Visual Aesthetic Analysis for Handwritten Document Images

Anshuman Majumdar, Praveen Krishnan and C.V. Jawahar
CVIT, IIIT Hyderabad, India
anshuman.majumdar@students.iiit.ac.in, praveen.krishnan@research.iiit.ac.in, jawahar@iiit.ac.in

***Abstract*—We present an approach for analyzing the visual aesthetic property of a handwritten document page which matches with human perception. We formulate the problem at two independent levels: (i) coarse level which deals with the overall layout, space usages between lines, words and margins, and (ii) fine level, which analyses the construction of each word and deals with the aesthetic properties of writing styles. We present our observations on multiple local and global features which can extract the aesthetic cues present in the handwritten documents.**

***Keywords*-Handwritten document images, aesthetic analysis, local features**

I. Introduction

What qualifies as a neat handwritten document? Often our definition of the neatness or the beauty of handwritten text and document is subjective. However, most human beings subjectively agree on these. In this paper, we study this subjective aspect [1] and explore how one can capture this with a set of features. Figure 1 shows a sample page image from our corpus where we are interested in giving a score in terms of neatness of layout and handwriting style.

Visual aesthetic analysis is getting popular in many domains. Researchers have been trying to understand the notion of aesthetics with the hope of reverse engineering these definitions with machine learning solutions [2], [3], which try to validate the photographic rules and compute features using color information. Our work follows a similar strategy, but in the domain of handwritten textual documents.

A subjective estimate of the visual aesthetic features such as neatness of handwriting has many practical applications. For many decades, this has been used in education systems (in many countries) for promoting good handwriting by giving bonus points for neatly written solutions. In addition to this, there have been many studies and axioms that relate the quality and consistency of the handwriting to the personality of the individuals [4]. An automatic estimate of a neatness measure can also be of immense use in education in the online setting [5]. In the case of printed documents, there have been many studies on the quality of document [1], [6], [7], degradation and how they relate to the accuracy of the recognition systems. Even for handwritten documents, one could estimate such correlations. However, the notion of aesthetics and quality that are appealing to humans and that favor a machine recognition need not be the same.

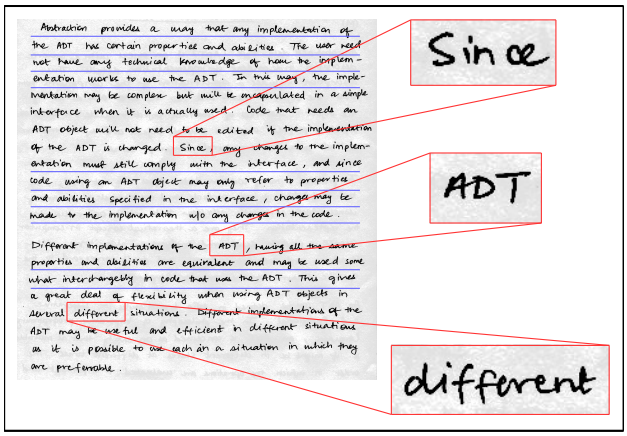

Figure 1. Given a sample handwritten page, we are interested in rating the quality in terms of aesthetic factors such as writing neatness, spatial arrangement of content, etc. We demonstrate the effectiveness of the proposed method in rating handwritten documents.

Ours is possibly one of the first attempts to capture the human notion of neatness and aesthetics in handwritten documents. While, in a way, aiming at imitating a task that humans do quite comfortably, with an automated process, we also provide a simple computational solution which provides reasonable results.

We believe that aesthetics gets defined at two levels in handwritten documents. First is at a fine level, where the local properties of the writing plays a critical role. This is somewhat similar to the degradation models that were used in printed documents. However, due to the nature of handwriting, these features are not purely pixel flips or erosion on the boundaries. Figure 2(a) and 2(b) show the important aesthetic cues which discriminate one style from the other. The second property is at a coarse level, where one captures the distribution of the ink throughout the page. This dimension is relatively absent in the printed documents since they are typeset with software or electro-mechanical systems which manage the space between characters, words and paragraphs, and aligns reasonably well with the paper and the neighboring lines.

