4

Document Image Processing for Hospital Information Systems

Hiroharu Kawanaka[1], Koji Yamamoto[2], Haruhiko Takase[1] and Shinji Tsuruoka[3]
[1]*Graduate School of Engineering, Mie University*
[2]*Suzuka University of Medical Science*
[3]*Graduate School of Regional Innovation Studies, Mie University*
Japan

1. Introduction

In this chapter, we introduce document image processing methods and their applications in the field of Medical (and Clinical) Science. Though the use of electronic health record systems are gradually spreading especially among big hospitals in Japan, and e-health environment will be thoroughly available in near future [1–3], a large amount of paper based medical records are still stocked in medical record libraries in hospitals. They are the long histories of medical examinations of each patient and, indeed, good sources for clinical research and better patient care. Because of the importance of these paper documents, some hospitals have started to computerize them as image files or as PDF files in which the patient ID is the only reliable key to retrieve them, however most hospitals have kept them as they are. This is due to the large cost of computerization and also the relatively low benefit of documents that can only be retrieved by patient ID. Indeed, the true objective of computerization of paper records is to give them functionality so that they can be used in clinical research such as to extract similar cases among them. If we cannot find out any practical solutions to this problem, large amounts of these paper based medical records will soon be buried in book vaults and might be discarded in near future. Thus we are confronted with a challenge to devise a good system which is easy to run and can smoothly incorporate the paper based large histories of medical records into the e-health environment.

In an e-health environment, health records are usually treated using an XML format with appropriate tags representing the document type. Here the document type means the scope or rough meaning of contents. Therefore, a good system might have such functions as to create XML files from paper documents that also have appropriate tags and keys representing the rough meaning of contents. Fortunately, most paper based medical records have been written on fixed forms depending on the clinic or discipline, such as diagnoses placed in a fixed frame of a sheet, and progress notes in another frame, etc., and these frames usually correspond to the document types. It would seem rather easy to assign an appropriate XML tag to each frame if we could determine the form or the style of the paper. And if such a frame can be determined and the scope of the contents in it is fixed, then translation into text from the document in that frame might be accurately performed by using dictionaries properly assigned to the scope. Also, as collaborative medicine spreads, many recent medical records have been typed so that they can easily be read among the team members; which

also improves the accuracy of translation. Therefore, if we can determine accurately the style of the document, many stylized documents which will have XML tags corresponding to the frames in it may be fixed and the contents will be rendered into a text file with good accuracy. With this premise we started to investigate a new indexing system for paper documents [4–7]. The key elements of our investigation are document image recognition, keyword extraction, and automatic XML generation. This chapter will be devoted to the introduction of our work and some recent topics about the document image processing method used in the healthcare sector.

After describing experimental materials used for this study in section 2, the proposed method will be presented in section 3. The results and discussions about the proposed method will come in section 4, other related topics in medical sector in section 5, and lastly, concluding remarks and future scope will come in section 6.

2. Materials

As it is very important to know the power and limitations of the idea we first used typed and stylized documents with frames or tables archived at the medical record library at Mie University Hospital. These documents were scanned by an optical image scanner with gray scale of a resolution of 300 dpi. The images thus obtained are the target of this research. The resolution of 300 dpi is the minimum requisite to satisfy the law for digital archiving of medical documents in Japan. To know the extensibility and power of our system, handwritten medical records were also tested, which we only discuss in the discussion. Extension of our method to atypical medical documents though stylized but without frames is also discussed in that section.

3. Document image recognition method for resemble case search

3.1 Employment of a master information

As is stated in the introduction, almost all paper-based medical records are written in stylized sheets. When the use of computers was not so common as today, each clinical department designed the sheets carefully so as to fulfill their clinical examination requirements. Indeed, the styles used in each clinic were the result of a long history of contemplation. Titles and frame lines to the frames were in many cases printed out on blank sheets, and bundles of blank sheets were stocked at the medical record library and the medical affairs office, which were used for in many years. To save running cost, these blank sheets have gradually been output from computers where frame lines and/or ruled lines were sometimes omitted, but, the majority of medical records archived at the library were written on these stylized sheets with frame lines. In our research, these blank sheets were used to obtain master information to determine the type of the sheets of the images examined. The blank sheets were scanned and thus obtained images we call “master images”, and the master information is generated from the “master images”. The master database consists of information about the positions and types of the corners of each frame in a sheet and XML tags representing the contents of each frame. The method of determining the positions and the types of the corners of each frame of master information is the same as is used in analyzing images examined, but adding XML tags to each frame was done using knowledge about the sheets used. Using the master information clarifies the XML structure and makes the extraction of strings inscribed by users easy. As the XML tag in the master database is given from outside, it is quite robust in creating XML files. We employed the master database to the proposed system aggressively so as to cover all kinds of blank sheets used in Mie University Hospital. Generally speaking, employment of such a

master database will improve the reliability of our system when the processed objects are regarded as stylized documents with frame lines.

Here, to make the statement more clearly, we use the word “sheet type” as the type of sheet included in the master information and “document type” as the type of document of each frame which directly relates to the XML tag. The word “table” is also used to mean a frame with frame lines, while the term “cell” means the contents in one cell of the table, and “cells”, the plural of “cell”, also means the contents in the table.

3.2 Outline of our system

Figure 1 illustrates the outline of the proposed method. The images obtained from paper-based medical documents have some factors, such as noise, tilts and so on, that deteriorate the accuracy of the following processes. These factors are reduced (or removed) by pre-processing described in 3.3, and some features to determine the sheet type are extracted from them. After this, each cell in the documents is extracted using cell positions and master information. The extracted cell often has images, e.g. schema images, sketches, as well as character strings. Thus such images are also extracted from the cell images. The extracted character strings are converted into text data by OCR engine, and the obtained text data are stored into a database. The extracted schema images are also recognized by a schema recognition engine, and the recognition results, i.e. schema name, annotation and its position etc., are also registered into the database. After this, an XML file is generated using the master information.

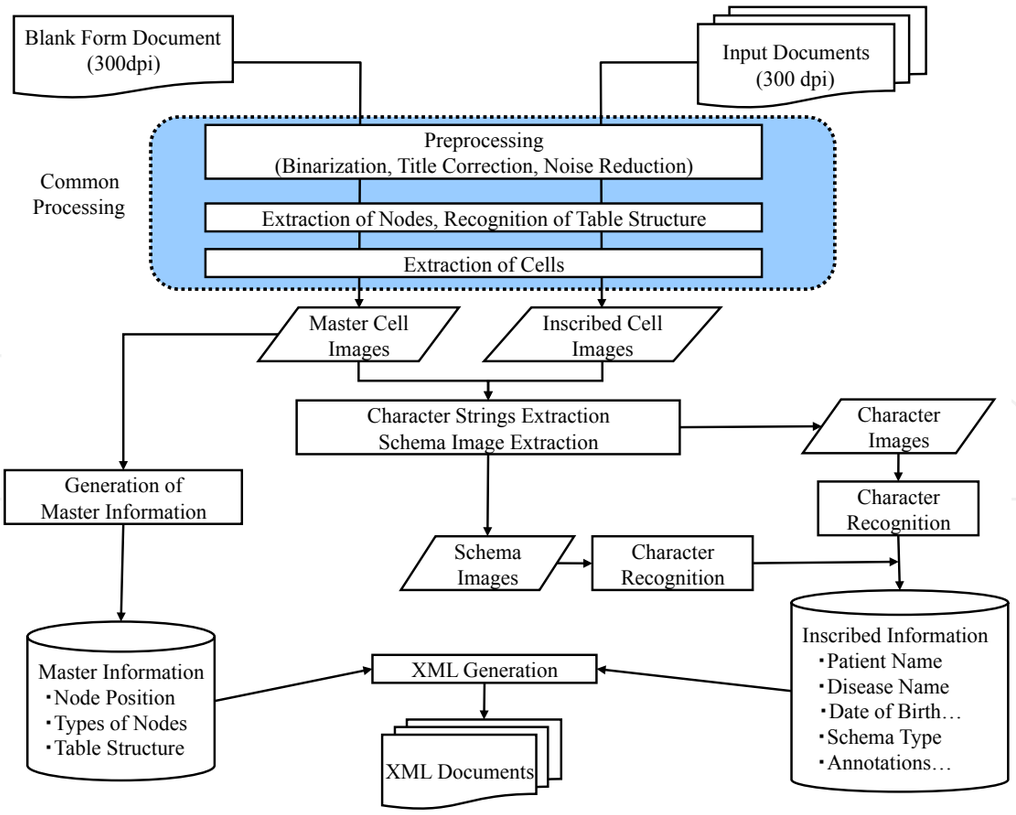


Fig. 1. Flow of the Tabular Document Recognition

3.3 Pre-processing

In this study, binarization, tilt correction and noise reduction techniques are applied to the input images as pre-processing. In the binarization process, Otsu's method was used [8, 9]. In his method, the threshold for binarization is determined by discriminant analysis using a density histogram of the input image. Therefore, no fixed threshold for each image is required. In the tilt correction process, the LPP method is used to correct the tilt of the images [10]. Figure 2 illustrates the rough image of LPP. In the LPP, the target image, i.e. the input image, is divided into n_s sub-regions and marginal distributions of each region are obtained. In this case, horizontal projection histograms are used as the marginal distributions. Next, correlations between each region (α_k) are calculated by

