Data Extraction from Exam Answer Sheets using OCR (Optical Character Recognition)

Mohammad Faaiz
KIET Group of Institutions,
Ghaiziabad, India

Samiksha Gupta
KIET Group of Institutions,
Ghaiziabad, India

Shreya Sharma
KIET Group of Institutions,
Ghaiziabad, India

Abstract- Manual Data Collection from a student's exam- sheets is always a tedious job which exacts ample amount of time and effort. This paper has suggested a novel approach for developing an automatic, adaptive, fast and reliable system capable of recognizing enrolment number and corresponding marks of student from answer sheet and storing it in the host computer. This system consist a hardware which picks out sheets one by one from a bundle and captures image of the front page of each answer script. This image is processed by proposed robust extraction and noise removal algorithm adaptive to environmental conditions. It is then passed through Optical Character Recognition (OCR) system which extracts characters using correlation. Accuracy of system depends on the sample space size of OCR system. In our experiment we have archived average 81 % accuracy in various light and paper (Exam-Sheet) condition. We had trained the OCR with 50 samples of numerals set (0-9). In this way developed system will not only replace the traditional tiring way of manual writing of marks in database but in addition can calculate average marks of all students, ranges of marks for assigning different grade and provide grade for each student automatically

Keywords- Application of Image processing, Optical character Recognition (OCR), Statistical lmage Processing

I. INTRODUCTION

Almost every college and university needs to maintain student marks in huge numbers. Storing these accurately and efficiently is paramount to evaluation of student grades etc. Manual handling of such tasks often leads to irregularities and discrepancies. An automated system capable of performing these tasks in a quicker, more efficient manner is discussed in this paper. To make the software user friendly, an interactive GUI has been prepared. Hardware also helps user in performing task with the help of fed voice instructions.

The generated Excel sheet with student's enrolment number and marks can be updated on the university' s server automatically. The hardware system and software application ' Exam paper reader (EPR) , implementing Optical Character Recognition (OCR) system will help teachers save a lot of time in carrying out the documentation and updating marks of each student [1] [2] . Updating server automatically can be done through the same GUI interface having the different tab for updating attendance. The applications are not limited to college, but with some modification same system can be used in post offices for separating letters according to zip code.

A template has been suggested for the front page of the answer script so as to make acquisition of the region of interest from the page more easy and reliable. Exam paper reader converts hardcopy of papers to soft copy with help of a web-camera.

The scanning is done by VGA camera mounted on hardware and not by traditional scanner to reduce image size and increase processing time. However, low quality of camera provides a lot of noise. The image is pre-processed which result in sufficient clarity or quality by smoothening filters.

Adaptive Threshold for Color Detection (ATCD) algorithm is used to extract the region of interest i.e. roll number and marks of student. The region of interest is passed to OCR which will convert it into machine-editable text which is updated to the documents.

The paper discusses the ATCD algorithm, a novel approach to identifY the color in different environmental conditions. This is done by defming an acceptable noise sphere in HSI coordinate system. The center of the sphere is at the pure color' s coordinate. The varying radius with environmental conditions brings the adaptive nature.

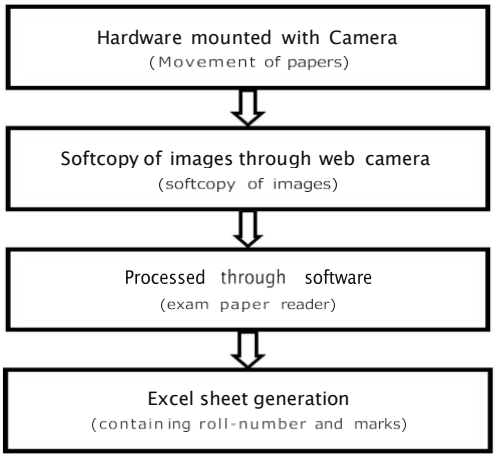

Fig. I. System architecture of exam paper reader

II. HARDWARE DESCRIPTION

The Hardware for EPR (Exam Paper Reader) processes the papers one by one. It has both rack & pin mechanism and Conveyor belt system. It consists of sharp sensors to ensure proper placement of paper on the conveyor belt and at exact position. The proper positioning of paper is necessary so that softcopy of photo must contain boxes having roll number and marks accurately. The images are taken by the webcam mounted above the conveyor belt .To maintain constant intensity of light the whole device is enclosed by a box with a light placed inside it. In order to make machine more interactive, a voice module APR9600 which can record and play the messages accordingly to provide instructions to user has been added [3].