We start by capturing the subjective definitions of neatness by taking inputs from multiple human beings. Note that this leads to a dataset which has multiple valid annotations. For each of the documents, we capture a score in the range of 1 to 5 with a subjective label of Poor, Fair, Average, Good and

Excellent. Our first observation has been that the evaluations by human beings is similar by and large. In Figure 5, we analyze the human agreements on the aesthetic property of each document and observe that these variations are small. We also observe that there are more deviations in the scoring of average quality documents as compared to the excellent or poor ones. Using this human assessment for the quality of documents, we are interested in training a SVM model that can be used to predict the human judgments on the aesthetic quality of handwritten documents. We also model the task of replicating human judgments on neatness as a comparative study between pairs of words and documents.

A. Related Works

One can broadly categorize the literature in document quality assessment into two categories: (i) content based [8]–[10], and (ii) degradation based [6], [11], [12]. In our work, the content deals with the inherent property of the document and relates to the aesthetic aspects. On the other hand, degradation models deal with the production process (scanning, page quality) of the documents. In most scenarios, the end user of a content based system is a human being while for degradation models it is a machine. For example, the performance of an optical character recognition (OCR) is highly dependent on the type of degradation a document undergoes. Works such as [8], [11] use heuristic features measured on connected components, stroke width distribution and so on. With the popularity of local features such as SIFT and bag of words framework, unsupervised techniques like [6], [12] gave promising results for predicting OCR accuracy. Since the criteria of evaluation of the above methods is in prediction of OCR accuracy of printed documents, these methods are not directly applicable for aesthetics analysis of handwritten document image. A recent survey paper from Ye et. al. [1] summarizes various works in the document quality assessment. More recently in the domain of natural scene images, visual aesthetic analysis for search and retrieval of high quality images is becoming popular [2], [3] which uses typical photographic rules and generic image descriptors.

The paper is organized as follows: In Section II, we explore and validate popular features from document image processing and computer vision for analyzing the aesthetic properties of a handwritten image at the level of word and page. In Section III, we present our in-house dataset, quantitative evaluation of various features and analysis. Finally, we conclude in Section IV and present the future scope of the current work.

II. Handwritten Document Image Aesthetics

Aesthetics is closely related to art and is a quite subjective topic where different individuals have different opinions. Even in this subjective scenario, one can observe that there are a few cues or patterns which are common among a large group and can be exploited for our purpose. In this work, we are interested in categorizing a handwritten document image into multiple groups based on its aesthetic nature. We divide our problem into two sub-categories referred to as fine and coarse grained analysis. Fine grained properties are related to the rendering of individual characters and words, and are typically restricted in a local neighborhood. Figure 2(a) and 2(b) show instances of word images where one can observe fine details such as the smoothness of strokes and details in the rendering of each character present in the word. Another interesting property is the nature of repeatability of similar strokes and its consistency. Similarly, coarse grained properties of a document image verify the consistency of the space usage in arrangement of paragraphs, line, word gaps and margins across the borders of the page, and in general, related to the layout of the page, which can be seen in Figure 2(c). Since we are attempting a novel problem and to our knowledge there has been no prior attempt to compare, we pose this problem in a typical supervised classification and regression setting and analyze the performance of various features.

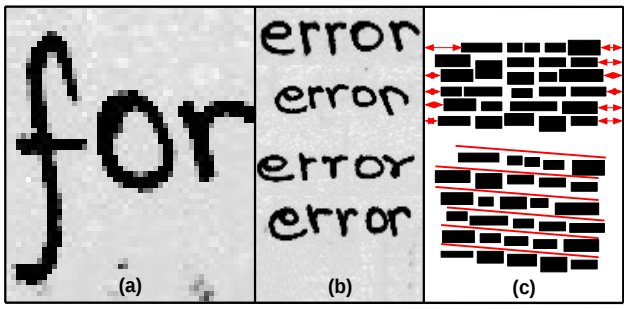

Figure 2. The aesthetic cues of a document image are: (a) smoothness of strokes, (b) nuances in character formation, (c) consistency of white spaces between lines, words, and paragraphs, along with the margin and layout of the text.

