SECURITY FILTERING OF MEDICAL IMAGES USING OCR

James Z. Wang
School of Information Sciences and Technology
The Pennsylvania State University
University Park, Pennsylvania 16802, USA
wangz@cs.stanford.edu
wang.ist.psu.edu

Abstract

Before digital medical images in computer-based patient record systems can be distributed online, it is necessary for confidentiality reasons to eliminate patient identification information that appears in the images. We present an automatic security filtering algorithm for on-line medical image distribution using Daubechies' wavelets [Dau92] and Optical Character Recognition (OCR). The system is practical for real-world applications, processing and coding each 12-bit image of size 512 x 512 within 2 seconds on a Pentium Pro. Besides its exceptional speed, the security filter has demonstrated high accuracy in detecting sensitive textual information within current or digitized previous medical images. The algorithm is of linear run time.

INTRODUCTION

With the advancements of the World-Wide Web, the Internet, and medical imaging technology, it is becoming increasingly difficult to maintain and retrieve digital health care information. Besides the traditional textual data such as patient reports, health care records are being filled with X-ray images, MRI scans, CT scans, 3-D volume reconstructions, and video streams. Efficient security filtering for digital medical images is desirable before medical images (Figure 1.) can be transmitted to researchers and external users.

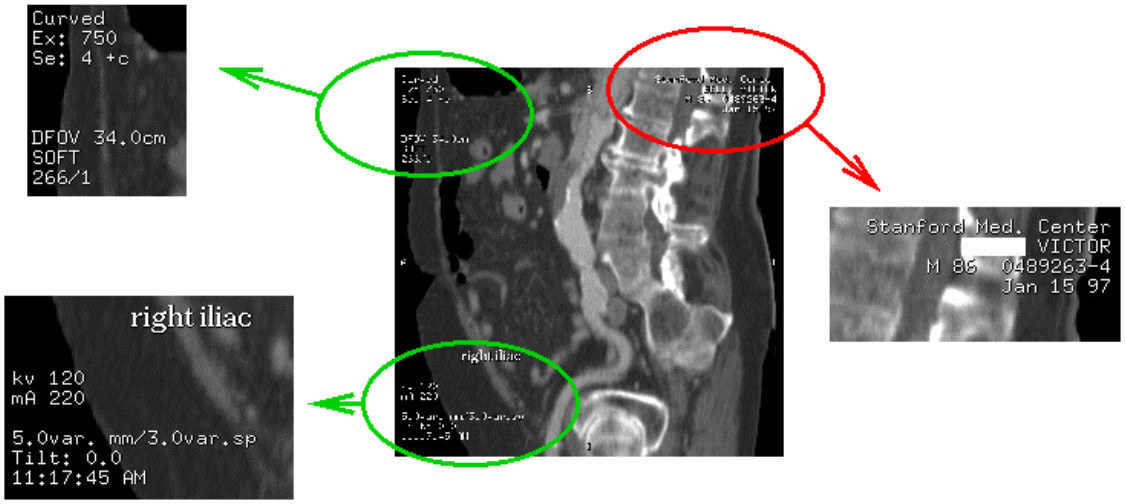

Figure 1. Text in medical images.

In this paper, we present a wavelet-based medical image-filtering algorithm that can detects textual information (including identifying information) from some current or digitized previous medical images. Optical Character Recognition (OCR) technology is used to covert pixel information to text. Textual terms not known to be innocuous are eliminated [Wi96]. The resulting processed images can then be made available to medical researchers, second-opinion physicians, students, and other legitimate users after being processed by our algorithm. Healthcare institutions and other medical image repositories may use such

systems in their medical image distribution systems. The system can be combined with other wavelet-based image analysis and retrieval systems [Wa01.1, Wa01.2].

