01 May 2012

# Character-based Automated Human Perception Quality Assessment In Document Images

Tayo Obafemi-Ajayi
*Missouri University of Science and Technology*, towd2@mst.edu

Gady Agam



## Recommended Citation

T. Obafemi-Ajayi and G. Agam, "Character-based Automated Human Perception Quality Assessment In Document Images," *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, vol. 42, no. 3, pp. 584 - 595, article no. 6062687, Institute of Electrical and Electronics Engineers, May 2012.
The definitive version is available at https://doi.org/10.1109/TSMCA.2011.2170417

# Character-Based Automated Human Perception Quality Assessment in Document Images

Tayo Obafemi-Ajayi and Gady Agam

***Abstract*—Large degradations in document images impede their readability and deteriorate the performance of automated document processing systems. Document image quality (IQ) metrics have been defined through optical character recognition (OCR) accuracy. Such metrics, however, do not always correlate with human perception of IQ. When enhancing document images with the goal of improving readability, e.g., in historical documents where OCR performance is low and/or where it is necessary to preserve the original context, it is important to understand human perception of quality. The goal of this paper is to design a system that enables the learning and estimation of human perception of document IQ. Such a metric can be used to compare existing document enhancement methods and guide automated document enhancement. Moreover, the proposed methodology is designed as a general framework that can be applied in a wide range of applications.**

***Index Terms*—Document imaging, feature extraction, human–machine interactions, image enhancement, learning systems, perception quantification, quality metrics.**

## I. Introduction

THE preservation of information is a key part of human society. There exist numerous large archives of historical document collections, the majority of which have deteriorated for various reasons. Documents become degraded due to various factors, such as the aging of the paper, poor ink quality, physical deterioration, and limitations in the scanning process. Enhancement of such document images is needed to improve readability and, subsequently, optical character recognition (OCR) performance. Enhancing the readability of document images is necessary in degraded documents to increase their legibility while preserving the authenticity of the historical document so that the document can be read in its original format.

Readability is a subjective quality measure which lacks a mathematical definition. There is a great need for a figure of merit whose value can guide and assess automated enhancement systems in gauging the quality of enhancement done with respect to readability. Binarization and restoration methods (for example, [1]–[4]) usually use objective metrics, such as OCR accuracy and misclassification error, for quantifying their performance. Sezgin and Sankur [5] present a survey of different thresholding algorithms and various performance criteria used. However, we lack an objective means of quantifying readability and estimating the document image quality (IQ), as perceived by a human. Enhancement methods such as that in [4] and [6] that attempt to evaluate their work by human judgments of quality have to include manual inspection of the output images by some human experts which can be tedious and time consuming.

Current IQ metrics are not consistent with human perception, as shown by the evaluation results presented in [7]. IQ metrics [8], [9] have been defined to determine OCR accuracy which is often viewed as an indicator of IQ. Cannon *et al.* [9] defined some quality measures, initially introduced by Blando *et al.* [8] for quantifying typewritten document text image degradation with respect to OCR accuracy. However, OCR accuracy cannot be used as a sole indicator of document IQ, as perceived by a human, for the following reasons. Usually, OCR engines attempt to enhance the document images in the preprocessing stage to better recognize degraded characters. Many engines also incorporate a language model and/or could be trained on degraded data to further improve the recognition rate. The aim is to make OCR engines more immune to these degradations via training and thus result in a better performance when applied. Moreover, OCR accuracy varies from engine to engine and is dependent on the quality of the OCR software, not just the level of degradation of the text image. Given the same document collection, there is a significant difference in the performance of an OCR engine from a decade ago to a more recent engine. Hence, there is a need to design measures that can predict/estimate human perception of text IQ.

In this paper, we present a framework for learning a perception-based evaluation metric. Our goal is to develop a user-centered objective function for evaluating the level of degradation in a document image, as perceived by a human user. We link human perception to a concrete metric by applying machine learning techniques. Using a neural network multilayer perceptron (MLP) regression model, we train a predictor to estimate the level of degradation, as perceived by a human, for a given image input. We focus on binary document images, which implies that some adaptive thresholding algorithm has already been applied to these images.

Our system computes a set of features from the character images in the document. The document images used in the context of our work are typewritten text images, a subset of the Frieder collection [10] of historical Holocaust documents. We are not concerned with interpreting the text of the document

Manuscript received June 25, 2010; revised December 1, 2010 and March 21, 2011; accepted August 12, 2011. Date of publication October 28, 2011; date of current version April 13, 2012. This paper was recommended by Associate Editor Q. Ji.

T. Obafemi-Ajayi is with the Department of Computer Science, University of Missouri, Columbia, Columbia, MO, 65211 (e-mail: obafemiajayit@missouri.edu).

G. Agam is with the Illinois Institute of Technology, Chicago, IL 60616 USA (e-mail: ihimtay@gmail.com; agam@iit.edu).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TSMCA.2011.2170417

images. Based on the set of features extracted, the predictor yields a numeric value that estimates the perceived level of degradation. This paper further extends the initial evaluation of human perception of degradation quality reported in [7]. Using the fact that there exists a correlation between degradation parameters (such as percentage of brokenness in a character) and human perception of quality [7], we develop a large set of synthetic data to use for training our predictor.

The proposed system is general and easily adaptable to various degradation models and types of document images. It is a general framework that can be applied in a wide range of applications such as predicting OCR accuracy, as demonstrated in Section VII. Moreover, the proposed methodology provides a general framework for quantifying human perception. Quantifying human perception can be used for building computer systems that assess machine performance and guide automated algorithm development.

This paper is organized as follows. The state of the art related to the research addressed in this paper is reviewed in Section II. The proposed methodology is described in Section III. In Section IV, we discuss the set of features extracted to predict the perceived level of degradation. The process of generating our data set and a description of the subjective quality experiment used to obtain the perceptual ranking information is presented in Section V. Finally, results are presented and discussed in Section VI, while we demonstrate in Section VII how our proposed framework can also be applied to predicting OCR accuracy, before drawing the conclusions in Section VIII.

## II. Related Work

The majority of work done in visual IQ assessment [11]–[15] has been focused on nondocument images. Automated IQ assessment aims to provide an objective measurement for the quality of a given image which is consistent with the result given by human observers. Bouzerdoum *et al.* [11] and Narwaria and Lin [14] both proposed methods for IQ assessment based on formulating effective features and fusing them into a single number to predict the quality score using machine learning algorithms. The former utilizes a neural network MLP model, while the latter uses a support vector regression model. Similar to our objective for document IQ assessment, they both attempt to replace human judgment of perceived IQ with a machine evaluation. Our proposed approach is unique in that it targets directly text images.

The INCITS W1.1 [16] project was chartered to develop an appearance-based IQ standard for text images from gray-level and full-color printing systems. They concluded that a psychophysical evaluation method utilizing reference samples will be necessary to design such a standard. This work focuses on bilevel document images that contain varying degradations that impeded readability based on human psychophysical experiments. The work done on deriving IQ metrics specifically for document text images has been mainly for the purpose of predicting OCR accuracy. Govindaraju and Srihari [17] attempted to measure readability of a document image to predict OCR performance. As stated in Section I, Blando *et al.* [8] introduced some image features, such as white speckle factor and broken character factor, that could be used for prediction of OCR accuracy. Their prediction system classifies the input image as either good (i.e., high OCR accuracy expected) or poor (i.e., low OCR accuracy expected). Cannon *et al.* [9] used the image features defined by Blando *et al.* to define five quality measures (white speckle factor, small speckle factor, broken character factor, touching character factor, and font size) for quantifying typewritten document text image degradation. These measures are then applied to a linear classifier to select a restoration algorithm for improving the OCR performance of the document image.

