An OCR based automated method for textual analysis of questionnaires

Devendra K. Tayal

Department of CSE, IGDTUW, Kashmere Gate
New Delhi- 110006
dev_tayal2001@yahoo.com

Sonakshi Vij

Department of CSE, IGDTUW, Kashmere Gate
New Delhi- 110006
sonakshi.vij92@gmail.com

Garima Malik

Department of CSE, IGDTUW, Kashmere Gate
New Delhi- 110006
annu.2353@gmail.com

Amishapriya Singh

Department of CSE, IGDTUW, Kashmere Gate
New Delhi- 110006
amisha.sh22@gmail.com

Abstract

In real world scenarios, most applications rely on questionnaires to capture accurate and informative data. In order to increase the efficiency of questionnaire based feedback evaluation, we propose an automated system to generate a summary of the data collected by the questionnaire through statistical plots. This paper presents a system built on OCR that marks and evaluates questionnaires with minimum human intervention. For the purpose of testing, we prepared a questionnaire to record student feedback for a certain coaching institution and fed the student response collected to our system. The system automatically marked the questionnaires and recorded question-wise score in a database. Once the marking of questionnaires is completed, the system displays the opinion of the class through statistical plots for identification of general opinion. The system has been tested for 20 feedback forms. This platform is targeted towards the institutional authorities that analyze their course structure through forms.

Keywords – OCR, Questionnaire data, Statistical Analysis

1. Introduction

For survey purposes, offline paper-based questionnaire is still considered a reliable and effective method for recording opinions/feedback. Even though computer based surveys are more convenient, it adds some overheads in the form of the need for a portable device and a remote internet access. Hence, whenever there is a need to carry out small-scale surveys, there is an evident inclination towards paper-based questionnaires due to its simplicity. However, analysing these questionnaires to draw patterns pertaining to the received feedback is a cumbersome process for any human.

Optical character recognition or Optical Character Reader (OCR) describes a system that performs the mechanical or electronic conversion of images. They are capable of converting typed and handwritten texts into "machine-encoded text". The source of the document fed to the OCR includes scanned document or image of document or even a scene-photo or text superimposed on an image. OCR finds a variety of applications and is a widely used form of information entry owing to its ability to digitize printed documents. This covers printed data records, invoices, cheques, bank statements, computer generated receipts, postal mail, documentations, printouts, business cards, etc. These digitized documents can then easily be electronically edited, checked, copied, displayed, searched. It can also be used in machine processes such as data mining, keyword extraction, summarizations, cognitive computing, automatic license plate recognition, text-to-speech etc. In pattern recognition, AI and computer vision, OCR has been tagged as a field of research. Most of the advanced OCRs are capable of giving a high degree of recognition accuracy for most of the fonts independent of the image format. Optical Mark Recognition (OMR) is similar to OCR with the exception that an OMR does not require a

pattern recognition engine. They recognize human marked data from documents and work with a dedicated scanner device. This scanner reads input by shining a beam of light onto the paper. The marked areas are detected by the contrasting reflectivity at predetermined positions on the page as the marked positions reflect less light as compared to the blank areas. But OMR sheets are comparatively costly and hence in this paper we have utilized the OCR technology in order to evaluate students form in a better and cost effective manner.

This paper tries to bridge the gap between the simplicity of paper-based questionnaire surveys and the arduous task of evaluating them by proposing an automated system that uses OCR to digitize the feedback and provide suitable data for online statistical computation and assessment. The analysis performed on the survey data for retrieving student opinion include bar plotting, histogram generation and heat map distribution. Patterns collected from these plots have been combined to generate an informative review of student opinion about the coaching institution. The proposed platform will be beneficial for those organizations that collect data through offline paper-based modes such as surveys, feedback forms etc. Moreover, in the field of education, it can be applied for evaluating faculty performance forms and lead to automation of all marking and evaluation processes altogether, thereby decreasing man power and time consumption.

