QUAD: Quality Assessment of Documents

Deepak Kumar and A G Ramakrishnan
Medical Intelligence and Language Engineering Laboratory
Department of Electrical Engineering, Indian Institute of Science
Bangalore - 560012, India
deepak,ramkiag@ee.iisc.ernet.in

***Abstract*—We propose a set of metrics that evaluate the uniformity, sharpness, continuity, noise, stroke width variance, pulse width ratio, transient pixels density, entropy and variance of components to quantify the quality of a document image. The measures are intended to be used in any optical character recognition (OCR) engine to a priori estimate the expected performance of the OCR. The suggested measures have been evaluated on many document images, which have different scripts. The quality of a document image is manually annotated by users to create a ground truth. The idea is to correlate the values of the measures with the user annotated data. If the measure calculated matches the annotated description, then the metric is accepted; else it is rejected. In the set of metrics proposed, some of them are accepted and the rest are rejected. We have defined metrics that are easily estimatable. The metrics proposed in this paper are based on the feedback of homely grown OCR engines for Indic (Tamil and Kannada) languages. The metrics are independent of the scripts, and depend only on the quality and age of the paper and the printing. Experiments and results for each proposed metric are discussed. Actual recognition of the printed text is not performed to evaluate the proposed metrics. Sometimes, a document image containing broken characters results in good document image as per the evaluated metrics, which is part of the unsolved challenges. The proposed measures work on gray scale document images and fail to provide reliable information on binarized document image.**

***Keywords*-Quality metrics; document images; multi-script; document image quality analysis; Optical character recognition;**

I. Introduction and Motivation

The current commerically available optical character recognition (OCR) engines generally do not infer about the document image quality. Quality of a document image has not been adequately studied in the document image analysis literature. Even though we can rate the quality by a simple calculation of number of correctly detected characters against total number of characters in the document image using best performing OCR engines [10], we require human intervention for calculating the correctly recognized characters and the actual number of characters. How to we obviate the need for human intervention and make the system by itself capable of predicting the quality of a document image? We propose quality metrics in an attempt to solve these important questions.

Now a days, camera captured document images are popular. We cannot take many objects on which text is present to a scanning machine. Here object may refer to a historical book or notice board. The quality metrics proposed are simple and can be implemented on cameras. The quality of a document image can be predicted while capturing a document using camera. With the predicted quality, an user has option to choose proper binarization algorithm on that document image.

Nagy [3] presents an overview of document image analysis. Nagy explains the evolution of document image analysis into text, graphics, page layout and retrieval. In the earlier image analysis systems, priorities were improved binarization and recognition of a document image. Roger et al. [4] presented a system to estimate the quality of a document image with different degradation procedures such as noise addition and blurring before passing it through an OCR engine. We cannot expect similar degradation process to occur on a document because a document image from one book will have different degradation compared with a document image from another book. Zramdini et al. [5] present a study of mechanical degradation of a document and its effect in font recognition.

Blando et al. [10] use correctly identified text regions for predicting OCR accuracy. Blando et al. developed two metrics namely white speckle factor (WSF) and broken character factor (BCF) for OCR evaluation. Cannon et al. [11] include small speckle factor (SSF), touching character factor (TCF) and font size factor (FSF) for predicting quality and choose appropriate filter to restore the degraded document. Souza et al. [12] propose a method for automatically selecting the best among the restoration filters. The above works [10], [11], [12] observe only the noise present in the document image to propose metrics and OCR engines are required to evaluate the quality of document image. In our case, we avoid the use of OCR to a great extent and ensure quality metrics are script independent.

Improving degraded document images was a focused topic for several years. With enhancement of degraded document image, the recognition rate of a document increases. Sauvola et al. [6] demonstrate a document image binarization based on local histogram threshold for each pixel in the document. Gatos et al. [7] introduce adaptive document image binariza-

tion for degraded documents and an improved method using Sauvola's binarization algorithm. Yang et al. [8] present another adaptive binarization technique using the runlength of strokes, since connected characters or broken characters have strokes comparable to normal characters. Banerjee et al. [9] report that Markov random fields are effective in severely degraded document image restoration. In all these papers, the quality of a document image is referred in terms of improved binarization. We do not find any relevant contribution towards a qualitative analysis of document image.