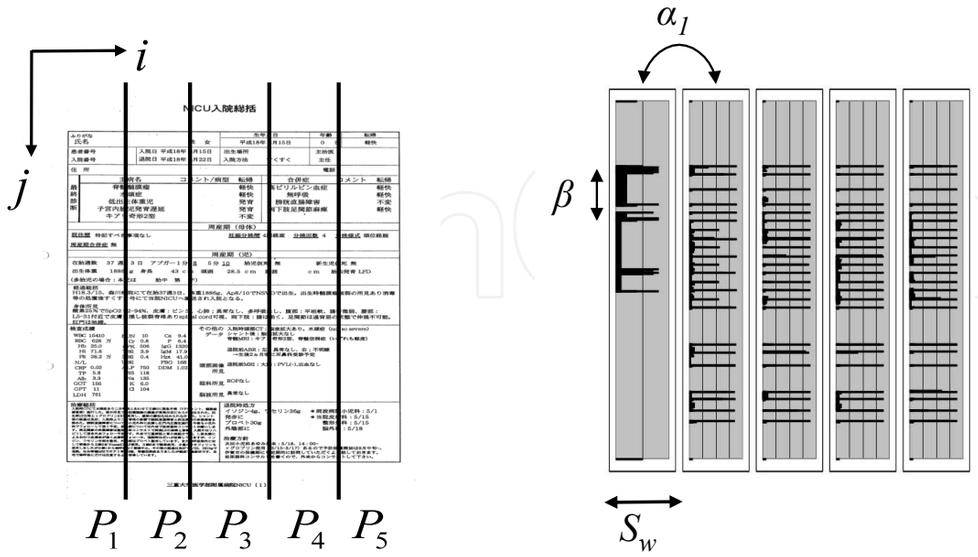
$$\alpha_k = \max_{-\beta \le y \le \beta} \left[\sum_j P_k(j) P_{k+1}(j-y) \right] = \sum_j P_k(j) P_{k+1}(j-\alpha_k). \quad (1)$$

$$(k = 1, 2, 3, \cdots, n_s - 1)$$

Here, $P_k(j)$ means the j-th value of the horizontal projection histogram of the k-th sub-region, and β does the range of calculation. These value indicate misaligns of the phases in each region, which are equivalent to the ratio of the tilt. As the result, the tilt angle of the paper θ is given by

$$\theta = tan^{-1} \frac{\alpha_m}{S_w}. \quad (2)$$

Here, α_m is the average of α_k, and S_w is the width of each sub region. the LPP method can detect the tilt of images with high accuracy with low calculation effort. For an image size of 1024×1024 pixels, the theoretical detection accuracy is about 0.06 degree and the detection



(a) Input Image (b) Obtained Marginal Distribution

Fig. 2. Tilt Correction using the LPP Method

range is from –10 to 10 degrees. We use the LPP method only, because images tilted by more than 10 degrees do not occur in practical cases. As a final step of pre-processing, a median filter is applied to the images to reduce speckle noise and salt and pepper noise.

3.4 Sheet type recognition using node information

Generally speaking, a tabular form document has at least one table, and its form and location heavily depend on sheet type. In other words, features of the table in the document would be the key information for sheet type recognition. Thus we extract crossover points of ruled lines, which we call "Nodes", from the document, then positions and types of these nodes are used for the sheet type recognition.

3.4.1 Feature extraction

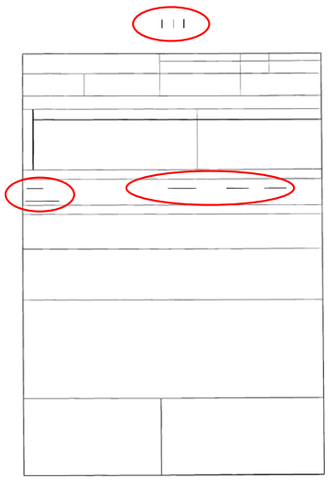
Figure 3 shows the outline of feature extraction for sheet type recognition. As a first step, ruled lines in the input images are extracted using black pixels forming a straight line. When there is a horizontal connected component that consists of n_h black pixels, it is regarded as a horizontal solid line. The same process is also applied to extract vertical ruled lines. In this study, the value of n_h is decided experimentally as 50. The length of 50 pixels is equivalent to about 4.2mm when the resolution of the input images is 300 dpi. Of course, the value of n_h affects the extraction accuracy of ruled lines. And in some cases, partial lines of characters or underlines in the image are also obtained as shown the circular parts in the figure 3(b). These parts may influence the processes follows. But in the proposed method, the detection of crossover points can remove these surplus lines, and the determination of value of n_h is not so significant. As a matter of fact, these surplus lines are removed by adjusting the value of n_h. Now the next step is to decide the types and positions of nodes. Since ruled lines usually have some width, the node where these ruled lines crossover usually form a rectangle. We set the node position as the center of gravity of such a rectangle. Then, from the node position, ruled lines are traced toward outside until they reach other lines. All ruled lines which failed to meet other lines are discarded. By doing so, the pattern of the node is decided.

Figure 3(c) shows the outline of the classification method. Generally speaking, a table consists of nine types of crossover points, which are called "Node" in this paper, and non-crossover points [11–13]. We express the table in the document using these features. In our method, ruled lines around the target nodes are searched first. In the case shown in Figure 3(c), when a ruled line exists above the target node, then, node No.1, 2 and 3 are excluded as candidates. In the next step, ruled lines are also searched for on the left, right and bottom of the target node. As a result, the target node is identified as node type 4. The same process is applied to all nodes in the image. The extracted nodes' numbers and their positions are stored into the database for sheet type recognition and cell image extraction. These features can express the structure of the table, and elements in the table can be extracted by using the nodes' types and their positions.

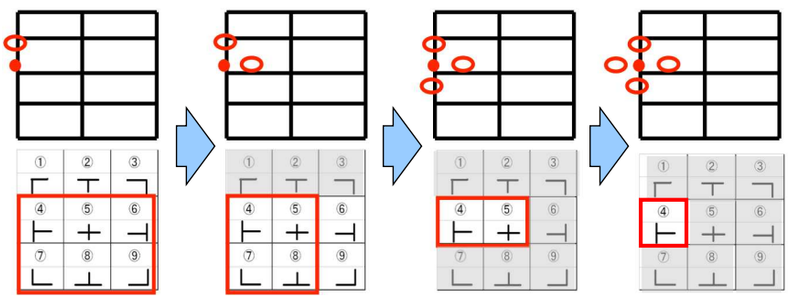
3.4.2 Determination of sheet type

Figure 4 illustrates the outline of our sheet type recognition technique. We first set a ROI of size $n_{roi} \times n_{roi}$ pixel to each node thus obtained in the above section. Then, we search whether the same type of node exists in the same ROI of a sheet in the master database, and count up the successful cases and calculate the degree of coincidence to the sheet in the master database as the ratio of the number of successes to the total number of nodes of that sheet registered in the master database. Lastly we determine the sheet type of the image as that which has the

(a) Input Image



(b) Detected Ruled Lines



(c) Node Type Detection

Fig. 3. Extraction of Node Type and Position from Input Images

highest degree of coincidence among the master database. As the master database contains all types of sheets used at Mie University Hospital, and the occurrence of an irregular type of sheet will be very rare if at all, the proposed method can determine the sheet type with quite good accuracy.

3.5 Detection and extraction of character strings in each cell

3.5.1 Cutout of cell images using node matrix

The elements of the table which we call “cells”, are extracted using node information. In this study, we use a matrix using the node’s number called “Node Matrix”. Figure 5 illustrates the generation process of the node matrix. The node matrix expresses the structure of table, thus we can extract cells from the table by using the matrix and the positions of these nodes. Figure 6 shows the outline of the cell extraction method. The node located on the top-left in the document is set as the starting point of the extraction. Then the matrix is scanned from the start point left to right until the nodes with a downward element, i.e. node 1 – 6 in Figure 5, appear. In this case, node 2 appears first as the node with a downward element. The node is the top-right point of the cell and the matrix is scanned from this point to the bottom again.

