Patient Information Extraction in Digitized Radiography

Hsien-Huang P. Wu

Digital imagery is gradually replacing the traditional radiograph with the development of digital radiography and film scanner. This report presents a new method to extract the patient information number (PIN) field automatically from the film-scanned image using image analysis technique. To evaluate the PIN field extraction algorithm, 2 formats of label acquired from 2 different hospitals are tested. Given the available films with no constraints on the way the labels are written and positioned, the correct extraction rates are 73% and 84%, respectively. This extracted PIN information can link with Radiology Information System (RIS) or Hospital Information System (HIS), and the image scanned from the film then can be filed into the database automatically. The efficiency this method offers can simplify greatly the image filing process and improve the user friendliness of the overall image digitization system. Moreover, compared with the bar code reader, it solves the automatic information input problem in a very economical way. The authors believe the success of this technique will benefit the development of the PACS (Picture Archiving and Communication System) and teleradiology.

KEY WORDS: Document image analysis, film scanner, digitizer, picture archiving and communication system, patient information extraction, radiograph

IN RECENT YEARS, the popularity of picture archiving and communication systems (PACS) and teleradiology¹⁻³ for medical imagery has made film scanning,⁴ which is used to digitize radiograph, more common. Current practice in using the film scanner generally follows 3 steps: (1) input patient information, (2) scan the film, and (3) save the digitized image into the database using the patient information input. Although this procedure is simple and straightforward, it has 2 disadvantages. The first is that before the film was printed and scanned, the patients' information (eg, name, ID) had been recorded in the Hospital Information System (HIS) or Radiology Information

System (RIS) database when their radiograph was taken. Therefore, the step (1) is redundant. From our experience, this repeated step usually makes the technologists feel annoyed when they use the scanner. The second disadvantage is that the user needs to enter patient's information right before or after each scan to file the digitized image; therefore, batch scanning of the films is impossible.

Because of these inconveniences, new systems have been developed for specific models of scanner to automatically read in the patient information. These systems can read the patient information printed on a bar code attached to the film. For example, Lumisys LS200 laser scanner is designed by installing a bar code reader inside the scanner to read the bar code sticker on the film while the film is scanned. This method does improve the efficiency of scanning and makes batch scanning by auto-feeder possible. However, it also increases the workload of the technologist for printing and attaching the bar code sticker. This disadvantage, together with high cost of the reader and the difficulty in installation, has hindered the popularity of this approach.

To solve the above problems, we have developed a method based on image analysis technique to automatically extract the patient information from the film. This new software

From the Department of Electrical Engineering, National Yunlin University of Science and Technology, Taiwan, China.

Correspondence to: Hsien-Huang P. Wu, Department of Electrical Engineering, National Yunlin University of Science and Technology, 123 University Rd. Section 3, Yunlin, Touliu, Taiwan, China; tel: 886-5-5353061; fax: 886-5-5312065; e-mail: wuhy@yuntech.edu.tw

Copyright © 2002 by SCAR (Society for Computer Applications in Radiology)

Online publication 20 May 2002

doi:10.1007/s10278-001-0006-z

approach not only reduces the technologist's workload, but also makes the cost of installation very low. Moreover, because it can be applied to all types of scanners, the availability will be much higher than that of the bar code reader.

Image analysis techniques have been used widely in complicated document application, for example, text reading system,⁵⁻⁸ newspaper text/image separation,⁹ mail address block location extraction,¹⁰ and license plates extraction.¹¹ In medical imaging application, although much work has been done on the computer-aided diagnosis (CAD)¹²⁻¹⁵ to assist radiologists in analyzing the images, little work has been published on applying image analysis technique to increase the film handling efficiency of the technologist. In this report, we tried to apply the image analysis technique to reduce the workload of the technologist in handling the digitization of radiograph. Hopefully, the proposed method will enhance the user friendliness of the scanner and boost the overall efficiency in converting the films into digital images.

