BUILDING AN INTEGRATED OMR-AND-PAPER-BASED DATA AUTOMATIC COLLECTION SUPPORT SYSTEM

W.H. Tseng*, P. Chang* and T.H. Ma**

* Institution of Health Informatics and Decision Making, National Yang-Ming University, Taipei, Taiwan/ROC
** Institute of Information Science, Academia Sinica, Taipei, Taiwan, ROC

vncnt@ms93.url.com.tw

Abstract: The purpose of our research is to design an automatic questionnaire recognition system to support entire life cycle of survey. Results showed the lowest error rate of questionnaire recognition is 0.00057 per checkbox, and the lowest time of questionnaire recognition which include gathering and recognition of questionnaire images is 1.17 seconds per page. This system will be a junction system between traditional paper-based survey and current electronic survey.

Keywords: OMR, automatic, paper-based, survey, questionnaire recognition, pattern recognition

Introduction

'Paperless offices' were anticipated in the decade ahead. But the concept of a paperless office may seem a distant dream. In 2004, the Wall Street Journal had reported that only 5% of physicians are using electronic medical records [1]. Therefore not all information in an organization is in paperless today. The materials in digital form will not displace traditional sources; instead both paper and digitized resources will coexist [2]. And even Robert Weideman, senior vice president at ScanSoft Corp., Peabody, Mass., had said "We're seeing a growing interest in allowing paper and digital content to more efficiently coexist inside a hospital." [3]. Paper documents continue to exist for two reasons. First, many companies have mountains of documents that contain useful data but rendering all of them to native Web formats would be expensive. Secondly, some applications require greater detail and typographic quality than Web formats can provide [4-10]. Therefore the most of organizations still archive information electronically and papers too coexist now. According to Eric [11], the problem of how to handle paper patient-related documents in EMR-based practices will not go away.

Paper-based Questionnaire is a still the main mechanism of survey for people to collect and exchange data in healthcare because it is the most popular approach to collect data. But when a lot of questionnaires surge, human errors become another serious issue to worry about. Human beings simply do not perform tasks with the precision and repeatability of robots. As a result, workplace activities are peppered with errors [12]. For this reason, the survey approach should include a well-systematized data collection technique [13-22].

Materials and Methods

This system consists of three primary modules which are "Document Positioning Generator (DPG)", "Questionnaire Recognition Engine (QRE)" and "Correction and Statistic Tools (CST)". The system architecture of this system showed in Figure 1.

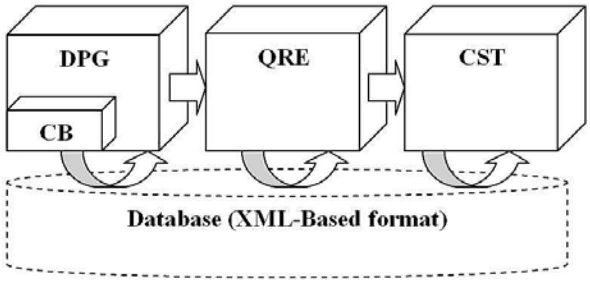

Figure 1: System architecture

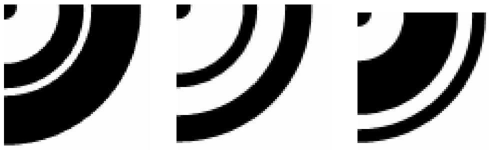

Figure 2: three special fan-shaped wavelike figures

ISSN: 1727-1983 © 2005 IFMBE

The basic format of questionnaire for recognition is to answer by checkboxes (Font: DFKai-SB; Size: 10; Bold), and in order to raise the recognition accuracy, we put three special 90^o fan-shaped wavelike figures (see Figure 2) on the left-top corner, right-top corner and left-foot corner respectively of every questionnaire. The purpose of the three special fan-shaped wavelike figures is to make system could recognize the three wavelike figures automatically before to recognize checkboxes. And then system could correct all positions of checkboxes by compare old positions of three wavelike figures of questionnaire draft with new positions of three wavelike figures of answered-questionnaire. All questionnaires will be transformed into monochromatic 200 dpi (dots per inch) BMP (bitmap) format image files by scan or fax. The each image of questionnaires is about 1700 X 2330 pixels and 500 KB (kilobyte).