Degradation model parameters have been designed by Baird [18] to model degradation features in scanned/photocopied uniform text document images with the goal of using the model to improve OCR engine performance. Given that the defined quality metrics are intended to predict OCR accuracy, Reed and Smith [19] explored the correlation between the proposed IQ metrics with these degradation model parameters. Their study concluded that some of the parameters did show a strong correlation while some did not show any, as expected. It should be noted that though our proposed method is based on predicting human perception of document IQ, we also demonstrate that it is a general framework that can be extended to other applications such as predicting OCR performance for a specific OCR engine on a given document collection.

Some human evaluation studies have also been carried out to evaluate correlation of human preferences to different degradation models/features. Hale and Smith [20] investigated the correlation between the perceived IQ by nondocument specialists and quantifiable degradations based on Baird's degradation model. Their goal was to determine if human preferences coincide with the categorization that leads to improved OCR performance, which could imply that untrained human operators may make good decisions about how to acquire an image for input into an OCR package.

Our previous work [7] focused on evaluating human perception of degradation of character images and correlated it to known degradation parameters and existing IQ metrics. The conclusion derived was that these metrics that have successfully predicted OCR accuracy, according to literature, do not perform relatively well in predicting a human's perception of IQ. Hence, we need IQ metrics that specifically attempt to estimate that. This work extends the information derived from the study to design a system that can learn and thereby estimate human perception of document IQ. We propose to design a new quantitative metric for performance evaluation that takes into account how human users perceive quality.

## III. Human Perception Learning System

### A. Overview

When a human expert is presented with two document images of varying degradation levels and asked to determine which is of better quality, many complex factors affect the final decision. Normally, the document image selected to be of better quality is chosen because in the overall context, it looks better than the other document. Thus, the decision is made on

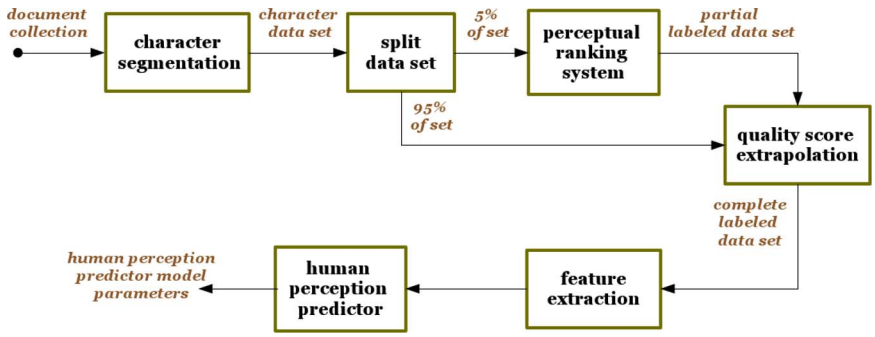


Fig. 1. Overview of the proposed methodology.

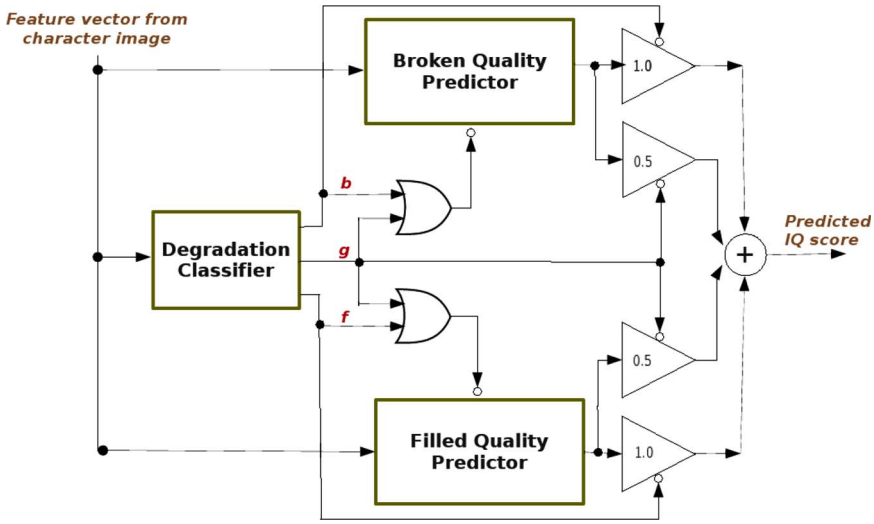


Fig. 2. Diagram of the ensemble system to predict IQ score.

a global level. It is intuitive that a degraded document image implies a high occurrence of degraded characters. Thus, to learn and estimate human perception of degradation, we perform our evaluation and analysis at the character level instead of page or paragraph block level. In this way, it is easier to obtain a consistent user discrimination at the character level. To obtain the final quality score for the entire image, we compute an average sum of the scores of the individual characters.

Fig. 1 shows a general overview of the proposed framework. The data set consists of extracted degraded character images from a given document collection via character segmentation. The process of generating our data set is described in Section V. The subjective quality experiment carried out to obtain perceptual ranking information using the online perceptual ranking system [7] is discussed in Section V-A. This information is used for training and validating our system. As shown in Fig. 1, only a subset of the training data set is labeled by the ranking system because of the expensive nature of subject experiments. However, using the quality score extrapolation method proposed in Section V-B, we generate labels for the entire training data set.

The proposed human quality perception predictor operates on a character image, which is represented by a feature vector $\vec{f}$. We are interested in developing a system that can estimate the perceptual ranking of the level of degradation in a document image without interpretation of the text. Thus, we compute a set of objective features from these character images, which is independent of the underlying character. We discuss these features in detail in Section IV. A full description of the process of training the predictor to generate the model parameters is presented in Section III-B.

### B. Ensemble Classifier–Predictor System

We propose an ensemble system composed of two stages to estimate human perception of level of degradation. The ensemble system consists of a degradation classifier in the first stage, and a set of two predictors (regressors) in the second stage, as shown in Fig. 2. The first stage is a classification stage to determine the main type of degradation in the image, since the framework does not make an assumption that the type of degradation is known *a priori*. In the second stage, the appropriate predictor is selected, based on the type of degradation identified in the former stage, to compute the IQ score. A two-stage framework is necessary because it is more accurate to train a predictor to estimate the level of degradation for a specific type of degradation than for a combination of different types of degradation. This is also demonstrated experimentally in Section VI. (This two-stage framework is similar to the one proposed by Moorthy and Bovik in [15] for no-reference IQ assessment based on natural scene statistics which also entails both classification and quality score prediction stages using support vector machines.) In this paper, we classify degraded

characters into two classes (filled and broken), as discussed in Section V, according to the predominant type of degradation exhibited in the image.

Each character image $i$ is represented by a feature vector $\vec{f_i}$, which is described in Section IV. The degradation classifier uses an MLP classification model to determine the main type of degradation in the image. The MLP [21] is a feed-forward neural network with one or more layers that are hidden from the input and output nodes. The model of each neuron in the network includes a nonlinear activation function that is differentiable such as the sigmoid. The units each perform a biased weighted sum of their inputs and pass this activation level through the transfer function to produce their output. For K-class classification, the MLP uses back propagation to implement nonlinear discriminants. There are $K$ outputs with softmax as the output nonlinearity.

We train our MLP classifier with the feature vectors of the labeled data set of good-quality, broken, and filled characters, employing one hidden layer. The classifier labels the input character image according to the main type of degradation, given its feature vector, and channels it to the appropriate degradation regression model in the second stage to predict the quality. If the character is identified as a good-quality character, i.e., having little or no degradation, then it is channeled to both predictors and the final assigned quality score is an equally weighted sum of the scores from both predictors. The proposed degradation classifier has $K = 3$ output labels: $g$ for the good character images, $b$ for the broken degraded images, and $f$ for the filled degraded images. Only one output can be activated per $f_i$ of the input image. If the classifier output label is $g$, then both predictors in the second stage are activated. A gain of 0.5 is applied to the output scores from each predictor, so the final sum is an average of both scores. If the classifier output label is either $b$ or $f$, then only one corresponding predictor is activated in the second stage. The final quality score is simply the active predictor's output score.