During digitization of books, suppose some intermediate pages in the document are severely degraded; then we cannot have any information or hint to look for those pages in the book. Improving binarization will not provide basic quality criteria when an user is evaluating the digitized book. If quality of each page in the book is predicted, a user can give careful attention to specific pages where the quality of the page is degraded. Our proposed metrics are meant to predict document image quality. Proposed quality metrics are evaluated against user annotated data for a set of document images. This set involves possible degradation and also cover multi-script. Multi-script based evaluation handles a document irrespective of the script, to make its quality predictable.

II. Factors that impact the quality of a document

A document image is obtained through scanning the hard copy of a document or using a camera. The quality of a document is determined by four major factors namely age of the document, scanner, printer and paper. In the case of a camera, scanner factor is replaced by the focus of the camera. We are visually capable of assessing the quality of a document image without the knowledge of printer, scanner or paper used in the preparation of the document. We can also recognize and read the text present in the document. However, in case of machine, it is a difficult task to assess document image quality.

We can pose a set of questions: How does a scanning device cause deterioration? What kind of paper is used to get a print of the document? What is the effect of the printing device outputting the textual data onto the paper? Each device acts independently in the creation of document image. To assess the quality we can write up a naive Bayesian formula as shown below,

$$p(\mathcal{D}|\mathcal{P},\mathcal{S},\mathcal{R},\mathcal{A}) = p(\mathcal{D}|\mathcal{P})p(\mathcal{D}|\mathcal{S})p(\mathcal{D}|\mathcal{R})p(\mathcal{D}|\mathcal{A}) \quad (1)$$

where $\mathcal{D}$ is actual document information, $\mathcal{P}$ indicates quality of the paper, $\mathcal{S}$ indicates scanner quality, $\mathcal{R}$ indicates printer quality and $\mathcal{A}$ is age of the document.

The skew correction in document images is exhaustively researched. Its effect is far less compared to other factors and we do not consider it in this paper. It is not considered as bad quality even though skew may have a role in the recognition of characters.

III. Degradation levels

We approach this problem by proposing quality metrics and validating the metric against user annotated degradation levels. A database of 132 annotated multi-script scanned document images comprising all forms of degradation is used for our experiment and each document image is annotated by 6 users according to the provided scale. Quality of the scanned image of a printed document is graded in the scale of 1-5. We have grouped all possible conditions of the document image degradations to be assigned by user in the form of a subjective score.

- Scale value 0: An image, not yet annotated by the user.
- Scale value 1: **Image with degradations caused while scanning**. Document image with bends or folds (so that part of the characters are lost), at the corners (eg. hard bound book) or sides of the document. One is sure that a sizable number of characters in the document image cannot be recognized. Smear in the document image falls into this category. A document image with skew (rotated with respect to scanner axis while scanning) does not fall into this category.
- Scale value 2: **Highly degraded document image**. Characters in the document image have very less contrast and separating the characters from the background may require a complex algorithm. The gray values of the background in some areas are comparable to the gray values of the characters in other areas. There will be uncertainty in the OCR performance.
- Scale value 3: **Background degraded document image**. Background of the image has patches in all places from top to bottom and from left to right side of the image. However, the characters in the image maintain clarity.
- Scale value 4: **Slightly degraded document image**. Characters in the document are distinct and clear. There are few local gray areas in the background of the document image.
- Scale value 5: **Good document image**, satisfies the following conditions: the characters in the image are distinct, sharp and clear; a clear background without any form of glitches or patches in the form of gray areas.

Document images annotated by six users are used in calculating quality metrics. The mean subjective score for a document image is computed for validation of quality metrics.

