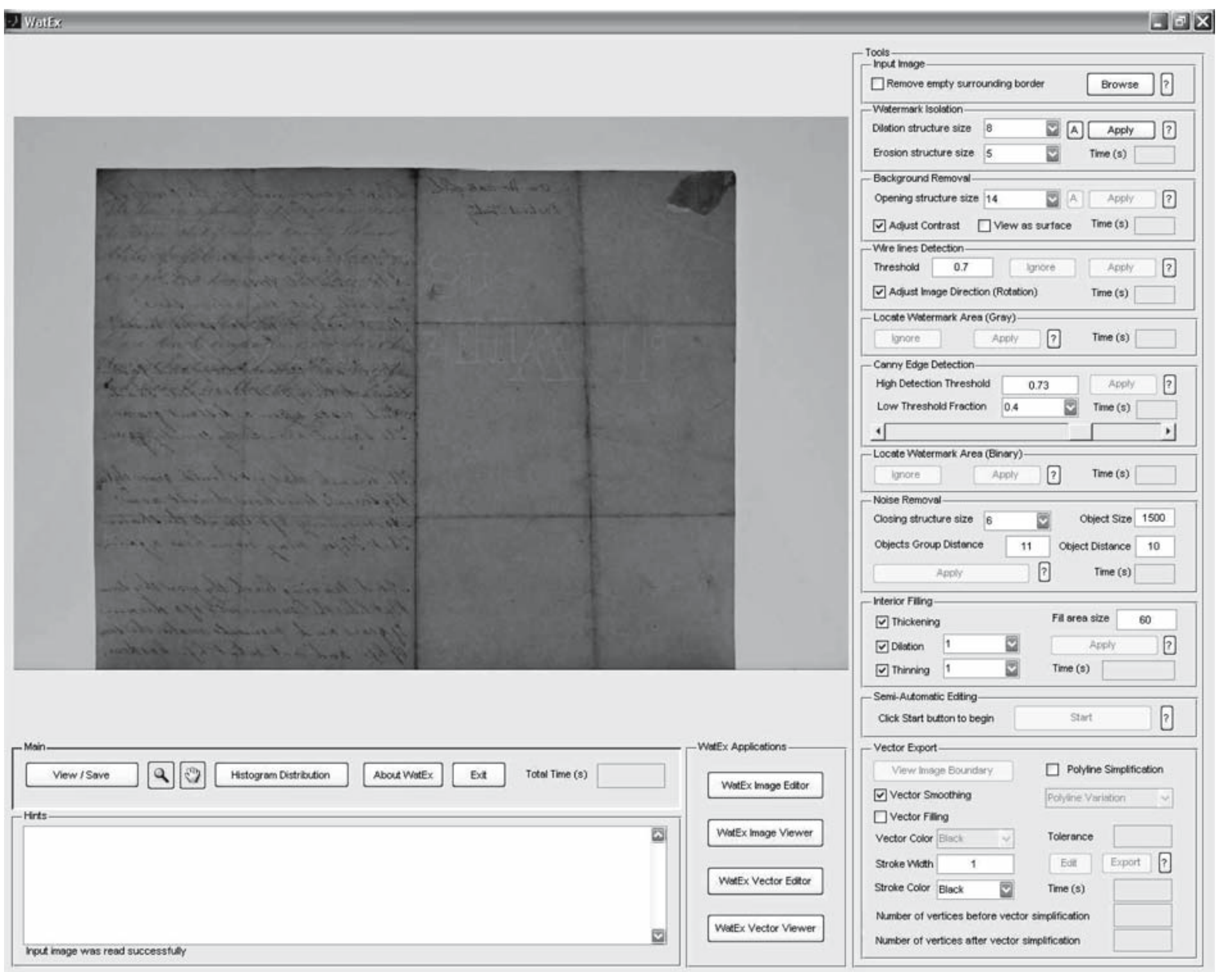

Fig. 11 Main graphical interface of the system

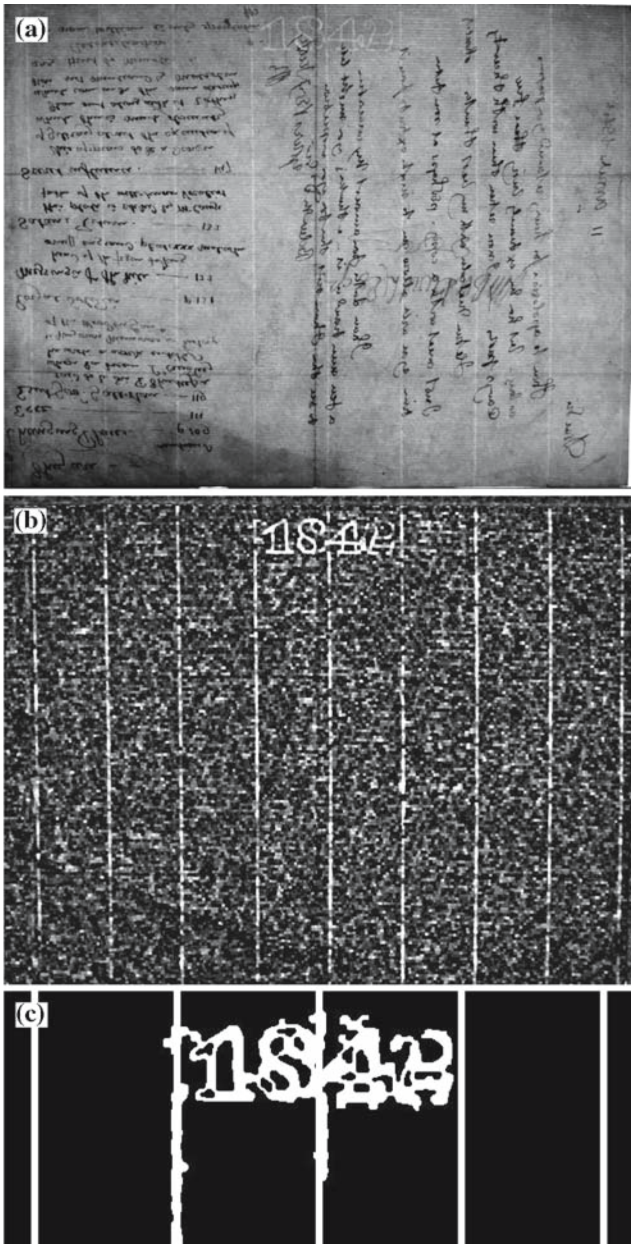

Fig. 12 Sample input 1 with handwritten watermarked paper **a** input source image digitised with back-lighting, **b** pre-processed intermediate output, **c** segmented and vectorised watermark design

a set of controls on the right-hand panel. The prototype can be operated in a fully automated manner or manually executing a step at a time to trace all the main stages of the processing and this is particularly useful for the development stage.

Figures 12, 13, 14, and 15 illustrate a selection of the results obtained with the current prototype system. For each sample, we present the key processing stages with the digitised input image and the intermediate and final results.

Figure 12a shows an example of a historical watermarked paper with handwriting (ink) on both front and back of the paper, noisy, and non-uniform background. It has been observed that the watermark and its chainlines are brighter than other features in the paper structure, the watermark signal becomes clear in the intermediate result after removing the foreground and background interferences as illustrated in Fig. 12b, where Fig. 12c demonstrates the output watermark pattern (zoomed for better visualisation) with the detected chainlines.

Another example of a historical paper with low foreground interference is provided in Fig. 13a. The paper has a noisy background which obstructs the watermark design, but these interferences were successfully removed after pre-processing stage as illustrated in Fig. 13b. The final output can be found in Fig. 13c. The segmentation is clean and contains only the extracted watermark pattern.

Figure 14a illustrates another example of a historical watermarked paper, which has a low watermark signal. This example is a musical manuscript with handwritten music notation, expressive symbols; text and signature, with both foreground and background interferences (mainly hand-drawn horizontal stave-lines). Figure 14b demonstrates the intermediate result after interference removal. The final result of the watermark design segmentation is presented in Fig. 14c (zoomed to provide better visualisation).

An example of a currently available watermarked paper is shown in Fig. 15a. In this example, the paper does not contain any writing. It has a uniformly textured background. The watermark pattern is partially interfered with the background pattern and cannot be clearly seen (by eye), and hence the quality and completeness of the segmented watermark design is hindered as demonstrated in Fig. 15b. Figure 15c shows the segmented watermark design segmented, and Fig. 15d illustrates the vectorised representation.

5 Conclusions and future plans

This paper presented a prototype to extract paper-based watermarks; this prototype has the ability to identify and remove different kinds of foreground and background interference. This is an important feature since this project is targeted at processing manuscript papers that have been written on. The prototype presented has been used to extract and trace watermark patterns successfully using a set of test data with a various range of watermark patterns (e.g., name of the manufacturer, geometric designs).