A. Fine Grained Assessment

In fine grained analysis, we deal with local features which operate on raw pixels, gradient of an image and the texture properties. Since most of these local features will lead to variable length feature representation, we used the standard BoW [13] encoding scheme using Gaussian Mixture Models (GMM) for learning the visual vocabulary and Fisher vectors (FV) [14] to aggregate the features.

Connected components (CC) [8]: These features are computed on a binary image by using Otsu thresholding [15]. These include mean height and width of CC's, mean pixel density and the total number of CC's.

Stroke based features [10], [16]: We computed stroke based features using: (i) stroke density distribution (SDD) and (ii) stroke width transform (SWT). SDD has been used in the past for measuring the quality of handwritten character samples [10]. In order to use it for document pages, we apply a fixed dimension sliding window over the text regions, and

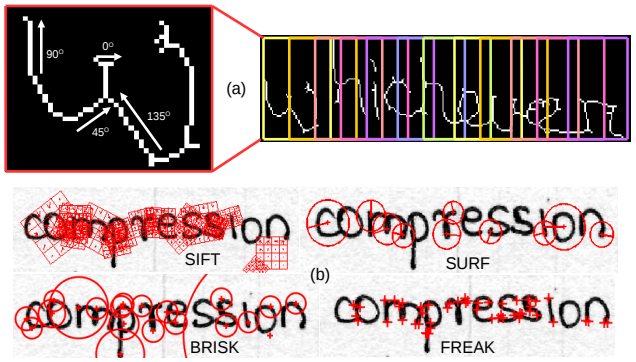

Figure 3. (a) Computation of stroke density distribution feature on a sliding window from word image. (b) Keypoint detection over a sample word image using SIFT, SURF, BRISK and FREAK.

for each window, a histogram of pixel density is computed in four different orientations ($0°$, $45°$, $90°$ and $135°$). Figure 3(a) shows the feature descriptor calculation on a particular word image. These local features were appended with the mean and variance of corresponding stroke width of the window using SWT [16]. We aggregate the features by forming a histogram of visual words. Here visual words correspond to the cluster centers of the above mentioned local features.

Gradient features [17]–[19]: Scale invariant feature transform (SIFT) is one of the most popular local feature which has been successfully used in numerous vision applications. SIFT uses difference of Gaussian (DoG) as the interest point detector which returns spatial x, y location, scale and orientation from the image. The SIFT descriptor is calculated as a histogram of gradients from spatial bins centered at the detected keypoint. We also used GIST descriptor which derives a holistic representation from an entire document image by convolving Gabor filters at multiple scales. They also summarize the gradient information from different regions of the image. Note that the SIFT keypoints are aggregated using Fisher vectors while there is no aggregation in the GIST descriptor which by default gives the global description. Another related feature that we used is speeded up robust features [19] (SURF), which is not exactly a gradient feature but uses a fast Hessian detector and sum of wavelet responses as a feature descriptor.

Binary features [20], [21]: Another class of local descriptors are binary descriptors such as BRISK and FREAK which encode the information of a patch as a binary string by comparing the intensity distribution. Figure 3(b) shows keypoint detection using various detectors such as SIFT, SURF, BRISK and FREAK. Here, we observed the SURF keypoints to be more sparse as compared to other detectors.

Texture features [22]: In texture features, we use local binary pattern (LBP), which is extremely fast to compute and uses the central pixel of a patch as a threshold for its surrounding neighborhood pixels.

B. Coarse Level Assessment

The intuition of coarse level features is to analyze the space usages in the entire document image, which includes space between lines, words, paragraphs and margins across the top and bottom. Here also, we will explore some statistical features and a global feature.

Statistical features: These features are computed on word-segmented images. We have used a multi-stage bottom-up approach for segmentation similar to [23] which joins adjacent neighboring CC's present in document image. The following features are extracted from the word bounding boxes (BB's): (i) mean and variance of left and right most bounding box which signifies margin, (ii) mean and variance of word gaps which is interpreted as the consistency among the spacing of words, (iii) mean and variance of height of BB's, and (iv) mean and variance of horizontal profiles [24] on left and right sides of the document.

