**Proceedings of Women in Science and Technology for the Developing World**

**Fifth General Assembly and International Conference**

**May 16-19, 2016, Kuwait**

**Structured Data OCR for the Dynamic Online Questionnaire System**

|  |  |
| --- | --- |
| Buhamad, M.  Kisr  Kuwait | Naseeb, A.  Kisr  Kuwait |

|  |  |  |
| --- | --- | --- |
| Al-Khayat, A.  Kisr  Kuwait | Aljaber, A.  Kisr  Kuwait | Al-Ali, M.  Kisr  Kuwait |

Abstract

**Background**

In past years a software system was developed to address a number of issues related to the questionnaire process. The system was successful in the data-collection of all the electronic answer sheets. Whereas the conventional hard copy answer sheet is more demanding when it comes to data entry. This paper talks about the approaches to automate the data entry of the hard-copy answers into an indexed database. Hence the need to develop a software that is aware of the design of the layout and the nature of the options.

**Material and Methods**

The system contains two layers, the first is the foundation layer, and second is the answers layer from the population sample. The first layer comes from the meta information composed of locations and dimensions, which is built by the form design subsystem. The second layer is a filtered process of scanned image minus the foundation layer. Afterwards the system can distinguish the logical areas in the answers layer, each logical area can be processed by a specialized algorithm.

**Results**

The developed system did show a remarkable improvement in the data collection stage compared to conventional paper-based data collection. The system proved to be a productive framework for building professional solutions to many frequent issues. The built-in automated calculations of frequency tables and the generation of charts was very impressive and fairly satisfactory to many study teams that have applied their study using this system.

**Conclusion**

It has been concluded that people with numerous experiences in conducting different surveys are able to formulate a formal process. Furthermore, crystalizing the formal process by using custom developed software based on modern IT practices. Hence, the experience of the formal process leads to the need to develop an intelligent algorithm to recognize pre-defined data schema.

**Keywords**

Optical Character Recognition, Electronic Answer Sheets, Structured Data, Indexed Database, Information Technology Practices, Intelligent Algorithm, Data schema.

INTRODUCTION

In past years a software system was developed by Techno-Economics Division (TED) in Kuwait Institute for Scientific Research (KISR) to address a number of issues related to the questionnaire process. The objective of TED software system is to develop an online survey answer sheet, that allow users to answer electronically. The system was successful in the data-collection of the electronic answer sheets. However, some users prefer using conventional Hard Copy Answer Sheet (HCAS), which is more demanding when it comes to data entry. That issue created some drawbacks in the system which are time consuming and high cost. On the other hand, the electronic form usually requires less than a second for the submitted answers to be stored in the database. The estimated ratio between the electronic form and HCAS is almost 1:259,200. (Mori, S.; Nishida ,H.; Yamada, H.,1999)

Furthermore, Optical Character Recognition (OCR) software packages are available in the market. The main issue of the OCR software packages is that they can only process generic printed text. Furthermore, it is very difficult for these generic packages to process tabular data, even though they are feed a large amount of inputs. In addition, the packages are not programmed to learn any data schema, and is almost impossible for the software to reproduce an electronic copy of HCAS. (Trier, O.D; Jains,A.K.; Taxt,T.,1996).

This paper talks about the Process to Automate the Data Entry of the Hard-copy answers (PADEH) into an indexed database. The structure of the paper is as follows; 2) methodology; 3) Results. 4) Conclusion.

methodology

The objective of the PADEH is to recognize scanned images and automate the data entry into an indexed database. To implement this research idea, the system will conduct a process that has seven stages. Stage One: Construction of Metadata; Stage Two: Scanning HCAS Images of Answers; Stage Three: Subtraction Bitmap[[1]](#footnote-1). Stage Four: Creating Cut-Out Images from the Subtraction Bitmap. Stage Five: Conversion to Vector Graphics. Stage Six: Mapping Graphics into Textual Information. Stage Seven: Insertion into Indexed Database. (White, J. M.; Rohrer,G. D.,2010).

Stage One: Construction of Metadata.