IV. Quality metrics and Experiments

We describe each of the quality metrics devised for document images. The calculated values of the quality metric are plotted against mean subjective scores of documents present

in the database. We fit a straight line using least squares approach and obtain its slope.

$$J_i = \min \parallel \mathcal{Q}_i - \mathcal{U} \parallel^2 \quad \forall i \in 1, 2, 3...m \tag{2}$$

where $\mathcal{Q}_i$ = i^{th} quality metric, $\mathcal{U}$ = annotated score, m = number of quality metrics. We have also calculated the correlation coefficient for each metric with the annotated mean score.

$$\rho_i = \frac{\mathcal{Q}_i^T \mathcal{U}}{\|\mathcal{Q}_i\| \|\mathcal{U}\|} \quad \forall i \in 1, 2, 3...m \tag{3}$$

In this paper, some quality metrics require the binarized image, which is generated by Otsu thresholding [1] of the gray document image. The binarized image is labelled for individual connected components (CC).

A. Foreground and Background uniformity

A document image quality varies with degradations. This measure is the standard deviation of estimated mean foreground and background values using Otsu threshold. To capture global and local uniformity, image is split into equal parts. In the first level, entire image is used. In the second level, document image is split into four equal parts and in the third level, image is split into 16 equal parts. Mean foreground and background values are calculated for each part image at each level. Calculated mean foreground and background values are grouped together. The foreground uniformity (U_f) and background uniformity (U_b) are calculated as shown below,

$$U_f^i = std(\mu_{fg1}^i, ..., \mu_{fg21}^i) \quad \forall i \in 1, 2, ...N \tag{4}$$

where 'N' is the total number of annontated document images.

$$U_b^i = std(\mu_{bg1}^i, ..., \mu_{bg21}^i) \quad \forall i \in 1, 2, ...N \tag{5}$$

In the second and third levels, due to localized text positions, sometimes background patch is segmented without any text. Therefore, we calculate a single peak threshold before estimating the metric,

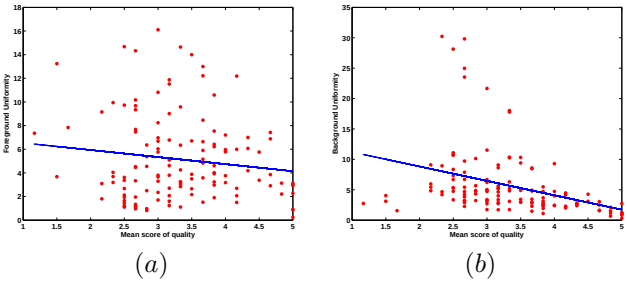
$$Single\ peak\ thresh_j = \frac{\mu_{bgj} - \mu_{fgj}}{\mu_{fgj}(255 - \mu bgj)} \quad \forall j \in 1, 2, 3, ..21 \tag{6}$$

If the single peak threshold value is less than 0.01 then that patch is removed else retained.

Figures 1 (a) and (b) show the plots of uniformity for foreground and background. The estimated correlation coefficient indicate that the background uniformity is suitable for predicting the document image quality.

B. Sharpness

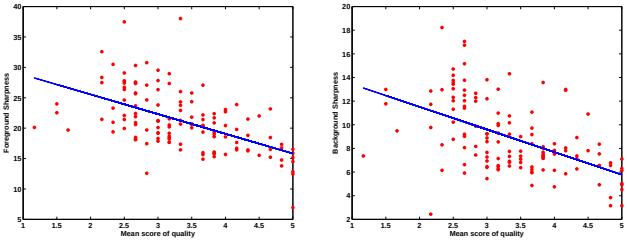
A document image is clustered into three parts using fuzzy clustering. These clusters in the document form foreground, background and transient regions. The standard deviation of gray values for foreground and background clusters are calculated. These indicate the sharpness of that document



(a) (b)

Figure 1. Uniformity indices for document images as a function of mean subjective scores. (a) Standard deviation of foreground values with estimated slope = -0.60 (b) Standard deviation of background values with estimated slope = -2.37.

image. Figures 2 (a) and (b) show the plots of sharpness values for foreground and background respectively. For degraded images the sharpness reduces and number of gray levels occupied by the foreground and background clusters increases. This can be inferred from the plot and also with the estimated slopes.