MATERIALS AND METHODS

The main difference in processing film document and ordinary document lies in the transparency of the film. That is, unlike paper documents, the film always can be scanned correctly regardless of whether it faces up or down (2 sides). Furthermore, to make the system more user friendly, we do not restrict the orientation of the film when it is fed to the scanner (4 orientations). Therefore, each film can be scanned with 8 possibilities (4 orientations by 2 sides). The proposed automatic patient information extraction process is summarized in Fig 1 and consists of 5 steps: (1) image acquisition, (2) patient information block (PIB) search and extraction, (3) label extraction, (4) label orientation correction and PIN (patient identification number) field extraction, and (5) PIN numeral extraction and recognition. The final step for number extraction and identification has been studied thoroughly in pattern recognition field,¹⁶⁻¹⁸ and will be omitted from this report. To speed up the extraction process, the efficiency and speed of algorithms used in various steps will be taken into account.

Currently, to create a more general automatic extraction system, we need to put some constraints on the composition of the label that contains patient's information. That is, we require the label used in the radiology department to have lines to separate each field (eg, name, patient's identification number, name of hospital, date of birth) and the label to be attached to the film properly. Also, the patient's identification number must be written inside the field's limited area, because it is the number in this field that finally will be extracted, recognized, and used to file the image into the database. These constraints will simplify greatly the algo-

Fig 1. Processing steps for the automatic patient information extraction from the digitized radiograph.

rithms and increase the processing speed and accuracy as will be shown later. Some of the test films contain Chinese characters, but the proposed method is independent of these characters. Therefore, the method can be applied to all kinds of language as long as the label structure satisfies the aforementioned requirements.

Image acquisition

A typical digitized x-ray image is presented in Fig 2, in which 2 different examples of label that came from 2 local hospitals are shown. These 2 images were acquired using Lumisys LS200 laser scanner with 300dpi resolution. The film can be scanned from either side in any orientation, but it must be positioned to be parallel to the sweep direction of scanning beam. This usually has been put into consideration

Fig 2. Typical images used in the development of the automatic PIN field extraction system (resolution: top 650×1245 ; bottom 100×2550).

in designing the scanner and will not be a restriction. Conforming to this requirement will keep the scanned lines in the label straight and increase the robustness of processing. Given a scanned image like the one in Fig 2, our goal is to extract the PIN field automatically from the image.

Patient information block search and extraction

While using the scanner, the technologists usually just drop the film and do not pay attention to the orientation of the film. Also, when they attach the patient information label to the film before exposure, the label can be placed at any position. However, observation of various radiographs shows that the patient information label almost always is attached to the corner of the film. Combining these 2 common practices, the label then will appear on one of the

4 corners of the scanned image as shown in Figs 2 and 3a. This feature makes the search of patient information block become easier, and the algorithm proceeds as follows. First, 4 corners of the image, each occupying about 1/16 or larger of the original size, are extracted. Second, edge detection¹⁹ is applied to these 4 subimages separately resulting in 4 binary images. Finally, the number of horizontal and vertical edge pixels in each binary image are counted, and the one with highest value is claimed to be the block that contains the label. The procedure is summarized in Fig 3b. The reason this algorithm works is because human body is full of complicated texture and only the label area contains many straight field lines. These lines result in many horizontal and vertical edges that are to be detected. In this processing stage, the image was down-sampled to reduce the size and increase the detection speed.

Label extraction

Although a patient information block can be extracted easily, the extra details it contains can interfere with the final goal of extracting PIN field. Therefore, our next step is to remove all these extra details and leave only the area of the label. The main technique used here is horizontal and vertical projection. Given the extracted patient information block, which has been binarized,²⁰ we accumulate (project) all the white pixels on each row in the image to form a vector P_H , and collect (project) all the white pixels on each column in the image to form a vector P_v . That is, given a $M \times N$ (M pixels by N lines) binary image $f(x, y)$ with $f(x, y) = 1$ for white pixel and $f(x, y) = 0$ for black pixel, we define projection as:

$$P_H(y) = \sum_{x=1}^M f(x, y), \text{ and } P_v(x) = \sum_{y=1}^N f(x, y) \quad (1)$$

If there are any significant vertical (horizontal) line segments in the image, this projection will create significant high projection values at the corresponding positions in the vector $P_v(P_H)$. These line segments come from the lines separating each field of the label. Because these high values contain the position information of the field-separating lines, we need to mark them inside vectors P_v and P_H . The way to extract this information is to select 2 thresholds T_v and T_H , and any position x for which $P_v < T_v$, $P_v(x)$ is set to 0. For any position y where $P_H(y) < T_H$, $P_H(y)$ is set to 0. After this thresholding process, all the significant line segments are now stored as nonzero elements in P_v and P_H vectors. These nonzero elements will be used to identify the field separating lines.