Finally interaction between webcam and hardware is of utmost important. This synchronization is attained by serial communication between microcontroller and computer. The hardware is synchronized with processing unit for capturing image of front page of each exam sheet.

III. EPR IMAGE PROCESSING

A front page template for the answer scripts has been proposed for easy extraction of marks and roll number from answer script as shown in figure 2. Boxes for Roll Number and Marks have a red boundary.

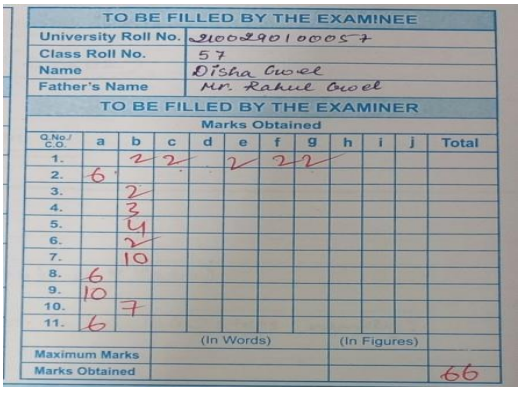

Fig. 2. Proposed template for front page of answer script Image processing algorithm for development of EPR

A. *Preprocessing a/Camera Images*

We use cameras that are affordable and easily available so as to simulate real-time scenarios. Such cameras are of low resolution and poor quality which induce random noise. Such noise is majorly pixels of abrupt color intensities. This is usually reduced by smoothening in RGB space [4], which brings a trade-off in sharpness of the image. We propose the adaptive threshold algorithm to reduce this trade-off by forming a sphere of 'acceptable intensities' - 'noise sphere' on HSI coordinates. The radius of this sphere is adaptive to the environmental conditions.