(a) (b)

Figure 2. Sharpness values of document images as a function of mean subjective scores. (a) Foreground values with estimated slope = -3.24 (b) Background values with estimated slope = -1.91.

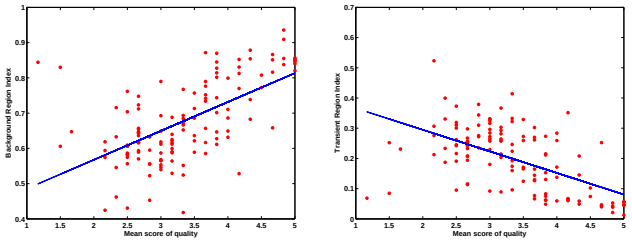
C. Transient region density

The number of pixels in each of foreground, background and transient clusters are normalized to the total number of pixels in the document image. As the document starts degrading, more pixels from background region fall into transient region. This quality measure indirectly captures complexity and non-uniform illumination in the document image.

Figures 3 (a) and (b) show the plots of normalized background and transient regions respectively. We can observe that degraded document images have more pixels in the transient region. The slopes of the two regions are almost similar and opposite.

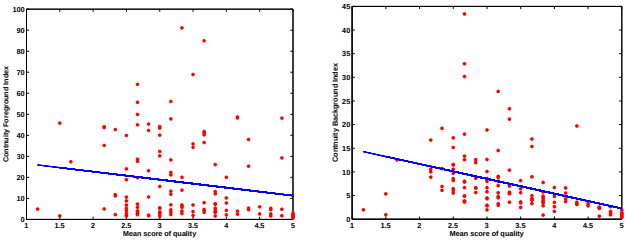
D. Continuity

A document image is binarized using Otsu threshold [1]. Each CC in the binarized document is labelled and height and width values are calculated. The mean values of height



(a) (b)

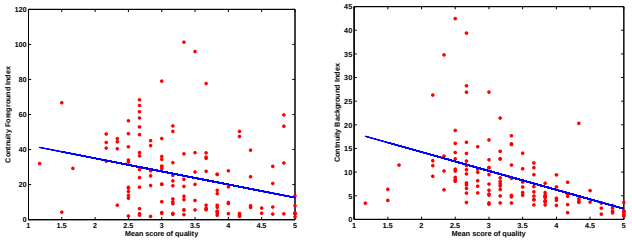
Figure 3. Densities of different regions for document images as a function of mean subjective scores. (a) Background region pixels normalized with estimated slope = 0.082 (b) Transient region pixels normalized with estimated slope = -0.071.



(a) (b)

Figure 5. Continuity index for document images with 8 times mean height and width values used as size of sub-block. (a) Foreground values with estimated slope = -3.81 (b) Background values with estimated slope = -3.14.

and width are estimated. Using 4 times the mean value of height and width as block size, the image is split into blocks and the mean foreground and background values are calculated from these blocks and also standard deviation. Single peak threshold is applied for blocks before calculating standard deviation. This quality metric is calculated again using 8 times the mean value of height and width as block size. Figures 4 (a) and (b) show the standard deviation plots of foreground and background for four times mean value of height and width of CC's. Figures 5 (a) and (b) show the same plots with block size of eight times mean value of height and width of CC's.