In the second stage, each model is trained to predict the quality of its specific degradation type using an MLP regression model. Essentially, we formulate IQ prediction in this stage as a regression problem based on its feature vector, using the MLP to find a mapping function between $\vec{f_i}$ and the quality score. In the MLP regression model, the output node approximates nonlinear functions of the input using a sigmoid. Training the MLP regressors involves finding the set of weight values that will minimize the prediction error made by the network using the backward propagation algorithm. Similar to the classifier, the MLP regression architecture consists of one hidden layer. By separating degradation models, our ensemble system is better suited to handle prediction of quality for a data set of mixed degradation types compared with a single predictor for a mixture of degraded characters, as found in a document image.

Our system predicts the level of degradation quality according to the main type of degradation. In this paper, we focus on the broken and filled degradation types. Both degradation types are not mutually exclusive; they both could possibly be found in a character image. However, we assume that one of them will be primarily dominant, so we make quality decisions based on the dominant degradation type.

## IV. Feature Extraction for Degradation Discrimination

Our goal is to compute a set of objective features $\vec{f_i} = \{f_n\}$ from each character image $i$ that can be fed into our learning system so that we can predict the level of degradation independent of the character image. These features are designed to capture different characteristics of the degradation. While a single feature may not be very discriminative, our premise is that the aggregation of these features will be discriminative. We leave it to the learning algorithm to determine the weight/contribution of each feature in predicting the IQ score.

We designed a set of 21 features as described hereinafter, which can be grouped under three categories: morphological-based features, noise-removal-based features, and spatial characteristic features. The following notations are used in defining the features. $I$ denotes the original $x \times y$ character image for which we desire to compute a given feature. $k_s^n$ represents a rectangular kernel of size $s \times s$ operated on itself $n$ times. The symbol $\#I$ denotes the number of foreground pixels in the image $I$. Background (white) pixels are assumed to have a value of zero, while foreground (black) pixels are assumed to have a value of one. $I_{(i,j)}$ denotes the $(i, j)$th pixel of the image $I$.

### A. Morphological-Based Features

Morphological operations (dilation, erosion, opening, and closing) have been demonstrated [22], [23] to be useful for salt-and-pepper noise removal in document images. Thus, we designed features based on these operations to capture degradations in the images that may be eliminated by them.

*Erosion:* The erosion feature $f_1$ is an iterative feature which computes the number of erosion operations it takes to completely erode all the foreground pixels in the image. It attempts to measure how thick the characters are as fattened stroke width usually implies some level of degradation. The erosion feature is computed as

$$f_1 = \min\left\{m | \left(I \ominus k_2^m\right)_{(i,j)} = 0,\ \forall i \in [0, x),\ j \in [0, y)\right\}. \tag{1}$$

$[0, x)$ is a standard set notation which implies that the index 0 is inclusive while $x$ is not.

*Dilation:* The dilation feature $f_2$ computes the number of dilation operations it takes to fill all the foreground holes in the image. It attempts to measure how thick the holes in the image are. To identify foreground holes, characters are dilated iteratively until there is no set of connected white pixels that is completely surrounded by black pixels. We compute this by inverting the image. In the inverted image, foreground holes are sets of foreground pixels not connected to the boundary of the bounding box.

The dilation feature is computed as the number of erosion operations needed to eliminate the foreground holes in the inverted image. Let $\tilde{I}$ denote the inverse of an image $I$, and let $\{C_i\}$ denote a set of connected components in $\tilde{I}$. $B$ represents foreground pixels at the edge of the bounding box of $\tilde{I}$. Let

$I^* = \bigcup\{C_i | C_i \cap B = \emptyset\}$. Thus, $f_2$ is defined in terms of the erosion operation on the inverted image as

$$f_2 = \min\left\{m | \left(I^* \ominus k_2^m\right)_{(i,j)} = 0,\ \forall i \in [0,x),\ j \in [0,y)\right\}. \quad (2)$$

Similarly, the horizontal dilation feature $f_3$ computes the number of dilation operations it takes to fill all horizontal holes, while the vertical dilation feature $f_4$ focuses on vertical holes. A horizontal (or vertical) hole is defined as a horizontal (or vertical for $f_4$) strip of background pixels bounded at both ends by a foreground pixel.

*Closing:* The closing feature $f_5$ counts the number of pixels changed by a single closing operation, normalized by the number of foreground pixels originally present in the image. This, in effect, attempts to measure foreground noise in the image, given that a closing operation is intended to fill small holes in an image. It is given by the quantity

$$f_5 = \left(\#\left(I \bullet k_3^1\right) - \#I\right)/\#I. \quad (3)$$

*Opening:* The opening feature $f_6$ counts the number of pixels changed by a single opening operation and normalizes it by the number of foreground pixels originally present in the image. An opening operation with a small kernel is intended to separate thin connected components in an image. Thus, this, in effect, measures background noise in the image. It is computed as

$$f_6 = \left(\#I - \#\left(I \circ k_3^1\right)\right)/\#I. \quad (4)$$

### B. Noise-Removal-Based Features

The median filter and the Gaussian smoothing are commonly used noise reduction techniques in image processing [23]. The median filter is used sometimes as a preprocessing step in binarization of document images. The median filter is a nonlinear digital filtering technique, often used to remove noise. Gaussian smoothing is the result of blurring an image by a Gaussian function. We designed a set of features using these filters to measure the amount of degradation removed as a result of applying these operations on the images.

*Gaussian:* The Gaussian features $\{f_n | n = 7, \ldots, 10\}$ measure the number of pixels changed by a Gaussian smoothing operation after thresholding the grayscale output to a binary image at a specified threshold value $\tau$. We obtain a set of features by thresholding at different levels ($\tau \in [120, 150, 180, 190]$) in our system. Each feature is normalized by the number of foreground pixels originally present in the image. Each Gaussian ($\tau$) feature is computed as

$$\underset{n\in[7,\ldots,10]}{f_n(\tau)} = 1 - \left(\#\left\{(i,j) | (I * K)_{(i,j)} > \tau\right\}\right)/\#I \quad (5)$$

where $K$ denotes a $3 \times 3$ Gaussian kernel with $\sigma = 0.95$. As shown in [24], this operation is equivalent to a generalized morphological operation that spans between erosion and dilation.

*Median:* Similar to the Gaussian feature, the median feature $f_{11}$ computes the number of pixels changed after applying the median smoothing operation, normalized by the number of foreground pixels present in the image. The median smoothing operation applies a median filter to the image using a $3 \times 3$ pixel window. In contrast to the Gaussian filtering, the median filtering operation results in a binary image output, thus eliminating the need to threshold the resulting image. Letting $N_{(i,j)}$ represent a $3 \times 3$ neighborhood of pixels in image $I$ centered at pixel $(i,j)$, the median feature is defined as

$$f_{11} = 1 - \left(\#\left\{(i,j) | \text{median}\left(N_{(i,j)}\right) > 0\right\}\right)/\#I. \quad (6)$$

### C. Spatial Characteristic Features

This group of features is designed to provide insight about the dimensional characteristics of the images. The goal is to obtain spatial peculiarities that may aid in identification of degraded characters.

*Foreground Percent:* The foreground percent feature $f_{12}$ simply computes the percentage of foreground pixels in the image given by

$$f_{12} = \#I/(x \cdot y). \quad (7)$$

*Image Gradient:* The image gradient features $\{f_n | n = 13, \ldots, 16\}$ attempt to provide information about the edges in the image. Intuitively, we expect a good-quality image to be smooth with a small number of sharp edges. We derive a set of four edge features, each normalized by the number of foreground pixels present in the image. The Gradient$^x$ feature $f_{13}$ and Gradient$^y$ feature $f_{14}$ are computed from the first-order $x$ and $y$ derivatives of the image, respectively. The features measure the number of pixels that have a value of one in the derived image. The Gradient$^{xy1}$ feature $f_{15}$ and Gradient$^{xy2}$ feature $f_{16}$ are computed by summing the magnitudes of the first-order $x$ and $y$ derivatives of the image. Feature $f_{15}$ measures the number of pixels that have a value of one in the derived image, while $f_{16}$ measures the number of pixels that have a value of two.