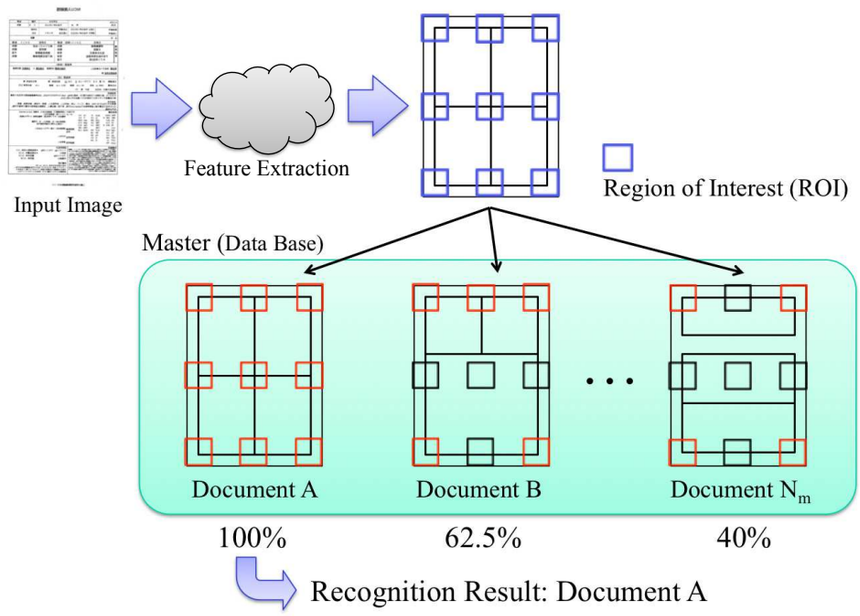


Fig. 4. Outline of Sheet Type Recognition

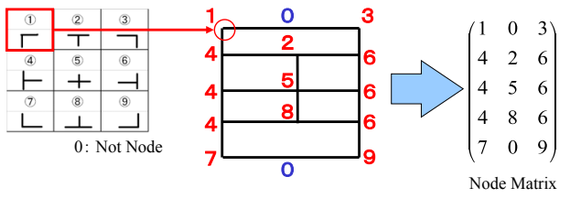


Fig. 5. Generation of Node Matrix

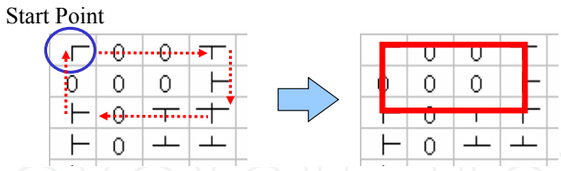


Fig. 6. Extraction of Cells from Table

When the node with a left element such as nodes 2, 3, 5, 6, 8 and 9 appears, the node is regarded as the bottom-right of the cell. The same process is repeated until the start point appears again. In this paper, the same process is applied to all nodes in the matrix to extract all cells in the table. Of course, the position of each node is stored into the database, thus we can finally cutout each cell from the table by using the information.

3.5.2 Detection of strings and character recognition

String regions in all the cells have to be extracted to recognize characters and generate an XML document. The proposed method extracts the regions using the master information. In this chapter, the cell image extracted from a blank table is called the "Master Cell Image", and the one from a table inscribed by users is called "Inscribed Cell Image", respectively. Since the

master cell image sometimes has images coming from the title printed in the blank sheet, the string regions inscribed by users in each cell are extracted by a subtraction between the master cell image and the inscribed cell image. However, when the position of the master cell image does not match that of the inscribed cell image, these regions cannot be extracted correctly. Therefore, our method calculates the ratio of difference between these images first, and then the position for the subtraction process is determined to solve the above problem. In this process, the ratio of difference is obtained by the sum of the number of pixels with different values in each pixel, and the string regions in the cell image are extracted by the subtraction process.

Figure 7 shows the outcome of the string extraction. The figure indicates that the inappropriate regions not inscribed by users are also extracted as well as the string regions inscribed. These results are caused by slight differences of tilt or input conditions between the master cell image and the inscribed cell images. But, it is very difficult to eliminate these differences completely. To solve this problem, the proposed method was changed to improve extraction accuracy. Specifically, the labeling process shown in Figure 8(a) was added. As a first step of the procedure, the labeling process is applied to the master cell image, and next the black pixels belonging to the large connected components are changed to white. After this, the same subtraction process is done again. Figure 8(b) shows a result of the improved method. It is obvious that characters in the master cell image are erased completely and strings inscribed by users are appropriately extracted compared with the result in Figure 7. Actually the extraction accuracy of the improved proposed method depends on that of the labeling process. In the case of the printed documents, variations of character size and distance between characters are not significant, thus the accuracy of the improved proposed method is high enough for practical use. In preliminary experiments, false extraction of string regions such in Figure 7 was not detected.

3.6 Schema image recognition method

3.6.1 Features for schema detection

Generally speaking, extracted cell images consist of some elements such as character strings, dotted (or broken) lines and schema images. In our method, as a first step, four features are extracted from the cell images to discriminate these elements. In this section, we focus on

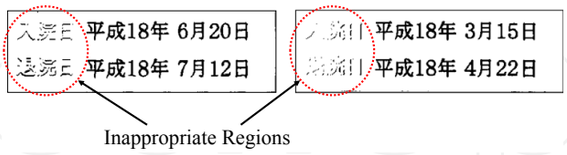
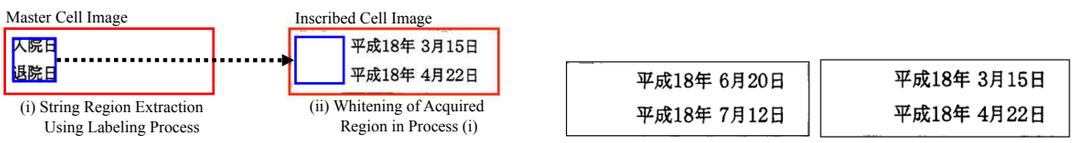


Fig. 7. Result Example of String Extraction



(a)Labeling Process for String Extraction (b)Extraction Result

Fig. 8. Extraction of String Regions (Improved Method)

the shape of dotted lines and schemas. It is supposed that dotted lines and schemas have the following characteristics:

1. The circumscription rectangle size of schemas is larger than that of a single component of dotted lines or character and the shape of schemas is vertically (or horizontally) longer than that of single component of dotted lines or character.
2. Each component of dotted lines is smaller than that of schemas and they are lined up on straight lines.

To express 1, we employ a variance of horizontal and vertical direction S_x and S_y and circumscribed rectangle area A of each connected component. For 2, the number of connected components lined up on straight lines is employed. We call this feature the horizontal (or vertical) connected level L. Figure 9 illustrates the rough image of a horizontal connected level. The center coordinates of each circumscription rectangle are obtained by labeling processing, and the center coordinate of the target rectangle is connected to that of other rectangles with straight lines. In the case that tilts of the lines are within $\pm t$ degrees, it is regarded that these circumscription rectangles distribute on the straight line. In this study, the value of t was set to 0.5 experimentally, because the theoretical detection accuracy is 0.06 degree in the LPP method. The processing for discriminant of vertical dotted lines is not done because tabular form documents used in this study do not have such structures. As a matter of course, the features about discriminate vertical dotted lines can be calculated easily by extending the previous processing.

Figure 10 shows the ideal distribution of the features. In this figure, the connected components of schema images will have large values of S_x, S_y and A as shown in Figure 10(a) and the components of dotted lines will come on the region with a large value of L. But the character components will appear in the region with small values of S_x, S_y and L (Figure 10(b)). It is expected that dotted lines, schemas and characters can be discriminated by using appropriate thresholds to these features.