The region of the label is bounded by 4 clearly defined edge segments that delineate the label, and positions of these line segments can be found in the vectors P_v and P_H . That is, left/right boundary corresponds to the first/last nonzero element of P_v , and top/bottom boundary corresponds to the first/last nonzero element of P_H . These 4 position values can be used to obtain the exact region of the label. Examples of the results before and after this proce-

Fig 3. (A) Corners of image are examined for searching the patient information block. (B) Algorithm of PIB search. Function of MAX is used to choose the corner block which contains the label.

Figure 4a, 4c (before) and Figs 4b, 4d (after), respectively.

In the extracted region obtained above (Fig 4b), there is extra white zone surrounding the label. Because we need to extract the exact region of the label itself for further processing, this extra part needs to be removed. Compared with the label body part, this extra zone usually has a very high bright contrast and can be separated easily from the label itself. The label surrounding can be removed as follows: (1) thresholding the image that we extracted above (Fig 4b), resulting in a binary image (Fig 5a); (2) from the center of the binary image, search to the left, right, top, and bottom for continuous white pixels; (3) if there are more than 3 consecutive white pixels, we mark the one nearest to the center as the new position of the label boundary. Results of this boundary refinement for Fig 4 can be found in Figs 5b and 5c.

Label orientation correction and PIN field extraction

Up to this point, except for the particular format of the label, the procedure that we just described is no different than the processing of a normal paper document. The main difference between paper document and film lies in the orientation correction step that we are about to describe below. As we have mentioned above, regardless of whether it is a single film scan or batch film scan, the technologist can digitize the film in 8 possible ways. In these 8 different scan

conditions, each one of them will cause the image to have a different combination of *90° rotated*, *180° rotated*, or *left-right-mirrored* geometrical transformation. Therefore, even though we now can isolate the label region quite accurately, there is no way to identify the orientation of the label. Because correct orientation is necessary to identify the PIN field, further processing is needed.

For the correction algorithm to be general, we utilize a method based on template matching. In a radiology department using a film scanner, a blank label that is commonly used will be digitized, converted into a binary image, and then saved as a template (Fig 6). Given a scanned label image that we extracted above, one way to recognize its orientation is to correlate the scanned image with its template image in 8 different known orientations. From these 8 different matchings, the one with the highest matching score will indicate the real orientation of the input image. The input label image then can be corrected geometrically by rotation or left/right flip.

Although this method is quite straightforward, there are 3 problems with it. The first is that the resolution of input image and the template image might be different, so direct matching is impossible. The second is that the image matching by correlation will be too time consuming. Finally, the contents of the patient information might corrupt the matching result. Normalizing (or scaling) the input image to have the same resolution and size as that of the stored template can solve the first problem. Matching the projection of the image instead of the image itself can unravel the second problem. The third problem can be solved by

Fig 4. (A) Patient information block extracted from the input image using the algorithm shown in Fig 3(B). (B) Label area extracted from Fig 4(A). (C) Another example of extracted patient information block. (D) Label area extracted from Fig 4(C).

carefully choosing a suitable subregion for matching. These 3 suggested solutions are elucidated further below.