Let $G_x(i,j)$ and $G_y(i,j)$ denote the horizontal and the vertical first-order derivative of a pixel at position $(i,j)$ in image $I$ obtained using a $1 \times 3$ or $3 \times 1$ kernel, respectively. The image gradient features are computed as

$$f_{13} = \left(\#\left\{(i,j) | G_x(i,j) = 1\right\}\right)/\#I \quad (8)$$

$$f_{14} = \left(\#\left\{(i,j) | G_y(i,j) = 1\right\}\right)/\#I \quad (9)$$

$$\underset{n\in[15,16]}{f_n(\tau)} = \left(\#\left\{(i,j) | \left(G_x(i,j) + G_y(i,j)\right) = \tau\right\}\right)/\#I \quad (10)$$

where $\tau$ is set to one and two for $n = 15$ and 16, respectively.

*Connected Components:* The foreground connected-component feature $f_{17}$ computes the sum of connected components in the image given by

$$f_{17} = \#\{C_i\} \quad (11)$$

where $\{C_i\}$ is the set of connected components as before.

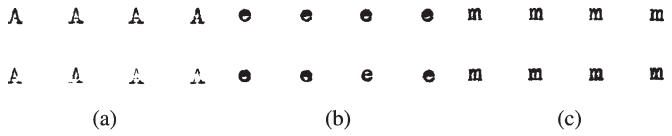


Fig. 3. Examples of common types of degradation that plague degraded documents. These are real degraded characters extracted directly from a document collection. (a) Broken: Faded ink. (b) Filled: Blotted ink. (c) Fattened stroke width.

*Length of Segments:* We compute the maximum and minimum lengths of both horizontal and vertical foreground segments along a line through the center of mass of the image. We obtain a set of four features: $\{f_n | n = 18, \ldots, 21\}$. Each feature is normalized by the number of foreground pixels present in the image.

Let $(r, c)$ be the center of mass of the image $I$. Let $R(i, j_1, j_2)$ be the sum of the pixel values on the $i$th row between columns $j_1$ and $j_2$: $R(i, j_1, j_2) = \sum_{j=j_1}^{j_2} I_{(i,j)}$. As already stated earlier, background pixels have a value of zero, while foreground pixels have a value of one. Thus, using $R(i, j_1, j_2)$, we define the maximum vertical segment feature $f_{18}$ and minimum vertical segment feature $f_{19}$ as

$$f_{18} = \max_{j_1 \le j_2} \{j_2 - j_1 + 1 | R(i, j_1, j_2) = j_2 - j_1 + 1\} / \#I \quad (12)$$

$$f_{19} = \min_{j_1 \le j_2} \{j_2 - j_1 + 1 | R(i, j_1, j_2) = j_2 - j_1 + 1 \wedge I_{(i,j_1-1)} + I_{(i,j_2+1)} = 0\} / \#I \quad (13)$$

where $i = r$. The maximum horizontal segment feature $f_{20}$ and minimum horizontal segment feature $f_{21}$ are defined similarly by setting $j = c$ and maximizing/minimizing over $i_1, i_2$.

## V. Data Extraction

The data set utilized consists of both real and artificially degraded character images extracted from different document images in our typewritten document collection [10] to span a wide range of degradations from good, clear, and legible characters to completely degraded and illegible characters. Typewritten documents have characteristic degradations of uneven text intensity. Some text is blurred and faint due to uneven typewriter key pressure or faded ink, while some appear blotted and filled due to the amount of force used in striking the typewriter keys [25] or overflow of the ink. We focus on three main types of degradation that commonly plague segmented typewritten document images: broken, filled (from shrinking white connected components), and fattened stroke width, as shown in Fig. 3. The fattened degradation is similar to the filled one, so we group both together under the filled degradation label.

Manual quality specification by a user is an expensive and labor-intensive process. Consequently, we limit the characters in the initial training set to a subset of characters which exhibit a large range of degradations. We describe in Section V-B how we expand this set to include the complete alphabet. The pool of real degraded character images selected include $\{a, e, s, f, g, m, n, A, E, F, M\}$, while the pool of artificially degraded characters include $\{s, A, F\}$. The set of real degraded character images is utilized for the validation of the proposed system, as described in Section VI-C. Each character

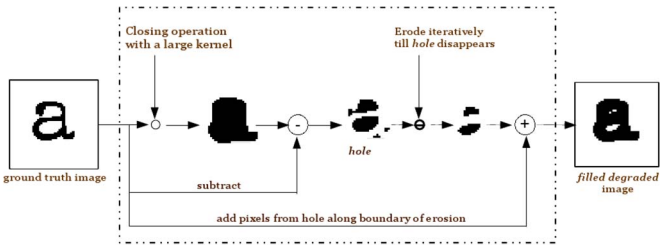


Fig. 4. Process of generating synthesized data sets of FILLED degradation.

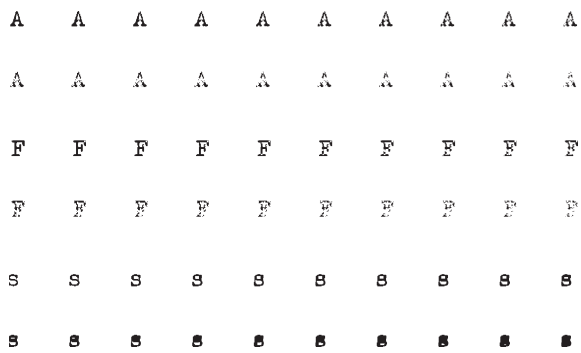

Fig. 5. Examples of synthesized data sets in order of increasing degradation. "A" and "F" represent *broken*-character degradation, while the "s" ones represent FILLED-character degradation.

set consists of 20 different images of the same character of varied degradation including the ground-truth character template. The ground-truth character template was extracted from ground-truth document images generated in our expert labeled data set using the semi-interactive document enhancement software [26].

We generated artificially degraded broken and filled character images from ground-truth images using the following degradation models.

1) *Broken characters*: To simulate broken-character degradation, we randomly select several windows of varying sizes in the ground-truth character image and flip the foreground pixels. The level of degradation is measured by the percentage of foreground pixels in the image relative to the total number of foreground pixels in the ground-truth image.
2) *Filled characters*: The process of obtaining filled characters of varying degree of degradation is shown in Fig. 4. Initially, we perform a closing operation with a large kernel on the ground-truth character until the entire image is completely closed. Then, the closed image is deducted from the original one to get a set of regions that can be filled which we define as *holes*. The *holes* are eroded iteratively with a small kernel until it is completely eroded. We store each set of pixels eroded for each iteration. To generate a filled character, we add pixels to the original character along the boundary of the hole to obtain a degree of filledness. The level of "filledness" is obtained by calculating the percentage of foreground pixels added relative to the total number of pixels that can be filled (i.e., the number of pixels in the *holes*).

Fig. 5 shows artificially degraded characters obtained for *broken* "A" and "F", and for *filled* "s" arranged in order of

increasing degradation from left to right, top to bottom. The first image in each set is the ground-truth image.

### A. Online Perceptual Ranking System

The purpose of our work is to design a system that can estimate a human perception quality score of the level of degradation in a document image using machine learning techniques. Thus, to obtain human input about perceptual quality to use to train our system, we conduct a subjective quality experiment. This provides the labeled data needed for training our system in addition to a labeled test set to evaluate and validate the proposed method. The subjective quality experiment was carried out via a web-based perceptual ranking system [7]. According to Sprow *et al.* [27], the Internet is very useful as a test platform in obtaining IQ judgment required by the human visual system. The study reveals close agreement between observers' preferences from the lab and the web, thus validating the use of the web as a time-effective approach compared to a lab-based environment.