Fourier features: We use discrete Fourier transform of an image and extract features from the corresponding power spectrum. We use it as a means of measuring the regularity in the arrangement of words in the document. Therefore, as a pre-processing step, the document image is binarized and the words in the document image are masked by using tight bounding boxes. The spectrum obtained is then divided into multiple bins and the histograms obtained from each of the bins are concatenated.

C. Modeling and Prediction

We model the aesthetic features as a simple linear vector, $\mathbf{x}$ with its elements as the features described in the previous section. The task of predicting the word and page level scores was carried out using two experiments: (i) regression, for predicting the mean score given by the users for each document, and (ii) classification, for characterizing each document into the five neatness scales given by the users. The performance of regression is quantified using mean squared error (MSE) and linear correlation coefficient (LCC). Here MSE captures the deviation of our predictions from the human rating, while LCC compares the strength of association between our predictions and the human ratings, respectively. The performance of classification is quantified using average accuracy of characterizing the documents in their respective neatness bins.

In another setting, we tried to formulate the problem as a ranking task where we perform a relative comparison among two document image representations and predict the more aesthetic document among them. Here we use the ranking loss function [25] instead of $0/1$ loss where the ranking pairs are induced from the annotations given by the users. The motivation behind this experiment is to further reduce the subjectivity in annotation by showing a pair of documents and asking the question, "Which is a better document in terms of aesthetics?"

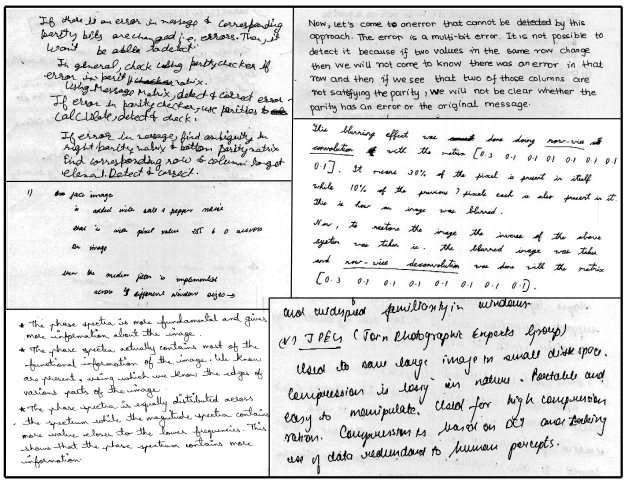

Figure 4. Sample page images from our dataset. Notice the variation of handwritten documents in terms of the clarity of word formation, layouts and usage of spacing, which makes the problem challenging.

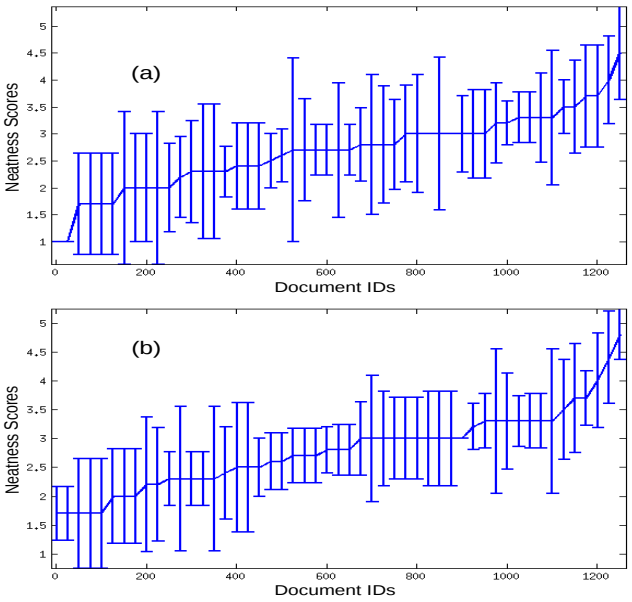

Figure 5. Analysis of human subjectivity for document aesthetic annotation at the level of (a) word neatness, and (b) page neatness.