The rest of the paper has been structured as follows: section 2 summarizes the related work, section 3 describes our proposed work, section 4 shows the implementation of the work, results and discussions are covered in section 5 and finally, section 6 concludes the work and underscores the future scope.

2. Related Work

The use of OCR for questionnaire handwritten document evaluation has been an active area of research. In this section the literature work related to existing questionnaire evaluation systems have been discussed.

Alvaro *et al.* developed a software application using an OCR as an evaluator to aid students in the age group of 4 to 6 in learning how to write with the help of stylus and touch screen monitor of computers [1]. The letters traced out by children on their tablets were evaluated by an OCR to determine the correctness of the letters. The conducted survey showed that the proposed software provided a reliable, functional and user friendly alternative to teaching children basic handwriting. P. Sanguansat proposed a system for automated data entry using Optical Mark Recognition (OMR) [2]. In it a method has been proposed for preparing questionnaire consisting of only close-ended questions as well as a provision is there for including proper report of the output that can be opened and edited in spreadsheet software. The experimental results included in the paper show that the proposed system has accuracy as high as 93.36% for choice selection by given three marker patterns.

T. Idé *et al.* addresses informative prediction for ordinal questionnaire data [3]. The paper quantitatively evaluated the informativeness of questions in the questionnaire by leveraging the predictability of individual samples' outcome. By extending an existing theory in psychometrics, they presented a framework based on outcome-aware item response theory, also known as oIRT, thereby proposing two new ideas.

- In order to include multiple states, extend the prior distribution for latent variable.
- To improve *k*-NN prediction, define Riemannian metric based on outcome-aware item response theory.

For collecting and analyzing survey responses, K.Y.Yigzaw *et al.* proposes a privacy-preserving method [4]. Based on a semi-honest adversarial model, the method is highly secure. It performs statistical computations using developed secure protocols. The results show that the method can be applied to tasks involving categorical data collection. The proposed method has been verified as efficient and scalable and available for use in practical scenarios. Y. Ohira *et al.* outlines the development of a web-based survey system that allows the general public, those having no knowledge of HTML and CSS to create surveys equipped with graphs and other visualization methods and confirm response rates [5]. Additionally, it has the provision to limit answers to a certain audience. A case study of the system reflects that it is easy to use and can effectively underscore the opinion of different factions of the survey-takers.

In [6], a novel natural language based system has been proposed for short-answer free-text authoring and marking. The system generates a detailed feedback on incorrect/incomplete answers after online question attempt submission. The deployment of the system in the real world demonstrated answer matching with similar or greater accuracy as compared to human assessment. Eftimie *et al.* in [7] proposed an automated system called Korect for test generation, processing and grading of paper quizzes. Using auto feed scanners and Optical Character Recognition (OCR) detection, it provided an inline answer marks and mark sheet processing. It minimizes human intervention in quiz grading to automating the entire process. In [8], Ghogare *et al.* presented an interactive approach for evaluation of paper based documents, books etc. Performance evaluation is carried out based on keywords on an electronic device using an adaptive predefined database. The system finds

application in query terms, error-grams etc. This system attempted to reduce the time consumption and man power associated with the process of evaluation.

[9] described a software system for automatic evaluation of free-text answers generated in response to open ended questions called AutoMark. The software is based on the concept of Information Extraction techniques. It takes into account spelling errors, typos, syntactical and semantic-based errors as well when marking responses. The system uses mark scheme template to look for specific content in the answer. The paper also exemplifies the system using blind experiment and moderation experiment.

The literature survey shows that feedback recording by the use of paper-based questionnaires and OCR for reading textual images has been extensively researched in the recent years. However, the two objects have not been combined to generate a system to perform automated questionnaire evaluation. In this paper, we have described the designing of such a system.

3. Proposed Work

This section presents the proposed work of this paper in the form of an automated system that will evaluate the questionnaires without the need for human supervision. It will generate an electronic copy of the questionnaire, read and store the answers in a database and perform analysis on this data to present truly interesting trends. Figure 1 shows the execution of the system.