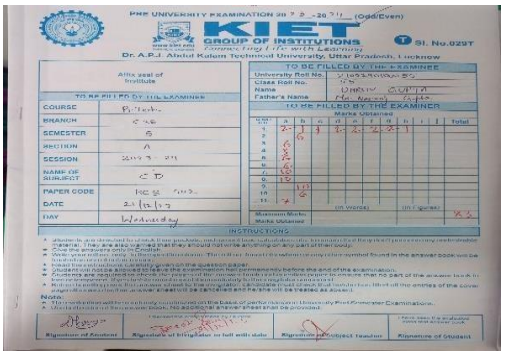

Fig. 3. Real time image of answer script

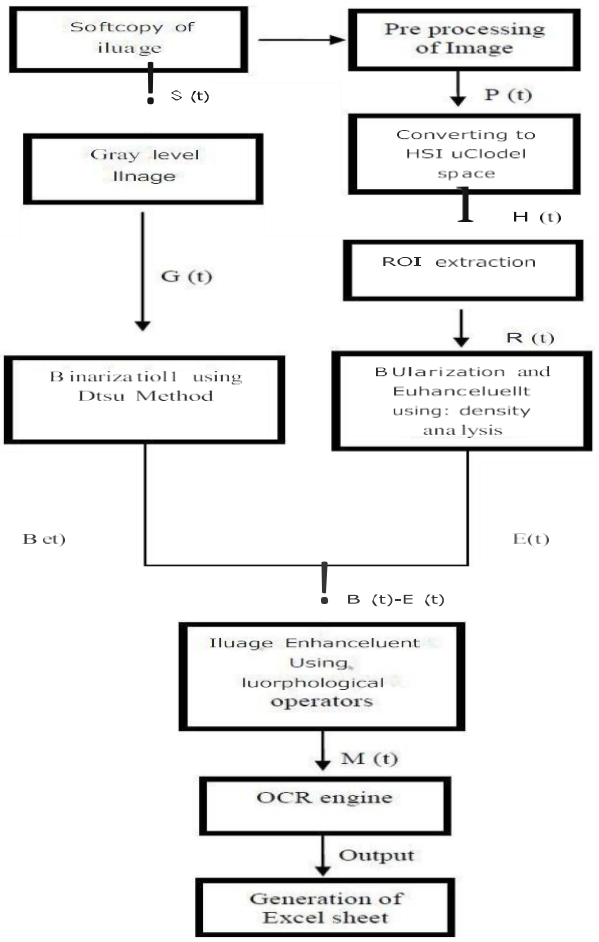

Fig. 4. EPR image processing algorithm

B. *Adaptive Threshold/or Color Detection (ATCD) Algorithm*

Our aim is to segment the box in which the characters are written. The algorithm is based on the principle that a segment in image will occupy same number of pixels in any environmental condition provided the distance between the camera and object is constant. In controlled environment, we calculate the actual number of pixels that are occupied by the 'box' in the image - P0.

A trivial method is to form a set of values that are acceptable as the color of the box; identifying that color, we fmd the corner co-ordinates. With distance between the camera and paper always constant, we resize this sphere to accommodate the previously calculated number of pixels.

Since the system should be adaptive to different light intensities, paper and printing quality, it cannot depend on some fixed threshold for the extraction of Enrolment or Marks Box. The algorithm is made such that it can handle all such irregularities efficiently. The main aim is to recognize corner coordinates ofthe box.

The transformation from the RGB space to the HSI space is performed with the already established formulas (R, G and B are in the range 0-1)

$$I = \frac{R+G+B}{3} \tag{1}$$

$$5 = 1 - \frac{3}{R+G+B} \times \min(R,G,B) \tag{2}$$

$$H = \arccos\left(\frac{\frac{1}{2}\times[(R-G)+(R-B)]}{\sqrt{(R-G)^2+(R-B)(G-B)}}\right) \tag{3}$$

$$if\ B > G, then\ H = 2n - H$$

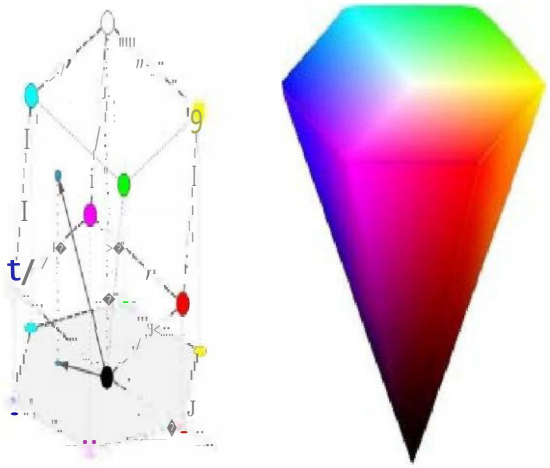

Fig. 5. Conversion of RGB space to HSI space

HSI space is conventionally visualized as an inverted cone or hexagonal pyramid, in which the vertical axis represents intensity, and the angle of divergence from the vertical axis represents saturation (Figure 5).

The main motive for transforming the image from RGB to HSI space is the fact that red color can be extracted efficiently. Also the mathematical representation is much easier than in the RGB space.

In Fig. 6, 0° represents center of pure red where H=O, S=255, I= 125 .We denote del(h), del(s), del(i) the variance in hue, saturation and intensity values respectively. This defines the area that bounds permitted red color.

We have assumed a spherical space for defining the variances, naming it noise sphere. The radius r is given by Equation (4)

$$r = del(h) + del(s) + del(i) \tag{4}$$

Noise sphere can be defmed as in Equation (5).

$$(x - 255)^2 + y^2 + zz = del(s)^2 + 255 \times \sin z\left(del(h)\right) + del(i)z \tag{5}$$

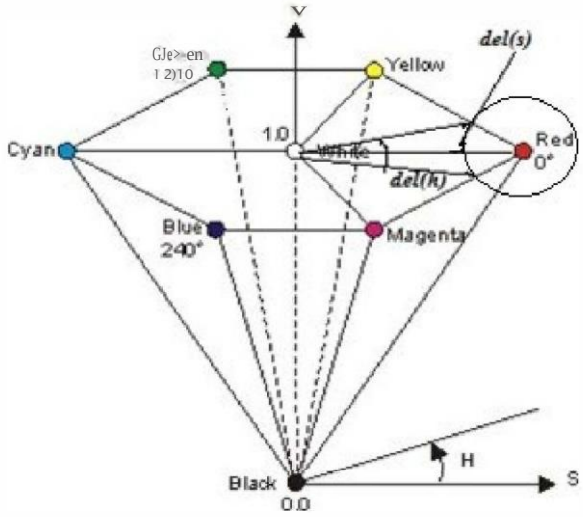

Fig. 6. Description of change in value of H,S,1

Let the number of red pixels in the captured box be p. The radius 'r' is varied to accommodate as many pixels as P0.
This variability of the radius of noise sphere brings the adaptive nature to the system.
Once the red color is defined, corner points of the red box can be found out. The region of interest (ROI) is extracted using these corner coordinates.

Fig. 7. ROl extraction using extreme coordinate

Fig. 8. Red color extraction using proposed algorithm with radius of noise sphere as 0.59

C. *Binarization andImage Enhancement*

This ROI R(t) is then converted into binary image so that time taken for processing decreases [5] .The output is shown in Fig. 9. The image is then enhanced using density analysis of the binarized image. We then calculate number of black pixel in both horizontal and vertical direction at each pixel. If any white pixel laying at any point where vertical or horizontal black density is larger than average indicate this point must be black similarly black point laying in low vertical and horizontal density coordinate, is changed into white. Density analysis The output E(t) thus obtained is clear binary image of outer box. The result is shown in Fig lO.

Fig. 9. Binar image with irregularities

Fig. 10. Binary image after noise and irregularities removal using density analysis.

D. Binarization ofOriginal Image Using Otsu Method:

Meanwhile, one more process is carried out in parallel to the whole above process for extraction of only box from the original image. In this process the original image is first converted into gray scale using following transformation as :

$$Y = 0.3 *R + 0.59* G + 0.11 *B \quad (6)$$

Here the out image is the result of 30% of red component, 59% of green and 11 % of blue. The gray scale image obtained is then binarized using Otsu algorithm [6] .

Fig. 11. Binarization of the segment image

The image obtained B(t) using Otsu algorithm is subtracted from E(t) to obtain image having only characters and some noise due to difference in Box shape of E(t) and R(t).The noise is then removed using Morphological operators open and close respectively which leaves out some spots. These spots are removed by Hit and miss transform with a variant structuring element as shown below:

[1 1 1 1 1 1 1 1 1 1
I x x x x x x x x i
l x x x x x x x x i
l x x x x O x x x l
l x x x x x x x i
l x x x x x x x i
1 1 1 1 1 1 1 1 1 1]

Here x can be any value in the mask and we are capable of removing noise which we are not possible through median filtering. The image obtained after this noise removal is skeletonized, so that recognition of text is easy from image using OCR engine.

Fig. 12. After subtration of outer boundry by doing B(t)-E(t)

Fig. 13. After process through standed skeletanization and hit and miss transform using above structuring element

IV. OPTICAL CHARACTER RECOGNITION

The goal of optical character recognition (OCR) [1] [2] is to classify optical patterns (often contained in a digital image) corresponding to alphanumeric or other characters. The process of OCR involves several steps including segmentation, feature extraction, and classification. The template matching technique is used for recognition of characters. Database of about 5000 numerals is stored to carry out the correlation process. The final result of OCR engine designed using steps shown in figure is editable text.