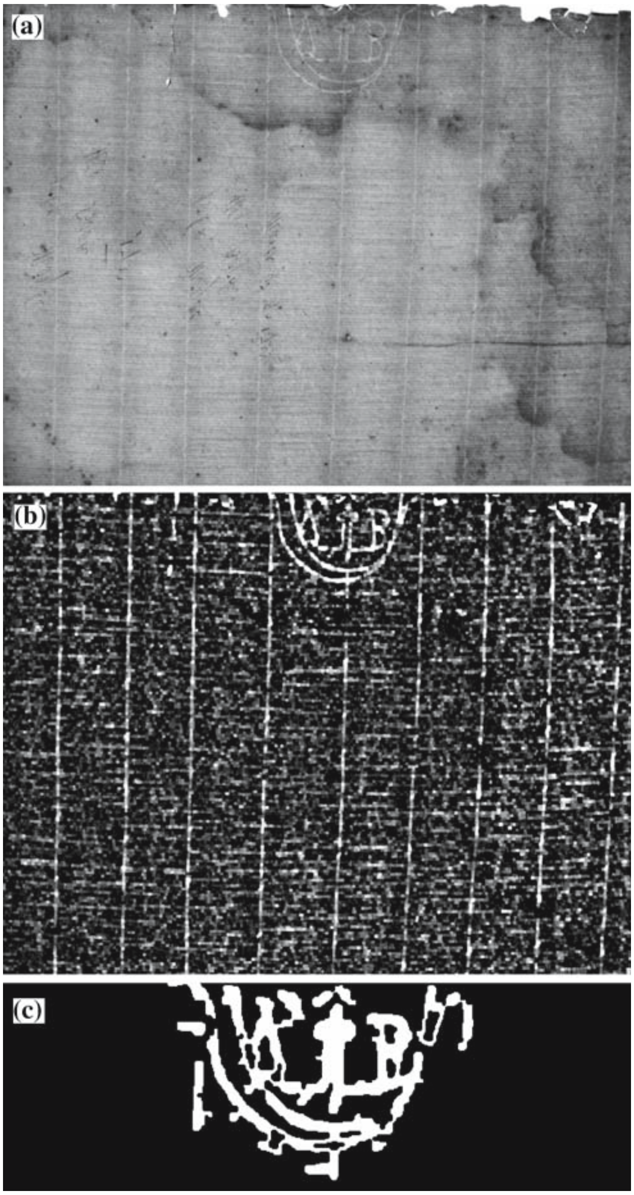

Fig. 13 Sample input 2 with low foreground interference **a** input source image digitised with back-lighting, **b** pre-processed intermediate output, **c** segmented and vectorised watermark design

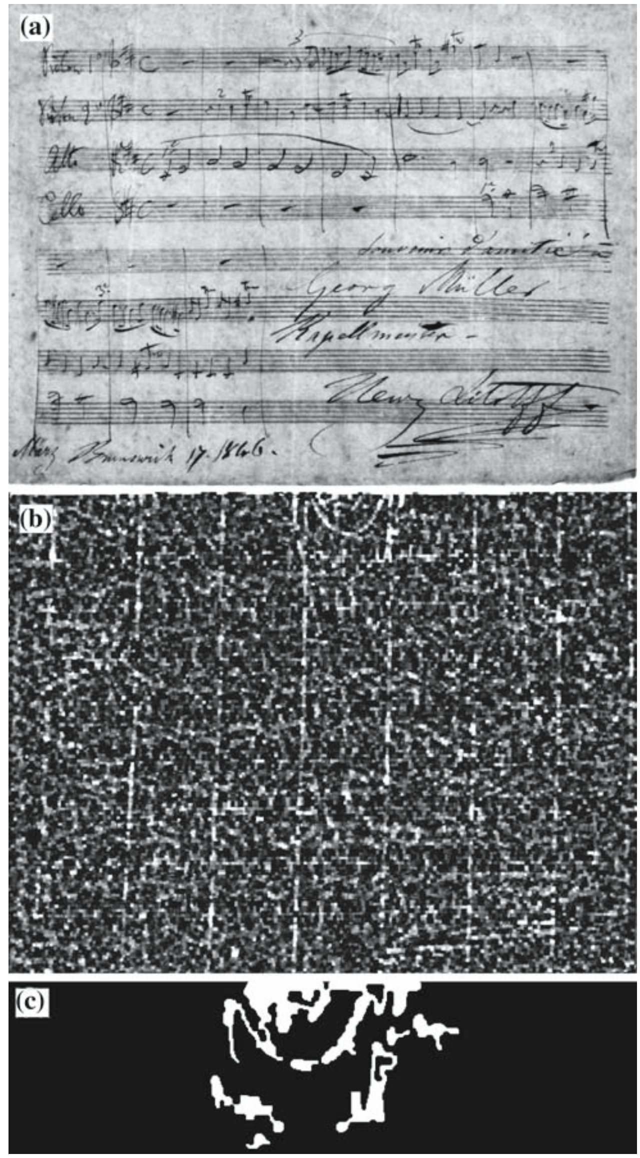

Fig. 14 Sample input 3 with handwritten music manuscript **a** input source image digitised with back-lighting, **b** pre-processed intermediate output, **c** segmented and vectorised watermark design

The paper focus was on designing a system to automatically trace and extract paper watermarks in order to offer its digital preservation. The paper also presented the detection of watermark chainlines and dynamic adaptation of the algorithms to automatically determine optimal structuring size and threshold values.