The system takes in hard copy of the questionnaires marked with the feedback. The questionnaire is scanned and a digital copy of the same is generated for evaluation. This digital copy is taken for the evaluation process. The evaluation process identifies the marks made by the person as his/her response and assigns a score to each question based on the appropriateness of the corresponding answer. The marks corresponding to the questionnaires are recorded in the database. Once the marking process is complete, the evaluation of the response begins. This involves generation of statistical data to provide a general opinion captured by the feedback recorded in the questionnaire. The system automatically generates a bar plot, histogram and heat map of the feedback and displays it post evaluation. All of these steps are performed by the system automatically with minimum human intervention.

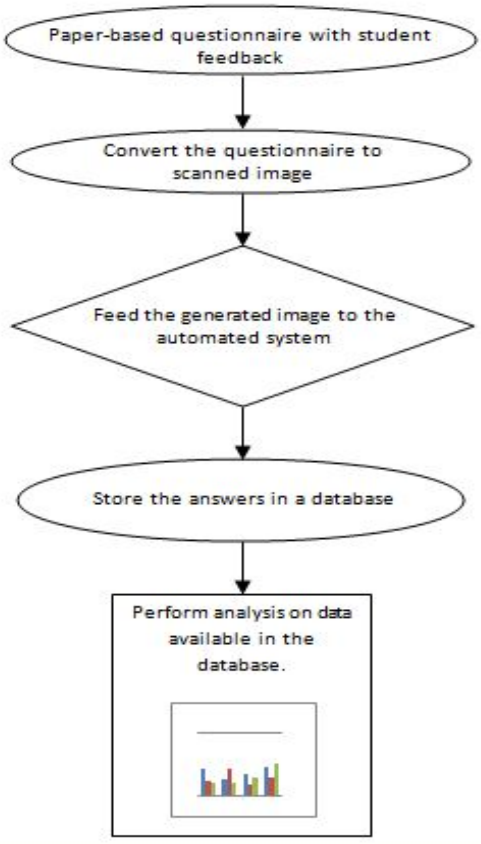
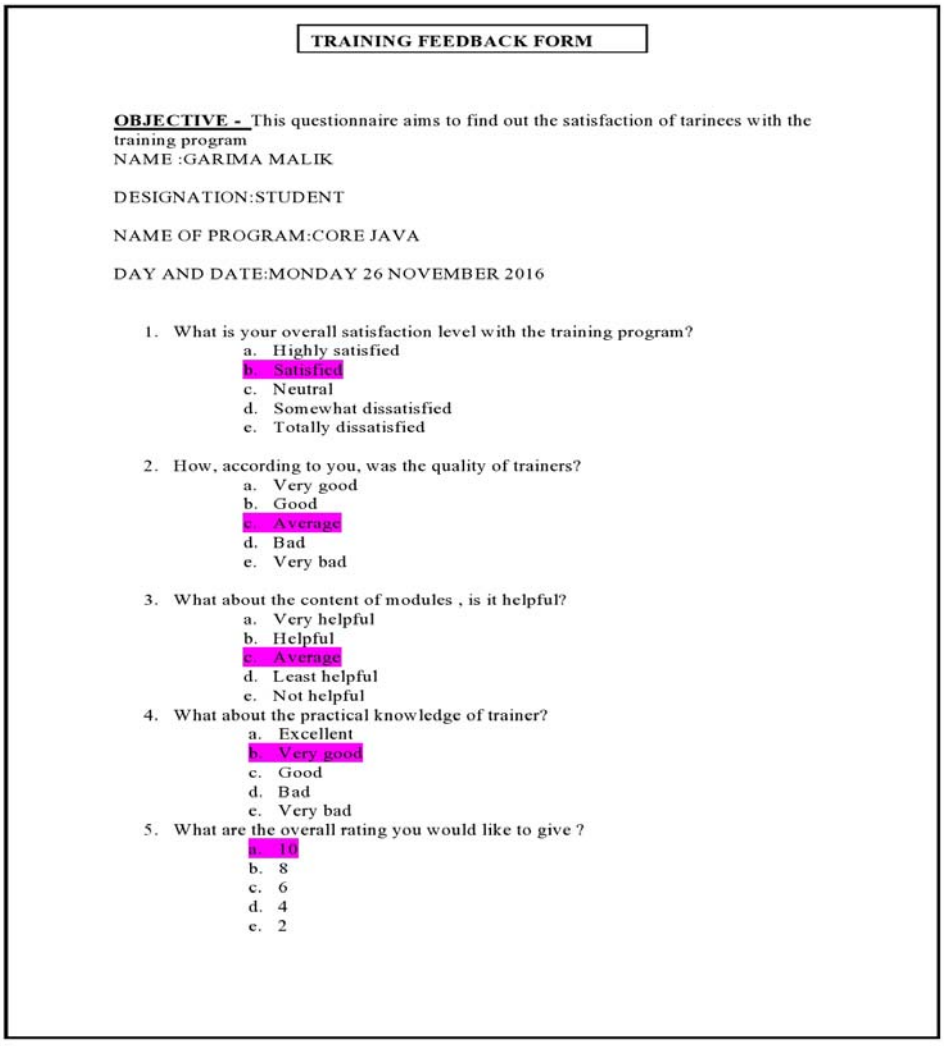