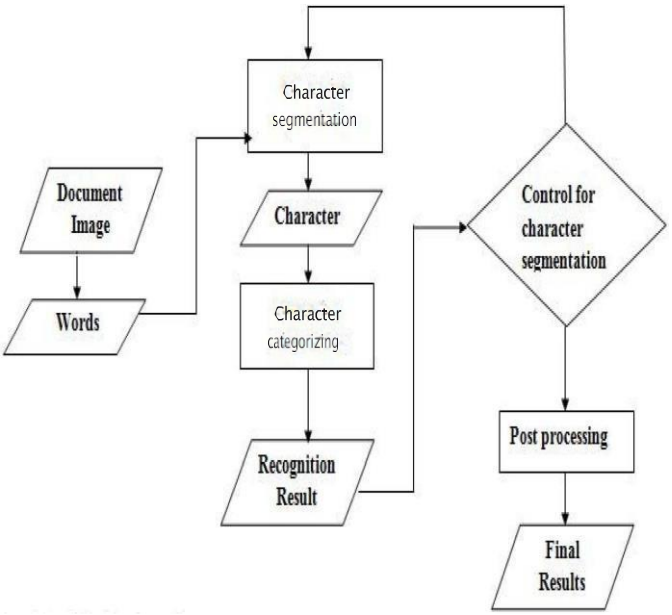

Fig. 14. ODCR algorithm

V. RESULTS

Accuracy of system depends on the sample space size of OCR system. In our experiment we have archived average 81 % accuracy in various light and paper (Exam-Sheet) condition. We had trained the OCR with 50 samples of numerals set (0-9).

VI. CONCLUSION

In this paper, we have discussed automated exam marks entry system in excel sheet format from hardcopy of student papers. The image processing technique used in Exam paper reader is adaptive to different noise parameter and gives accurate results for prediction of numbers in roll number and marks using OCR .OCR algorithm has been discussed and correlation is used for template matching to give accurate result . OCR is one of the most emerging technologies and its reliability is continually improving. Soon OCR will become a powerful tool for data entry applications which will lead to automated data entry by OCR thus reducing labor. Incorporating OCR will be an attractive feature of any Data Entry System. However in past due to limited availability of a capital and short environment was restricting the growth of this technology, but today more and more enterprises are working on this technology and that will definitely lead to 100% accuracy in this technology thus making the dream of paperless world true.

REFERENCES

[I] Cuhadar A., "Scalable parallel processing design for real time handwritten OCR: Pattern Recognition", *Signal Processing, Proceedings ofthe 12th iAPR International Conference,* vol. 3, 1994.

[2] Kameshiro T., Hirano T., Okada Y., Yoda F., "A document retrieval method from handwritten characters based on OCR and character shape information", *Sixth international Conference on Document Analysis and Recognition,* 2001.

[3] Bhatlawande S., Mukhopadhyay J . , Mahadevappa M., "Ultrasonic Spectacles and Waist-belt for Visually Impaired and Blind Person", Communications (NCC), *National Conference on Kharagpur,* pp no. 978-1 -4673-08 15 - 1, 2012.

[4] Deng, G. Cahill, L.W., "An adaptive Gaussian filter for noise reduction and edge detection", *IEEE Nuclear Science Symposium and Medical Imaging Cotiference,* pp. 1615 - 16 19, 1993 .

[5] Sangshin Kwak , Yeongwoo Choi, Kyusik Chung, "Video caption image enhancement for an efficient character recognition", *IEEE I5th international Conference on Pattern Recognition,* pp. 606-609, vol. 2, 2000.

[6] Zhang Zhi Yong, Song Yang, "The License Plate Image Binarization Based on Otsu Algorithm and MATLAB Realize", *International Cotiference on industrial Control and Electronics Engineering,* pp. 1657 - 1 659, 2012.