Performance Assessment of A Questionnaire Input Software by Computer Vision

Yong-Ren Pu, Su-Hsing Lee and Yu-Tung Dai
Department of Occupational Safety and Health
Chang Jung Christian University
Tainan City 71101, Taiwan (R.O.C.)
yrpu@mail.cjcu.edu.tw

Ying-Che Kuo
Department of Electrical Engineering
National Chin-Yi University of Technology
Taichung 41170, Taiwan (R.O.C.)
kuoyc@ncut.edu.tw

Abstract - Questionnaire survey is one of the important research tools for scholars to conduct quantitative information collections. For achieving fast data input, the authors have developed an automatic questionnaire input software by computer vision. This image processing system captures the questionnaire images by a high-resolution camera, and recognizes the answers inside the checkboxes in a graphical user interface swiftly. This study is designed to realize how this software performs when processing the questionnaires for data input under various conditions. In the experiments, we obtained the accuracies for the software to recognize the checkboxes filled with the check marks in commonly seen colors, as well as the accuracies to recognize the checkboxes answered with four commonly used marks. One-way ANOVA and Scheffe's test were used as the statistical analysis to find the differences between various conditions. The results can provide useful suggestions for future administration of the questionnaire survey.

Index Terms – Questionnaire. Image processing. Statistical analysis.

I. INTRODUCTION

Questionnaire survey has been widely utilized by scholars in many research fields, especially in the areas of public health and social science, to collect quantitative information [1]. In a large-scale survey, researchers collect a huge amount of information through questionnaires. The questionnaire input becomes an important step in the analytical process. For convenience and simplicity, most questionnaires are designed to have many checkboxes so that respondents can answer according to the printed questions. People often leave the checkboxes blank to mean “NO” or fill in with the marks like check, tick, cross, circle or even complete color-in for “YES”. Complete color-ins are often required when read by Optical Mark Read (OMR) scanners and the corresponding questionnaires must be ordered by design and print services. The other marks are often used when filling out regular papers with popular writing instruments and need either researchers to record manually, which is really time-consuming, or some key punch operators trained to key in them one by one in a speedy manner. No matter how it takes through machines or key punch operators, the data input costs much and takes time.

For the purpose of saving the time of questionnaire input, the authors are developing an automatic questionnaire input system by integrating a number of hardware and software modules. Each questionnaire can be captured by a CCD camera, and processed in a graphical user interface by computer vision. The data acquired after image processing are then saved as a text file.

Along with the progress of computer capabilities, the technologies of image processing have been developed for about 50 years [2]. To recognize an image file it is required to segment the image into objects, each of which usually called a region of interest (ROI), containing the important attributes for processing. If the attributes are handwritten, their variations will make it very difficult to recognize. This is obvious and important especially to identify bank checks [3].

In recent years, there are some developments for the method of information input. A US-patented system utilizes a capture device which is a flat panel and is able to digitally capture a respondent’s pen strokes in real time [4]. By printing various shaped checkboxes on different pages, the system may detect each answer’s shape and associate it to the specific page of questionnaires, since the respondent is asked to color in all checkboxes completely. However, it needs to operate on site, which makes it impossible to administer a large amount of questionnaires at one time. Scholars in Taiwan have developed an automatic collection support system by integrating three primary modules, which are actually functions of some existing commercial software [5]. This system can process either electronic files or scanning images. However, it does not perform well in positioning electronic questionnaire drafts and still uses a scanner or a fax machine to acquire scanning image files that may occupy certain storage memories.

The authors have developed an automatic questionnaire input software that is functioned with image capturing, positioning, checkbox labeling/locating, check/cancel recognizing, and text file output. It is capable of executing a speedy data input for those questionnaires answered with various marks and colors [6]. The results therein were primitive since some parameters in the software were set as a result of conjectures by merely the authors’ experiences. To acquire the more precise parameters, the authors performed the calibration of the camera used in the system [7]. From there the optimal values of various focal lengths can be determined and implemented in the software, which is able to amend the skews and aberrations in the images and to locate all checkboxes more accurately. Based on the above, this study is designed to realize how this software performs when filling out a questionnaire using the writing instruments with different colors, as well as using different marks except complete color-ins.
II. SYSTEM FRAMEWORK

The framework of the developing automatic questionnaire input system is shown in Fig. 1. The system is composed of several devices including a CCD camera, an image interface, a computer, a controller and a paper feeder. Unlike the high-priced IO controller used in [6], this system adopts a programmable system on chip instead. This replacement not only reduces the research cost, but also can be integrated into the paper feeder to form a single device.

