Document Image Quality Assessment Using Discriminative Sparse Representation

Xujun Peng†, Huaigu Cao†, and Prem Natarajan§

†Raytheon BBN Technologies, Cambridge, MA, USA
§Information Sciences Institute, Univ. of Southern California, Marina del Rey, CA, USA
Email: {*xpeng, hcao*}*@bbn.com, pnataraj@isi.edu*

***Abstract*—The goal of document image quality assessment (DIQA) is to build a computational model which can predict the degree of degradation for document images. Based on the estimated quality scores, the immediate feedback can be provided by document processing and analysis systems, which helps to maintain, organize, recognize and retrieve the information from document images. Recently, the bag-of-visual-words (BoV) based approaches have gained increasing attention from researchers to fulfill the task of quality assessment, but how to use BoV to represent images more accurately is still a challenging problem. In this paper, we propose to utilize a sparse representation based method to estimate document image's quality with respect to the OCR capability. Unlike the conventional sparse representation approaches, we introduce the target quality scores into the training phase of sparse representation. The proposed method improves the discriminability of the system and ensures the obtained codebook is more suitable for our assessment task. The experimental results on a public dataset show that the proposed method outperforms other hand-crafted and BoV based DIQA approaches.**

I. Introduction

After decades of research and development, automated document processing systems which convert the original document images into digital format and analyze the content of the documents have attained considerable attention from researchers. However, the success of document processing is highly dependent on the quality of document images. Normally, the document images with bad quality tend to have lower readability, which also decrease the performance of Optical Character Recognition (OCR), layout analysis, document image retrieval and other document analysis tasks [1]. Thus, it is of great value to estimate the quality of document images and utilize this information to facilitate other higher level document processing/analysis tasks. Furthermore, with the vast use of digital cameras and smart-phones, it is critical to provide real-time feedback to users according to the document images' quality by integrating quality assessment models into hand-held devices.

Similar to the natural scene images, the quality of document images is affected by many factors, including: (1) blurs and distortions caused during image acquisition process, such as out-of-focus blur and motion blur. (2) physical noise introduced from the environment or capture devices' defects, such as uneven/bad lighting. (3) degradations created during digitization, such as compression loss. Besides these common defects, the document images can suffer particular degradations that are only applied on them. These degradations include bad quality or flaw of document medium during creation and storage, such as inadequate/heavy printing from low quality printers, paper yellowing and bleed-through ink from old documents, and inferior papers for producing documents. Another main source which causes degraded document images is happened during document digitization. For instance, the document image captured by hand-held devices might suffer from warping distortion. Furthermore, salt-and-pepper noise and other artifacts are easily produced during document binarization, de-skew, line-finding and other post-processing steps.

Thus, in order to model the document degradation process and to estimate the quality of document images, many assessment algorithms are developed by researchers. Based on the availability of high quality reference images, these methods can be categorized as three groups: full-reference assessment, reduced-reference assessment and no-reference assessment. When the high quality ground-truth (reference) images are available, both full-reference and reduced-reference assessment methods can be carried out by comparing the test image with the reference image. However, because the high quality reference images are not always obtainable in most applications, the no-reference based DIQA approaches dominate the current research field.

To discover the underlying principles that affect the quality of document images without references, Kanungo and Zheng simulated degraded images using a generative model where the local neighborhood pattern distributions were applied to estimate the model's parameters [2]. This model can be easily implemented for quality assessment tasks. In [3], Nayef and Ogier claimed that non-learning-based quality assessment is suitable for heterogeneous documents. So they designed a metric-based no-reference approach, where the sharpness was measured followed by the estimation of character's quality. Similarly, Peng *et al.* used the edge gradient and height-to-width features to represent character's quality, which indicates the overall quality of the document image [4]. Originated from this idea, Kieu *et al.* proposed an algorithm which simulated the degradation process of characters [5]. This approach can

[1]The views expressed are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government.

[2]Distribution Statement "A" (Approved for Public Release. Distribution Unlimited.)

[3]This research was developed with funding from the Defense Advanced Research Projects Agency (DARPA) MADCAT Program.