3.6.2 Extraction of schemas from cell image

To extract schemas from a cell image, we must decide the threshold values for S_x, S_y and A. Since the objective of this section is to extract schemas from the cell, only the threshold values for S_x, S_y and A are used. (The threshold value for L is necessary to discriminate characters and dotted lines.) These threshold values were decided by considering the shape of histograms of each feature. As expected in the above section, the histograms will show bimodal patterns and the threshold values will easily be determined at the bottom of valley

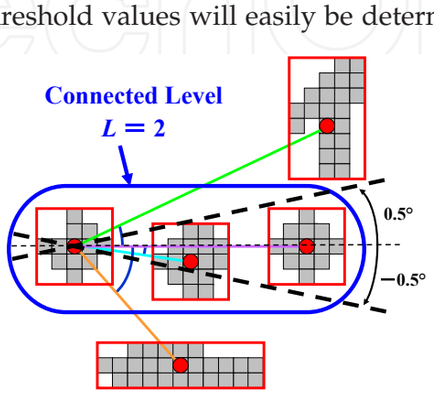
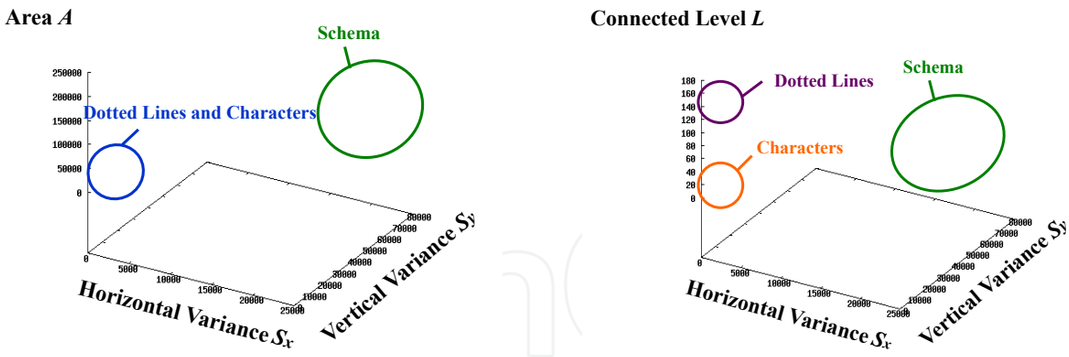


Fig. 9. Horizontal Connected Level L



(a) Area A (b) Connected Level L

Fig. 10. Ideal Distribution of Features

between two peaks. To get statistically meaningful histograms we use Sturges' formula, given as:

$$n_c = 1 + \log_2 n_d \approx 1 + 3.32 \log_{10} n_d. \tag{3}$$

Here n_c is the number of classes and n_d means the number of data, respectively. And the threshold value is determined at the bottom of valley between two peaks. With this method, all data having schema characteristics is extracted. In other words, all data having characteristics of dotted lines or characters are not extracted even when they are located in the schema area. These should be recovered.

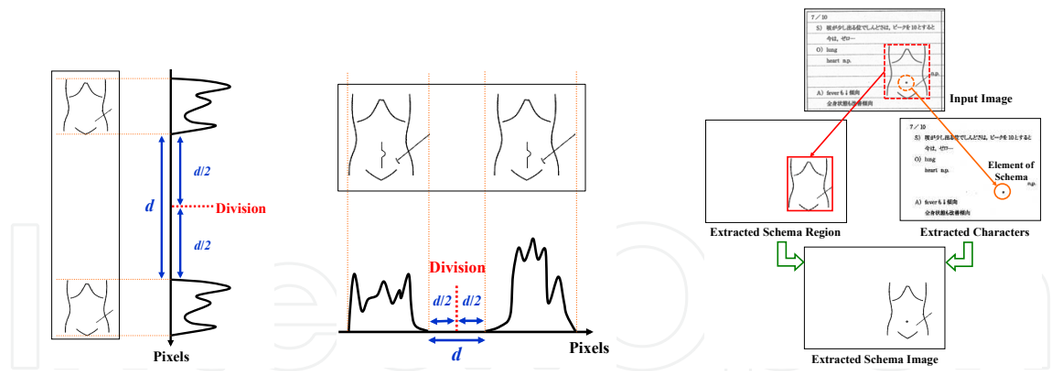
3.6.3 Extraction of schemas from schema area and recovery

In some cases several schemas are placed closely in a document. In such cases the schema area obtained in the above section might have several schemas which should be divided and extracted from the cell image appropriately. For this we prepare a dividing process in the system using the shape of histogram.

Figure 11(a) and (b) illustrate the outline of the dividing process. As a first step of this process, we obtain a projection histogram of vertical direction for the schema. In the obtained histogram, the part that consists of d_0-continuous elements with zero value is regarded as the boundary of each schema, and the image is divided on the middle point of the part. The same processing is applied to the image for division on horizontal direction. By this processing, schema regions are divided into several mutually independent ones. In this paper, the value of d_0 is given experimentally. Finally, the connected components in the schema, which were classified as characters, are added to the original schema image (Figure 11(c)).

3.6.4 Schema recognition using weighted direction index histogram method

Weighted direction index histogram method (WDIHM) is one feature extraction method. It is often used in handwritten character recognition systems [14–16]. Figure 12 illustrates the rough image of this method. As you can see, the method traces the contour of the character image first, and direction index histograms in each sub-region are generated using chain codes. After this, the spacial weighted filter based on Gaussian distribution is applied to the obtained histograms to generate a feature vector. WDIHM has enough robustness to local shape variations of input character images. As the accuracy of this method is extremely high compared with other character recognition algorithms, this method is employed in many



(a) Horizontal Division (b) Vertical Division (c) Extraction of Schema Images

Fig. 11. Division and Extraction of Schema Images from Documents

fields such as commercial OCR software, e-learning systems, factory automation systems and so on [17–20].

Figure 13 shows the outline of schema image recognition method using WDIHM. For schema image recognition, we first have to make a dictionary for recognition. In this method, many images are required to make the dictionary. Since this method divides the input images into some sub-regions and calculates covariance matrix among them for feature vectors, the dimension of feature vector is very large. We used not only basic schema images employed in the hospitals shown in Figure 14(a) but also some additional images, e.g. rotated and shifted ones etc., to make the dictionary (Figure 14(b)). Actually there are more than 120 kinds of schema images used in HIS, but in this study we picked up only five kinds of typical schema images as shown in Figure 14(a) to examine the effectiveness of the proposed method. For the recognition of input schema images, we employ the following discriminant function called Modified Bays Discriminant Function (MBDF)[14, 15].

$$d^l(x) = \sum_{i=1}^{k_1} \frac{\{{}_l\varphi_i{}^t\,(x - {}_l\mu)\}^2}{{}_l\lambda_i} + \sum_{i=k_1+1}^{n} \frac{\{{}_l\varphi_i{}^t\,(x - {}_l\mu)\}^2}{{}_l\lambda_{k_1+1}} + ln(\prod_{i=1}^{k_1} {}_l\lambda_i \cdot \prod_{i=k_1+1}^{n} {}_l\lambda_{k_1+1}) \qquad (4)$$

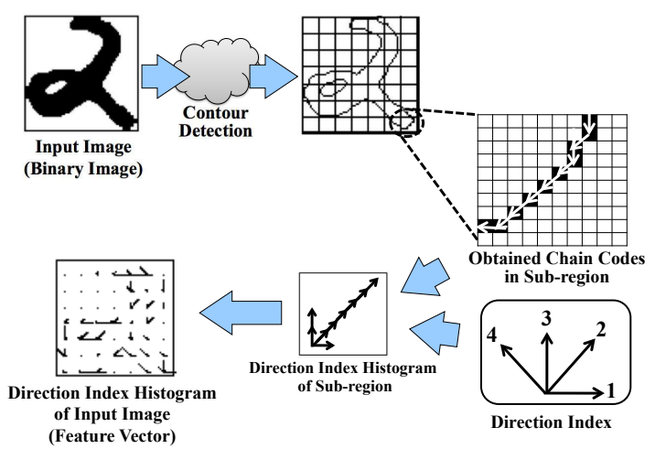


Fig. 12. Rough Image of Direction Index Histogram Method

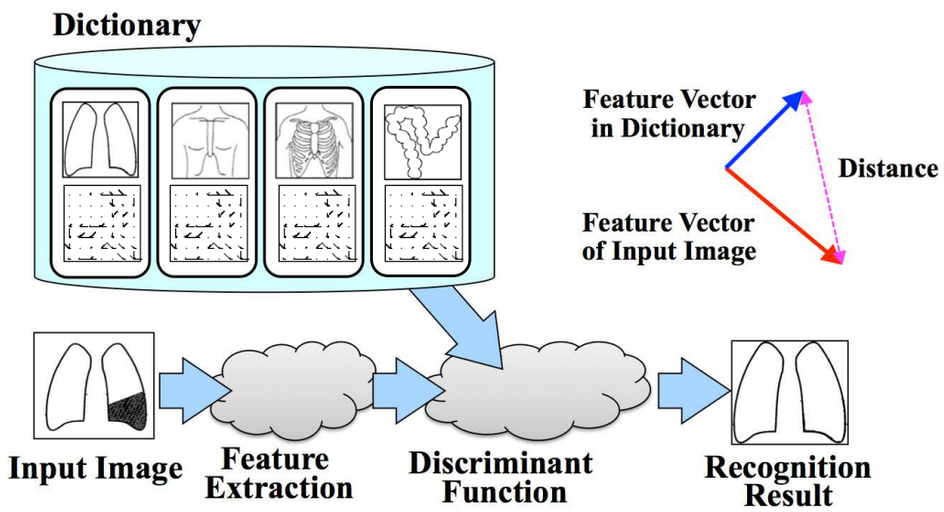
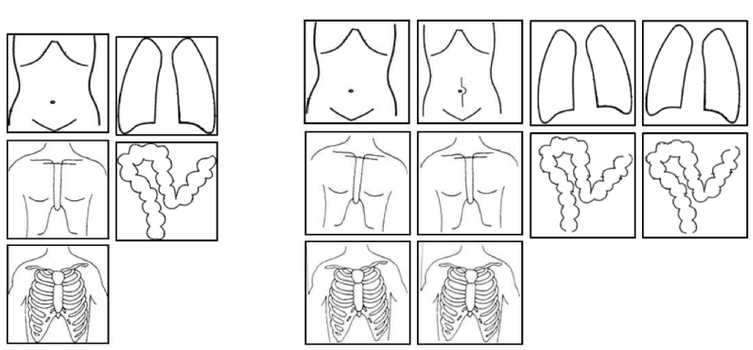


