A Textual Information Detection and Elimination System for Secure Medical Image Distribution *

James Ze Wang, M.S., Michel Bilello, Ph.D., Gio Wiederhold, Ph.D.
School of Medicine, Stanford University, Stanford, CA 94305

Before medical images can be distributed online, it is necessary for confidentiality reasons to eliminate text that appears in the image. This paper describes TIDE (Textual Information Detection and Elimination) system for secure medical image distribution. The algorithm uses Daubechies' wavelets to detect and eliminate areas of textual information within digital medical images. The system is practical for real-world applications, processing each 512×512 *image in about 10 seconds. Besides its exceptional speed, the algorithm has demonstrated a remarkable accuracy when tested on various types of medical images, including X-ray and CT scans.*

INTRODUCTION

Health care is exceptionally information intensive, and the United States spends hundreds of billions of dollars each year in processing and managing such information. However, it is becoming increasingly difficult to maintain and retrieve health care information manually as more and more advanced medical equipment is used in diagnosis and management of disease. Besides the traditional textual data such as patient reports, health care records are being filled with image scans, 3-D volume reconstructions, and video streams.

As the demand for greater accessibility to health care information grows, medical institutions are being urged to make information available to legitimate external parties in a timely fashion (*e.g.*, on-line) while protecting the privacy of patient data. It is therefore crucial that health care institutions be provided with on-line tools that allow them to disseminate medical information without compromising data privacy. In this paper, we present an algorithm that strips textual information (including identifying information) from medical images such as digitized x-rays and CT scans. The resulting processed images can then be made available to medical researchers, physicians, and other legitimate users. Such a tool could be used by health care institutions and other repositories of medical images as part of a data security system.

* *Correspondence to:* wangz@cs.stanford.edu

Related Work

Recent advances in content-based image retrieval and digital library management have made it possible to retrieve multimedia data efficiently and effectively. Interested readers are referred to [7, 11, 15, 16, 17, 19].

Most of the work in medical database security has focused on authentication and encryption. For example, the security mechanisms of Telemed [8]—a system that allows sharing among physicians of multimedia patient information across remote sites—are limited to RSA encryption and access control lists. Similarly, WebReport [13], a medical multimedia reporting system, implements a traditional security scheme using registered usernames and passwords.

In contrast, our work fits in a security framework based on *content* of information. It is not sufficient to grant or deny access to information solely on the basis of access control. Researchers or physicians, for example, could benefit from accessing medical images from remote sites, even in cases when they should not see patient names. Our algorithm allows such images to be readily shared without compromising patient privacy.

The problem of text identification [10] arises in many applications other than medical security. Document understanding systems locate text and figure captions on a page for processing by optical character recognizers. The detection of text in scanned maps and mechanical, electrical, and piping drawings is important for converting the paper form to computer-analyzable form. Work done by University of Maryland [5, 6] uses neural network, texture and multiresolution analysis to segment the documents into areas of text and areas of image or graphics. However, the algorithms used in such systems are not designed to handle superimposed text. Moreover, neural network algorithms are not suitable for real-time processing.

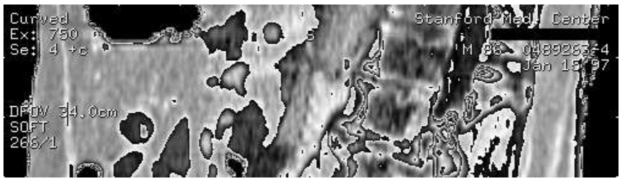
Overview of Our Work

In the TIDE project, we developed an efficient and accurate algorithm to distinguish areas with and without textual information in medical images. Because variations in the diagonal directions can be found in almost all Roman characters or Arabic numbers, we use Daubechies' wavelets and post-processing techniques to detect the high frequency variation in the diagonal direction that is indicative of text. Promising results have been obtained in experiments using a set of real-world medical images, many with superimposed text.

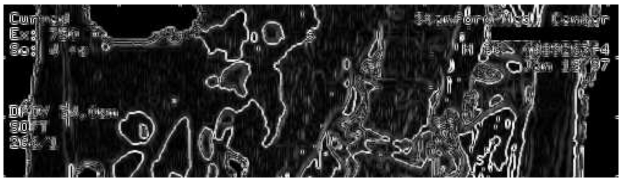
BACKGROUND

Wavelets, developed in mathematics, quantum physics, and statistics, are functions that decompose signals into different frequency components and analyze each component with a resolution matching its scale. Applications of wavelets to signal denoising, image compression, image smoothing, fractal analysis and turbulence characterization are active research topics.