Our system could pre-define the checkbox positions of questionnaire by controlling Microsoft Office Word like using the VBA (Microsoft Office Visual Basic® for Applications) programming language, because of our system could communicate with the Microsoft Office Word by calling the Microsoft.Office.Interop.Word.dll and Interop.Microsoft.Office.Core.dll.

The two main methods to design QRE for getting and adjusting the position of every checkbox before start of questionnaire recognition are offset and corrective matrix. The problems could not be solved by offset, for instance, the size change and the spin of questionnaire images will be done by using the corrective matrix. The algorithm to produce the corrective matrix of our study is to adopt the affine mapping functions (2D affine transformation) [23]. A general affine transformation from 2D to 2D as in Equation E.4 requires six parameters and can be computed from only 3 matching pairs of points $([m_j,n_j],[p_j,q_j])_{j=1.3}$.

An example to explain is as below. Before questionnaire recognition, system will to automatically find the three special fan-shaped wavelike figures of questionnaire image we want to recognize. And after finding all positions of the three special fan-shaped wavelike figures, system will use another three position points $A(x_1,y_1)$, $B(x_2,y_2)$, $C(x_3,y_3)$ of the three special fan-shaped wavelike figures of questionnaire draft from database and the three position points $A'(u_1,v_1)$, $B'(u_2,v_2)$, $C'(u_3,v_3)$ we got just now to be the six parameters of Equation (E.1) to gain the anti-matrix (E.2). And then system will rearrange the anti-matrix (E.2) we got just now to be a new matrix (E.3) which is just the corrective matrix we want. Finally, system will use the old position point D (m,n) of each checkbox from database to be the parameter of Equation (E.4) respectively, and then we could get and adjust all new position points D′(p,q) of checkboxes of all questionnaire image we want to recognize.

The performance evaluation of this system was tested by using three kinds of questionnaires, and the total numbers of each kind of questionnaires was 30. We respectively used the word-processor-based method and the paper-image-based method to pre-position the questionnaire drafts. And then we respectively used scan and fax to get the images of each kind of questionnaires for recognition. Of course we also entered data by hand, and we checked the result again and again in order to enable this result could be the gold standard.

Results

The overview of system flow was shown in Figure 3. Before start of survey, users must to build the code book and design the questionnaire first. And users must embed the background picture includes the three special 90^o fan-shaped wavelike figures in the questionnaire draft before pre-position questionnaire, but users could also use the word-processor-based-preposition function of DPG to embed the background picture. And then users could use the DPG to pre-position all checkboxes of questionnaire by inputting electronic file (.doc) or scanned image of the original questionnaire draft.

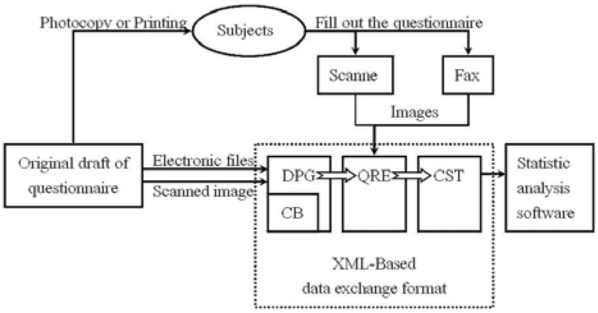
Photocopy or Printing
Subjects
Fill out the questionnaire
Scanne
Fax
Images
Original draft of questionnaire
Electronic files
Scanned image
DPG
QRE
CST
CB
Statistic analysis software
XML-Based data exchange format

Figure 3: system flow chart

After finishing the work of pre-position, the position database will be produced. And then users could start to survey by using the questionnaire draft to photocopy or print. After completing the survey, users could get the images of questionnaire which subjects had filled out by scanner or fax. And then users could use the QRE to recognize all questionnaire images to replace the traditional hand-data-entry. After questionnaire recognition, the system will automatically pass the results on to the CST. And then users could use the CST to view the statistical data of survey, and to correct the results of questionnaire recognition with GUI. In the end, users could also export the results which include statistical data and code book to text-based file which can be used by other statistical software to do advanced statistical analysis and produce the related statistical tables or charts, e.g. Microsoft® Office Excel and SPSS (Statistical Package for the Social Sciences.