BACKGROUND

With the DICOM standard, it is easy to eliminate textual information such as patient name and ID. However, for digitized films or previous history images, a computerized detection and elimination algorithm is needed. The problem of text identification[Ja92,Ta97] 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 [Do94,Et94] 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 because it is difficult to differentiate the edges of text from the edges of the medical objects in the image.

The security filtering process in our system consists of an efficient and accurate algorithm to distinguish areas with and without textual information in digital or digitized medical images. Areas with text can then be blurred or striped. Because variations in the diagonal directions can be found in almost all Roman characters or Arabic numbers, we use Daubechies' wavelets and analysis techniques to detect the high frequency variation in the diagonal direction that is indicative of text. A mask is used to preserve the losslessness of non-textual areas. With some basic knowledge of the machine used to create the image, we are able to eliminate only sensitive patient identification information while retaining the medical information in the image. Excellent results have been obtained in experiments using a large set of real-world medical images, many with superimposed text.

THE SYSTEM

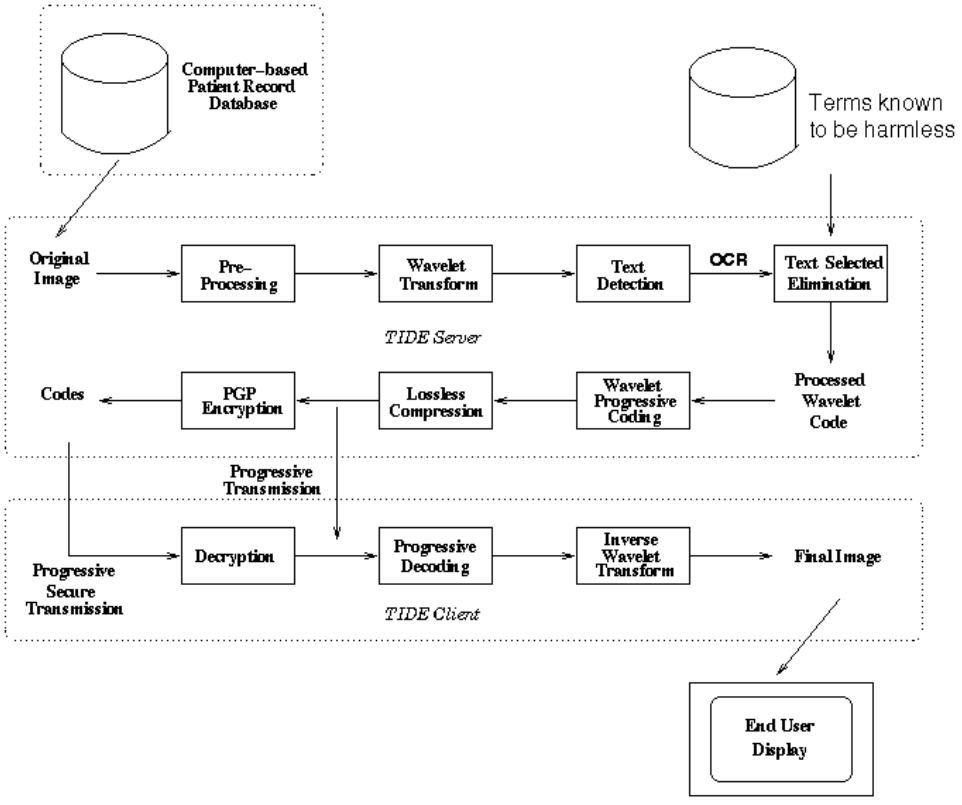

Figure 2. Architecture of the system.

The architecture of the system is shown in Figure 2. In this paper, we focus on the security filtering process in the system.

We apply an N-level fast wavelet transform (FWT) with Daubechies-4 wavelet to each medical image, where N is determined adaptively by the image size. If the image is of DICOM standard, we may eliminate the patient identification information without processing the image content.