III. Experiments & Analysis

A. Dataset

We test on an in-house dataset of 1.2K handwritten pages from nearly 100 writers. Every writer was given some topics to write upon, from a set of selected topics, without any constraints on time, amount of text or the content. As a result, we obtained a dataset which captured different handwriting styles, irrespective of the content. The documents were annotated on the basis of their neatness using a 5-point Likert scale given as poor, fair, average, good and excellent, with poor denoting the worst and excellent denoting the best quality for assessment. For each document, two neatness annotations were collected (each using the aforementioned Likert scale): (i) word neatness label, describing the quality of strokes and the handwriting (for a fine level assessment of strokes), and (ii) page neatness label describing the overall alignment of words, lines and paragraphs in the document (for a coarse level assessment of the document). In total we employed 18 human annotators and each page in the dataset was annotated by at least 3 different individuals. Figure 4 shows some sample page images from our corpus. One can observe the variation among different documents which makes the problem challenging.

The dataset so obtained was split in a 60-40 ratio, mutually exclusive of the writers, for training and testing respectively. During the annotations, we faced some human subjectivity issues in the labeling of the documents using the absolute 5-point scale. Figure 5 shows the distribution of mean scores for each document along with their standard deviations depicted as vertical bars. These variations arose mainly due to the differences in human perceptions and the absolute nature of rating the documents, using the 5-point neatness scale.

Table I
Quantitative evaluation of various features for analyzing word level neatness: Fine level assessment

Type	Features	Accuracy	MSE	LCC
Raw	CC [8]	44.11%	0.656	0.221
Raw	SDD [10] + SW [16]	43.13%	0.518	0.346
Binary	BRISK [20]	48.03%	0.466	0.508
Binary	FREAK [21]	46.47%	0.351	0.585
Texture	LBP [22]	50.19%	0.342	0.632
Gradient	SIFT [17]	50.58%	0.338	0.641
Gradient	GIST [18]	40.39%	0.425	0.513
Gradient	SURF [19]	**55.88%**	**0.308**	**0.682**

B. Experimental Setup

In all our experiments, we have extracted a global feature vector from the document image. The features that are computed locally are aggregated using Fisher vector on the visual vocabulary learned using Gaussian Mixture Models (GMM). We have kept the number of clusters fixed at 80, which was set empirically. For classification experiments, we used linear support vector machines (SVM). For regression, we used kernel SVM using RBF kernel. The SVM parameters are learned using cross validation.

C. Fine Level Assessment

Table I shows the performance evaluation using various feature descriptors as mentioned in Section II-A. The baseline results obtained using raw pixel level features such as connected components (CC) features, stroke density distribution (SDD) and stroke width (SW) were inferior to

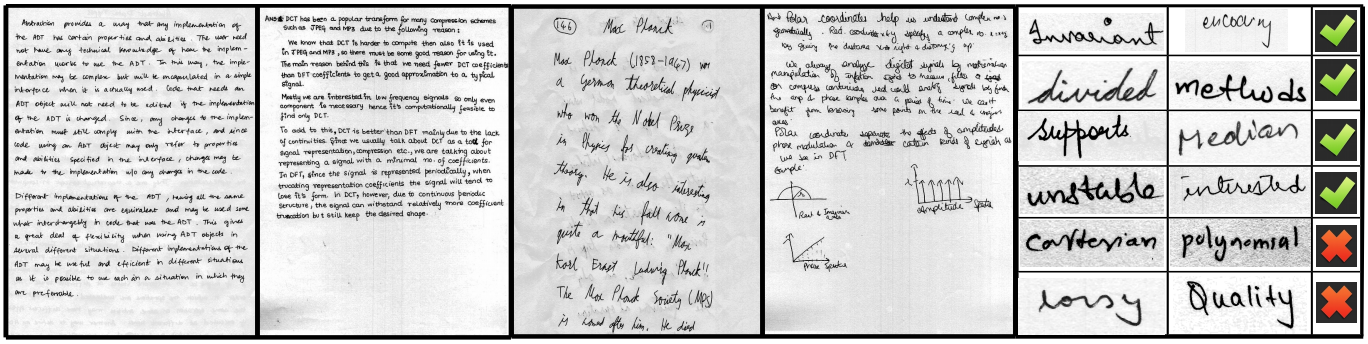

Figure 6. Qualitative analysis of the proposed method in predicting the aesthetic measure of the document. (a) The top scoring document images, having properly aligned and beautiful handwritten text. (b) The lowest scoring document images where one can observe inconsistent word spacing, skew and highly irregular word formation. (c) Sample pairs of word images from the human verification experiment where the words in the first column are predicted better than the words in the second column whereas the third column denotes whether the prediction agrees with human judgment.

other generic feature descriptors. CC features gave a MSE of 0.656 for regression and an accuracy of 44.11% for classification. The combination of stroke density distribution along with stroke width gave a MSE of 0.518 for regression and an accuracy 43.13% for classification. The observed LCC values are also quite low in both the cases.

The binary features such as BRISK and FREAK, which are appreciated for fast computation of keypoints, gave an improvement over raw features on all evaluation measures, especially the LCC, value which is nearly 0.58. Using a texture descriptor such as LBP, the performance improved to 0.63 in terms of LCC and the accuracy rate to nearly 50%. Here the patch size for LBP is empirically set to 16.

To capture better information at a stroke level and at different scales, dense SIFT features were extracted from the text regions in the document using multiple scales of 2, 4 and 6. We observed that the performance of SIFT and LBP are nearly the same. The best results are obtained using SURF keypoints and the corresponding descriptors which gave an accuracy of 55.88%, MSE of 0.308 and LCC of 0.682. Considering the complexity of the task, we think the accuracy is reasonably good. We further observed that GIST descriptors, which are preferred in natural scene images, couldn't perform well for computing local information from word images. In Figure 7(a), we show the confusion matrix of the predicted labels using SURF features. As one can observe, the maximum confusion is among the labels 2-3 and 3-4, which are near the mean value where the subjectivity among the users' annotation is maximum and thereby introducing noise in the training data. Also the performance of class 5 having a subjectivity label "excellent" performed low since in the training data we only had very few annotations with label 5.

D. Coarse Level Assessment

In coarse level assessment, we used only three types of features as shown in Table II. The performance of statistical features computed on word bounding boxes which captured the rough layout of margins, word spacing and horizontal profiles gave better performance than GIST and Fourier features. As compared to the GIST features for word neatness prediction, which reported an accuracy of 40.39%, GIST for page neatness made more sense and reported a higher accuracy of 46.66%.

Fourier features gave a similar accuracy of classification as given by the statistical features, i.e., 48.03% but, failed to improve the MSE and LCC. Hence, we observe that even though the statistical features were quite naive, they are efficient in capturing the overall layout and arrangement of text in the documents. Figure 7(b) shows the confusion matrix for predicting the neatness at a coarse level. Here also, the trend is almost the same as word level neatness, where maximum confusion is near the mean labels. Figure 6(a & b) present

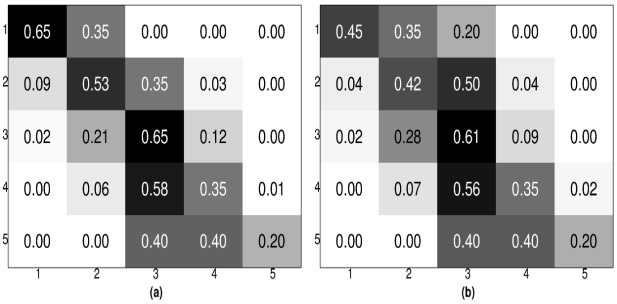

Figure 7. Visualization of confusion matrix obtained from the classification experiments for predicting the neatness label of documents from (a) word level, and (b) page level assessment.

Table II
QUANTITATIVE EVALUATION OF VARIOUS FEATURES FOR ANALYZING PAGE LEVEL NEATNESS: COARSE LEVEL ASSESSMENT

Features	Accuracy	MSE	LCC
Statistical	**48.03%**	**0.463**	**0.414**
GIST	46.66%	0.518	0.368
Fourier	48.03%	0.541	0.341

a few top and lowest scoring document pages on the basis of aesthetic properties using the neatness features proposed in this work.

E. Document Aesthetic Verification

For the verification of features at word level neatness, we trained a RankSVM model using the Fisher vector representation using SURF features. We verified the model with 30K randomly selected document pairs using the word neatness labels and got a verification rate of 78%. Similarly, for the verification of the features for page neatness, we trained a RankSVM model using the aforementioned statistical features and a similar 30K randomly selected document pairs. Here, we got a verification rate of 73%.