Before the system goes into the automatic mode, the user needs to place one blank questionnaire for initial positioning that is to locate checkboxes manually and orderly. This can generate the coordinates of all checkboxes which will be applied to all succeeding completed questionnaires. When in automatic operating mode, a cycle of vast data input is as follows. The developed software commands the controller through RS232 to send an output signal to the paper feeder. One page from the stack of questionnaires is then fed forward. The high-resolution CCD overhead continuously grabs images and sends them back to the computer via the image interface. The software subsequently snaps a frozen picture which is composed of pixels in a two-dimensional array that forms a still image. To speed up the processing, the image is segmented into small ROIs, each of which contains a checkbox. Once the image processing for all ROIs is done and stored, it goes to the next cycle.

III. IMAGE PROCESSING

It is noted that the snapped image of each questionnaire undergoes translation, rotation and distortion. During image processing, certain corrections have to be made so that all of the stored coordinates of the checkbox centers from the page of initial positioning can map to those of the following questionnaires. The developed software has to be coded to encompass the techniques that will be introduced in this section.

A. Positioning of Locating Marks

Each questionnaire is a printed page of A4 size. For the purpose of proper affine translation, two L-shaped locating marks are printed on both upper-left and upper-right corners of the page, as shown in Fig. 2. Through image processing of the edges finding as discussed in [6], the inner corners of both locating marks can be determined. The origin is set at the left one where the local coordinate axes are chosen to be horizontal to the right and vertical downward. These two points will later be mapped to the corresponding positions by fisheye correction.

B. Fisheye Correction

Every camera has certain range of view angle that causes an effect called “fisheye” which makes the questionnaire image look distorted, as shown in Fig. 2. To accurately locate the checkboxes, it is necessary to correct this effect. The actual page tends to shrink toward the center O. The longer distance of a point has from the center, the more shrinking effect of that point does towards O. It’s because the fisheye image is a projected image from the actual page onto the visual sphere of radius $H$. Let $P$ is one of checkboxes and $OP$ is on CCD image. The actual location of $P$ on the actual page is denoted by $P'$, so that

$$OP' = OP \cdot \tan \left( \frac{OP}{H} \right) \cdot \frac{OP}{H}.$$  

It is noted that $H$ is also called the shooting range from the CCD lenses to the questionnaire. $H$ can be measured in the unit of length. But when it comes to the CCD image, different set of lenses with different focal length actually makes $H$ vary in the unit of pixels. Since some inherent properties of the lenses are unknown, it is feasible to find $H$ by experiment instead of derivations. The authors have performed the calibration experiment for the camera used in this system as discussed in [7], so that $H$ had been pre-determined and can be set as the operator’s wish in the user interface.
C. Rotation of Coordinates

When each questionnaire is placed and ready for image grabbing, it goes skew inevitably. There are a translation and a rotation between the actual (reference) page and the global plane. Our goal is to locate the corresponding positions of all checkboxes for each questionnaire sent to the distorted computer screen. The positioning of locating marks followed by the fisheye correction for each questionnaire provides both the origin and the inclination angle. The translation can be solved by putting together the origins of each questionnaire and the reference one. The feature of rotation can be realized by the rotation matrix \( \Lambda \) (direction cosine) through the inclination angle \( \theta \) from the reference coordinates to the coordinates of each page. That is

\[
\Lambda = \begin{bmatrix}
\cos \theta & -\sin \theta \\
\sin \theta & \cos \theta
\end{bmatrix}.
\] (2)

This is also the case for processing the first blank questionnaire for initial positioning. If it needs to map all checkbox positions back to each computer image, simply use the inverse of the rotation matrix for each page.

D. Interior Extraction of a Checkbox

To extract the interior area of a checkbox, we first transform the grayscale values of pixels into the binarized ones. Black pixels are categorized as the foregrounds that contain printed and handwritten marks, while white pixels are left as backgrounds. Next, one point inside a checkbox is assigned either by hand in initial positioning or by automatic positioning. Starting from that point, the corresponding ROI is then defined by expanding some pixels along four directions. This assures that the ROI must encompass the specific checkbox. Further, the projections of foreground pixels are made onto both horizontal and vertical axes. The dramatic changes of foreground projections designate those four edges. It is obvious that complete color-ins made by some respondents are not suitable in our case. Finally, the interior area can be extracted to produce a rectangle in which image analysis may take place subsequently.

Fig. 3 shows examples of the extractions from the original ROIs to the checkbox interiors, where the checkboxes are (a) blank, (b) checked and (c) canceled.