In the early stages of designing the questionnaire, the project team has developed an editor to aid the user in building an interactive questionnaire design. Behind the scenes the Form Designer Sub-system (FDS) has a special database for storing data about data, referred to as metadata. The FDS provides many tools, one of the tools is to append a new question. When a question is created, the metadata is updated to store the (index/order) of the questions. The metadata is also updated to store the location of the questions in terms of x and y coordinates on the page. Additionally, the metadata is updated with the width and height of the box of each question. Only then the metadata will have a logical representation of the questions added, and will have a bitmap representation of the question.

Another tool that is provided in the FDS is the opportunity of adding Answer Choice Boxes (ACB). With each ACB appended to the question, the metadata is updated with (x, y) coordinates, (width, height) dimensions, and a bitmap (picture) of the option. Besides the questions and options, the FDS provides other elements like header-text, horizontal lines, tables, and many other items that would be expected from a text-editor/word-processor. The coordinates and dimensions are logical representations of the items in the page. Hence, the metadata has dissected (the parts of the page) the logical information of each page, in the final form of the (questionnaire/survey) sheets. Furthermore, the meta data has sub-images of each item on the page distinguishable from each other.

(A)

Figure (1) Scanned Image of the HCAS

Stage Two: Scanning HCAS Images of Answers

Figure (1) above demonstrates an example of HCAS scanned image, where it contains three questions, and the first ACB of each question is marked.

Stage Three: Subtraction Bitmap

In this stage a matrix subtraction operation is applied on the two bitmaps, as seen in figure (2) below. The first bitmap (A) is the HCAS scanned image, whereas the second bitmap (B) is the unanswered questionnaire, that comes from the metadata. This metadata is composed of locations and dimensions, which is built by the FDS. As it can be seen in figure (2) below, the subtraction bitmap is obtained by applying the matrix subtraction operation (MSO). This MSO is applied on each element of the two matrices, and setting the result into a third matrix. The elements of the bitmap are pixels, which are scalars that represents a gray color value, as seen in figure (2) below.

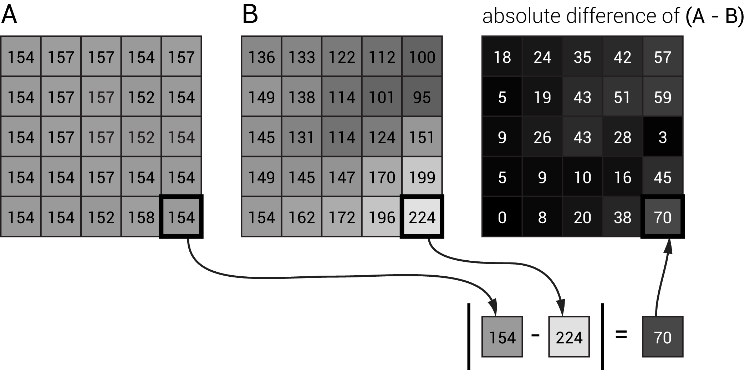


Figure (2) a subtraction bitmap operation, by applying Matrix Subtraction.

(C)

(A)

Figure (3) Two layers

Stage four: Creating Cut-Out Images from the Subtraction

Bitmap.

From the result of stage two (computed the subtraction bitmap), the page would conceptually look like as though it erased what was produced. Therefore, the System leaves only what the user has written, as seen in figure (3) above.

This stage will go through each question to create Cut-out Images from the Subtraction Bitmap (CISB). These CISB’s are the answers that the applicant provided in the HCAS. There are three methods the applicant can answer: 1) Multiple Choice; 2) Hand Writing; 3) Writing in Unexpected Locations.

1. Multiple Choice:

This stage will retrieve from the metadata the numbers of ACB and their locations and dimensions. Then it will create ᶇ number of images for ᶇ number of ACB. The created images are copies from the subtraction bitmap. In addition, the size of the newly created images are the same dimensions as the ACB.

1. Hand Writing:

This stage will retrieve from the metadata the location and dimension of the text box where the user is expected to write in. Hence, creating a new image which is a copy from the subtraction bitmap of the same dimension.

1. Writings in Unexpected Locations.

Another kind of erasing occurs due to cut-outs, is writing in unexpected locations. The system can identify different regions on the sheet. For example, if the writing is on page margins or on the regions of one of the questions. Hence, the system can associate the unexpected writing to a certain question or to a section of the study at a broader level. ( Plamondon, R. ; Srihari, S. N,2002).

Stage Five: Conversion to Vector Graphics

In general, the different strategies in the system rely on the preparation of the cut-out. Part of the preparation is to convert the raster image into a vector-graphics representation. Basically the conversion reduces the pixel data into more recognizable data and higher leveled information. For example, a vertical line written by the user would be presented in vector graphics as a start point of the line and an end point of the line. In this representation it’s easier to find a pattern opposed to matching bitmaps.

For example, a curve would have four points, two points for the start and end, and two points as control points of the curve in the case of a quadratic curve. This stage makes it easier to find a match between the users writing as four points opposed to a grid of pixels. The quality of the conversion to vector-graphics plays an enormous role in pattern matching, and as a consequence in the recognition process. A critical remark is in the case of recognizing connected letters of a word, especially with the Arabic writing fashion.( Badr Al-Badr,B.;Mahmoudb, S.A., 1995). There are almost six levels of heights of the base line. Additionally, there are almost 9 modes how letters are merged together, or how letters can be reformed in shape. With the current processing-power of CPU’s, this should take a few seconds to complete for each word. Because there are a finite number of letters, an algorithm has a finite number of patterns to match. The closest match would be based on an evaluation of the structure of points(start/end/control-points), and ratios of the features of letters. (Lorigo, L. ; Govindaraju, V. , 2005) Rather than a strict resemblance to letters(alphabets) of computer fonts and glyphs. As examples of features, curved letters vs straight line letters, slope lines vs non slope lines, or line intersections in a letter. (Saleem,S.;Cao,H.;Subramanian,K.;Kamali,M.;Prasad,R.;Natarajan,P.,2009),(Al-Hajj,R;Sulem,L.L.;C. Mokbel,C.,2005).

Stage Six: Mapping Graphics into Textual Information

This stage will go through each question to process the cut-outs from stage three. There are two methods in this stage; one for multiple choice; and the other for hand writing. As for multiple choice each group of cut-outs for one question are processed as follows: 1) an average is computed for each cut-out image from the (Width x Height) pixels values. To elaborate more, the average of a blank box will have a low average value, due to most of the pixels in the cut-out image being blank. Whereas the average of a marked box will have a high average value. 2) picking the highest average representing the user`s answer, and from the metadata we will be able to output the code for the choice.

For the hand writing answers, there are 6 strategies, 1) expecting pre-defined words; 2) expecting numbers; 3) expecting a number of words; 4) writings that are not aligned on a horizontal line; 5) highly none-regular strokes, but when taking into account other recognized words in the same text box, can be predicted from a language dictionary; 6) hard to determine the structure of strokes.

Stage Seven: Insertion into Indexed Database.

The results of stage six are in a format that is ready to be stored into the indexed database. Therefore, each answered choice (or text of hand writing) will be stored in the database in its corresponding question, and the PADEH is completed.

Results and Discussions

The developed system did show a remarkable improvement in the data collection stage compared to conventional paper-based data collection. The system proved to be a productive framework for building professional solutions to many frequent issues. The built-in automated calculations of frequency tables and the generation of charts was very impressive and fairly satisfactory to many study teams that have applied their study using this system.

From literature review, the research in the field of vision recognition is a highly specialized research area, and requires going up a very high learning curve. The team has gained experience in the literature, programing skills, and the diverse concepts of computer graphics and disciplines. The team built many software stacks, because in the field of vision recognition, the starting foundational software libraries are complicated to properly link all the dependency components.

Furthermore, the manner in which the FDS built the metadata was relatively simple, and yet had strategic information and data, this metadata puts critical guidelines for the architecture of the automation process, and also plays a critical role in simplifying the endless possibilities of the execution of the algorithms.

Additionally, as far as the software system production goes, stage one was successful, Stage two has many case studies, and the project was successful to establish an implementation, Stage three has a big number of approaches, a few were established, and in the process helped capacity building. Stage four is a very critical software stack, the project choose a narrow feature set in order to proceed with the sequence of the automation process. Stage five built a specification in order to set a scope of work in order to tackle the vast complexity of the field of vision recognition. Stage six only works with a specific set of strategies.

conclusions and recommendations

It has been concluded, that people with numerous experiences in conducting different surveys are able to formulate a formal process. Furthermore, crystalizing the formal process by using custom developed software based on modern IT practices. Hence, for PADEH, the experience of the formal process leads to the need to develop an intelligent algorithm, to recognize pre-defined data schemas.