Fig. 13. Outline of Schema Image Recognition Method



(a) Standard Schema Images (b) Additional Schema Images

Fig. 14. Example of Schema Images for Generating Dictionary

In the above formula, x is the n-dimensional feature vector of the input schema image, and ${}_l\mu$ is the average vector of schema image l in the dictionary. ${}_l\lambda_i$, and ${}_l\varphi_i$ are the i-th eigen value and eigen vector of schema image l, respectively. And k_1 is determined by the number of learning sample $m(1 \leq k_1 \leq m, n)$. These higher-order eigen values are in many cases not used due to the increase of calculation time while contributing little to the improvement of recognition accuracy. But in our case, the higher-order eigen values and vectors will be necessary components to improve recognition accuracy, since the construction of characters (or schema images) are very complex. As the absolute value of higher-order eigen values are very small and the true values of them are difficult to obtain, λ_{k_1+1} are used as the approximation of $\lambda_i (i = k_1 + 1, \cdots, n)$. In this study, the number of sub-regions and the value of k_1 were determined based on the literature [14, 15]. After this process, inscribed annotations are detected by subtracting the input image and the recognition result, i.e. the master image stored in the dictionary. The subtraction result indicates the position of annotations inscribed by medical doctors. To identify the anatomical position of them, we use an anatomical dictionary. The anatomical positions of the annotations are identified after matching between the detected annotations and the dictionary.

4. Experimental results and discussion

4.1 Accuracy of sheet type recognition

To make our system robust in the case of the misalignment of medical records to the scanning machine, we introduce ROI of size $n_{roi} \times n_{roi}$ pixels in 3.4.2 But, if the misalignment error exceeds this range due some reason, say, distortions caused by anthropogenic factors or by a mechanical error of the copying machine, a further improvement will be necessary. We used the following three techniques in the recognition method, and examined their accuracy and the processing time by using 325 sheets.

1. Using Absolute Coordinate System Based on the Top-left Pixel
2. Using Relative Coordinate System Based on the Position of each Node
3. Using Relative Coordinate System Based on the Position of the Top-left and Bottom-right Nodes

Table 1 shows experimental results of sheet type recognition. The table shows that all documents were recognized correctly in cases of relative coordinate systems of 2 and 3. But, when using the absolute coordinate system (case 1), the recognition rate was 96.3%. But the method of case 2 requires a lot of calculation time because of large number of nodes. Since a few thousands paper-based documents are generated in the hospital every day, case 2 might not be a practical solution. From these results, we can conclude the following.

1. The methods using relative coordinate systems are effective for determining the sheet type.
2. From the view point of processing time, we should use as few nodes as possible for sheet type recognition.

Coordinate System Based on...	Recognition Accuracy [%]	Processing Time [msec/sheet]
the Top-left Pixel	96.3	17
the Position of each Node	100	16961
the Position of the Top-left and Bottom-right Nodes	100	17

Table 1. Results of Document Type Recognition

4.2 Result of schema image recognition

4.2.1 Features for schema image extraction

Figure 15 shows an example of distribution of the features extracted from an input image. The obtained distribution of the features was similar to the ideal one as shown in Figure 10. In this experiment, we also applied the extraction method to 6 kinds of printed discharge summary documents in print [21], which have dotted (or broken) lines and schema images. The obtained distributions for these 6 cases were almost same as those of the ideal one. These results indicate that these elements can be divided by using linear discriminant functions with these features in good accuracy.

Figure 16 and 17 are the examples of experimental results from the input images (located at left side in each figure). The result of the extracted ruled lines is shown in the middle, and characters and schemas images are on the right mostly. Figure 17 is a result for an example having plural number of schema images. The extracted dotted lines are not shown in the

figures, but the images can easily be acquired by the subtraction of (b), (c) and (d) from the input image (a). To know the effectiveness of the proposed method for cases of handwritten summary documents, we applied the method to such cases. Figure 18 shows an example of the results. Figure 18(a) is a summary for gynecology with some schema images. In this case medical records were written on the sheet with ruled lines. The result shows that each schema can be extracted even for such case of a handwritten summary. But, characters were regarded as ruled lines because they were located on the original ruled lines (Figure 18(b)). In addition, some characters were also extracted with the schema (Figure 18(d)) as the obtained circumscription rectangle has these characters. A method to eliminate them has to be added to the current extraction method.

4.2.2 Accuracy of schema image recognition

Table 2 shows the obtained results of schema image recognition. In this table, each row means the schema type of the input image and each column is that of the recognition result. This table shows that the recognition accuracy of the proposed method was more than 90%.

	Recognition Result						
Input Image		a	b	c	d	e	Accuracy
	a	20	0	0	0	0	100% (20/20)
	b	0	20	0	0	0	100% (20/20)
	c	0	1	19	0	0	95% (19/20)
	d	0	0	5	14	1	70% (14/20)
	e	1	0	2	0	17	85% (17/20)

Table 2. Result of Schema Image Recognition

Figure 19 shows results of success cases of correctly recognized images. These figures were recognized appropriately by using the proposed method even if there are marks, comments, lead lines for explanations in them. These results indicate that the dictionary with various schema images may not be necessary for recognition if input images do not have many annotations. On the other hand, the schema images with large marks or many annotations were not recognized correctly (Figure 20). Table 3 shows the obtained difference values

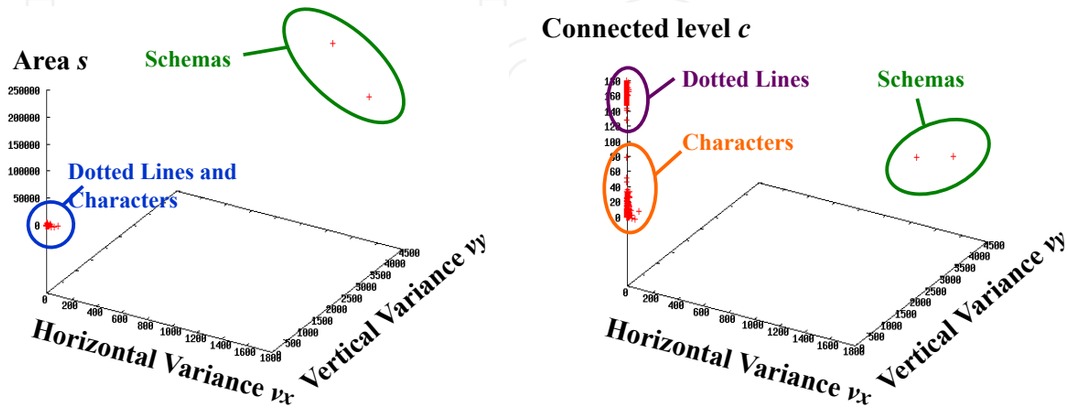
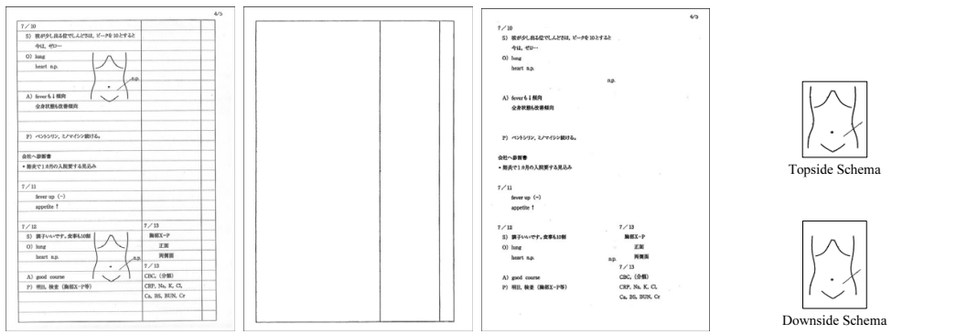