To make the input image and the template image have the same resolution, length of the shorter side of the extracted label is normalized to have the same length as that of the template image. The longer side is then modified by the same factor. Assume the size of the label template is $X_t \times Y_t$, and the label extracted from the input image is $X_{in} \times Y_{in}$, where Y represents line number in the image, and X is the pixel number per line. If X_{in} is less than Y_{in} , then normalization will scale X_{in} to become X_t , and Y_{in} is modified by

setting it to new value $Y_{in} \cdot X_t / X_{in}$. However, if X_{in} is greater than Y_{in} , then normalization will scale Y_{in} to become Y_t , and X_{in} is modified to become $X_{in} \cdot Y_t / Y_{in}$. After this normalization process, the input label would have the same resolution as that of the label template, and further processing now can proceed.

As we mentioned above, the input label image might have gone through a different combination of 90° rotated, 180° rotated, or left-right-mirrored geometrical transformation during scanning. We will check them in the above order sequentially, and correct them if any of these

Fig 5. (A) Segmentation of label image by thresholding. (B) Refined label image of Fig 4(B). (C) Refined label image of Fig 4(D).

3 problems occurred. The 90° rotated condition is checked first. Because we already have the aspect ratio information from the label template (width/height, or X_t/Y_t), any 90° rotation will change this value dramatically. We calculate the aspect ratio of the rectangular region that frames the label image, and if this value is different from that of the label template, then the input image has been scanned with a 90° rotation. Rotating the input image by 90° at either direction will correct this problem. The input image now will be checked for the other 2 orientation problems by template matching.

To further simplify the algorithm, the matching for orientation identification is done only on a particular part of the label image instead of the whole label image. To avoid the interference from the written content, the first field of the label is used for matching. This field usually contains only the hospital's name and is ideal for orientation identification. That is, we will extract the image rows at which first field lies then apply a horizontal and vertical projection to obtain 2 vectors $FP_H(y)$ and $FP_v(x)$. Apply the same procedure to the template image, and we obtain the other 2 projection vectors $TP_H(y)$ and $TP_v(x)$. The results of this projection process for template label image are presented in Fig 7.

The vectors $FP_H(y)$ and $TP_v(x)$ are used for checking if the scanned image needs a left/right flip correction. The vectors $FP_v(y)$ and $TP_H(y)$ are used for examining if the scanned label has been gone through a 180° rotation. Both of these 2 analyses are based on cross correlation for matching. Given 2 vectors $f(\cdot)$ and $g(\cdot)$, the cross correlation is defined as:

$$r_V(m) = \frac{\sum_x [f(x) - \bar{f(x)}][g(x - m) - \bar{g(x)}]}{\sqrt{\sum_x [f(x) - \bar{f(x)}]^2 \sum_x [g(x - m) - \bar{g(x)}]^2}} \quad (2)$$

A

B

Fig 6. Binarized template label image of two different hospitals. (A) Template label image with size 788×310 . (B) Template label image with size 227×430 .

$$r_H(n) = \frac{\sum_y [f(y) - \bar{f(y)}][g(y - n) - \bar{g(y)}]}{\sqrt{\sum_y [f(y) - \bar{f(y)}]^2 \sum_y [g(y - n) - \bar{g(y)}]^2}} \quad (3)$$

where $m = -(M - 1), -(M - 2), \dots, 0, 1, 2, \dots, (M - 2), (M - 1)$, and $n = -(N - 1), -(N - 2), \dots, 0, 1, 2, \dots, (N - 2), (N - 1)$, and $\bar{f(\cdot)}$ and $\bar{g(\cdot)}$ are the averages of vectors $f(\cdot)$ and $g(\cdot)$, respectively.

The correlation coefficients $r_V(m)$, $r_H(n)$ are in the range from -1 to 1 . Based on these 2 equations, the 180° rotated condition is examined first. In the extracted label image, if a 180° rotation does exist, then the first field will become the last field. To check this rotation, the first field and the last field of the label image are both extracted. These 2 subimages are projected horizontally to obtain 2 vectors $FP_{H1}(y)$ and $FP_{H2}(y)$ as shown with one example in Fig 8. Then $r_{H1}(n)$ are computed between vectors $FP_{H1}(y)$ and $TP_H(y)$. Also $r_{H2}(n)$ are computed between $FP_{H2}(y)$ and $TP_H(y)$ using equation (3). If $\max_n \{r_{H1}(n)\} < \max_n \{r_{H2}(n)\}$, then the label image was scanned with a 180° rotation, and correction is applied; otherwise, the image is left unchanged.