Fig. 1: Execution of the automated system

The user is only required to provide the system with the scanned images of the questionnaires. The system will automatically read the answers in the questionnaire including text type answers as well as MCQ. This data will be converted into an electronic copy and will be immediately available in the database. Performing statistical computation on the available data is also the work of the system. Different plots are generated to allow the users to visualize the data obtained from the questionnaire. Summaries and patterns are then retrieved from the answers read to assess the general opinion of the survey-takers, in our case the students.

4. Implementation

This section presents the implementation details and strategies for the proposed system in order to provide an insight into the working of the automated system. For the purpose of conducting the survey, we have designed a feedback form to act as the input questionnaire to the system. This form consists of 5 multiple choice type questions for evaluating the training feedback. The students were asked to mark the option (with the help of a highlighter) that they most related to. A sample questionnaire is given in Figure 2.



TRAINING FEEDBACK FORM

OBJECTIVE - This questionnaire aims to find out the satisfaction of tarinees with the training program
NAME :GARIMA MALIK

DESIGNATION:STUDENT

NAME OF PROGRAM:CORE JAVA

DAY AND DATE:MONDAY 26 NOVEMBER 2016

1. What is your overall satisfaction level with the training program?
   a. Highly satisfied
   b. Satisfied
   c. Neutral
   d. Somewhat dissatisfied
   e. Totally dissatisfied
2. How, according to you, was the quality of trainers?
   a. Very good
   b. Good
   c. Average
   d. Bad
   e. Very bad
3. What about the content of modules , is it helpful?
   a. Very helpful
   b. Helpful
   c. Average
   d. Least helpful
   e. Not helpful
4. What about the practical knowledge of trainer?
   a. Excellent
   b. Very good
   c. Good
   d. Bad
   e. Very bad
5. What are the overall rating you would like to give ?
   a. 10
   b. 8
   c. 6
   d. 4
   e. 2

Fig. 2: Sample questionnaire

The OCR in the system reads the user input and stores the data in a database. While testing our system, we provided as input, 20 training feedback forms collected after recording student reviews. The student highlights the answer and the answer that was not readable by the OCR due to uniform color change was selected as the answer. For instance if the student marks “option b” as the answer then this text won’t be recognized by the system but since the system has an original copy of the feedback-form hence it would compare it to it to see that “option a” and “option c” are recognized along with other options but not “option b” and hence it would be the option that was selected. Each question is marked on a scale of 10 to 2. The most positive response is marked as 10 while the worst is rated as 2. The options are marked with the interval of 2 starting from 10. The system was

able to successfully identify each of the answers for all the provided forms with 100% accuracy. Based on the cumulative score given by the 20 students, the system assigned a grade to all the questionnaires. Table 1 contains the value read by the system and stored in the database.

