Current Trends in Technology and Science
ISSN: 2279- 0535. Volume: 3, Issue: 3(Apr-May 2014)

Marksheet Image Processing

Richa Muke
B.E. Computer, MMCOE, Pune, Email ID: richamuke@gmail.com
Sharvari Patil
B.E. Computer, MMCOE, Pune, Email ID: sharvari.amazing@gmail.com
Janhavi Acharya
B.E. Computer, MMCOE, Pune, Email ID: janhaviacharya777@gmail.com
Sankirti Shiravale
M.E. Computer, Pune, sankirtishiravale@mmcoe.edu.in

***Abstract* – Image analysis is the extraction of meaningful information from images; mainly from digital images by harnessing image processing techniques. Digital image processing allows the use of complex algorithms, and offers more sophisticated performance at simple tasks. We are developing a system for retrieving information from digital document images which is later stored on a database. Preprocessing of an image includes binarization and for this, we are using Maximum Entropy method. Optical Character Recognition (OCR) is a technology for converting scanned papers, digital photos and PDF files to text documents which can be edited. OCR is used to detect text from images which is then processed.**

***Keywords* – Binarization, Digital document image, Database, Image processing, Image analysis, Maximum Entropy, OCR.**

1. Introduction

Traditionally, document storage was done by paper work and traditional file systems, however, this form of storage is immune to degradation due to time and natural decaying. Another alternative is to digitalize all the information manually. This is not only tedious but is also prone to human errors. This is the reason why Digital Document Image Processing is emerging as the vital function for any organization.

Increasing demand for Digitalization of data needs an automated tool for converting hard coded data into digital format. Data cannot be retrieved directly from images, this has to be done manually. An OCR alone can only detect text, however, preprocessing of the images is very crucial before using an OCR because an image in raw form cannot be processed by OCR. Also, after OCR detects the text from images, the data obtained should be stored in a database where it can be handled and processed easily. Currently, marksheet details are manually entered into the database. This requires lots of human efforts and is time consuming as well. Moreover there is a risk of human errors because it is a tedious job. Hence, using methods of image processing we attempt to automate the whole process of creating a student database from marksheets.

1.1 Stepwise Flow of our System:

1. The input provided is scanned marksheet image.
2. Maximum entropy algorithm is applied onto this image.
3. The binarized image results (step 2) are given as an input to OCR.
4. OCR detects text out of these binarized image.
5. This text is fetched from OCR and bifurcated in .csv (comma separated values) form.
6. This .csv file is transferred into MySQL database.

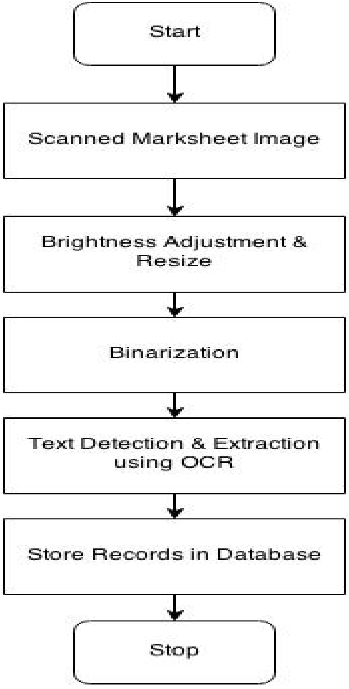

Start
Scanned Marksheet Image
Brightness Adjustment & Resize
Binarization
Text Detection & Extraction using OCR
Store Records in Database
Stop

Fig. 1. System Flow

Binarization is a significant step for this system, as, text can be detected only from binarized images. Different methods have been tried and tested for binarization of which, Otsu's method[11], Adaptative Smart-Binarization Method[5] and Effective Hybrid Thresholding Technique[3] give acceptable results. In Otsu's method[11], Otsu divides the image into two classes of pixels(foreground and background) and calculates the optimum threshold such that the combined spread between the two classes is minimal. In [5], an adaptive binarization method for document images is presented that takes into account unique characteristics of documents. Initially, image phase congruency (IPC) is calculated for the image and then, connected component analysis is carried out on the IPC edges. This segments out a local window for each symbol, thus, a local threshold is calculated for each window to binarize the corresponding portion of gray scale image. In [3], two types of thresholding are applied, that is, global thresholding and local thresholding, hence it can deal with various types of complex images. Taking into account all these methods, we have analysed that Maximum Entropy method for binarization[1], is the most effecient and suitable method for our needs.

2. Algorithm for Binarization

Thresholding is an important technique for image segmentation. It tries to identify and extract the object of interest from its background based on the grey-level distribution or texture in image areas. Entropy distribution of the grey-levels in an image is the most efficient techniques for image thresholding.

We are using Maximum entropy binarization method which has following steps :

[1] The Maximum Entropy is an automatic thresholding method. In this the optimal threshold value can be found by maximizing the entropy of the resulting classes (foreground and background).

[2] This thresholding technique is classified as bi-level approach, in which a unique threshold value is obtained.

[3] The bi-level segmentation techniques give satisfactory results on the images with clear foreground-background differentiation.

[4] A multi-thresholding technique converts the various regions of the image into regions having the optimal number of grey-level values.

[5] The Maximum Entropy approach is one of the most important threshold selection methods.

[6] Suppose that h(i) is a value in a normalized histogram, where *i* takes integer values from 0 to 255 (for 8-bitdepth images). It is assumed that h(i) is normalized, as

[7] $\sum_{i=0}^{i(max)} h(i) = 1$

[8] The entropy of black pixels is defined as follows:

[9] $H_B\,(t) = -\sum_{i=0}^{t} \frac{h(i)}{\sum_{j=0}^{t} h(j)} \log \frac{h(i)}{\sum_{j=0}^{t} h(j)}$

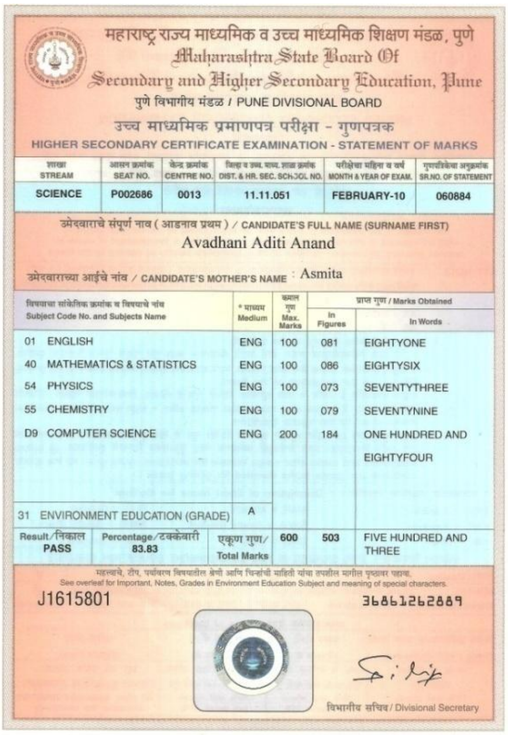
[10] The entropy of white pixels is defined as below:

[11] $H_W\,(t) = -\sum_{i=t+1}^{i_{max}} \frac{h(i)}{\sum_{j=t+1}^{i_{max}} h(j)} \log \frac{h(i)}{\sum_{j=t+1}^{i_{max}} h(j)}$

[12] The optimal threshold can be selected by maximizing the sum of foreground and background entropies as below:

$$T_{opt} = ArgMax_{t=0\ldots i_{max}}[H_B\,(t) + H_W\,(t)]$$

4. Experimental Results



महाराष्ट्र राज्य माध्यमिक व उच्च माध्यमिक शिक्षण मंडळ, पुणे
Maharashtra State Board Of
Secondary and Higher Secondary Education, Pune
पुणे विभागीय मंडळ / PUNE DIVISIONAL BOARD
उच्च माध्यमिक प्रमाणपत्र परीक्षा - गुणपत्रक
HIGHER SECONDARY CERTIFICATE EXAMINATION - STATEMENT OF MARKS

शाखा STREAM	आसन क्रमांक SEAT NO.	केन्द्र क्रमांक CENTRE NO.	जिल्हा व उच्च. माध्य. शाळा क्रमांक DIST. & HR. SEC. SCHOOL NO.	परीक्षेचा महिना व वर्ष MONTH & YEAR OF EXAM.	गुणपत्रिकेचा अनुक्रमांक SR.NO. OF STATEMENT
SCIENCE	P002686	0013	11.11.051	FEBRUARY-10	060884

उमेदवाराचे संपूर्ण नाव (आडनाव प्रथम) / CANDIDATE'S FULL NAME (SURNAME FIRST)
Avadhani Aditi Anand

उमेदवाराच्या आईचे नांव / CANDIDATE'S MOTHER'S NAME : Asmita

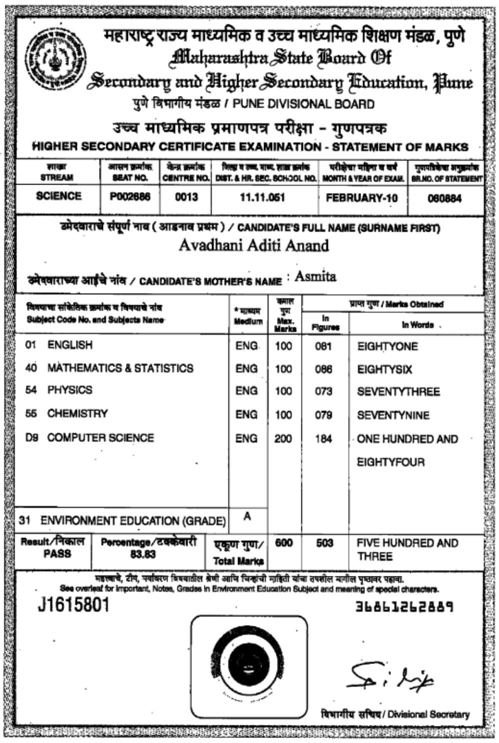
विषयाचा सांकेतिक क्रमांक व विषयाचे नांव Subject Code No. and Subjects Name	* माध्यम Medium	कमाल गुण Max. Marks	प्राप्त गुण / Marks Obtained In Figures	 In Words
01 ENGLISH	ENG	100	081	EIGHTYONE
40 MATHEMATICS & STATISTICS	ENG	100	086	EIGHTYSIX
54 PHYSICS	ENG	100	073	SEVENTYTHREE
55 CHEMISTRY	ENG	100	079	SEVENTYNINE
D9 COMPUTER SCIENCE	ENG	200	184	ONE HUNDRED AND EIGHTYFOUR
31 ENVIRONMENT EDUCATION (GRADE)	A			
Result/निकाल PASS · Percentage/टक्केवारी 83.83	एकूण गुण/ Total Marks	600	503	FIVE HUNDRED AND THREE

महत्वाचे, टीप, पर्यावरण विषयातील श्रेणी आणि चिन्हांची माहिती यांचा तपशील मागील पृष्ठावर पहावा.
See overleaf for Important, Notes, Grades in Environment Education Subject and meaning of special characters.

J1615801 3686126288 9

विभागीय सचिव / Divisional Secretary

Fig. 2. Original Image



महाराष्ट्र राज्य माध्यमिक व उच्च माध्यमिक शिक्षण मंडळ, पुणे
Maharashtra State Board Of
Secondary and Higher Secondary Education, Pune
पुणे विभागीय मंडळ / PUNE DIVISIONAL BOARD
उच्च माध्यमिक प्रमाणपत्र परीक्षा - गुणपत्रक
HIGHER SECONDARY CERTIFICATE EXAMINATION - STATEMENT OF MARKS

शाखा STREAM	आसन क्रमांक SEAT NO.	केन्द्र क्रमांक CENTRE NO.	जिल्हा व उच्च. माध्य. शाळा क्रमांक DIST. & HR. SEC. SCHOOL NO.	परीक्षेचा महिना व वर्ष MONTH & YEAR OF EXAM.	गुणपत्रिकेचा अनुक्रमांक SR.NO. OF STATEMENT
SCIENCE	P002686	0013	11.11.051	FEBRUARY-10	060884

उमेदवाराचे संपूर्ण नाव (आडनाव प्रथम) / CANDIDATE'S FULL NAME (SURNAME FIRST)
Avadhani Aditi Anand

उमेदवाराच्या आईचे नांव / CANDIDATE'S MOTHER'S NAME : Asmita

विषयाचा सांकेतिक क्रमांक व विषयाचे नांव Subject Code No. and Subjects Name	* माध्यम Medium	कमाल गुण Max. Marks	प्राप्त गुण / Marks Obtained In Figures	 In Words
01 ENGLISH	ENG	100	081	EIGHTYONE
40 MATHEMATICS & STATISTICS	ENG	100	086	EIGHTYSIX
54 PHYSICS	ENG	100	073	SEVENTYTHREE
55 CHEMISTRY	ENG	100	079	SEVENTYNINE
D9 COMPUTER SCIENCE	ENG	200	184	ONE HUNDRED AND EIGHTYFOUR
31 ENVIRONMENT EDUCATION (GRADE)	A			
Result/निकाल PASS · Percentage/टक्केवारी 83.83	एकूण गुण/ Total Marks	600	503	FIVE HUNDRED AND THREE

महत्वाचे, टीप, पर्यावरण विषयातील श्रेणी आणि चिन्हांची माहिती यांचा तपशील मागील पृष्ठावर पहावा.
See overleaf for Important, Notes, Grades in Environment Education Subject and meaning of special characters.