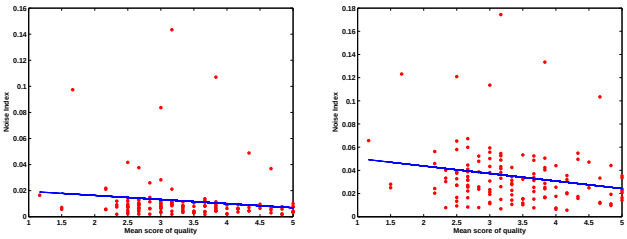
(a) (b)

Figure 4. Continuity index for document images with 4 times mean height and width values used as size of sub-block. (a) Foreground values with estimated slope = -7.48 (b) Background values with estimated slope = -4.00.

E. Document noise measure

Median filtering is performed on the binarized image for estimating noise measure. We count the number of pixels converted from both 1 to 0 and 0 to 1. This count gives an idea about the amount of salt and pepper noise, which indirectly points to the quality of a document. Cannon et al. [11] and Souza et al. [12] use median filtering to remove speckles and improve degraded document image.

Figures 6 (a) and (b) show normalized counts for number of pixels converted by median filter of size 3x3 and 5x5, respectively. The correlation coefficient indicate that this quality metric is not suitable for prediction. Even mean height and width of CC's were used for size of median filtering. The result with mean size method is not promising due to high degree of variation in the size of CC's of same document image quality.

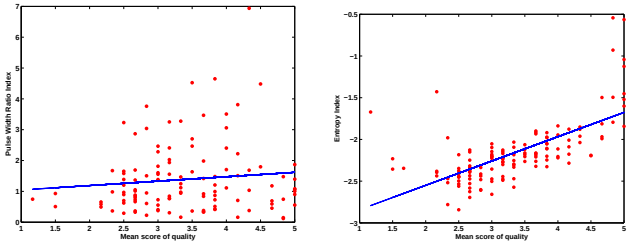


(a) (b)

Figure 6. Noise measure for document images as a function of mean subjective scores. (a) Median filter of size 3x3 is used with estimated slope = -0.003 (b) Median filter of size 5x5 is used with estimated slope = -0.006.

F. Pulse width ratio

The labeled CC's in binarized document are replaced by a patch of black pixels with height and width same as CC's. The newly formed logical image is used to calculate horizontal and vertical run lengths. We treat run length as similar to pulse width ratio of a pulse. Mean value is calculated over the document except with zero pulse width. The product of horizontal and vertical pulse width is used as quality measure. Figure 7 (a) shows the plot of the product. The correlation coefficient indicates that this metric is not valid for prediction. This metric relies on the inter character and inter word gaps for quality measurement. In any document type, these gaps are of almost standard sizes based on the font and the font size. Fourier transforms of each horizontal and vertical line were also calculated for quality measurement. Due to spurious peaks in the transformed vector, the idea of using the transformed vector as quality metric was dropped.



(a) (b)

Figure 7. (a) Plot of pulse width ratio product as a function of mean subjective scores with estimated slope = 0.14 (b) Entropy values with estimated slope = 0.29.

G. Entropy

Entropy is the measure of information at the source. Kapur et al. [2] use entropy for binarization of document image. In general cases, Otsu's binarization method was superior to Kapur's entropy method. The number of levels in a gray document image is 256. Histogram of the image is used as probability of source symbols and then entropy is calculated. Document image can be compressed to bi-level image. The entropy of the document decreases with the number of levels. In this paper, to measure entropy index as quality metric, the number of gray levels is halved iteratively until the number of gray levels is 2. At each stage, histogram and entropy are calculated for the given document image. Entropy is a global approach; the level of details expressed in the reduced number of quantization levels infer complexity of the document. Mean value of entropy index is calculated combining all quantized image entropies. Figure 7 (b) shows the plot of mean entropies against mean score of document images. The estimated correlation coefficient of 0.65 indicates that this quality metric is useful in predicting the document image quality as well as complexity.

The equation for calculating entropy index is shown below,

$$Entropy_j = -\sum_{i=1}^{N_j} p_i \log\ p_i \quad \forall i \leq N_j \tag{7}$$

where N_j is the number of gray levels for downsampled document images. In our experiment, the range of N_j is (2, 4, 8, 16, 32, 64, 128, 256).