The perceptual ranking system is designed such that the user can rank a set of degraded characters in order of decreasing IQ. Each set ranked consists of images of varying degradations from none to very poor for the same character image. To simplify the process for the user and thus ensure a more reliable judgment, the system displays one pair of character images in the set at a time during the run. For each pair of images, the user is asked to make a decision about their IQ relative to each other, i.e., decide whether the left or right image is better (or slightly better) or if they seem to be of identical quality. The images were randomly positioned in the left/right windows to prevent a bias toward a specific side. Based on the comparison information from each iteration during the run, the system sorts the list of characters in the set from best to worst quality using the bubble sort algorithm.

At the end of the sorting run, for verification purposes, the user is shown an evenly sampled subset consisting of half of the character images from the ranked character set arranged in increasing order of degradation according to the user's grading input. Again, to ensure a more reliable outcome from the user, only half of the set, evenly sampled, is displayed so that the user is not overwhelmed with verifying 20 images at once. The system automatically assigns each of the images a rank score between 10 (best) and 1 (worst) based on the user's sorting. The user is asked to modify these assigned scores to reflect his/her judgment. More than one image can share the same rank score.

The ranking system allows us to obtain a quantitative human perception metric of IQ with respect to a set of characters of varying degradation. Each character set consists of 20 different images of the same character of varied degradation including the ground-truth character template. The ground-truth character template was extracted from the ground-truth document images generated in our expert labeled data set using the semi-interactive document enhancement software [26]. The solicitation for users was done twice. In the first stage, we had 73 user inputs for our web-based ranking system obtained over a period of three weeks, while in the second stage, we had 57 user inputs, thus a total of 130 entries. The anonymous human users came from a pool of student volunteers from the Computer Science Department at the Illinois Institute of Technology. For each user's run, the character data set was randomly drawn from the pool.

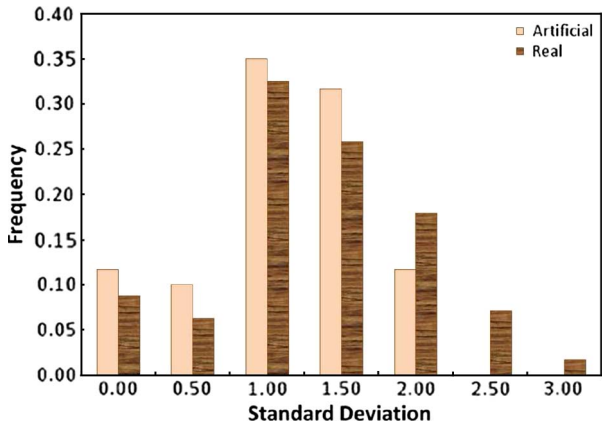


Fig. 6. Normalized histogram of the standard deviation of the rank scores assigned by the human experts for each of the degraded characters. We can observe that for both sets of degraded characters, the standard deviation peaks at one. This demonstrates that user judgments were consistent with both real and artificially degraded data sets.

Fig. 6 shows the distribution of the standard deviation of all user rankings obtained in our experiment, prior to removal of the outliers. We defined outliers as rank scores outside the $\pm 2\sigma$ range. The final user rank score for each character image is the **mean** rank score over the individual scores (excluding outliers) assigned by each user. We removed the outliers in the computation of the final user rank scores to improve the reliability of the mean score as the final user rank score. Seven rank scores were discarded as outliers out of a total of 39 000 scores. We can observe that for both sets of degraded characters, the standard deviation peaks at one. This demonstrates that user judgments were consistent with both real and artificially degraded data sets. We can also observe that the correlation among user rankings for the artificially degraded characters is stronger. It is easier to learn human preference which we can translate to a more complex combination that occurs in the real degraded characters. This demonstrates an additional advantage of using synthetic data, which is that the parameters can be controlled to span an intended range of degradation.

### B. Synthesized Labeled Data

It is well known that subjective quality evaluation experiments are very expensive and time consuming. According to the study done by Nonnemaker and Baird [28], the difficulty and cost of acquiring large training sets on real data can be alleviated by generating artificial synthetic samples. They show how synthetic data can be effective and useful for generating training data sets.

To overcome these obstacles and limitations inherent with subjective experiments, we build on the preliminary results [7] of human perception evaluation of text images where we demonstrated that there exists a strong correlation between degradation parameters (such as percentage of brokenness in a character) and human perception of quality (see Fig. 7). This

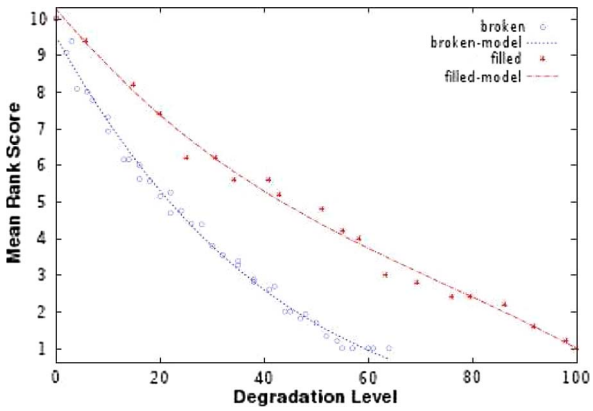


Fig. 7. Correlation of human perception of degradation to degradation parameters. We use polynomial models to best fit the data set for expansion of the training set.

allows us to generate a much larger training data set by using the labeled data to build a model that predicts human perception of quality based on the synthesized degradation parameter. Thus, using this model, we develop labeled data of the entire alphabet for broken and filled characters without having to solicit more human input which would be expensive to do for the entire alphabet. The broken and filled character sets each consist of both uppercase and lowercase characters.

The synthesis of training data is performed as follows. First, we compute a simple polynomial model to predict human perceptual rank based on a known degradation parameter using the current synthetic labeled data we have for each class of degradation. Next, we synthesize new training data for both broken and filled characters by applying the degradation models on all possible characters in the alphabet of our document collection. Finally, we use the polynomial model to associate a ranking with each example to generate a new set of labeled data.

Thus, using this model, we expand the *broken* labeled data set from 40 to 3190 instances. Likewise, we expand the *filled* set from 20 to 1031 instances. The experimental results shown in Section VI demonstrate the effectiveness of the synthesized labeled data. However, it should be noted that the proposed approach to expanding the data set by extrapolating quality scores using a regression model is a coarse approximation to conducting a human study. Although the technique has been justified by the literature [28], nevertheless, it is ideal, whenever possible, to conduct human studies to cover the entire data set. Hence, to verify and validate our proposed system in Section VI-C, we use test sets for which IQ scores have been obtained directly via human users with the perceptual ranking system.

## VI. Experimental Results and Analysis

### A. Experimental Setup

In this section, we evaluate the performance of our proposed perceptual metric estimator. To the best of our knowledge, there are no other existing human-perception-based document IQ evaluation metrics reported in literature that we can compare with. For the experiments, we generated the model parameters for the human perception predictor by training each component on the artificial data set described in Section V-B. We used a tenfold cross-validation method for training. The labeled training data set for broken characters consists of 3190 examples which include the ground-truth character images (also referred to as good-quality characters), while the labeled training data set for filled characters consists of 1031 instances. All these examples have perceptual IQ scores that range from 1 to 10. A score of 10 implies a very good legible character, while a score of 1, the minimum score assigned by the system, implies a character of very poor quality which is highly degraded.