The graphical user interface was coded with LabVIEW, as shown in Fig. 4, to follow those procedures and techniques mentioned above. The questionnaire image is on the largest window, in which a frame covering the left locating mark is processed to determine the origin at the inner corner of that mark. The size of this frame can be adjusted in advance in the controls labeled “ROI for locating marks” on the left hand side. The next control below is to set the focal length of the lenses in the unit of pixels determined by former article [7]. The “Initial positions” button initiates the procedure, when a blank questionnaire is in place, to locate the designated checkboxes manually so that they are labeled in order. In this case, the number of chosen checkboxes is 24 which is shown in the indicator named “Label number”. The “Auto recognition” button activates the subsequent procedure to process all ROIs of those 24 checkboxes in the following completed questionnaire, as shown in the lightened checkboxes. Each recognizing result appeared on the labeling number right next to the checkbox is revealed by the specific color according to its Boolean value. Therefore, the operator can monitor the colors to make sure they are correct.

E. Graphical User Interface of the Software

Because of the need for clearer checkbox images, the questionnaire images are captured by an AVT Guppy F146B CCD camera, which has four times of resolutions compared with the regular cameras. These images are continuously transferred to a computer through an NI PCI-8252, IEEE 1394 interface. The computer system with Intel Core 2 Duo CPU E8300, 2833 MHz is utilized to process the questionnaire input in real time.

Fig. 4 The graphical user interface of Automatic Questionnaire Input System.
It is noted that there is a dot inside each lightened checkbox in Fig. 4. It is actually the suppositional center mapped from the reference coordinates of the corresponding checkbox center on the actual page. As long as this dot arrives at the proximity to the genuine center of the current checkbox, it assures that each ROI must contain the checkbox correctly and the software is easy to succeed in extracting the interior.

**F. Information Output**

An extracted interior of a checkbox is firstly segmented into the background and the foreground areas. The latter is composed of all the dark pixels where the written strokes passed by. In order to recognize the marks in the checkbox, the foreground area percentage \( A \) is calculated. A blank interior is defined by the thresholding of \( A < 10\% \). A choosing mark (either check, tick, cross or circle) is determined by \( A \geq 10\% \). If \( A \geq 46\% \), the software goes into the determination of a canceled answer indicating that the respondent has changed his/her mind. For this we assume that there are two or more horizontal strokes overlaying any mark in the checkbox so that the horizontal projections along each row of the interior are applied to calculate the foreground pixel projecting percentages of all rows \( (R_1, R_2, \ldots) \) accumulating onto the vertical axis. A canceled answer is, according to [8], determined by the following criteria,

\[
\begin{align*}
& \text{the largest value of } R_i \text{'s } \geq 85\% , \\
& \text{the second largest value of } R_i \text{'s } \geq 78\% , \\
& \text{the third largest value of } R_i \text{'s } \geq 63\% .
\end{align*}
\]

For those chosen answers that do not match these criteria, the software determines that they are checked and assigns them the Boolean values of TRUE. Others, both blank and canceled answers, are given the FALSE ones.

**IV. EXPERIMENT SETUP**

For this section, we conducted two experiments to assess the performance of the questionnaire input software. The effects of colors and strokes of the marks that the respondents might use upon the recognizing accuracies were investigated. In the first experiment a one page questionnaire containing 75 checkboxes was used. 280 subjects were classified into 7 groups evenly, and each of them filled out this one page questionnaire using different writing instruments, including black, red, purple, blue, light blue, and green rollerball pens, along with 2B pencils. The subjects were told to fill in checks, cancel marks and blanks that were distributed evenly. After image processing, the recognizing data of each questionnaire were saved in a file and compared with the actual results. The accuracy rate \( AR \) of a single questionnaire is defined by

\[
AR = \frac{75 - NF}{75} \times 100\% ,
\]

where \( NF \) is the number of failures in recognizing all the checkboxes, of which there are 75 in a page.

In the second experiment we tested whether the input software was able to recognize all the checkboxes answered with different marks. 160 subjects were classified into 4 groups evenly. Each group of subjects was asked to fill out the aforementioned questionnaires with one type of the specified marks from either check, tick, cross or circle. Fig. 5 shows three of these four examples in extracting the checkbox interiors. The example for recognizing a check mark does not appear in this figure, but was already shown in Fig. 3(b) instead. The accuracy rate for processing each questionnaire is, again, computed by (4).

For the statistical analysis, apart from calculating the mean accuracy and the standard deviation for each group, one-way ANOVA was used to test the overall differences among groups in either experiment. If there was a statistical significance, Scheffe’s test was further performed to find the difference between any two groups.