The result of performance evaluation was shown as Table 1. And from this table, we could discover that the best performance is the combination of using paper-image-based method to pre-position and recognizing images by scan, this combination will have the best accuracy and use the least time.

Discussion

The cost of manpower and the human error of electronic survey (e-survey) all are very low. Therefore e-survey (i.e. web survey and e-mail survey) is popularized now [24]. But e-survey really is not fit for any situations. It has some problems such as the problem of insufficient sample representativeness, the problem of being filled out again by the same subject, and the problem of only subjects in network community. Besides, e-survey needs extra equipments such as PC (Personal Computer), notebook, PDA (Personal Digital Assistant) and network. Therefore it is not fit for outdoor survey.

The answer sheet has the feature of traditional paper-based questionnaire. It is light, portable and does need electric power in addition. But answer sheet also has an extra advantage which traditional paper-based questionnaire does not have is the high speed of data entry. Therefore it has been used on large scale for examinations now. However, the UI (Users Interface) of answer sheet for writing is not friendly. The focus of users always needs to round between examination questions and answer sheet. And in consequence, users frequently fill out the wrong field because of unclear and tired sight. Besides, the answer sheet usually needs the high price equipments in addition, and also needs to use exclusive software and special paper. The position of our system is lay in between traditional paper-based questionnaire and answer sheet. The UI of it for writing is friendly like traditional paper-based questionnaire, because users can directly answer questions by using the general traditional paper-based questionnaire. However, our system also like answer sheet has the advantage of high speed of data entry. The data entry by computer is more stable than work done by hand. The recognition results of same questionnaires at the same environment are always the same every time, but the results of work done by hand will be possible not the same every time.

Our system has the lower cost of manpower. But it also has the higher cost of equipments such as scanner. And the depreciation of equipments is also a problem.

The system has still some limitations or defects now. For example, when we scan a questionnaire, maybe it will be stick in the scanner. The quality (more or less dots) of questionnaire images will affect the accuracy of recognition. If we raise the sensitivity of QRE to make the mark of loss dots could be identify, but maybe it will also identify some little stains as marks. Besides, if the questionnaire images slant or revolve too much, it is possible to fail the questionnaire recognition.

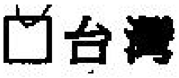

Figure 4: overlap I

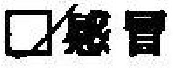

Figure 5: overlap II

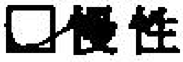

Figure 6: overlap III

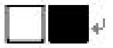

Figure 7: encode by checkboxes

The overlap is the oftenest situation of recognition failure. Sometimes the check overlaps the checkbox exactly such as the Figure 4, Figure 5 and, Figure 6; therefore it maybe can not be identified. We had tried to identify surrounding environment of checkbox, but it was easy failure because of being affected by other surrounding content of checkbox (i.e. words or serial numbers of questions). And we had also tried to identify the pattern of check, but it was difficult and needed complex algorithm. Furthermore, it will also limit users to write the questionnaire only by check. And the complex algorithm will also extend the recognition time. Besides, we had also tried to identify check and checkbox by colors. But it will consume large system resource, the memory especially. Because the image type of questionnaire must be changed from monochrome to color, and then the image size will rise sharply from 500 KB (kilobyte) to 20 MB (megabyte).