Daubechies' wavelet has been used in content-based image retrieval [19] and a system for screening objectionable images [20]. Various experiments and studies have shown that because Daubechies' wavelets can better represent continuous functions with continuous derivatives, they are better suited for natural signals or images than other wavelets.



Original image



Traditional edge detection

Figure 1: **Traditional edge detection does not distinguish areas with and without textual information.**

In textual information detection, especially for superimposed text, we want to distinguish areas with and without text information as effectively as possible. When using the Haar wavelet, we obtain too much noise in the high-pass bands within the non-text areas. Traditional edge detection algorithms have the same problem, as illustrated in Figure 1. For both of the two algorithms, it would be difficult to differentiate the edges of text from the edges of the objects in the image.

ALGORITHM

Overview

We have developed a new textual information detection and elimination algorithm for digital medical images using Daubechies' wavelet transforms. Figure 2 shows the basic structure of the algorithm.

We apply a 1-level fast wavelet transform (FWT) with Daubechies' Symlet-8 wavelet or Daubechies-8 wavelet to each image. Then we extract and post-process the lower right-hand corner of the transform matrix, where the diagonal directional high frequency information is located, to obtain a mask containing only the areas with textual information. Once such a mask is computed, we may apply it to the original image to eliminate the areas with textual data.

Our design has several immediate advantages.

1. Unlike traditional approaches, such as the neural network, our algorithm does not depend on the actual font and style of the text in the medical image. Preliminary experiments indicate that the algorithm is capable of handling images with superimposed hand-written text. Figure 9 shows one example.
2. We used Daubechies' wavelets rather than a traditional edge detector to capture the high frequency information in the images. This reduced the dependence of the results on the quality or the sharpness of the images.
3. It does not rely on the color of the image or the text. It also has minimum dependence on the contrast between text and background objects.
4. It has potential to be much faster than other algorithms.

Pre-processing

Many medical image formats are currently in use, e.g., DICOM, PPM, GIF, JPEG and TIFF are the most widely used formats. Because the images may have different format, we must first normalize the data for computation. A gray scale PPM image of any size is adequate for our algorithm. A color medical image may be converted to a gray scale image using the equation $WB = (R+G+B)/3$, where R, G and B are values of a pixel in the RGB color space and WB is the value of this pixel in gray scale. The range of the values of each pixel, or the number of bits per pixel, for the PPM image is not limited for our algorithm. Usually it is adequate to use 8 bits per pixel to store a reasonably clear gray-scale medical image.

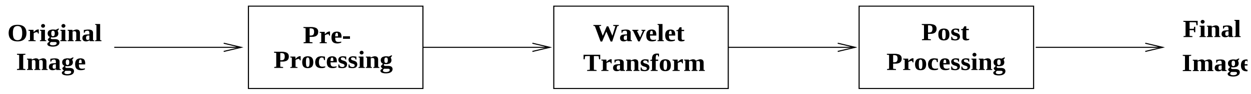


Figure 2: **Basic structure of the algorithm used in TIDE.**

Wavelet Transform

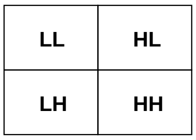
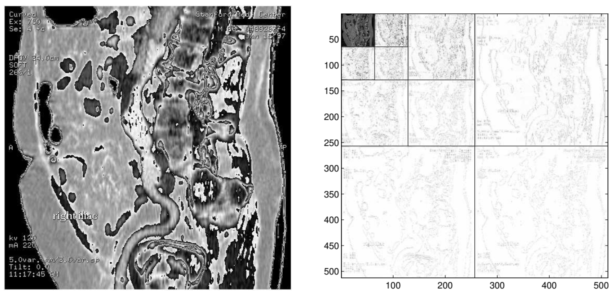