Next, the left-right-mirrored condition will be checked. After the last 2 analyses, the label should have been put in the correct orientation except that it might be left-right-mirrored. This can be checked using vertically projected vectors. That is, given the vector $FP_v(x)$ of the input image, which is defined above, we reverse the order of its components and create a new vector $FP_{VR}(x)$, shown with one example in Fig 9. Then $r_V(m)$ are computed between $FP_v(x)$ and $TP_v(x)$. Also, $r_{VR}(m)$ are computed between $FP_{VR}(x)$ and $TP_v(x)$ using equation (2). If $\max_m \{r_V(m)\} < \max_m \{r_{VR}(m)\}$, then the label image was scanned with a left-right-mirrored condition, and correction by left-right flip is needed; otherwise, the image is left unchanged.

Fig 7. Vertical and horizontal projection of the template label image. Use in rotation and left-right-mirrored condition checking (top right: $TP_H(y)$; bottom right: $TP_V(x)$).

After these 3 steps of checking and correction, the label image now is in the right orientation and is ready for field extraction. Assuming we can extract the first field, then by giving the relative position of the PIN field to the first field, it is easy to get to the PIN field and extract it. As long as the resolution and orientation of the scanned label image is the same as that of the template image, this relative position

information can be measured and used right after the template image is created. Examples of the extraction of the PIN field are available in Fig 10. Given the PIN field, the final step would be to recognize the numbers inside it. This PIN number will be used to link with the database and file the image data automatically. This final step is not considered in this report.

Fig 8. Horizontal projection of the first and the last field (right top/bottom: $FP_{H1}(y)/FP_{H2}(y)$) of the label image. Use in 180° rotation checking.

Fig 9. Vertical projection of the first field (right top) of the label image and its reverse vector (right bottom). Use in left-right-mirrored condition checking.

Fig 10. Examples of the extracted PIN field. Left, inside the label image. Right, extracted field.

RESULTS

To evaluate the PIN field extraction algorithm, 2 formats of the label acquired from 2 different hospitals are tested. Examples of label format are shown in Figs 4a and 4c. The films are scanned at 300 dpi using Lumisys LS200 laser scanner. The total amount of film under test is 150 for each label format. Each scanned image was processed in less than 0.2 seconds using a Pentium-III 450 MHz PC. This speed can be treated as real-time in filing the image to database. The results of PIN field extraction we obtained are summarized in Table 1, which are classified into 3 categories: (1) full extraction: successfully extracted the whole PIN field, (2) over/under extraction: extracted region larger/less than the PIN field, and (3) fail: cannot find the PIN field.

The results in Table 1 are obtained using currently available films, which are not specifically written or positioned for the automatic extraction purpose. The over/under extraction and fail problems mainly are caused by the following factors: (1) image blurring caused by the scanner makes the line segment extraction difficult, (2) too much tilting of the attached label caused the projection algorithm to fail, (3) the labels are not positioned at the corner. If restrictions are put on the way the label is written and positioned, better results can be obtained.

Table 1. Results of PIN Field Extraction for 2 Sets of 150 Scanned Images Acquired From 2 Hospitals

Extraction result	Hospital A	Hospital B
Full extraction	109 (73%)	126 (84%)
Over/under extraction	33 (22%)	19 (12.7%)
Fail	8 (5%)	5 (3.3%)

DISCUSSION AND CONCLUSIONS

Here we report a new method to improve the user friendliness of radiograph scanner. The proposed method can automatically extract the patient's vital information (ie, patient identification number, chart number) field from the scanned image, and therefore greatly reduce the workload of the technologists. Two formats of label acquired from 2 different hospitals are tested, and the results are evaluated. Given the available films with no constraints on the way the labels are attached and positioned, the correct extraction rates are 73% and 84%, respectively.