In some survey, users accept certain mistake of questionnaire recognition. They don't have demand for absolute accuracy. Therefore we just need to reduce the errors to the level users could accept. But in another situation, maybe users need the absolute accuracy. In former days users had a double check after data entry, and it were all done by person. As regards our system, after questionnaire recognition maybe we still need a double check by person, and make system to give users some error messages when recognition may be failure.

Conclusions

Our system will be a junction system between traditional paper-based survey and current electronic survey. And it is fit for the social survey has big samples and large checkboxes per page, because of the small error will have no effect on the result of survey (work done by hand also has the same problem). Furthermore, the high checkbox density will raise the accuracy, because the lower offset will be able to avoid the variation of image and adjust the position of checkbox more exactly. Oppositely, our system is not fit for the questionnaire type of needing 100% accuracy such as medical record, order and accounting, because any small error will cause the serious effect.

We have some recommendations of system use here. First, standardizing design of questionnaire will improve the recognition accuracy. Users could limit their subjects to write questionnaire by dark and bold ﹀ mark but not ✕ mark, because of the ﹀ mark is easier checked outside. Second, the questionnaire design is fit for high checkbox density. As regards the system operation, we suggest that pre-positioning directly by questionnaire image and using the scanned images for recognition.

This system is still insufficient for user’s needs. We shall endeavor to improve our system continually.

References

[1] Laura Landro, (2004): Push Grows for Online Health Data. http://online.wsj.com/ . 3-11-2004. The Wall Street Journal Online.

[2] Barbara B.Moran, (1999): Virtual Realists: Librarians in a Time of Transition. North Carolina Libraries 1999; 57(4):165-169.

[3] Mark Hagland, (2005): Document Management Systems. http://www.healthcare-informatics.com/issues/2005/04_05/snapshot.htm . 2005. Healthcare Informatics Online.

[4] Stephen R.Jones, (1999): Keynote: Dr. Charles Geschke at PC Expo 99 New York. http://www.reviewsonline.com/PCX99KY3.HTM . 6-24-1999.

[5] James T.Mulder, (2005): Hospitals Going Paperless. http://www.syracuse.com/business/poststandard/index.ssf?/base/business-6/111062036127250.xml . 3-13-2005.

[6] Pincus T., (1996): Documenting quality management in rheumatic disease: are patient questionnaires the best (and only) method? Arthritis Care & Research 1996; 9:339-348.

[7] Slack WV, Slack CW., (1972): Patient-Computer dialog. N Engl J Med 1972; 286:1304-1309.

[8] Hodsdon D.F., (1967): Remarks from a representation of farm workers, Ergonomics problems in agriculture. 1967. Meeting at Nottingham University School of Agriculture.

[9] Davis FD., (1989): Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology. MIS Quarterly 1989; 13(3):319-340.

[10] Michael G.Morris, Andrew Dillon, (1997): The Influence of User Perceptions on Software Utilization: Application and Evaluation of a Theoretical Model of Technology Acceptance. IEEE Software 1997; 14(4):58-76.

[11] Eric Rose, (2003): Life After Go-Live-Part 1: Paper in the Paperless Practice. Journal of Healthcare Information Management 2003; 17(1):24-26.

[12] Michael E.Wiklund, (1995): Medical Device and Equipment Design-Usability Engineering and Ergonomics. Interpharm Press, Inc., 1995.

[13] Lu Ann Aday, (1989): Designing and Conducting Health Surveys: A Comprehensive Guide. San Francisco, California: Jossey-Bass Inc., 1989.

[14] James J.Neutens, Laurna Rubinson, (2002): Research Techniques for the Health Sciences. 3 ed. San Francisco, CA.: Benjamin Cummings, Inc., 2002.

[15] Czaja S., (1990): Special issue preface. Human Factors 1990; 32(5):505.

[16] Diffrient N, A.Tilley, J.Bardagjy. Humanscale 4/5/6. Cambridge, MA: MIT Press. In press.

[17] LeGros Clark W.E., (1954): The anatomy of work. Floyd W.F., Welford A.T., editors. 1954. Lewis, London., Symposium on human factors in equipment design.