TABLE I
Evaluation of Features on Degraded Characters

| Filled Model | | Broken Model | |
|---|---|---|---|
| Feature | RAE | Feature | RAE |
| Gaussian(190) | 0.289 | Foreground | 0.351 |
| Erosion | 0.386 | Median | 0.362 |
| Gradient$^{xy1}$ | 0.410 | Gaussian(120) | 0.399 |
| Gradient$^{x}$ | 0.431 | Opening | 0.419 |
| Gaussian(180) | 0.527 | Gaussian(190) | 0.437 |
| Gradient$^{y}$ | 0.587 | Connected Components | 0.490 |
| Foreground | 0.605 | Gradient$^{xy1}$ | 0.535 |
| Horizontal Dilation | 0.607 | Gaussian(150) | 0.545 |
| Gradient$^{xy2}$ | 0.658 | Gradient$^{y}$ | 0.608 |
| Min. vertical segment | 0.780 | Erosion | 0.687 |
| Max. horizontal segment | 0.791 | Gradient$^{x}$ | 0.721 |
| Min. horizontal segment | 0.802 | Gaussian(180) | 0.936 |
| Max. vertical segment | 0.857 | Dilation | 0.947 |
| Median | 0.883 | Gradient$^{xy2}$ | 0.962 |
| Opening | 0.902 | Horizontal Dilation | 0.998 |
| Vertical Dilation | 0.907 | Max. vertical segment | 1.051 |
| Dilation | 0.908 | Closing | 1.052 |
| Gaussian(150) | 0.925 | Min. vertical segment | 1.055 |
| Closing | 1.019 | Min. horizontal segment | 1.056 |
| Connected Components | 1.027 | Vertical Dilation | 1.056 |
| Gaussian(120) | 1.028 | Max. horizontal segment | 1.056 |

We evaluated each predictor model and the overall ensemble system using the relative absolute error (RAE) as a performance criterion, while we evaluate the degradation classifier in the first stage of the ensemble system using the accuracy measure. RAE is defined as the average of the error between the predicted value and the known value normalized by the known value for all the instances. RAE is computed as

$$\text{RAE} = \frac{1}{n}\sum_{k=1}^{n}\frac{|S_k - \hat{S}_k|}{S_k} \tag{14}$$

where $S_k$ is the known quality score for example $k$, $\hat{S}_k$ is the predicted score, and $n$ is the total number of examples in the test set. Using the RAE as an error metric enables us to understand how large/significant the error is in relation to the correct value. The RAE metric gives different weights to the prediction error in relation to the distance from the actual score. The accuracy measure used to evaluate the classification stage is defined as the percentage of correctly classified instances over the entire set of instances classified.

Section VI-B evaluates the discriminative power of each image feature utilized in characterizing the degradation of character images. To evaluate the system's performance, as described in Section VI-C, we tested on the real labeled degraded data set obtained from the subjective experiments in Section V-A. We evaluate the performance of the predictor

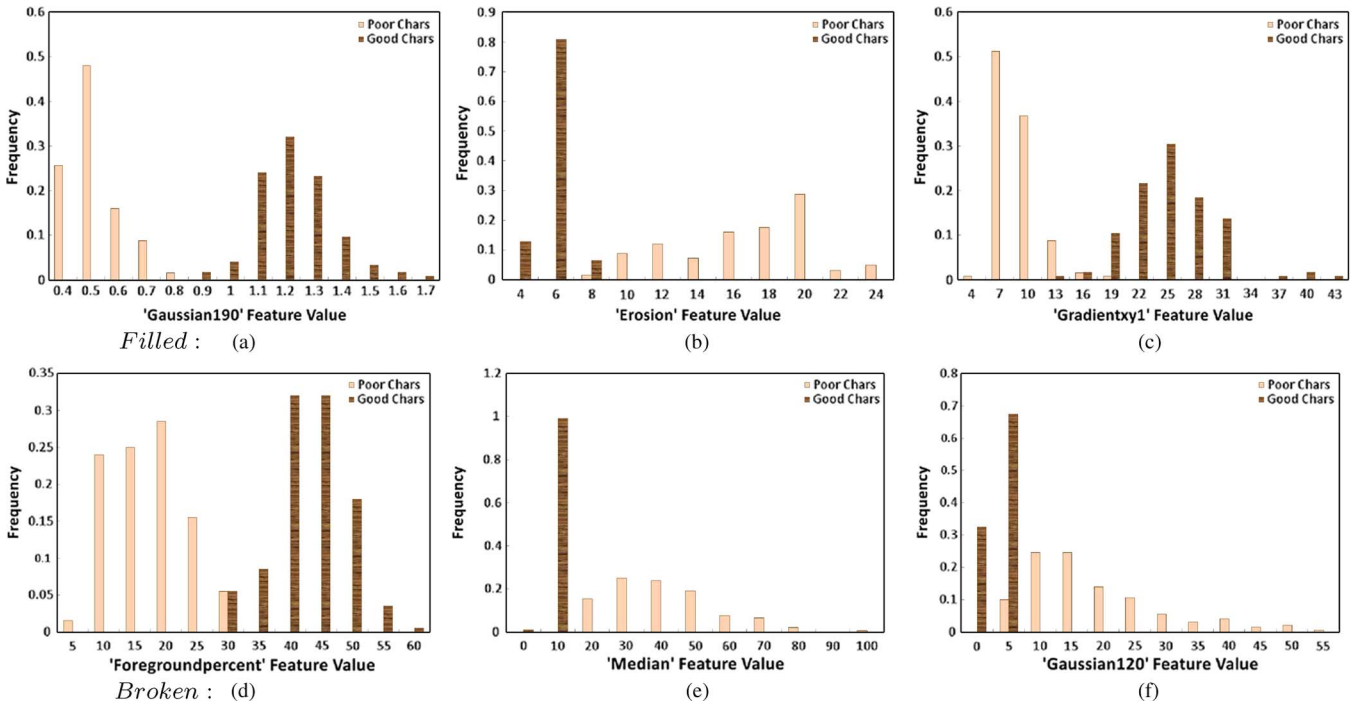


Fig. 8. Normalized histograms of the top three discriminative features for each class of degraded characters evaluated. (a) $f_{10}$. (b) $f_1$. (c) $f_{15}$. (d) $f_{12}$. (e) $f_{11}$. (f) $f_7$.

models and the classifier individually in addition to the overall ensemble system. To demonstrate the advantage of the two-stage classification and prediction framework, we also compare the performance of the ensemble estimator system to a single-layer predictor model. The single-layer predictor model was trained on both the filled and broken data sets.

TABLE II
PERFORMANCE OF PREDICTOR MODELS

| Data Set | Applied System's Performance | | | | |
|---|---|---|---|---|---|
| | $F_{RAE}$ | $B_{RAE}$ | $M_{RAE}$ | D-$C_{Acc}$ | $Ensemble_{RAE}$ |
| Filled | 0.250 | n/a | 0.274 | 92.00% | 0.249 |
| Broken | n/a | 0.233 | 0.286 | 96.43% | 0.237 |
| Perfect | 0.064 | 0.078 | 0.083 | 58.62% | 0.071 |

### B. Analysis of Degradation Features

A detailed description of each of the features extracted from the character images has been provided in Section IV. Our goal in this section is to provide a more meaningful insight of each of the features as per their discriminative power in modeling the degradation characteristics of the characters in a document image. This analysis is performed on the entire labeled training data set (i.e., the synthetic degraded data set along with the ground-truth images). We analyze the broken-character images separately from the filled ones.

Our objective is to analyze the order of importance of the discrimination power of each individual feature $f_n \in \vec{f}$ in estimating the IQ score of the degradation level of the character images. We generated 21 (given that $|\vec{f}| = 21$) sets of filled and broken predictor models, each built with a single-feature value $f_n \in \vec{f}$. These 21 pairs of models were trained and tested on the labeled training data set by a tenfold cross-validation method. Table I shows the performance of the single-feature models, as ranked by the RAE metric. (It should be noted that during training and testing, the IQ scores generated by the predictors were not bounded to [1, 10], as in the actual system, so an RAE score of $>1$ is possible.)