**V. RESULTS AND DISCUSSION**

In the first experiment, we investigated the effect regarding the colors of strokes upon the recognizing accuracy. The result of each group after processing 40 pages is sorted in Table I. It shows that the accuracy rates of the software recognizing different colors range from 99.63% to 99.77%. The ANOVA test shows no statistical significant difference in using any writing instruments \((F=0.20, p=0.98)\). Therefore, these colors have no effect upon accuracy when processed in the software. Since these are common colors people often use, this suggests that, in future, respondents will have the liberty of choosing their writing instruments. Although all seven groups have high accuracy rates in recognition, as you can see that they are not 100% accurate. The reasons, in viewing all the mistaken checkboxes, were related to the qualities of the questionnaires as well as how the respondents filled in their answers. For example, the printing quality of some questionnaires was so poor that the ink of the checkboxes is too light for the software to identify. Furthermore, the checkboxes might have stains and folds right inside their interiors, which rendered excessive and erroneous foregrounds to be processed. Finally, some marks were almost off the checkboxes due to the negligence of the respondents, so that the software mistook them for the blank cases.

![Fig. 5 Examples of the extractions from the ROIs to the checkbox interiors, where the checkboxes are answered with (a) a tick, (b) a cross and (c) a circle.](image-url)
TABLE I
ACCURACY RATES AND ANALYSIS OF RECOGNIZING QUESTIONNAIRES
COMPLETED WITH THE DESIGNATED WRITING INSTRUMENTS

<table>
<thead>
<tr>
<th>Recognizing pages</th>
<th>Accuracy rate/STD (%)</th>
<th>F/p *</th>
</tr>
</thead>
<tbody>
<tr>
<td>Writing instrument</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Black pen</td>
<td>40 99.63/0.67</td>
<td></td>
</tr>
<tr>
<td>Red pen</td>
<td>40 99.73/0.62</td>
<td></td>
</tr>
<tr>
<td>Purple pen</td>
<td>40 99.77/0.51</td>
<td></td>
</tr>
<tr>
<td>Blue pen</td>
<td>40 99.70/0.56</td>
<td></td>
</tr>
<tr>
<td>Light blue pen</td>
<td>40 99.73/0.81</td>
<td></td>
</tr>
<tr>
<td>Green pen</td>
<td>40 99.67/0.72</td>
<td></td>
</tr>
<tr>
<td>Pencil</td>
<td>40 99.67/0.72</td>
<td></td>
</tr>
</tbody>
</table>

*ANOVA test

In the second experiment, we investigated the effect regarding the written marks upon the recognizing accuracy. The result of each group after processing 40 pages is sorted in Table II. It shows that the accuracy rates of the software that recognized different marks range from 95.70% to 99.93%. The ANOVA test shows that there was difference between at least one group (F=41.57, p<0.0001). Further analysis, using Scheffe’s test, shows that there was no difference between either check and tick or cross and circle, but difference was found between check and cross, check and circle, tick and cross, and tick and circle. This indicates that using ticks and checks to answer the questions had higher accuracies in questionnaire recognition than using crosses and circles. The reason for this is that the latter two marks would have more strokes and foreground pixels appearing inside the checkbox interiors, which might cause the software to mistake few cases for the canceled ones, especially when the strokes were a little thicker. Ticks and checks, however, were simpler in shape for the software to process and identify. Therefore, it will be advisable to have the respondents answer their questionnaires with either ticks or checks for our developing system.

TABLE II
ACCURACY RATES AND ANALYSIS OF RECOGNIZING QUESTIONNAIRES
ANSWERED WITH THE DESIGNATED MARKS

<table>
<thead>
<tr>
<th>Recognizing pages</th>
<th>Accuracy rate/STD (%)</th>
<th>F/p * Scheffe’s test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mark</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(1) Check</td>
<td>40 99.93/0.29</td>
<td>(1) &gt; (3)</td>
</tr>
<tr>
<td>(2) Tick</td>
<td>40 99.90/0.36</td>
<td>(1) &gt; (4)</td>
</tr>
<tr>
<td>(3) Cross</td>
<td>40 96.50/2.71</td>
<td>(2) &gt; (3)</td>
</tr>
<tr>
<td>(4) Circle</td>
<td>40 95.70/3.40</td>
<td>(2) &gt; (4)</td>
</tr>
</tbody>
</table>

*ANOVA test

VI. CONCLUSIONS

In this paper, we introduced the development of the questionnaire input system and some associated techniques. The image processing software therein demonstrated the capability of a fast information input for most of the questionnaires to be answered using checkboxes, regardless of the paper materials and question layouts. For the experiments, we collected plenty of questionnaires answered with commonly used handwritten marks, and with different colors as well, in the checkboxes. The results by statistical analysis showed that the accuracies for the software to recognize the checkboxes in most groups were more than 99%, and in few cases less than that by still more than 95%. It also showed that the colors of marks had no effect upon the recognizing accuracy, but their shapes did. These preliminary results demonstrated that the software functioned in an accurate manner, which has laid the firm groundwork for the success of the automatic questionnaire input system in the near future.

REFERENCES