Human evaluation: This was performed to verify our features at a word level. We selected a random subset of pairs of words from our documents and human judgments were used to realize which of the two words was aesthetically more appealing than the other. A total of 300 random word pairs were selected for the task and the comparison results were obtained using the help of three human evaluators. Using this data and our RankSVM model, we performed the experiment to compare the predicted relative neatness of the word pairs with manual human judgments. Here we got a reasonable mean verification rate of 65%. We also observed that given two word images having different content, it is very difficult even for a human to compare their neatness. Due to the minute differences between the strokes and handwriting of most word pairs, it is quite hard to judge one over the other. Figure 6(c) shows some sample word image pairs along with their correctness with respect to human evaluations.

IV. Conclusion

In this work, we have attempted to solve a novel problem of predicting the aesthetic score of handwritten document images and tried to analyze what qualifies for a neat document which matches human judgments. We tested many standard features for measuring the neatness, both at the level of word (fine grained) and entire page (coarse grained). As part of this work, we would be releasing a subset of our dataset of 275 document pages along with the page and word level neatness annotations, which would be useful for the document research community. In the future, we plan to focus on more complex documents such as containing mathematical symbols, and also explore deep features for judging the aesthetic quality of words.

Acknowledgment

Praveen Krishnan is supported by TCS Research Scholar Fellowship.

References

[1] P. Ye and D. Doermann, "Document image quality assessment: A brief survey," in *ICDAR*, 2013.

[2] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Studying aesthetics in photographic images using a computational approach," in *ECCV*, 2006.

[3] N. M. Luca Marchesotti and F. Perronnin, "Discovering beautiful attributes for aesthetic image analysis," *IJCV*, 2015.

[4] R. Coll, A. Fornés, and J. Lladós, "Graphological analysis of handwritten text documents for human resources recruitment," in *ICDAR*, 2009.

[5] M. Miura and T. Toda, "Estimating writing neatness from online handwritten data." *JACIII*, 2014.

[6] P. Ye and D. Doermann, "Learning features for predicting OCR accuracy," in *ICPR*, 2012.

[7] L. Kang, P. Ye, Y. Li, and D. Doermann, "A deep learning approach to document image quality assessment," in *ICIP*, 2014.

[8] V. Govindaraju and S. N. Srihari, "Image quality and readability," in *Image Processing*, 1995.

[9] V. Kulesh, K. Schaffer, I. K. Sethi, and M. Schwartz, "Handwriting quality evaluation," in *ICAPR*, 2001.

[10] S.-L. Chou and S.-S. Yu, "Sorting qualities of handwritten chinese characters for setting up a research database," in *ICDAR*, 1993.

[11] D. Kumar and A. Ramakrishnan, "Quad: Quality assessment of documents," in *CBDAR*, 2011.

[12] P. Ye, J. Kumar, L. Kang, and D. S. Doermann, "Real-time no-reference image quality assessment based on filter learning," in *CVPR*, 2013.

[13] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *ICCV*, 2003.

[14] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *CVPR*, 2007.

[15] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, 1975.

[16] B. Epshtein, E. Ofek, and Y. Wexler, "Detecting text in natural scenes with stroke width transform," in *CVPR*, 2010.

[17] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 2004.

[18] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *IJCV*, 2001.

[19] H. Bay, T. Tuytelaars, and L. Van Gool, "SURF: Speeded up robust features," in *ECCV*, 2006.

[20] S. Leutenegger, M. Chli, and R. Y. Siegwart, "BRISK: Binary robust invariant scalable keypoints," in *ICCV*, 2011.

[21] A. Alahi, R. Ortiz, and P. Vandergheynst, "FREAK: Fast retina keypoint," in *CVPR*, 2012.

[22] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *PR*, 1996.

[23] P. Krishnan and C. Jawahar, " Matching Handwritten Document Images," in *ECCV*, 2016.

[24] T. M. Rath and R. Manmatha, "Word spotting for historical documents," *IJDAR*, 2007.

[25] O. Chapelle and S. S. Keerthi, "Efficient algorithms for ranking with SVMs," *Information Retrieval*, 2010.