Fig. 15. Example of Obtained Distributions

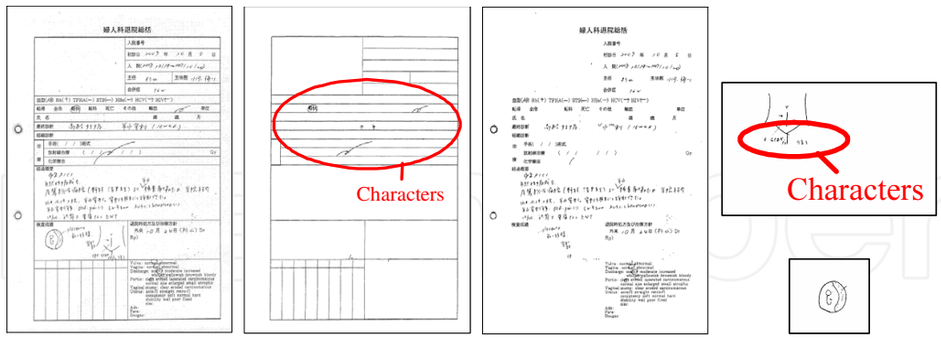
(a) Input Image (b) Ruled Lines (c) Characters (d) Schema Image

Fig. 16. Example of Extraction Results (1)



(a) Input Image (b) Ruled Lines (c) Characters (d) Schema Images

Fig. 17. Example of Extraction Results (2)



(a) Input Image (b) Ruled Lines (c) Characters (d) Schema Images

Fig. 18. Example of Failure Case

given by the discriminant function. In these cases, the large marks (or lead lines) made the contour shape of the input image change drastically, as a result the distance between the input image and the original schema image was larger than that between the input image and the recognition result. In addition, the proposed method outputs the schema type with the smallest distance as a recognition result. Thus it is difficult to detect schema images not

registered in the database. To solve these problems, additional techniques considering the obtained distance values will be required.

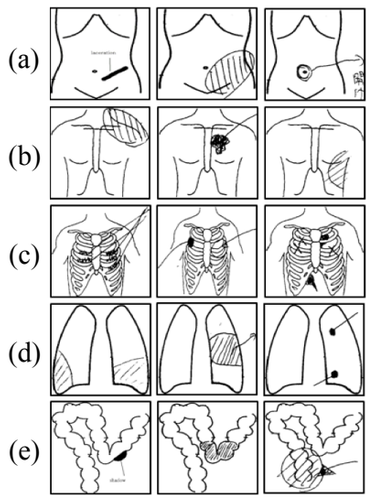


Fig. 19. Result Examples of Schema Recognition (Successful Case)

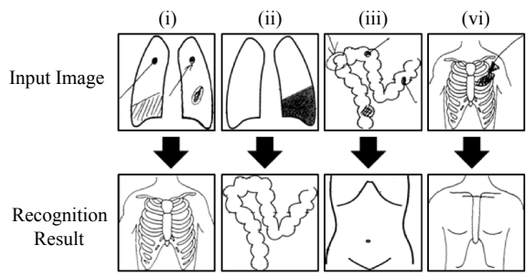


Fig. 20. Result Examples of Schema Recognition (Case of Failures)

Distance between... \ Input Image	(i)	(ii)	(iii)	(iv)
Recognition Result and Input Image	1.24	604	1375	-50
Correct Schema Type and Input Image	897	1944	2774	-25

Table 3. Distance Values Given by Discriminant Function

4.3 Generated XML documents

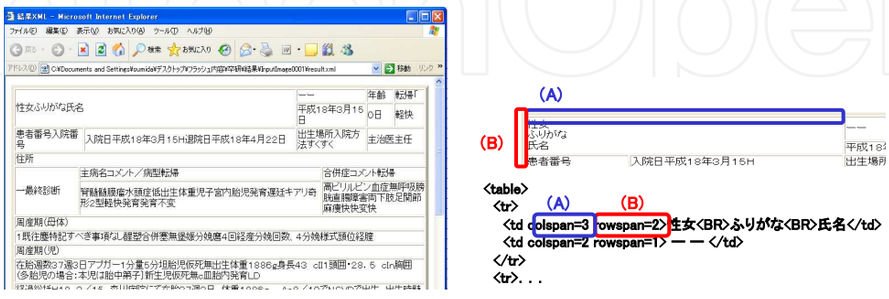
Characters in extracted strings have to be recognized and converted to text data by an Optical Character Reader (OCR) engine. The very strength of our method is that we can define the document type of each frame before the start of recognition of cell images by using the master database and can use any type of OCR engine pertinent to that type. It was found, however, some work is necessary to create interfaces between various OCR engines and our system. At present, we use a commercially available OCR library, developed by "Panasonic Solution Technology, Inc." [22]. The table structure and characters acquired by the proposed method are used to generate an XML file. In this study, an XSL, i.e. defining the table structure of the document, is generated from the acquired node matrix first. The table structure is defined by

table tags in XSL. In the next step of the process, an XML document is generated using XSL and converted text data corresponding to the contents of each cell.

Figure 21 and 22 show examples of generated XML files. In the experiments, the table structures of all input images were recognized correctly. In the case of the document with a schema image, the recognition results, i.e. schema type and annotation part, were inserted to the generated XML file (schema tag in Figure 22). In the present experiments, some parts of the characters were misrecognized. These errors may come from the OCR engine itself. To reduce such errors, it would necessary to use an OCR engine pertinent to the scope of the documents analyzed.

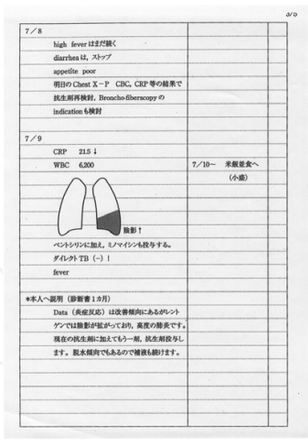
4.4 Developed system for resemble case search

As stated in the introduction, the objective of developing our system is to create a system actively used at healthcare sectors, so that a large volume of paper-based medical records can be included in the e-health environment. For this objective it is necessary to show quickly the usability and/or capability of the method for clinical requirements. Though the research is ongoing, we have developed a prototype system to demonstrate what we can do using this system. We developed a system to search similar cases using Microsoft Visual C# .NET. Figure 23(a) and (b) show the photograph and screenshot of the developed system, respectively. In the system, we used a wizard form with icons to improve the usability of the system. When the system is started, then the wizard window appears at the top left of root window and navigates users who are not experts of information systems. The wizard window of the system consists of some components such as “Image Input”, “System Configuration” and “Scanning”, “Generated XML Viewer” and so on. The image input component supports various input methods. For example, we can input document images from TWAIN devices as well as image files such as Bitmap, JPEG, or PDF files and so on. The system configuration component is so designed as to guide users to set up system parameters easily. When the scanning processes are finished, the structure (and contents) of the input document image are recognized, and a XML file is generated. It takes several tens of seconds before the XML document is generated. The generated XML file is shown in the viewer window (Figure 23(c)). After this, we can search similar cases from the stored documents by using keywords like Figure 23(d). Since the generated XML documents have high compatibility with relational databases, the documents can easily be imported to hospital information systems. If data mining software (or systems) such as data ware house OLAP tools, and so on can be used, these XML documents would be used more effectively for clinical and medical study.



(a) Generated XML Document (b) Source of XML File

Fig. 21. Example of Generated XML File (1)

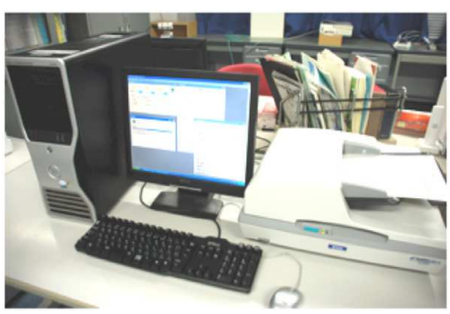


(a) Generated XML Document

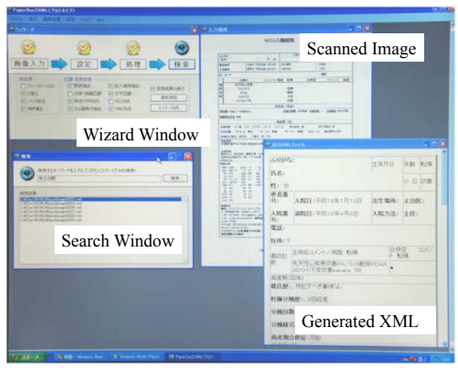
```
<セル no="0" colspan="1" rowspan="1">
  <内容>7／8high fever はまだ続く diarrhe
ite poor 明日の Chest x - P CBC, CRP等
roncho・fn) erscopy の indication も検討 7
BC 6, 200 陰影↑ベントシリンに加え，ミノマイシンも投与する
fever＊本人へ説明（診断書1ヵ月）Data（炎症反応）は改善
は陰影が拡がっており，高度の肺炎です。現在の抗生剤に加えてもう一
水傾向でもあるので補液も続けます。</内容>
  <schema>
    <name>肺</name>
    <annotation-part>左肺下葉</annotation-part>
  </schema>
</セル>
<セル no="1" colspan="1" rowspan="1">
  <内容>7／10～ 米飯並食へ（小盛）</内容>
</セル>
<セル no="2" colspan="1" rowspan="1">
  <内容 />
</セル>
```