[18] MB Weigner, CE Englund, (1990): Ergonomic and human factors affecting anesthetic vigilance and monitoring performance in the operating room environment. Anesthesiology 1990; 73(5):995-1021.

[19] Anonymous, (2000): The next level of Internet applications: providing quality care to high-risk patients. Qual Lett Healthc Leaders 2000; 12(5):2-3.

[20] Scott E Umbaugh, (1998): Computer vision and Image Processing: A Practical Approach Using CVIPtools. Prentice Hall PTR, 1998.

[21] Sing-Tze Bow, (1992): Pattern Recognition and Image Preprocessing. Marcel dekker, Inc., 1992.

[22] Neil Collings, (1988): Optical Pattern Recognition Using Holographic Techniques. Addison-Wesley Publishing Company, 1988.

[23] Linda G.Shapiro, George C.Stockman, (2001): Computer Vision. Prentice Hall, Inc., 2001.

[24] IT Discovery: Web Survey Project Final Report, (2001): http://web.mit.edu/is/discovery/web-surveys/report/ . 9-5-2001. Massachusetts Institute of Technology.

$$\begin{pmatrix} \Sigma x_j^2 & \Sigma x_j y_j & \Sigma x_j & 0 & 0 & 0 \\ \Sigma x_j y_j & \Sigma y_j^2 & \Sigma y_j & 0 & 0 & 0 \\ \Sigma x_j & \Sigma y_j & \Sigma 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & \Sigma x_j^2 & \Sigma x_j y_j & \Sigma x_j \\ 0 & 0 & 0 & \Sigma x_j y_j & \Sigma y_j^2 & \Sigma y_j \\ 0 & 0 & 0 & \Sigma x_j & \Sigma y_j & \Sigma 1 \end{pmatrix} \begin{pmatrix} a_{11} \\ a_{12} \\ a_{13} \\ a_{21} \\ a_{22} \\ a_{23} \end{pmatrix} = \begin{pmatrix} \Sigma u_j x_j \\ \Sigma u_j y_j \\ \Sigma u_j \\ \Sigma v_j x_j \\ \Sigma v_j y_j \\ \Sigma v_j \end{pmatrix} \qquad \text{(E.1)}$$

$$\begin{pmatrix} a_{11} \\ a_{12} \\ a_{13} \\ a_{21} \\ a_{22} \\ a_{23} \end{pmatrix} \qquad \text{(E.2)}$$

$$\begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{pmatrix} \qquad \text{(E.3)}$$

$$\begin{pmatrix} p \\ q \\ 1 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} m \\ n \\ 1 \end{pmatrix} \qquad \text{(E.4)}$$

Table 1: performance evaluation

Type	position	image source	accuracy	time
questionnaire of airline company (77 checkboxes)	Image	scan	accuracy: 99.8701%	scan: about 1 spp* QR**: about 0.17 spp
	Word	scan	accuracy: 99.8268%	
	Image	fax	accuracy: 99.2641%	fax: about 22 spp QR: about 0.17 spp
	Word	fax	accuracy: 96.2771%	
medical record of race (117 checkboxes)	Image	scan	accuracy: 99.9430%	scan: about 1 spp* QR: about 0.23 spp
	Word	scan	accuracy: 88.4615%	
	Image	fax	accuracy: 99.3447%	fax: about 25 spp QR: about 0.23 spp
	Word	fax	accuracy: 97.6923%	
questionnaire of voice system (160 checkboxes)	Image	scan	accuracy: 99.9167%	scan: about 1 spp* QR: about 0.3 spp
	Word	scan	accuracy: 98.8542%	
	Image	fax	accuracy: 97.6250%	fax: about 25 spp QR: about 0.3 spp
	Word	fax	accuracy: 95.6875%	

*spp: seconds per page
**QR: average time of questionnaire recognition (three special wavelike figures: about 0.545 second; each checkbox: about 0.0015 second)