As can be observed from Table I and as expected, the top discriminating features differ for both the filled and broken models. Their order of importance of the features is not the same because the degradation patterns are manifested differently. Therefore, to capture various degradations, it is important to use a broad set of features. Although each feature itself may not be quite discriminative, it is expected that an aggregation of the features will perform much better. The learning algorithm is also capable of ignoring the redundancy that occurs in overlapping features.

To visually observe the quality of the features, we construct two pruned sets from the training data set: one for characters with a filled degradation and the other for characters with a broken degradation. The poor-quality characters were selected from one end of the degradation spectrum (with a score below 2.5), while the good-quality images were from the other end of the spectrum (with a score above 8.5). Each test collection was constructed so that the number of degraded characters matches the number of good-quality characters. For each of the top three best performing features, we plot the feature distribution in Fig. 8. We can observe in Fig. 8 that fundamentally, the features have different values, depending on the level of degradation of the character image.

### C. Component Evaluation of the Ensemble System

We validate the performance of our system using a test collection obtained from subjective experiments on real degraded characters. The test collection contains real degraded


Fig. 9. Performance of the ensemble system: The real degraded data set is arranged in order of decreasing IQ according to the users' ranking. The lower score is the estimated perceptual ranking according to our system.

characters (220 images: 20 instances of 11-character images: $\{a, e, s, f, g, m, n, A, E, F, M\}$), which were ranked by human experts and assigned a perceptual IQ score between 1 and 10. The first seven characters suffered predominantly from filled degradation, while the latter four images underwent mainly broken degradation. Table II summarizes the performance of individual components in our system tested on filled, broken, and undegraded (perfect) characters. In this table, $\mathrm{F_{RAE}}$ and $\mathrm{B_{RAE}}$ denote the RAEs of the filled and broken quality predictors, respectively. $\mathrm{M_{RAE}}$ denotes the RAE of a mixed quality predictor that is trained on a mixture of broken and filled degradations. D-$\mathrm{C_{Acc}}$ denotes the accuracy of the degradation-type classifier, and $\mathrm{Ensemble_{RAE}}$ denotes the RAE of the complete two-stage system.

From Table II, we can observe that the proposed ensemble system performance ($\mathrm{Ensemble_{RAE}}$) is comparable to that of the individual predictor models ($\mathrm{F_{RAE}}$ and $\mathrm{B_{RAE}}$), which assume that the degradation in the character image is known beforehand. This is because the degradation-type classifier in the first stage of the ensemble system, as demonstrated by D-$\mathrm{C_{Acc}}$ of over 90%, is very accurate in identifying the type of degradation in the character image. Although the degradation-type classifier is not as accurate in identifying good characters (D-$\mathrm{C_{Acc}}$ is 58.62%), both predictor models in the second stage make up for its weakness. The final IQ score for good characters is an average of the scores from both predictor models. The relatively low classification accuracy of the degradation-type classifier for good/perfect character images is probably due to the very limited training samples of perfect-quality character images compared to the number of samples available for the degraded images. Table II also demonstrates, as expected, that


Fig. 10. Sample of synthesized combination images processed by the Tesseract OCR engine. Each character image is adjoined to "Figure" to overcome the engine's inability to process individual character images.

it is more effective to apply a two-step process, as done in the proposed ensemble system ($\mathrm{Ensemble_{RAE}}$), in predicting the IQ score compared to build a general predictor model in one layer ($\mathrm{M_{RAE}}$).

Some visual/qualitative results given by our ensemble system are illustrated for a subset of the test set in Fig. 9. The top portion of each diagram illustrates the visual ranking and scores assigned by the users, while the lower portion shows the predicted scores. As can be observed in the figure, the predicted results are close to the known result and correlate to the actual perception when looking at the characters. We also observe that though our system tended to rank the IQ scores lower than the actual, it was consistent with majority of the images. It predicted the score of the "F" good-quality characters poorly. This is probably due to the sharp edges in the character. We plan, in future work, to broaden our feature vectors to account for these types of characteristics to improve the predictor's performance.

## VII. Adaptation of the Proposed Model to Predicting OCR Accuracy

A key strength of our proposed methodology is that it can be extended to other prediction domains. We demonstrate an example in this section by applying our model to the prediction

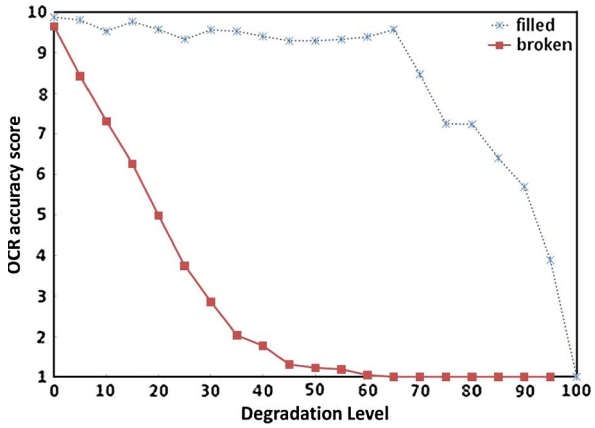


Fig. 11. Performance of the OCR engine with respect to the level of degradation in the synthesized characters.

of OCR accuracy using the Tesseract OCR engine [29]. The purpose of the experiment is to show that the features extracted in Section IV can also be used to learn and predict the OCR accuracy of character images. The experiments reported in this section were carried out using a tenfold cross-validation method on the synthesized data set used in Section VI-B experiments. Given that the Tesseract OCR engine is not fully equipped to function on single-character images, we adjoined each character to the same word image ("Figure" in our experiments), as shown in Fig. 10, and evaluated the OCR performance on the adjoined character alone. To disable the dictionary model in Tesseract, we ran the experiments using the settings suggested by Sturgill and Simske [30] for OCR without dictionary-based corrections.

We use the measure $OCR_{\text{score}}$ to quantify OCR recognition accuracy per level of degradation for the character images. It is defined as follows: $OCR_{\text{score}} = 1 + 9 * R/T$, where $R$ is the number of characters recognized correctly by the OCR engine among a group of $T$ characters, in total, that share the same level of degradation. Thus, a score of 10 implies that the engine accurately predicted all instances of the character images with the same level of degradation, while a score of 1 implies that all the predictions of the engine for that group were false. Fig. 11 shows the performance of the OCR engine as a function of the character's degradation level. We can observe that the engine is more resilient to filled degradation compared to broken degradation, so it often reaches the maximum value for OCR accuracy on filled characters. It has already been observed in the literature [8] that OCR engines are more susceptible to broken-character degradation and not necessarily to filled-character degradation. Given that we thicken and fill each character separately, we preclude the cases where fattened stroke widths lead to touching-character degradation which usually degrades OCR performance. (To account for this using our model, special character segmentation for touching characters should be applied and $\vec{f}$ extended with additional relevant features.) However, the sole purpose of Fig. 11 is to demonstrate that the relationship between OCR performance and degradation level is nonlinear.

Our goal is to extend our proposed model to predict the OCR accuracy of character images as defined by the $OCR_{\text{score}}$. We used a tenfold cross validation on each set (broken and filled) to train and test the OCR prediction model with the same MLP architecture and feature vector used for the proposed system. Table III shows the performance of the system in predicting OCR accuracy. As can be observed, the system performs very well in predicting OCR accuracy. The performance of our model in predicting OCR accuracy is higher compared to prediction of the IQ scores shown in Table II. This is because the data tested in this paper is synthesized rather than the actual degraded data set utilized in the experiments analyzed in Table II. The proposed system works well because the degradation-type classifier is very accurate in channeling the images to the proper degradation-specific predictor model.