Table 1: Database for 20 forms

S. No	Q1	Q2	Q3	Q4	Q5	Total	Grade
1	8	10	8	10	8	44	A
2	8	8	6	8	6	36	B
3	10	10	8	8	8	44	A
4	8	6	8	6	8	36	B
5	6	4	4	4	6	24	C
6	10	8	10	8	10	46	A
7	6	8	6	6	8	34	B
8	8	6	8	8	8	38	B
9	10	6	8	8	10	42	A
10	6	8	8	10	10	42	A
11	10	10	6	8	8	42	A
12	8	10	6	10	8	42	A
13	4	6	4	4	6	24	C
14	10	8	8	10	10	46	A
15	8	8	10	8	8	42	A
16	8	10	8	10	8	44	A
17	6	4	2	4	4	20	D
18	4	4	2	4	4	18	D
19	2	2	2	2	2	10	E
20	4	4	2	4	4	18	D

The graphical representation allows for better visualization and provides an overall picture of the result of the survey. The plots rate the question based on the assigned grade. In this manner, one quick look at the plots can tell us the areas that received appreciation and those that very criticized by the students.

5. Results and Discussion

This section provides the graphical plots returned by the system on the basis of its survey evaluation. Figure 3 shows the bar plot, Figure 4 depicts the histogram and heat map is given in Figure 5.