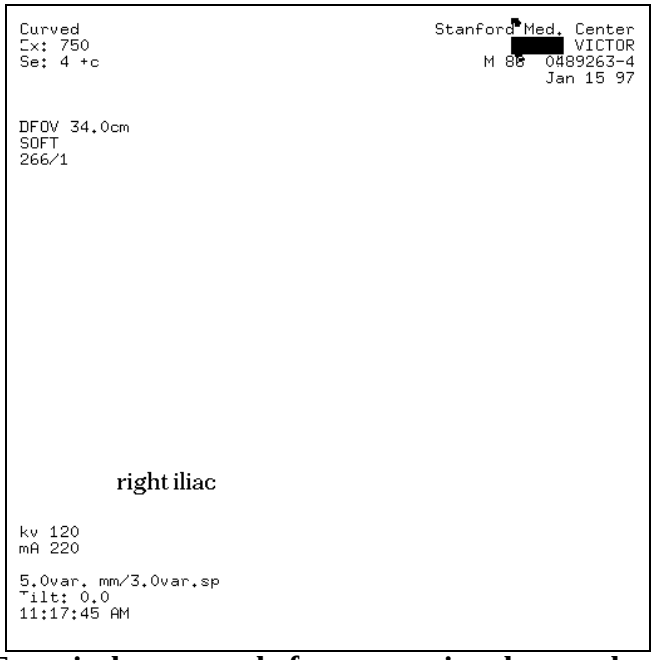

Figure 3. Text pixels extracted after processing the wavelet transform.

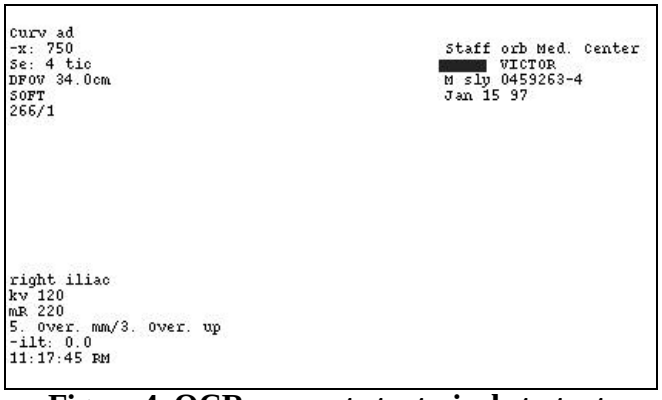

Figure 4. OCR converts text pixels to text.

For non-DICOM images, we extract and analyze the lower right-hand corners of each level 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 apply it to all the high-frequency bands to eliminate the text within areas with textual data. Or, we may apply the mask selectively to all the frequency bands to block the areas with text. Knowledge of the rough location (e.g., which one of four corners) of the critical patient identification information of certain type of medical images or the TIHI (Trusted Interoperation of Health care Information) system [Wi96] is used to eliminate only information needed to be deleted while preserving the rest. When we do not have knowledge of the rough location of patient identification information, we may apply the mask to eliminate all textual

information within the medical image. Figures 3, 4, and 5 show the application of the process to one medical image.

Figure 5. Patient identification information is eliminated.

To achieve secure transmission, we may also apply a PGP encryption [Ga95] to the code segments before sending the data to the client via public network.

Our text detection algorithm has several immediate advantages.

1. Unlike traditional approaches, such as the neural network, our algorithm does not depend on the actual font size, font type and style of the text in the medical image. Experiments indicate that the algorithm is capable of handling images with superimposed hand-written text and even foreign languages.
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. The algorithm 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 is faster than other algorithms due to our adaptive multiresolution approach.
5. Wavelet-based algorithm using Daubechies' wavelets can be easily integrated with cutting-edge image compression, compressed-domain indexing and processing algorithms.

RESULTS

This algorithm has been implemented on a Pentium Pro 200MHz workstation. We have tested about 100 medical images of different modalities, 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.