(b) Source of XML File

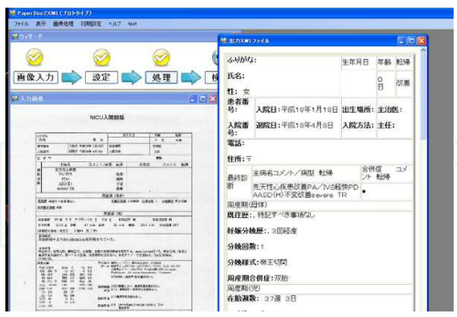
Fig. 22. Example of Generated XML File (2)



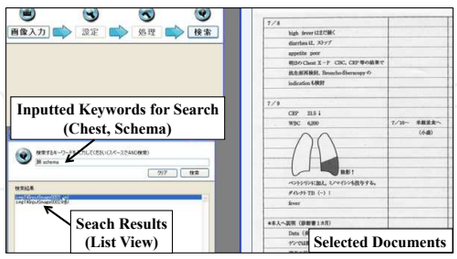
(a) Overview



(b) Schreenshot of Main Window



(c) Scanning and XML Viewer



(d) Resemble Case Search

Fig. 23. Developed System

5. Related works

Studies and research for document image analysis systems have been reported [23]-[31]. As related works to ruled line extraction, the detection methods using the Hough transform technique are reported by literature [23]-[27]. Particularly in literature [23] and [24], complex line shapes can be extracted using a pattern-matching method and Hough transform method.

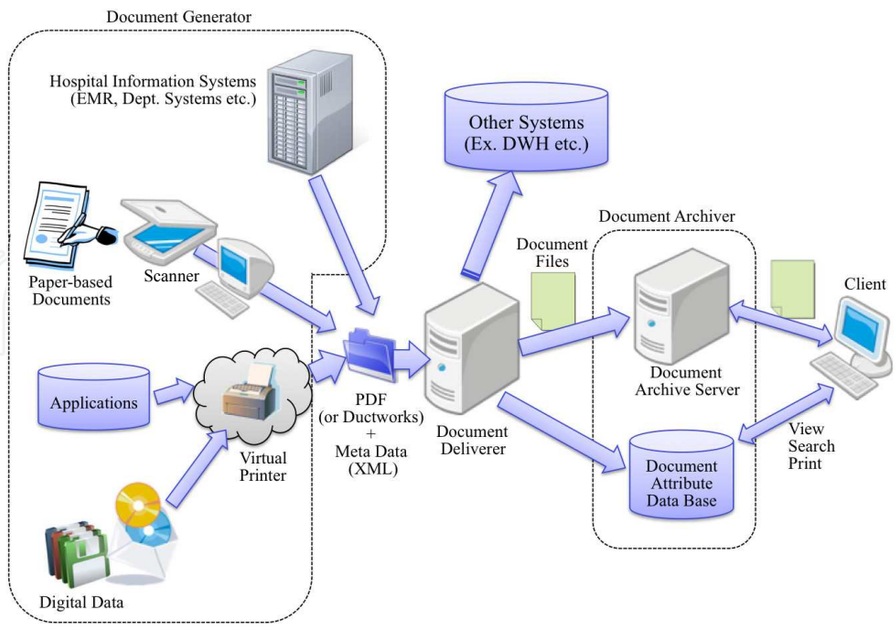


Fig. 24. Basic Structure of DACS

In Literature [25] - [27], authors propose detection methods for character patterns, general curving lines, quadratic curving lines, circular patterns using the concept of [23] and [24], and discuss their effectiveness. These methodologies may have higher extraction accuracy when compared with the proposed method, but they require a large amount of calculation time, because these algorithms are so complex. In practical situation, processing time is the most important factor to evaluate systems. Therefore, it is not realistic to employ them in cases where large number of documents are being processed.

As for related methods for document layout and structure recognition, literature [28] reports the table structure recognition method based on the block segmentation method and literature [29] tries to extract the contents from printed document images using model checking. The method of literature [28], however, depends on the output of commercial OCR systems. On the other hand, our proposed method identifies table types, i.e. document types, using a node matrix and positions of nodes. The node matrix can be acquired easily by using the extracted ruled lines and the lines themselves are obtained by very simple image processing techniques. The proposed method does not depend on an external library in image processing. In the case of [29], only the logical structures in the documents are detected using image analysis but the system is not developed to reuse the information. In a different field, methods to analyze cultural heritage documents are reported by Ogier et al. [30]. In this literature, document analysis techniques are employed to preserve and archiving cultural heritage documents. Literature [31] reports a prototypical document image analysis system for journals. Most of these studies mainly describe the methodology and processing for typical business letters. According to the authors' survey, only a few articles propose document image recognition method for medical documents, such as patient discharge summaries to search similar cases.

In medical fields, many novel information systems have been studied. As one of such examples, we introduce here a new concept and systems to assure lifelong readability for Medical Records in HIS. Figure 24 illustrates the outline of the concept, called Document

Archiving and Communication Systems (DACS), proposed by Prof. Matsumura et al. in 2010 [32]. Since the lifespan of computer systemss is usually very short compared with the need for medical records of a patient, great care is necessary to shift paper-based toward computer based society. DACS is such a system which covers this problem. Because of the very nature of rapid progress of medical science, all the electronic health record system used now will never mature, and, indeed, the system architecture itself is changing. It is sometimes very difficult to retrieve data created by a system previously used. Though electronic health record systems offer us utilities to retrieve any type of data in the database, they loose functionality to grasp many features at a glance which the paper systems had. Prof. Matsumura et al. deliberately introduces the combination of these two concepts. In the DACS, all medical records are not treated as data but as an aggregation of documents. The medical documents generated by the electronic health system are converted to PDF (or JPEG, TIFF, Docuworks) and XML files. By converting the data to such files, the readability of the data are guaranteed, and the meta-data of the documents, e.g. timestamp, patient ID, document type etc., are used as key information of search. After this, these files are delivered to Document Archive Server of DACS, and then system users can view and search the stored documents easily. As a matter of course, the document deliverer of DACS can also deliver the generated files (and XML data) to other systems such as Data Ware House (DWH), and we can use the data for clinical analyses and studies. DACS also supports not only the data stored in HIS but also other data types, e.g. paper-based documents, other applications' data, PDF files generated by other systems and so on. In the case of a paper-based document, the target document is scanned by the optical scanning device and transferred into a PDF file. The meta-data of the documents are also obtained by a scanning sheet with the QR code. This sheet is generated using stored clinical data in the HIS (or input data to the DACS by hand) before scanning. The generated PDF file and its meta-data are delivered by the document deliverer and stored into the database. As you can see, DACS can keep readability of medical records and supports various data types. One of the problems that DACS has now will be the problem of creating meta-data manually. Our method can cover this problem as much of these meta-data are automatically extracted from the images, which would contribute to improve DACS.

6. Toward the future

In this chapter we introduced document image recognition, keyword extraction and automatic XML generation techniques to search similar cases from paper-based medical documents. These techniques were developed for practical use at healthcare sectors, so as to help the incorporation of vast volumes of paper-based medical records into the e-health environment. Good usability and speed, robustness, low running cost and automated execution will be the key requisite for such a system to practically be used, and our system will satisfy many of these requirements. These characteristics of our system mainly come from the use of master information which covers almost all type of medical documents. However, there remain many problems unsolved. One of the largest drawbacks of our system might be the anxiety whether we can get similar accuracy and effectiveness of such documents without tables. As is stated in 3.1 there are many paper based medical documents without tables. But, even in such cases, they are not written randomly in free format. Since medical records are the most important documents for physicians to keep continuity of healthcare, the format itself has been deliberately designed and used. Therefore it is quite plausible that any medical documents without tables will match one of the master information if we can insert frame lines in it. If so, it may not so difficult to improve the algorithm of determining the best suited sheet to include mass or area information.