J1615801 3686126288 9

विभागीय सचिव / Divisional Secretary

Fig. 3. After Binarization

3. SYSTEM ARCHITECTURE

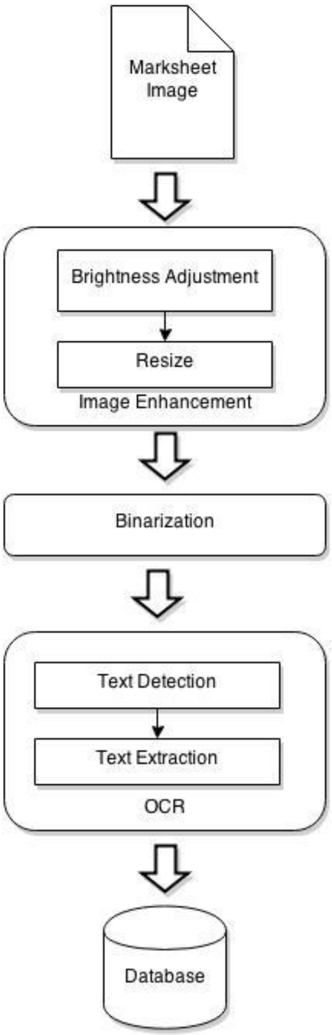


Fig. 4. System Architecture

5. CONCLUSION AND FUTURE SCOPE

Document image processing is burgeoning as a critical domain in computer engineering. This system revolutionalizes the conventional approach by automating document image processing. It minimizes the need to do any manual work. This system can be further extended for various other documents having fixed formats and bold faced text.

6. REFERENCES

[1] Magdolna Apro, Szabolcs Pal, Sandra Dedijer, "Evaluation of single and multi-threshold entropy-based algorithms for folded substrate analysis" Journal of Graphic Engineering and Design, Volume 2011

[2] K. Ntirogiannis, B. Gatos and I. Pratikakis, "An Objective Evaluation Methodology for Document Image Binarization Techniques" 2008 IEEE.

[3] Mohamed Zayed, Asma Ouari, Meriem Derraschouk and Youcef Chibani, "An Effective Hybri Thresholding Technique for Degraded Documents Images Binarization" 2011 IEEE.

[4] B. Gatos, I. Pratikakis and S.J. Perantonis, "Improved Document Image Binarization by Using a Combination of Multiple Binarization Techniques and Adapted Edge Information" , 2008 IEEE

[5] Djamel GACEB Frank Lebourgeois, Jean Duong, "Adaptative Smart-Binarization Method" 2013 IEEE

[6] Chang, C.-I, Du, Y., Wang, J., Guo, S.-M., Thouin,P.D. (2006) Survey and comparative analysisof entropy and relative entropy thresholdingtechniques, IEE Proceedings of Vision, Image and Signal Processing, 153 (6), pp. 837-850.

[7] Yin, P.-Y. (2002) Maximum entropy-based optimal threshold selection using deterministic reinforcement learning with controlled randomization, Signal Processing 82, 993-1006.

[8] Tabbone, S., Wendling, L. (2003) Multi-scale binarization of images, Pattern Recognition Letters 24, 403-411.

[9] Sezgin, M., Sankur, B, (2004) Survey over image thresholding techniques and quantitative performance evaluation, Journal of Electronic Imaging, 13 (1), 146-165.

[10] Liao, P.S., Chen, T.S., Chung, P.C. **(**2001**)** A fast algorithm for multilevel thresholding. Journal ofInformation Science and Engineering 17, 713–727. Otsu, "A Threshold Selection Method from Gray-Level Histograms", IEEE 1979

[11] Sezgin, M., Sankur, B, (2004) Survey over image thresholding techniques and quantitative performance evaluation, Journal of Electronic Imaging, 13 (1), 146-165.

[12] Tabbone, S., Wendling, L. (2003) Multi-scale inarization of images, Pattern Recognition Letters 24, 403-411.

[13] Zhang, Y., Wu L. (2011) Optimal Multi-Level Thresholding Based on Maximum Tsallis Entropy via an Artificial Bee Colony Approach, Entropy, 13 (1), 841-859.

[14] Xiao, Y., Cao, Z., Zhang, T. (2008) Entropic Thresholding Based on Gray-level Spatial Correlation Histogram, Proceedings of the19th International Conference on Pattern Recognition, ICPR 2008, 8-11 December 2008, Tampa, Florida, USA, pp. 1-4.

[15] Yin, P.-Y. (2002) Maximum entropy-based optimal threshold selection using deterministic reinforcement learning with controlled randomization, Signal Processing 82, 993-1006

.