Figure 3: **Naming convention for a 1-level wavelet transform.**

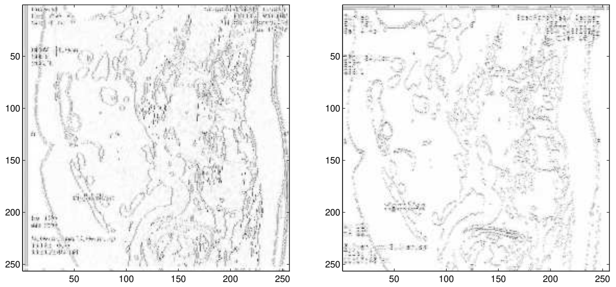
In this step, we apply a wavelet transform to the image obtained from the pre-processing step. Our purpose is not to obtain a high quality edge detection algorithm for this application. Rather, since the goal here is to effectively distinguish the areas with and without textual information, it is not necessary to produce a perceptually pleasant edge image. Consequently, we try to keep the algorithm simple to achieve a fast computation speed.



Original image *Wavelet transform*

Figure 4: **Daubechies' wavelet transform in TIDE.** Name of the patient is blackened in the original image.

We start the process by transforming the the gray-scale PPM image converted from the pre-processing using the Daubechies' Symlet-8 or Daubechies-8 wavelet basis. Figure 4 shows the wavelet transform on a sample medical image. The image is decomposed into four frequency bands with corresponding names marked in Figure 3. The notation is borrowed from the filtering literature [18]. The letter 'L' stands for low frequency and the letter 'H' stands for high frequency. The left upper band is called 'LL' band because it contains low frequency information in both the row and column directions. We avoid the details of explaining the filtering terminologies here; interested readers are referred to [18]. An even number of columns and rows is required in the image due to the downsampling process of the wavelet transform. However, if the number of rows and columns is odd, we simply delete one column or one row of pixels from the boundaries.



HL band *LH band*

Figure 5: **HL band and LH band are not suitable for detecting textual information within medical images.**

The LH frequency band is sensitive to the horizontal edges, the HL band is sensitive to the vertical edges, and the HH band is sensitive to the diagonal edges [4]. For the medical images that our system is designed for, the HH band is much better dealing with the distinctions between areas with and without text. In fact, variations in the diagonal directions can be found in almost all Roman characters or Arabic numbers. Such variations are detected much more frequently in areas with textual information than those with only medical objects, if we make a reasonable assumption that the text in the medical image is small, in general, compared to the objects in the image. The LH bands and the HL bands are not useful for distinguishing areas with and without textual information. As shown in Figure 5, vertical clusters of points can be found in both text areas and non-text areas in the HL band, and horizontal clusters of points can be found for the LH band. Therefore, we discard both of these bands and retain only the HH band for post-processing.

Post-Processing

Post-processing is required to avoid the incorrect elimination of diagonal-wise variations in areas of the image without text. Without loss of generality, we as-

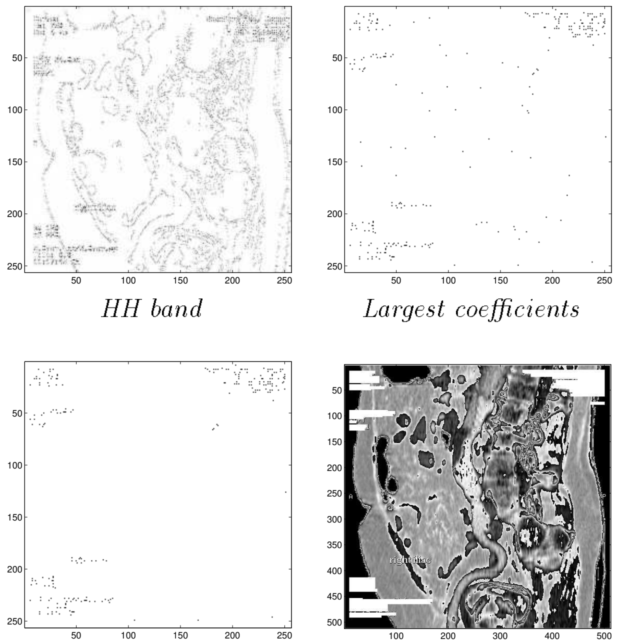


Figure 6: **Post Processing Step in the TIDE algorithm.** Note that only corners of the original images are considered.