7. References

[1] H. Harold Friedman, Ed., *Problem-Oriented Medical Diagnosis 5th edition*. Lippincott Williams & Wilkins, 1991

[2] K. Seto, T. Kamiyama, H. Matsuo, "An Object-Modeling Method for Hospital Information Systems," *The 9th World Congress on Medical Informatics*, 52 Pt.2, pp.981–985, 1998

[3] HJ. Lowe, I. Antipov, W. Hersh, CA Smith, M. Mailhot, "Automated Semantic Indexing of Imaging Reports to Support Retrieval of Medical Images in the Multimedia Electronic Medical Record," *Methods of Information in Medicine*, vol. 38, no. 4, pp. 303–307, 1999

[4] H. Kawanaka, Y. Otani, T. Yoshikawa, K. Yamamoto, T. Shinogi, S. Tsuruoka, "Tendency Discovery from Incident Reports with Free Format Using Self Organizing Map," *Japan Journal of Medical Informatics*, vol. 25, no. 2, pp. 87–96, 2005

[5] Y. Otani, H. Kawanaka, T. Yoshikawa, K. Yamamoto, T. Shinogi, S. Tsuruoka, "Keyword Extraction from Incident Reports and Keyword Map Generation Method Using Self Organizing Map," *Proc. of 2005 IEEE International Conference on Systems, Man and Cybernetics*, pp. 1024–1029, 2005

[6] H. Kawanaka, T. Sumida, K. Yamamoto, T. Shinogi, S. Tsuruoka, "Document Recognition and XML Generation of Tabular Form Discharge Summaries for Analogous Case Search System," *Method of Information in Medicine* (Schattauer), vol. 46, no. 6, pp. 700–708, 2007

[7] H. Kawanaka, Y. Shiroyama, K. Yamamoto, T. Shinogi, S. Tsuruoka, "A Study on Document Structure Recognition of Discharge Summaries for Analogous Case Search System," *Proc. of International Workshop on Document Analysis Systems* (DAS2008), pp. 423–430, 2008

[8] N. Otsu, "Discriminant and Least Squares Threshold Selection," *Proc. of 4IJCPR*, pp. 592–596, 1978

[9] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," *IEEE Trans. Systems, Man, and Cybernetics*, SMC-9, no.1, pp. 62–66, 1979

[10] T. Akiyama, I. Masuda, "A Segmentation Method for Document Images without the Knowledge of Document Formats," *The IEICE Transactions on Information and Systems*, vol. J66-D: pp. 111–118, 1983

[11] T. Tanaka, S, Tsuruoka, "Table Form Document Understanding Using Node Classification Method and HTML Document Generation." *Proc. of third IAPR Workshop on Document Analysis Systems*, pp. 157–158, 1998

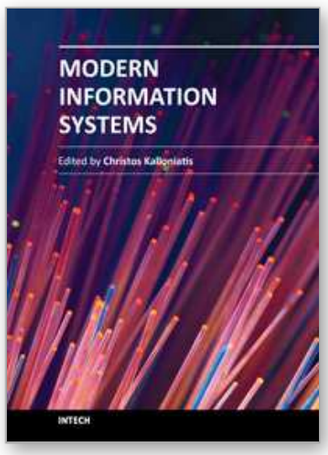
[12] Y. Ito, M. Ohno, S. Tsuruoka, T. Shinogi, "Document Structure Understanding on Subjects Registration Table," *Proc. of the fourth International Symposium on Advanced Intelligent Systems*, pp. 571–574, 2003

[13] S. Tsuruoka, C. Hirano, T. Yoshikawa, T. Shinogi, "Image-based Structure Analysis for a Table of Contents and Conversion to XML Documents," *Proc. of Document Layout Interpretation and its Application*, pp. 59–62, 2001

[14] F. Kimura, T. Wakabayashi, S. Tsuruoka, and Y. Miyake, "Improvement of handwritten Japanese character recognition using weighted direction code histogram" *Pattern Recognition*, vol. 30, no. 8, pp. 1329 – 1337, 1997

[15] S. Tsuruoka, M. Kurita, T. Harada, F. Kimura, Y. Miyake, "Handwritten KANJI and HIRAGANA Character Recognition Using Weighted Direction Index Histogram Method," *The Transactions of the Institute of Electronics, Information and Communication Engineers*, vol. 70-D, no. 7, pp.1390-1397, 1987

[16] S. Tsuruoka, H. Morita, F. Kimura, Y. Miyake, "Handwritten Character Recognition Adaptable to the Writer", *Proc. of IAPR Workshop on CT – Special Hardware and Industrial Applications*, pp. 179 – 182, 1988
[17] H. Takebe, "Pattern recognition apparatus and method using probability density function," *United States Patent*, no.7003164 B2, 2006
[18] H. Takebe, Y. Hotta, S. Naoi, "Word recognizing apparatus and method for dynamically generating feature amount of word," *European Patent Specification* (European Patent Office), no.EP0997839, 2005
[19] S. Tsuruoka, N. Watanabe, N. Minamide, F. Kimura, Y. Miyake, M. Shrindhar, "Base line correction for handwritten word recognition," *Proc. of the Third International Conference on Document Analysis and Recognition*, vol.2, pp. 902 – 905, 1995
[20] S. Hirose, M. Yoshimura, K. Hachimura, R. Akama, "Authorship Identification of Ukiyoe by Using Rakkan Image," *The Eighth IAPR International Workshop on Document Analysis Systems*, pp. 143 – 150, 2008
[21] All Japan Hospital Associations, Ed., *Text Book for Generation of Standard Medical Records and their Administration*. Jiho Inc., 2004
[22] Panasonic Solution Technologies Co., Ltd. Color OCR Library. Color OCR Library: "Yomitori Kakumei" SDK, http://panasonic.co.jp/
[23] D. Casasent, R. Krishnapuram, "Curved Object Location by Hough Transformations and inversions," *Pattern Recognition*, vol. 20, no. 2, pp. 181-188, 1987
[24] R. Krishnapuram, D. Casasent, "Hough Space Transformation for Discrimination and Distortion Estimation," *Computer Vision, Graphics and Image Processing* vol. 38, no. 3, pp. 299–316, 1987
[25] D. Pao, H.F. Li, R. Jayakumar, "Detecting Parametric Curves Using the Straight Line Hough Transform," *Proc. of 10th International Conference on Pattern Recognition*, pp. 620-625, 1990
[26] K. Fujimoto, Y. Iwata, S. Nakata, "Parameter Extraction of Second Degree Curve from $\theta - \rho$ Hough Plane," *The IEICE Transactions on Information and Systems*, vol. J74-D2, no. 9, pp.1184–1191, 1991
[27] J. Yan, T. Agui and T. Nagao, "A Complex Transform for Extracting Circular Arcs and Straight Line Segments in Engineering Drawings," *The Trans. of the Institute of Electronics, Information and Communication Engineers*, vol. 75, no.8, pp.1338–1345, 1992
[28] T. G. Kieninger, "Table Structure Recognition Based on Robust Block Segmentation," *Proc. Document Recognition V, SPIE*, vol. 3305, pp. 22-32, 1998
[29] M. Aiello, "Document Image Analysis via Model Checking," *AI*IA Notizie*, vol.1, 200–2, 2002
[30] J.M. Ogier, K. Tombre, "Madonne: Document Image Analysis Techniques for Cultural Heritage Documents," in Digital Cultural Heritage, *Proceedings of 1st EVA Conference*, pp. 107–114, 2006
[31] G, Nagy, S. Seth and M. Viswanathan, "A Prototypical Document Image Analysis System for Technical Journals," *IEEE Computer*, vol. 25, no. 7, pp. 10-22, 1992
[32] Y. Matsumura, N. Kurabayashi, T. Iwasaki, S. Sugaya, K. Ueda, T. Mineno, H. Takeda, "A Scheme for Assuring Lifelong Readability in Computer Based Medical Records", *MEDINFO 2010* C.Safran et al. (Eds.) , IOS Press, pp. 91 – 95, 2010



Modern Information Systems

Edited by Dr. Christos Kalloniatis

ISBN 978-953-51-0647-0
Hard cover, 166 pages
Publisher InTech
Published online 13, June, 2012
Published in print edition June, 2012

The development of modern information systems is a demanding task. New technologies and tools are designed, implemented and presented in the market on a daily bases. User needs change dramatically fast and the IT industry copes to reach the level of efficiency and adaptability for its systems in order to be competitive and up-to-date. Thus, the realization of modern information systems with great characteristics and functionalities implemented for specific areas of interest is a fact of our modern and demanding digital society and this is the main scope of this book. Therefore, this book aims to present a number of innovative and recently developed information systems. It is titled "Modern Information Systems" and includes 8 chapters. This book may assist researchers on studying the innovative functions of modern systems in various areas like health, telematics, knowledge management, etc. It can also assist young students in capturing the new research tendencies of the information systems' development.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen