High speed and accurate OMR[1] using any typical scanner
By

Morteza Ghazisaedy*
mghazi@semnan.ac.ir

Hassan Hamzehloo
Monfared**

Abstract:

One of the main factors in data collection and processing is using computers and scanners as manageable and instantaneous tools, to make massive archives. Nowadays, questionnaire forms with multiple choice boxes can only be marked with pencil, and then by using special scanners, the system detects the carbon marked boxes for data processing. The designed system is a software package that prepares a report which is based on scanning a regular questionnaire form and detecting the marked boxes. Its detection is independent of pencil or any other dark writing pen. In addition, the boxes do not need to be filled in entirely, meaning that they can even be marked with an "X" or even a check mark. Every sheet is scanned by a typical fast scanner with an Automatic Feeder Device (AFD) and a minimum 200 (Dot Per Inch) DPI. After scanning the sheets in gray scale level and JPG format, each one will be saved as a picture file into a folder. Then the folder will be extracted and the files will be converted into binary as black and white according to a predefined algorithm. This algorithm opens the folder again and starts searching the marked area's region, if the filling area percentage is over 40%, it will be counted as marked. Another feature of this system is its capability to detect and correct the rotation of the sheet up to 30 degrees. After the correction is made, the sheets will be scanned and the marked locations are saved into the database for further report extraction. This system was tested with more than 10,000 survey sheets filled in by students. The scanned sheets were compared with the originals and the error deviation was less than 0.04 percent which the failure was due to badly (less than 40% filled) marked boxes and jammed paper. Semnan University is currently using this system for evaluating faculty teaching performance.

Keywords:

---

1. Introduction:

Information Technology (IT) has affected our daily lives. Computer and digital systems are the most important tools in business, research, military, industry and education. In the traditional method for surveying, the answer's sheet were marked by pen or pencil and then checking them by a mask answer's sheet[1]. Then preparing reports by extracting data from the corrected answers from the answer's sheet. This method is inaccurate, time consuming, expensive and not easy to save data and archive them. In today's fast-paced information driven society, the need for accurate, timely and cost-effective data collection is very critical [2]. Using computer and optical systems such as scanner and CCD will help to read answer's sheet. OMR is a device that reads answer sheets where marks appear on the paper in predefined areas [3]. The OMR data entry system contains the information to convert the presence or absence of marks into a computer. Some OMR systems need special devices to sense only pencil markers [4]. This paper is about an OMR system which is fully software based using general scanner with an Automatic AFD. The scanned forms are saved into a file and then they will be processed to extract the pre-marked areas. This OMR is fast, accurate and cost effective. At the beginning the questionnaire form format will be explained, then other features of the form and calculating paper orientation output of the scanner will be displayed. Due to the calculations and the questionnaire, marked areas will be found and will be saved into the SQL database. The process is divided into two parts; the first part is scanning and saving the forms into a file and the second part, processes the saved files and extracts the information and records it into a SQL database system. When the database is complete then it is ready for any report.

2. Questionnaire specification forms

The questionnaire form consists of a frame work and all the question mark boxes, questions and all other identification information are inside. Each of the 17 questions stands in a separate row with its multiple choice box in the same row on its left. Each of the questions are capable of being edited or replaced easily. On the left side of the form there are 17 black small boxes, indicating the row position of the multiple choice boxes. There is no indication box on the top and the other boxes on each row will be calculated based on the fixed location from its left and right boxes.

In the first step, a scanner with automatic feeder is suitable for fast scanning; the manual scanners are also acceptable. The scanning format should be in 256 gray scales and JPG format. The scanned saved forms must be filtered with a specific thresh hold value to convert it into binary (black-white). After this process, the gray scale level pixels were plotted and the following histogram in Fig 1 was formed.
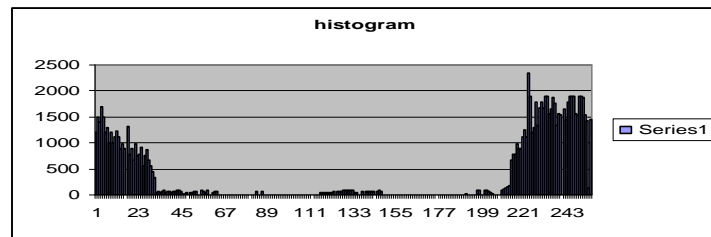
Figure (1): Histogram Gray scale

Based on the above histograms if the color pixel is less than 32, then it is white, if the color pixel is greater than 200 it will be black. The final form is completely black and white.