The paper also presented examples processing, sample results, and discussed the applications with different sources including old and modern watermarked papers and different types of writing including graphical notation.

This paper reported advancements in the digital extraction and classification of watermark design from paper. It provides technology-enhanced techniques and automation for digital watermark preservation and archiving. This approach has its advantages in watermark studies over many existing methods due to its simplicity and usability.

Future developments include enhancing the detail feature of the extracted watermark by computing the mean shape from a collection of watermarked documents that hold the same watermark design, and

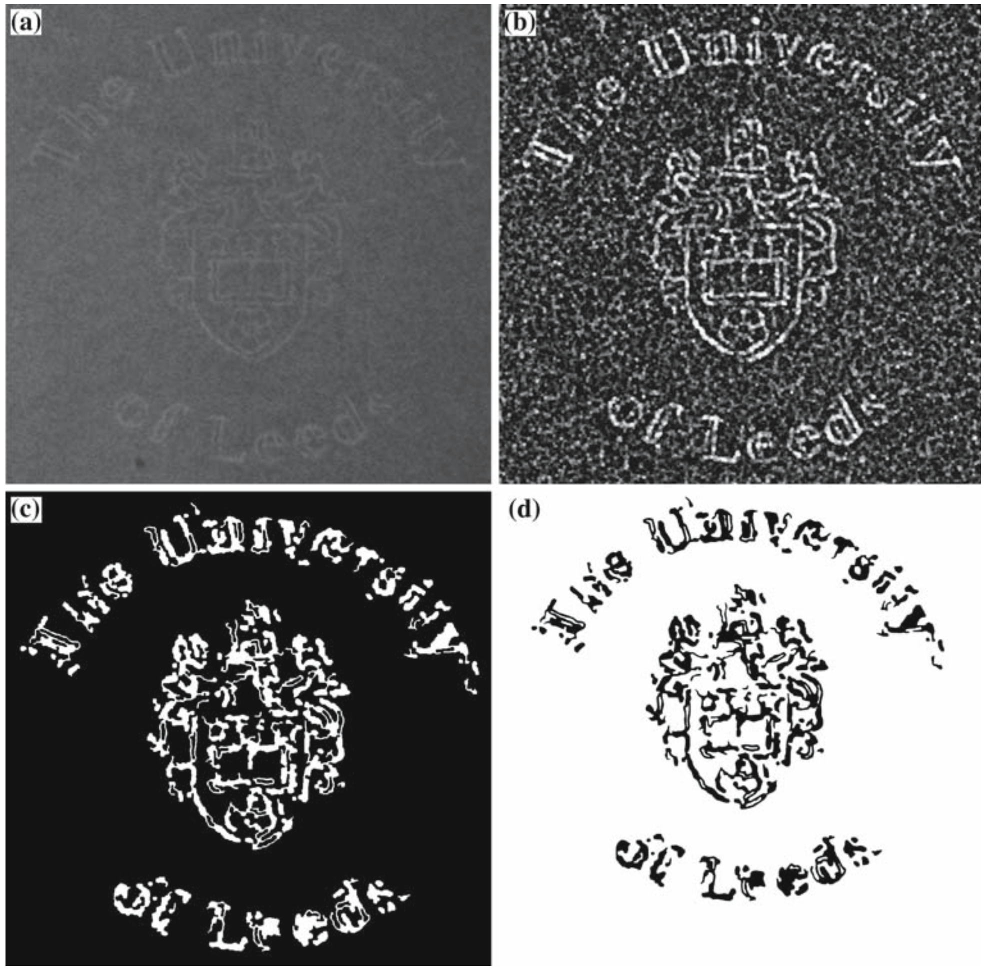

Fig. 15 Sample input 4 with currently available watermarked paper **a** input source image digitised with back-lighting, **b** pre-processed intermediate output, **c** segmented watermark design, **d** and its vectorised representation

combining partial watermarks from different documents back to its full design.

There are many different applications for this system, including digital embedding of the extracted watermark design from analogue paper-based materials into digital media, as an application for authentication and copyright protection purposes in digital libraries, watermark design cataloguing, archiving, searching, matching, and retrieval.