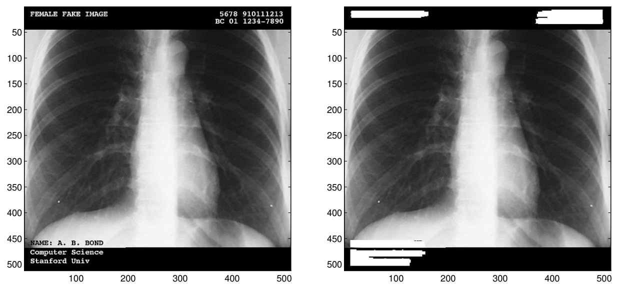


Figure 7: **Another image processed by the TIDE system.** Text Information simulated.

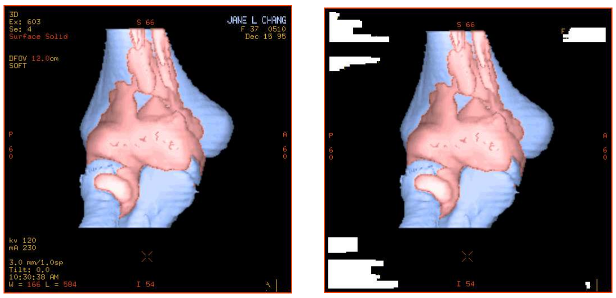


Figure 8: **A full color medical image processed by the TIDE system.** Note that only corners of the original images are considered.

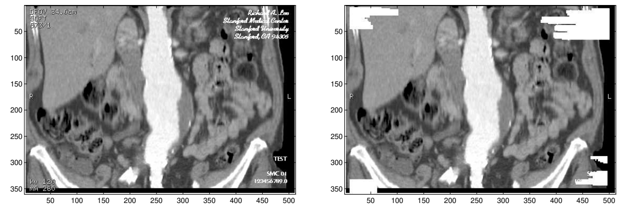


Figure 9: **A medical image with hand-written text processed by the TIDE system.** Text Information simulated.

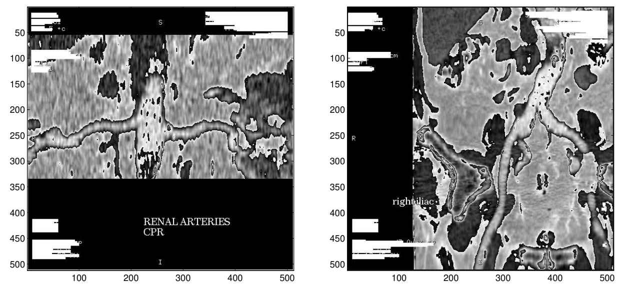


Figure 10: **Two more sample final images.** Note that only corners of the original images are considered.

sume that the original image is of size $2n \times 2n$. Then the wavelet transform is a matrix of size $2n \times 2n$. In the post-processing step, we process the $HH(1:n, 1:n)$ matrix[1] obtained from the previous step. A binary matrix, denoted as $B(1:n, 1:n)$, is constructed from the HH matrix so that the largest $M = O(n)$ coefficients in magnitude in the HH matrix are replaced by 1 and all other coefficients are replaced by 0. Then we use a moving square window matrix of about 20×20 pixels to determine the isolated points in the binary matrix by setting a threshold for the minimum number of non-zero points in such a moving window. These isolated points are then deleted because they represent diagonal-wise variations in the areas without text.

Denote $B'(1:n, 1:n)$ the matrix without these isolated points converted from $B(1:n, 1:n)$. Then we group up the remaining points in the matrix $B'(1:n, 1:n)$ to form a matrix $Mask(1:n, 1:n)$ containing detected textual areas. Finally, we rescale $Mask(1:n, 1:n)$ to $2n \times 2n$ and apply it to the original image to obtain the final image. Figure 6 shows the post-processing on a sample medical image.

[1] Here we use MATLAB notation. That is, $A(m_1:n_1, m_2:n_2)$ denotes the submatrix with opposite corners $A(m_1, m_2)$ and $A(n_1, n_2)$.

RESULTS

This algorithm has been implemented on a Sparc-20 workstation. We have tested about 30 medical images of different types, collected from different sources. Some of them are downloaded from the world-wide web and medical imaging newsgroups, while others are provided by the Stanford Medical Center.

It takes about 10 seconds of CPU time to process each medical image of size 512×512. Besides the fast speed, the algorithm has achieved remarkable accuracy. It successfully detected and eliminated all of the critical textual information within the corners of the medical images.

Figure 10 and Figure 7 show some sample results on gray scale medical images processed by the TIDE system. Figure 8 shows the results on a full color medical image.