TABLE III
PERFORMANCE OF OCR ACCURACY PREDICTOR MODEL

| Data Set | Applied System's Performance | | | |
|---|---|---|---|---|
| | $F_{RAE}$ | $B_{RAE}$ | D-$C_{Accuracy}$ | $Ensemble_{RAE}$ |
| Filled | 0.116 | n/a | 97.21% | 0.119 |
| Broken | n/a | 0.167 | 98.79% | 0.169 |

## VIII. CONCLUSION

Predicting user perception is an important task in numerous applications. We have presented an MLP-based ensemble framework for learning and predicting human perception of document IQ. Our system is trained using human input derived from an online perceptual ranking system. We defined a set of features that is used to obtain a measure of degradation quality for character images. The experimental results demonstrate the effectiveness of our proposed human perception predictor model. Our model works because the designed features capture diverse perceptual qualities of degradation and the learning algorithm successfully utilizes the information from the aggregation of these features to predict perceived quality. We plan to broaden our feature set to further improve the performance of the model. The proposed approach is general and can be easily adapted to a wide range of applications where predicting user performance is required. We demonstrate experimentally how it can naturally be extended to the prediction of OCR accuracy for character images.

## REFERENCES

[1] P. Stubberud, J. Kana, and V. Kallurit, "Adaptive image restoration of text images that contain touching or broken characters," in *Proc. 3rd ICDAR*, 1995, vol. 2, pp. 778–781.
[2] J. He, Q. Do, A. Downton, and J. Kim, "A comparison of binarization methods for historical archive documents," in *Proc. Int. Conf. Document Anal. Recognit.*, 2005, vol. 1, pp. 538–542.
[3] B. Gatos, I. Pratikakis, and S. J. Perantonis, "Adaptive degraded document image binarisation," *Pattern Recognit.*, vol. 39, no. 3, pp. 317–327, Mar. 2006.
[4] T. Obafemi-Ajayi, G. Agam, and O. Frieder, "Ensemble LUT classification for degraded document enhancement," in *Proc. SPIE—Document Recognition and Retrieval XV*, B. Yanikoglu and K. Berkner, Eds., 2008, vol. 6815, p. 681 509.
[5] M. Sezgin and B. Sankur, "Survey over image thresholding techniques and quantitative performance evaluation," *J. Electron. Imaging*, vol. 13, no. 1, pp. 146–165, Jan. 2004.
[6] E. Kavallieratou and E. Stamatatos, "Improving the quality of degraded document images," in *Proc. Int. Conf. DIAL*, 2006, pp. 340–349.
[7] T. Obafemi-Ajayi, G. Agam, and O. Frieder, "Evaluation of human perception of degradation in document images," in *Proc. SPIE—Document*

*Recognition and Retrieval XVII*, L. Likforman-Sulem and G. Agam, Eds., 2010, vol. 7534, p. 753 40T.
[8] L. Blando, J. Kanai, and T. Nartker, "Prediction of OCR accuracy using simple image features," in *Proc. 3rd ICDAR*, 1995, vol. 1, pp. 319–322.
[9] M. Cannon, J. Hochberg, and P. Kelly, "Quality assessment and restoration of typewritten document images," *Int. J. Document Anal. Recognit.*, vol. 2, no. 2/3, pp. 80–89, 1999.
[10] The Diaries of Rabbi Dr. Avraham Abba Frieder. [Online]. Available: http://ir.iit.edu/collections/
[11] A. Bouzerdoum, A. Havstad, and A. Beghdadi, "Image quality assessment using a neural network approach," in *Proc. 4th IEEE Int. Symp. Signal Process. Inf. Technol.*, 2004, pp. 330–333.
[12] H. Tong, M. Li, H. Zhang, C. Zhang, J. He, and W. Ma, "Learning no-reference quality metric by examples," in *Proc. 11th Int. Multimedia Modelling Conf.*, 2005, pp. 247–254.
[13] J. Kim, M. Cho, and B. Koo, "Experimental approach for human perception based image quality assessment," in *Proc. 5th Int. Conf. Entertainment Comput.*, 2006, vol. 4161, pp. 59–68.
[14] M. Narwaria and W. Lin, "Objective image quality assessment based on support vector regression," *IEEE Trans. Neural Netw.*, vol. 21, no. 3, pp. 515–519, Mar. 2010.
[15] A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," *IEEE Signal Process. Lett.*, vol. 17, no. 5, pp. 513–516, May 2010.
[16] E. N. Dalal, E. H. B. Smith, F. Gaykema, A. Haley, K. Kirk, D. Kozak, M. Robb, T. Qian, and M. Tse, "INCITS W1.1 standards for perceptual evaluation of text and line quality," in *Proc. SPIE—Image Quality and System Performance VI*, S. Farnand and F. Gaykema, Eds., 2009, vol. 7242, p. 724 203.
[17] V. Govindaraju and S. N. Srihari, "Image quality and readability," in *Proc. Int. Conf. Image Process.*, 1995, vol. 3, pp. 324–327.
[18] H. S. Baird, "Document image defect models and their uses," in *Proc. 2nd ICDAR*, 1993, pp. 62–67.
[19] D. K. Reed and E. H. B. Smith, "Correlating degradation models and image quality metrics," in *Proc. SPIE—Document Recognition and Retrieval XV*, B. Yanikoglu and K. Berkner, Eds., 2008, vol. 6815, p. 681 508.
[20] C. Hale and E. H. B. Smith, "Human image preference and document degradation models," in *Proc. ICDAR*, 2007, pp. 250–254.
[21] E. Alpaydin, *Introduction to Machine Learning*. Cambridge, MA: MIT Press, 2004.
[22] K. Chinnasarn, Y. Rangsanseri, and P. Thitimajshima, "Removing salt–pepper noise in text/graphics images," in *Proc. IEEE APCCAS*, 1998, pp. 459–462.
[23] L. G. Shapiro and G. Stockman, *Computer Vision*. Upper Saddle River, NJ: Prentice-Hall, 2001.
[24] G. Agam and I. Dinstein, "Regulated morphological operations," *Pattern Recognit.*, vol. 32, no. 6, pp. 947–971, May 1999.
[25] A. Antonacopoulos and D. Karatzas, "A complete approach to the conversion of typewritten historical documents for digital archives," in *Proc. IAPR Int. Workshop DAS*, 2004, pp. 8–10.
[26] G. Bal, G. Agam, G. Frieder, and O. Frieder, "Interactive degraded document enhancement and ground truth generation," in *Proc. SPIE—Document Recognition and Retrieval XV*, B. Yanikoglu and K. Berkner, Eds., 2008, vol. 6815, p. 681 50Z.
[27] I. Sprow, Z. Baranczuk, T. Stamm, and P. Zolliker, "Web-based psychometric evaluation of image quality," in *Proc. SPIE—Image Quality and System Performance VI*, S. Farnand and F. Gaykema, Eds., 2009, vol. 7242, p. 724 20A.
[28] J. Nonnemaker and H. Baird, "Using synthetic data safely in classification," in *Proc. SPIE—Document Recognition and Retrieval XVI*, K. Berkner and L. Likforman-Sulem, Eds., 2009, vol. 7247, p. 724 70G.
[29] R. Smith, "An overview of the Tesseract OCR engine," in *Proc. Int. Conf. Document Anal. Recognit.*, 2007, vol. 2, pp. 629–633.
[30] M. Sturgill and S. Simske, "An optical character recognition approach to quantifying thresholding algorithms," in *Proc. ACM Symp. DocEng*, 2008, pp. 263–268.

**Tayo Obafemi-Ajayi** received the Ph.D. degree in computer science from the Illinois Institute of Technology, Chicago, in 2010.

She is currently a Postdoctoral Fellow with the Computer Graphics and Image Understanding Lab, University of Missouri, Columbia. Her research interests include document imaging, computer vision, human–computer interaction, and machine learning.

**Gady Agam** received the Ph.D. degree in electrical and computer engineering from Ben-Gurion University of the Negev, Beersheba, Israel.

He is currently an Associate Professor of computer science with the Illinois Institute of Technology, Chicago, where he directs the Visual Computing Lab. His current research interests include document imaging, medical imaging, geometric modeling, human–computer interaction, and machine learning.