Acknowledgments This research is supported by the University of Jordan scholarship ref. no. 8/1918/10892. Thanks to the Special Collections of the Leeds University Library (http://www.leeds.ac.uk/library/spcoll/index.htm) for the manuscripts and other test samples used in this research project. Thanks to the University of Leeds Interdisciplinary Centre for Scientific Research in Music (http://www.icsrim.org.uk) for the hardware setup and support for the digitisation. The authors would also like to thank the anonymous reviewers and Prof Roger Boyle for their valuable comments.

References

1. Allison, R., Hart, J.: Paper watermarks, paper making, and dylux proofing paper. http://abacus.bates.edu/Faculty/wmarchive/wm-initiative/index.html. Last accessed: 12th November 2006
2. Ash, N.: Recording watermarks by beta-radiography and other means. In: The Book and Paper Group Annual, vol. 1. American Institute for Conservation of Historic and Artistic Works, Washington (1982)
3. Badke, D.: Paper-based watermarks and beta-radiography method. http://gateway.uvic.ca/lydgate/lydgatems/pages/manuscript/watermarks/watermarks.htm. Last accessed: 12th November 2006
4. Canny, J.: A computational approach to edge detection. IEEE Trans. Pattern Anal. Mach. Intell. **8**(6), 679–698 (1986)
5. Gants, D.: Pictures for the page: techniques in watermark reproduction, enhancement and analysis (1994). http://jefferson.village.virginia.edu/gants/BibSocUVa/paper.html. Last accessed: 12th November 2006
6. Gonzalez, R., Woods, R.: Digital Image Processing, 2nd edn. Prentice-Hall, Upper Saddle River (2002)

7. Hiary, H., Ng, K.: Watermark: from paper texture to digital media. In: Proceedings of International Conference on Automating Production of Cross Media Content for Multi-channel Distribution conference (AXMEDIS ’05), pp. 261–264. IEEE Computer Society Press, Florence, Italy (2005)
8. Karnaukhov, A., Merzlyakov, N., Milyukova, O., Karnaukhov, V., Wenger, E., Aizenberg, I.: Digital restoration of watermark images. In: Proceedings of EVA01, pp. 196–199. Centre PIC of Ministry Culture of Russia, STG, Moscow (2001)
9. Katzenbeisser, S., Petitcolas, F.: Information Hiding Techniques for Steganography and Digital Watermarking. Artech House, Boston (2000)
10. Lim, J.: Two-dimensional signal and image processing. Prentice-Hall, Englewood Cliffs (1990)
11. LIMA: Literary Manuscript Analysis. http://www2.warwick.ac.uk/fac/arts/ren/publications/lima/paper/. Last accessed: 12th November 2006
12. Rauber, C., O’Ruanaidh, J., Pun, T.: Secure distribution of watermarked images for a digital library of ancient papers. In: Proceedings of the Second ACM International Conference on Digital Libraries (DL’97), pp. 123–130. Philadelphia, PA, USA (1996)
13. Rauber, C., Tschudin, P., Pun, T.: Retrieval of images from a library of watermarks for ancient paper identification. In: Proceedings of EVA’97—Elektronische Bildverarbeitung und Kunst, Kultur, Historie, vol. 14 (1997)
14. Rauber, C., Tschudin, P., Startchik, S., Pun, T.: Archival and retrieval of historical watermark images. In: Proceedings of the International Conference on Image Processing (ICIP), pp. 773–776. IEEE Press Lausanne (1996)
15. SHREW: A Shape Retrieval System for Watermarking Images project. http://www.chart.ac.uk/chart2001/papers/noframes/mulholland.html. Last accessed: 12th November 2006
16. Sonka, M., Hlavac, V., Boyle, R.: Image Processing, Analysis, and Machine Vision, 2nd edn. PWS Publishing, Pacific Grove (1999)
17. Stevenson, A.: Watermarks are twins. Stud. Bibliogr. **4**, 57–91 (1951-2)
18. Vander Meulen, D.: The identification of paper without watermarks. Stud. Bibliogr. **37**, 58–81 (1984)
19. Wenger, E., Karnaukhov, V., Haidinger, A., Merzlyakov, N., van Thienen, G., Oukhanova, E., Erastov, D.: A distributed database and processing system for watermarks: an intas project. In: Proceedings of EVA01, pp. 200–206. Centre PIC of Ministry Culture of Russia, STG, Moscow (2001)
20. Whelan, P., Soille, P., Drimbarean, A.: Real- time registration of paper watermarks. Real-Time Imaging **7**(4), 367–380 (2001)