CONCLUSIONS

In this paper, we have demonstrated an efficient textual information detection and elimination system for secure medical image distribution. The algorithm uses Daubechies' wavelets to detect and eliminate areas of textual information within digital medical images.

We are working on applying this technique to large number of real-world medical images. We are also trying to improve this technique so that only texts related to patients' private information, e.g. patient name or patient identification number, are eliminated.

Acknowledgements

We would like to thank Oscar Firschein, Visiting Scholar, Stanford University Computer Science Department for valuable help in writing this paper. We would also like to thank the Stanford University Libraries and Academic Information Resources (SULAIR) for providing computer equipment during the development and testing process.

References

[1] Charles K. Chui, *An Introduction to Wavelets*, Academic Press, Inc., San Diego, 1992.

[2] Charles K. Chui, *Wavelets: A Tutorial in Theory and Applications*, Academic Press, Inc., San Diego, 1992.

[3] Ingrid Daubechies, Orthonormal bases of compactly supported wavelets, *Communications on Pure and Applied Math.*, 41(7):909-996, October 1988.

[4] Ingrid Daubechies, *Ten Lectures on Wavelets*, CBMS-NSF Regional Conf. Series in Applied Math., 1992.

[5] David Doermann, Document Understanding Research at Maryland, *DARPA Image Understanding Workshop Proc.*, Vol 2, p.817-826, 1994.

[6] K. Etemad, D. Doerman, R. Chellappa, Page Segmentation Using Decision Integration and Wavelet Packet Basis, *Proc. Int. Conf. on Pattern Recognition*, 1994

[7] C. Faloutsos et al, Efficient and Effective Querying by Image Content, *J. of Intelligent Information Systems*, 3:231-262, 1994.

[8] David W. Forslund et al, Experiences with a Distributed Virtual Patient Record System, *Proceedings of the 1996 AMIA (formely SCAMC) Conference*, Washington DC, October 1996.

[9] Rafael C. Gonzalez, Richard E. Woods, *Digital Image Processing*, Addison-Wesley Publishing Co., 1993.

[10] A.K.Jain, S. Bhattacharjee Text Segmentation Using Gabor Filters for Automatic Document Processing, *Machine Vision and Applications*, 5:169-184, 1992.

[11] C. E. Jacobs, A. Finkelstein, D. H. Salesin, Fast Multiresolution Image Querying, *Proceedings of SIGGAPH 95, in Computer Graphics Proceedings, Annual Conference Series*, pp.277-286, August 1995.

[12] Gerald Kaiser, *A Friendly Guide to Wavelets*, Birkhauser, Boston, 1994.

[13] Henry J. Lowe, Ilya Antipov, William K. Walker, Stacey E. Polonkey and Gregory J. Naus, WebReport: A World Wide Web Based Clinical Multimedia Reporting System, *Proceedings of the 1996 AMIA (formely SCAMC) Conference* Washington DC, October 1996.

[14] Yves Meyer, *Wavelets: Algorithms & Applications*, SIAM, Philadelphia, 1993.

[15] W. Niblack et al, The QBIC project: Query image by content using color, texture and shape, *Storage and Retrieval for Image and Video Databases*, pages 173-187, San Jose, 1993. SPIE.

[16] R. W. Picard, T. Kabir, Finding Similar Patterns in Large Image Databases, *IEEE ICASSP*, Minneapolis, Vol, V., pp.161-164, 1993.

[17] A. Pentland, R. W. Picard, S. Sclaroff, Photobook: Content-Based Manipulation of Image Databases, *SPIE Storage and Retrieval Image and Video Databases II*, San Jose, 1995.

[18] Martin Vetterli, *Wavelets and Subband Coding*, Prentice Hall, N.J., 1995.

[19] James Ze Wang, Gio Wiederhold, Oscar Firschein, Sha Xin Wei, Wavelet-Based Image Indexing Techniques with Partial Sketch Retrieval Capability, *Proc. of the 4th Forum on Research and Technology Advances in Digital Libraries*, Washington D.C., May 1997.

[20] James Ze Wang, Gio Wiederhold, Oscar Firschein, System for Screening Objectionable Images Using Daubechies' Wavelets and Color Histograms, *Proc. of the 4th European Workshop on Interactive Distributed Multimedia Systems and Telecommunication Services*, Darmstadt, Germany, September 1997.