H. Stroke width

The CC's and their bounding boxes are used in estimation of this metric. We estimate mean and variance of the stroke width calculated. A small value of variance indicates that the characters are distinctive from the background and can be recognized. We have designed two methods of stroke width estimation. They are based on bounding box values, area values and thinning of CC. In thinned CC method, the stroke width is calculated as below,

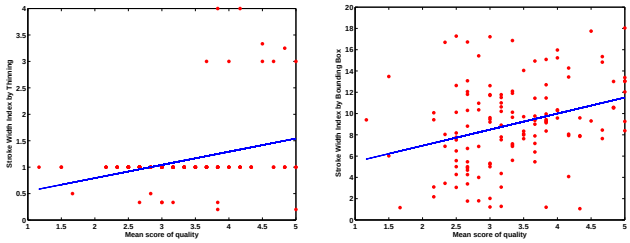
$$SW\ thin_i = \frac{Area\ of\ CC_i}{No\ of\ pixels\ in\ thinned\ CC_i} \quad \forall i \in 1, 2, 3, ...n \tag{8}$$

where 'n' is the total number of connected components in the document. In bounding box method, the stroke width is calculated as shown below,

$$SW\ bb_i = \frac{1}{2}(H_i + W_i) + \sqrt{\frac{1}{8}(H_i + W_i)^2 - \frac{Area\ of\ CC_i}{2}} \tag{9}$$

where H_i and W_i are height and width of i^{th} CC.

Figures 8 (a) and (b) show the plots of the two different estimates of stroke width. We can observe that in thinned CC stroke width calculation, the mean value is a horizontal line with some outliers. Thinned CC stroke width is not a reliable quality metric from this observation. Stroke width is used in [8] for improving degraded document images.



(a) (b)

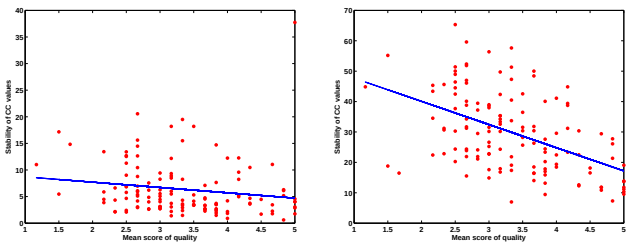
Figure 8. Stroke width measure for document images as a function of mean subjective scores. (a) Thinned CC values with estimated slope = 0.25 (b) Bounding box values with estimated slope = 1.51.

I. Stability of CC values

As number of characters, character size and document size varies in a document, there will be difficulty to get information of degradation. Here, we calculate minimum and maximum gray values in each labelled CC. A standard deviation is calculated for minimum and maximum gray values of all CC's in the document. If document has local or global degradation it will be captured in terms of standard deviation. Figures 9 (a) and (b) show the values of minimum and maximum gray value, respectively. The slope of standard deviation in maximum gray value of CC indicates a high variation with respect to degradation.

V. Results

In Table I, we have tabulated the quality metric, slope of the estimated line and correlation coefficient. For validation of the quality measure, the correlation coefficient should exceed ± 0.3. Higher correlation factor indicates that the corresponding quality metric is useful in predicting the document image quality.



(a) (b)

Figure 9. Stability of CC values for document images as a function of mean subjective scores. (a) Standard deviation of minimum gray value of each CC in a document image with estimated slope = -1.00 (b) Standard deviation of maximum gray value of each CC in a document image with estimated slope = -7.62.

Table I
QUALITY METRICS, ESTIMATED SLOPE AND CORRELATION COEFFICIENT FOR DOCUMENT IMAGES

Quality metric	**Slope**	**Correlation coefficient**
Foreground Uniformity	-0.60	-0.14
Background Uniformity	-2.37	-0.36
Foreground Sharpness	-3.24	-0.55
Background Sharpness	-1.91	-0.52
Transient region density	-0.07	-0.57
Background region density	0.08	0.60
Foreground Continuity4	-7.48	-0.30
Background Continuity4	-4.00	-0.47
Foreground Continuity8	-3.81	-0.17
Background Continuity8	-3.14	-0.40
Noise 3x3 median filter	-0.003	-0.14
Noise 5x5 median filter	-0.006	-0.20
Stroke width by thin	0.25	0.32
Stroke width by bounding box	1.51	0.31
Pulse width ratio	0.14	0.11
Entropy	0.29	0.65
Stability of CC min gray value	-1.00	-0.16
Stability of CC max gray value	-7.62	-0.49

VI. CONCLUSIONS AND FUTURE WORK

The quality metrics proposed are validated by fitting a straight line by least squares approach and also correlation coefficient. Table I indicates that some of the metrics can be used for predicting the document image quality and others fail. Broken or merged characters have not been considered in the experimental evaluation of quality metrics. If recognition module with confidence interval is used, then we can devise quality metrics, which exploit the recognition engine. The main drawback of recognition based quality metrics is script dependency. The metrics proposed can predict the quality of the document using fuzzy logic. Implementation of validated quality metrics on camera mobiles and camera captured document images are our future work.

REFERENCES

[1] Nobuyuki Otsu, *A threshold selection method from gray-level histogram*, IEEE Transactions on Systems, Man and Cybernetics, Vol.9, no.1, pp.62-69, 1979.

[2] J. N. Kapur, Prasanna K. Sahoo and A. K. C. Wong, *A new method for gray-level picture thresholding using the entropy of the histogram*, Computer Vision, Graphics, and Image Processing, Vol.29, no.3, pp.273-285, 1985.

[3] George Nagy, Twenty Years of Document Image Analysis in PAMI, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.22, no.1, pp.38-62, 2000.

[4] Roger T. Hartley and Kathleen Crumpton, *Quality of OCR for Degraded Text Images*, Proceedings of the Fourth ACM conference on Digital Libraries, August 11-14, 1999, Berkeley, CA, USA, pp.228-229, 1999.

[5] Abdel Wahab Zramdini and Rolf Ingold, *A study of document image degradation effects on font recognition*, Third International Conference on Document Analysis and Recognition, Motreal, Canada, August 14-16, pp.740-743, 1995.

[6] Jaakko J. Sauvola and Matti Pietikäinen, *Adaptive document image binarization*, Pattern recognition, Vol.33, no.2, pp.225-236, 2000.

[7] Basilios Gatos, Ioannis Pratikakis and Stavros J. Perantonis, *Adaptive degraded document image binarization*, Pattern Recognition, Vol.39, no.3, pp.317-327, 2006.

[8] Yibing Yang and Hong Yan, *An adaptive logical method for binarization of degraded document images*, Pattern Recognition, Vol.33, no.5, pp.787-807, 2000.

[9] Jyotirmoy Banerjee, Anoop M. Namboodiri and C. V. Jawahar, *Contextual restoration of severely degraded document images*, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA, pp.517-524, 2009.

[10] L. R. Blando, Junichi Kanai and Thomas A. Nartker, *Prediction of OCR accuracy using simple image features*, Proceedings of Third International Conference on Document Analysis and Recognition, Motreal, Canada, August 14-16, pp.319-322, 1995.

[11] Michael Cannon, Judith Hochberg and Patrick Kelly, *Quality assessment and restoration of typewritten document images*, International Journal on Document Analysis and Recognition, Vol.2, no.2-3, pp.80-89, 1999.

[12] Andrea Souza, Mohamed Cheriet, Satoshi Naoi and Ching Y. Suen, *Automatic filter selection using image quality assessment*, Seventh International Conference on Document Analysis and Recognition (ICDAR 2003), 2-Volume Set, 3-6 August 2003, Edinburgh, Scotland, UK, pp.508-, 2003.