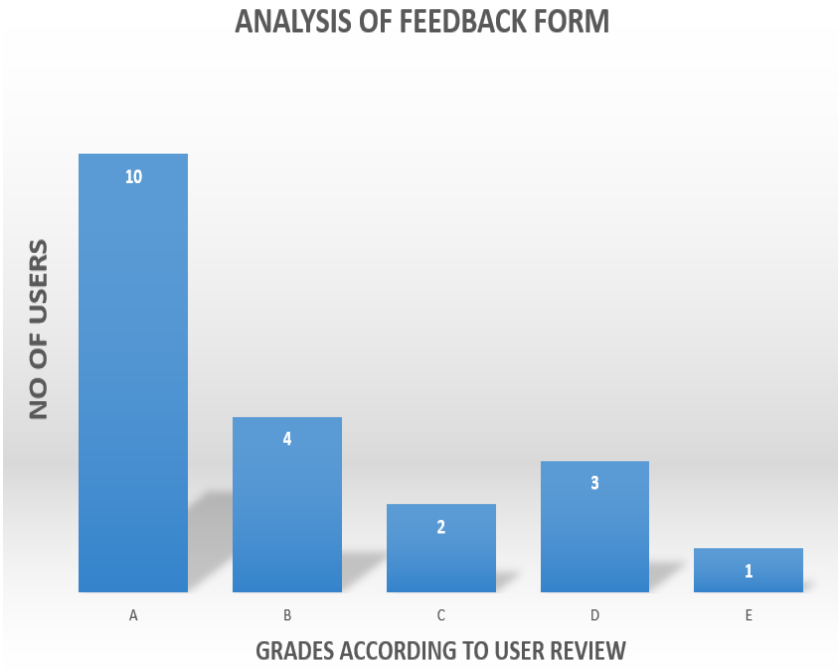


Fig. 3: Bar Plot returned by the system

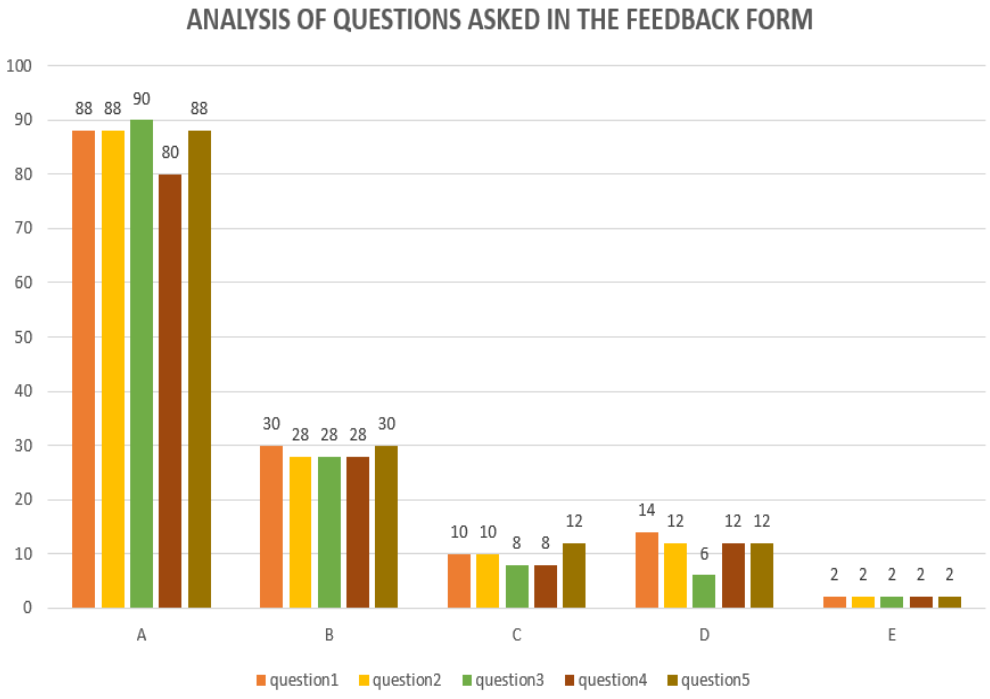


Fig. 4: Question-wise histogram returned by the system

According to Figure 3, 10 students gave the course an overall grade of A, 4 gave B, 2 students gave it a C, 3 for D and 1 student gave the course a grade of E. Figure 4 shows the question-wise analysis of feedback form. It groups cumulative question scores on the basis of grades. For example, Question 1 received a total score of 88 among the questionnaires that received a final grade of A.

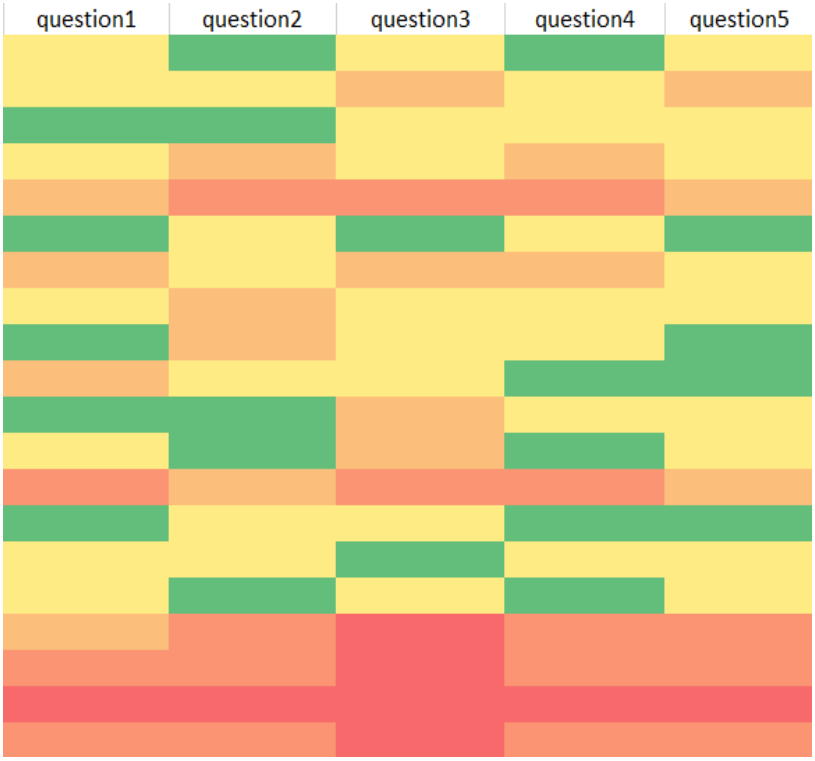


Fig. 5: Heat Map returned by the system

Figure 5 is the heat map returned by the system that analyses every question of the 20 forms under evaluation .The default color gradient sets the lowest value in the heat map to Red color representing a score of 2, the highest value to a Green with an associated score of 10, and mid-range values to Yellow, with a corresponding transition (or gradient) between these extremes. Through this figure we can analyze the weak zones, in our case form numbers 16 to 20. These forms lie in the red region which signifies that all the questions corresponding to these forms got low scores.