Tilting of the attached label and mispositioning of the label are 2 major reasons for the limitation on the correct extraction rate. However, these 2 problems can be avoided by requesting the technologists to attach the label properly in their routine practice. Currently, evaluation of the proposed method is only limited to Lumisys LS200 laser scanner; tests on other models and types (eg, CCD) of scanner are left for future study.

Even though different radiology departments might not have the same label format as was introduced here, the algorithm can be modified because its modular structure. This capability of adaptation will be investigated further in our future study. Because every hospital has its own label format and usually differs in a very significant way, developing an algorithm to take care of this modification efficiently will be a great challenge.

The approach presented in this report not only makes the batch scanning of film possible, but also solves the problem of redundant input on patient information. Although bar code readers are available for installation in certain models of scanner to achieve the same

purpose, our software-based approach can save the cost of the reader, greatly reduces the workload of the technologists, and is suitable for all the makers and models of film scanner. We believe the success of this technique will benefit the development of the PACS and teleradiology

ACKNOWLEDGMENTS

The author thanks Dr S.K. Chen of the Taiwan University Hospital and Dr W.C. Ye of the Provincial Yunlin Hospital for their kind assistance in providing and digitizing the radiographs.

REFERENCES

1. Cao X, Huang HK: Current status and future advances of digital radiography and PACS. *IEEE Engineering in Medicine and Biology Magazine* 19:80-88, 2000
2. Inamura K, Kondoh H, Takeda H: Development and operation of PACS/teleradiology in Japan. *IEEE Communications Magazine* 34:46-51, 1996
3. Special issue on Medical Communication. (1992) *IEEE Journal on Selected Areas in Communications*, 10:1158-1196
4. MacMahon H: Clinical experience with an advanced laser digitizer for cost-effective digital radiography. *Radiographics* 13:72-81, 1993
5. Tsujimoto S, Asada H: Major components of a complete text reading system. *Proc IEEE* 80: 1133-1149, 1992
6. Casey R, Ferguson D, Mohiuddin K, et al: Intelligent forms processing system. *Mach Vision and Appl* 5:143-155, 1992
7. Schürmann J: Document analysis-from pixels to contents. *Proc IEEE* 80:1101-1119, 1992
8. Taylor SL, Fritzson R, Pastor JA: Extraction of data from preprinted forms. *Mach Vision and Appl* 5:211-222, 1992
9. Fletcner L, Kasturi R: A robust algorithm for text string separation from mixed/graphics images. *IEEE Trans PAMI* 10:910-918, 1988
10. Palumbo PW, Srihari N, Soh J, et al: Postal address block location in real time. *IEEE Computer* 25:34-42, 1992
11. Cui Y, Huang Q: Extracting characters of license plates from video sequences. *Mach Vision and Appl* 10:308-320, 1998
12. Yu S, Ling G: A CAD system for the automatic detection of clustered microcalcifications in digitized mammogram films. *IEEE Transactions on Medical Imaging* 19: 115-126, 2000
13. Giger, ML: Computer-aided diagnosis of breast lesions in medical images. *Computing in Science & Engineering* 2:39-45, 2000
14. Penedo MG, Carreira MJ, Mosquera A, et al: Computer-aided diagnosis: A neural-network-based approach to lung nodule detection. *IEEE Transactions on Medical Imaging* 17:872-880, 1998
15. Lo SB, Freedman MT, Lin J, et al: Automatic lung nodule detection using profile matching and back-propagation neural network techniques. *J Digit Imaging* 6:48-54, 1993
16. Lee SW: Off-line recognition of totally unconstrained handwritten numerals using multilayer cluster neural network. *IEEE Trans PAMI* 18:648-652, 1996
17. Mori S, Suen CY, Yamamoto K: Historical review of OCR research and development. *Proc IEEE* 80:1029-1058, 1992
18. LeCun Y, Boser Y, Denker JS, et al: Backpropagation applied to handwritten zip code recognition. *Neural Computation* 1:541-551, 1989
19. Canny J: A computational approach to edge detection. *IEEE Trans PAMI* 8:679-698, 1986
20. Trier OD, Jain AK: Goal-directed evaluation of binarization methods. *IEEE Trans PAMI* 17:1191-1201, 1995