The field of computer vision recognition is a new and highly specialized area of research, with a very high learning curve. The team has gained experiences in literature review, programing skills, and the diverse conceptual computer concepts and disciplines. Through research, the team was able to create a hierarchy of software libraries. In order to build up the tools that can apply the theories in vision recognition, and executing computer graphics algorithms.

PADEH is dynamic enough to adapt to an ever changing questionnaires designs, each with a different number of pages and a different number of questions and options. unlike OCR software from the market where the user would have to revise the output text, and there after manually enter the textual data into the data base and manually sort out which parts of the outputted text, and manually identify which question fields in the database.

The PADEH would be robust to a certain extent, meaning even when using a camera of smart phone, where the image is inconsistently distorted. In the printing process of the survey the system in the back end will add visual markers on the HCAS for each page, each question, each option, and every other printed element on the paper. These visual markers make the system intelligent enough to recognize the input even with circumstances were the pages are not sorted and damaged to a certain extent. In addition, the PADEH knows exactly where to look, the system has a mental mind map of the layout of the printed page. Furthermore, the PADEH would have a pre-conceived knowledge of the nature of the format of the answers from the user, as hand writing.

Although the project has achieved many accomplishments and research advances. On the other hand, there still remains numerous and enormous challenges. In order to broaden the set of features of the PADEH, it will require further research in the following aspects:

1. The Arabic Hand Writing Recognition.
2. Mapping graphics into textual information
3. Selecting strategies methodology.
4. Practical quality of the strategies for hand writing

Nomenclature

Techno-Economics Division (TED)

Kuwait Institute for Scientific Research (KISR)

Hard Copy Answer Sheet (HCAS)

Form Design Subsystem (FDS).

Optical Character Recognition (OCR).

Form Designer Sub-system (FDS).

Process to Automate the Data Entry of the Hard-copy (PADEH).

Acknowledgments

We would like to thank OWSD for the opportunity to present this paper in their prestigious conference.

References

Mori, S.; Nishida ,H.; Yamada, H., “Optical Character Recognition” John Wiley & Sons, Inc,1999, ISBN:0471308196.

Trier, O.D; Jains,A.K.; Taxt,T. "Feature extraction methods for character recognition-A survey" Pattern Recognition,1996, Volume 29, Issue 4, April 1996, Pages 641–662.

White, J. M.; Rohrer,G. D., "Image Thresholding for Optical Character Recognition and Other Applications Requiring Character Image Extraction” IBM Journal of Research and Development,2010, Volume:27 , Issue: 4, pages : 400-411. ISSN :0018-8646.

Plamondon,R. ; Srihari, S. N." Online and off-line handwriting recognition: a comprehensive survey",IEEE Transactions on Pattern Analysis and Machine Intelligence,2002, Volume:22 , Issue: 1.Pages 63-84,ISSN:0162-8828.

Badr Al-Badr,B.;Mahmoudb, S.A., "Survey and bibliography of Arabic optical text recognition”, Signal Processing,1995, Volume 41, Issue 1, January 1995, Pages 49–77.

Saleem,S.;Cao,H.;Subramanian,K.;Kamali,M.;Prasad,R.;Natarajan,P., "Improvements in BBN's HMM-Based Offline Arabic Handwriting Recognition System", 10th International Conference on Document Analysis and Recognition,2009, Pages: 773 – 777, ISBN: 978-0-7695-3725-2.

Lorigo, L. ; Govindaraju, V. , "Segmentation and Pre-Recognition of Arabic Handwriting" , International Conference on Document Analysis and Recognition, Seoul, Korea, 2005.

Al-Hajj,R; Sulem,L.L.; C. Mokbel,C. "Arabic handwriting recognition using baseline dependant features and hidden markov modeling," International Conference on Document Analysis and Recognition, 2005, pp. 893-897.

1. Bitmap: Two dimensional grid of pixels, in the main Memory. [↑](#footnote-ref-1)