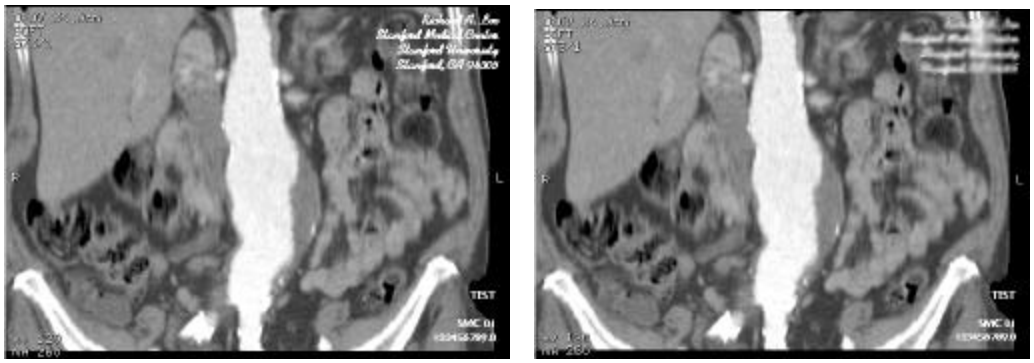

Figure 6. The system can handle hand-written text.

The textual information detection and elimination module takes about 1 second of CPU time to process a 12-bit medical image of size 512 x 512. The algorithm is a linear algorithm with respect to the size of the image. 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.

Figures 5 and 6 show some sample results on gray scale medical images processed by the system. The areas without text are maintained without loss. The algorithm can also be applied to color medical images.

CONCLUSIONS

In this paper, we have demonstrated an efficient wavelet-based security filtering algorithm for on-line medical image distribution. The algorithm uses Daubechies' wavelets and OCR to detect and eliminate textual information within digitized medical images, while maintaining non-textual areas lossless.

The system is practical for real-world applications, processing and coding each 12-bit image of size 512 x 512 within 2 seconds on a Pentium Pro. Besides its exceptional speed, the security filter has demonstrated a remarkable accuracy in detecting sensitive textual information within digital medical images.

Acknowledgments

This work was supported in part by the National Science Foundation Grant No. IIS-9817511. We would like to thank Gio Wiederhold, Jia Li, Oscar Firschein, Michel Bilello, and Stephen Wong for valuable help and discussions.

References

[Dau92] I. Daubechies, Ten Lectures on Wavelets, Capital City Press, 1992.

[Do94] D. Doermann, Document Understanding Research at Maryland, DARPA Image Understanding Workshop Proc. vol. 2, pp. 817-826, Science Applications, Inc, 1994.

[Et94] K. Etemad, D. Doerman, R. Chellappa, Page Segmentation Using Decision Integration and Wavelet Packet Basis, Proc. Int. Conf. on Pattern Recognition, IEEE, 1994

[Ga95] S. Garfinkel, PGP : Pretty Good Privacy, O'Reilly & Associates, Inc., CA, 1995.

[Ja92] A.K. Jain et al., Text Segmentation Using Gabor Filters for Automatic Document Processing, Machine Vision and Applications, vol. 5, pp. 169-184, 1992.

[Ta97] Y. Y. Tang et al., Quadratic Spline Wavelet Approach to Automatic Extraction of Baseline from Document Images, Proc. 4th Int Conf on Document Analysis and Recognition, IEEE, August 1997.

[Wa01.1] J. Z. Wang, J. Li, G. Wiederhold, ``SIMPLIcity: Semantics-sensitive Integrated Matching for Picture LIbraries," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 9, 16 pp., 2001.

[Wa01.2] J. Z. Wang, J. Li, R. M. Gray, G. Wiederhold, ``Unsupervised multiresolution segmentation for images with low depth of field," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 1, pp. 85-90, 2001.

[Wi96] G. Wiederhold, M. Bilello, V. Sarathy, X. Qian, A Security Mediator for Health Care Information, Proc. 1996 AMIA Conference, Washington DC, Oct. 1996, pp.120-124.