3. Calculating rotation angle

When the scanner feeder pulls the form it might rotate or miss feed it. Therefore all the incorrect information will be based on the Tang α, the rotation angle, and the following algorithm calculates the rotation angle. In figure 2, a sample form with its original size and the reference points are marked.
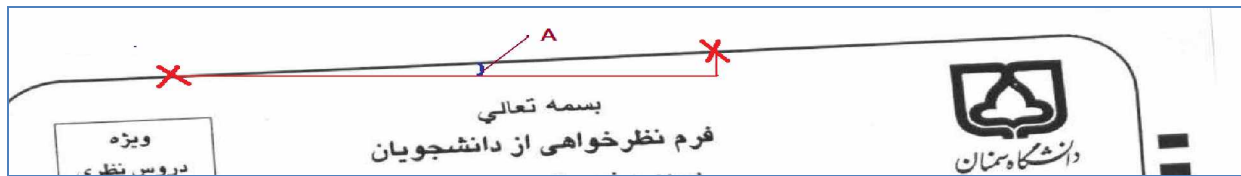

Figure (2): Rotation reference points

Algorithm with 30 pixels from the upper top left starts moving down until one pixel from the top border frame is found. Then it starts searching from location (50, 1) and moves down until the next point from the frame is found, B (50, y_2). Based on trigonometric calculations, the rotation angle is $\alpha = \text{Tang}^{-1} (y2-y1)/50-30)$.

If $\alpha > 0$ the scanned form must be rotated by $-\alpha °$, otherwise it must be rotated by $+\alpha °$, if $\alpha > 10 °$ the rotation is not acceptable and the process is done with 15 pixels one more time. If the result is the same and $\alpha > 10 °$ the form is rotated by 180 ° and the position is calculated again. If α is still out of range the form is rejected.

4. Calculating the central coordinate system:

With respect to point A & B the top line equation is : $(y-y1) = ((y1-y2)/(50-30))(x-x1)$.

Then finding another point from the left border, C (x з, yз), and using α as the rotation angle, the vertical left border equation will be $(y-y3)= ((y3-y1)/(x3-x1))(x-x3)$.   By solving these 2 equations the central point will be calculated as displayed in figure 3.
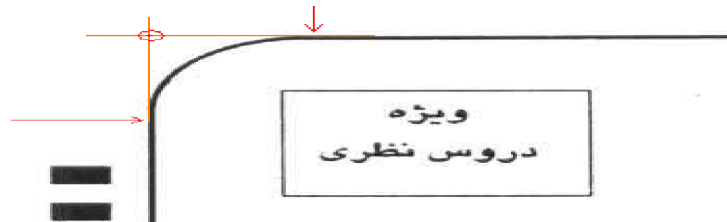

Figure (3): Finding the vertical border

5. Calculating the marked boxes as indicators

Based on step 2 and 3 the cursor must move to the first black box on the left. This box indicates the first row of multiple choice boxes. The location of the other rows are fixed but the rotation or relocation of the form is fully calculated. The new coordination of the first black box is (Sx, Sy) = (x + Wx, y+ Hy), where Wx and Hy are the horizontal and vertical differences from the original coordinate respectively. In order to find the center of the marked boxes we have to add half of the width and length of the box to the center of the first box in (Sx + center x, Sy+ Center y). With respect to the length and width spaces, the rotation angle and the location of all the boxes will be defined.

When the scanned multiple choice boxes are more than 40 % black then it will be counted as marked regardless of the filling shaped boxes. By scanning the forms based of the above algorithm the extracted info will be saved into the SQL database with specific fields for further processing.

6. Conclusions:

 Traditional scanning methods react towards carbon filled multiple choices. These scanners need special hardware devices to read the marked areas. The disadvantages of this method consists of its high cost, non flexibility, not secure for reading some erased boxes as marked boxes and not being a reportable system. The new system is based on any regular scanner and it is free from using any specific filling tools like pen or pencil, so no special hardware is needed. This system is flexible, reliable, printable, cheap, and accurate. Any questions with any chosen answer are definable. The accuracy of the corrected answers are 99.96%. This system is extended to scan any other multiple choice forms in 4 columns and up to 400 multiple choice questions. This system is currently working at Semnan University as instructor's evaluation system by the

students. This form is used for more than 4000 students, 5 course per students, and total 20,000 forms per semester. This OMR is fast, accurate and cost effective. This software is implemented to read any kind of questionnaires form which is predefined in the system and could read up to 400 questions with 4 multiple choice answers for each.

References:

 [1] P Aveyard, S Manaseki and C Griffin, *The cost effectiveness of including pencils and erasers with self completion epidemiological questionnaires*, Public Health, volume 115, Issue 1, January 2001 Pages 80-81

 [2]   Stephan Hussmann and Peter Weipig Deng, *A high speed optical mark reader hardware implementation at low cost using programmable logic*, Real-Time imaging Volume 11, issue 1, February 2005, page 19-30

 [3] http://microlab.uwyo.edu/Documentation/OMR
 [4] http://www.bristol.ac.uk/omr/OMR