The statistical results returned by the system summarizes the survey to a great extent and gives an accurate evaluation of the feedback provided giving clear demarcations between areas that have excelled in public opinion and those that have disappointed. Hence, it saves the time for having to peruse the questionnaires manually and identify these areas and uncover the hidden patters.

6. Conclusion and Future Scope

In this paper an automated method for questionnaire evaluation has been proposed using optical character recognition. The results are analyzed for 20 forms collected from students at a coaching institute. The statistical results for the same are obtained which includes the bar plot for user grading, histogram for question wise analysis and a heat map denoting different scores. The user can interpret using these visualizations that what is the overall assessment of the forms and what are the strong or weak areas for the institute. Till date all this work is done manually and no automated process is suggested for the same. The proposed system is an attempt to make the questionnaire evaluation system a smooth process with minimum human intervention.

This system will directly benefit those organizations and people that still rely on paper-based modes for data collection. This includes surveys, feedback, MCQ tests etc. The system can be used in remote areas where online review recording is not a viable option such as villages. Moreover, education institutions can use this system for faculty feedback evaluation, MCQ answer-sheet evaluations etc. This method will eliminate manual grading of questionnaires to recognize opinions of test-takers.

The system can be extended in the future to provide more extensive statistical evaluation of the collected data. This may include pie-charts, bubble plots etc. Additionally, the system can be further developed to include other types of feedback that do not involve marking of response but takes subjective answers.

REFERENCES

[1] Alvaro, A. K. S., Cruz, R. L. D. D., Fonseca, D. M. T., & Samonte, M. J. C. (2010, June). Basic handwriting instructor for kids using OCR as an Evaluator. In 2010 International Conference on Networking and Information Technology (pp. 265-268). IEEE.

[2] Sanguansat, P. (2015, June). Robust and low-cost Optical Mark Recognition for automated data entry. In Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology (ECTI-CON), 2015 12th International Conference on (pp. 1-5). IEEE.
[3] Idé, T., & Dhurandhar, A. (2015, November). Informative prediction based on ordinal questionnaire data. In Data Mining (ICDM), 2015 IEEE International Conference on (pp. 191-200). IEEE.
[4] Yigzaw, K. Y., Michalas, A., & Bellika, J. G. (2016). Secure and scalable statistical computation of questionnaire data in r. IEEE Access, 4, 4635-4645.
[5] Ohira, Y., Ogashiwa, K., Muranaga, S., Matsumoto, T., & Naitoh, H. (2016, July). A Development of a Questionnaire System for Institutional Research. In Advanced Applied Informatics (IIAI-AAI), 2016 5th IIAI International Congress on (pp. 505-508). IEEE.
[6] Jordan, S., & Mitchell, T. (2009). e-Assessment for learning? The potential of short-answer free-text questions with tailored feedback. British Journal of Educational Technology, 40(2), 371-385.
[7] Eftimie, I. A., Bardac, M., & Rughiniş, R. (2011, June). Optimizing the workflow for automatic quiz evaluation. In Roedunet International Conference (RoEduNet), 2011 10th (pp. 1-4). IEEE.
[8] Ghogare, S., Mahajan, C., & Mulay, P. (2015, October). Automation related to professor evaluation. In Applied and Theoretical Computing and Communication Technology (iCATccT), 2015 International Conference on (pp. 579-582). IEEE.
[9] Mitchell, T., Russell, T., Broomhead, P., & Aldridge, N. (2002). Towards robust computerised marking of free-text responses.