978-1-5090-1792-8/16 $31.00 © 2016 IEEE
DOI 10.1109/DAS.2016.24

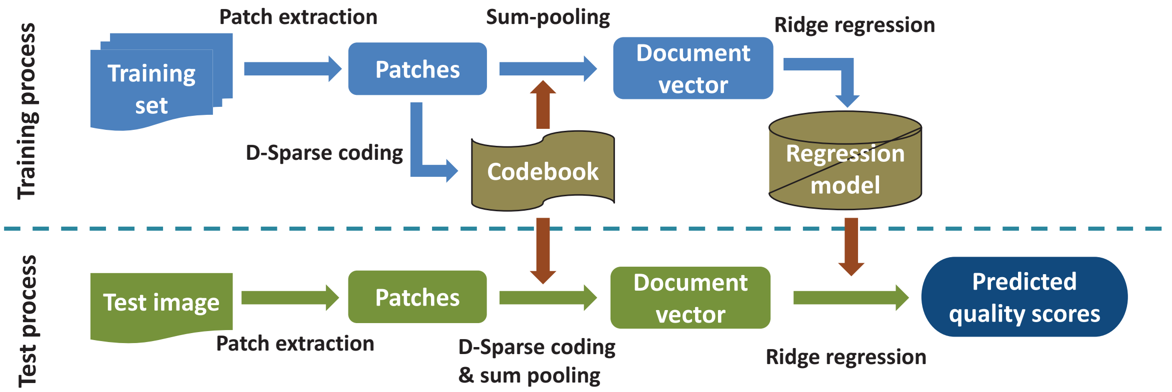


Fig. 1: The architecture of the proposed DIQA system.

be viewed as an extension of Kanungo's work. Also, Kumar *et al.* proposed a sharpness estimation based method to predict the document images' quality, which was more befitting the out-of-focus distortion [6]. To reveal the relationship between binarization degradation and human preference, Hale and Barney Smith presented a mathematical model in [7], where the edge displacement caused by the degradation was used as the clue to evaluate document images' quality.

One potential drawbacks of the abovementioned DIQA approaches is that they depend on heuristic rules or handcrafted features, which were designed for one or a set of particular distortions. Hence, to avoid the tedious efforts to collect empirically defined features, Ye and Doermann proposed a learning-based framework where BoV and max pooling techniques were employed to predict the document image's OCR accuracy [8]. Instead of using the K-means clustering approach to build the codebook for BoV, Peng *et al.* proposed a latent Dirichlet allocation based approach to learn the quality topics of each document, and then the distributions of those topics were used as features to evaluate the document's quality or OCR's accuracy [9]. In [10], Kang *et al.* implemented a deep learning method to evaluate document image's OCR accuracy, where the document images were initially divided into small patches and then were fed into a convolutional neural network to produce predicted quality scores. As a trade-off, Bhowmik *et al.* suggested an approach which collected a bag of pre-defined features (allograhic components) from documents to estimate documents' OCR quality, where the allographic components were derived from Canny edge detectors [11]. Because of their generalizability, these unsupervised learning feature based approaches have shown promising futures.

Unlike the scene images, document images not only have different characteristics which lead to their own features used for the quality assessment, they also utilize different evaluation protocol from scene images. Generally, the quality assessment for scene images prefer subjective criterion, which attempts to calculate the image quality score to match human perception. Whilst, as most document processing and retrieval systems' goal is to extract, organize and understand texts from document images, which is fully depended on the OCR's outputs, the OCR's accuracy is directly correlated to the readability and quality of document processing system. So, DIQA normally applies objective quality measure (such as OCR accuracy) to accomplish assessment tasks. This strategy is applied in [3], [4], [8], [9], [10] and [11]. Also, in [12], Salah *et al.* detected the missed text components in the document image to quantify the OCR defects, which can be considered as an assistant technique for the image quality controller. To predict the OCR accuracy for mobile-captured document images, Rusiñol *et al.* selected 24 measures from 6 families of focus operators to tackle the problem of out-of-focus for document images [13].

In this paper, we propose a no-reference DIQA method, where a semi-supervised learning technique is applied to represent document images. Particularly, a discriminative sparse representation approach is developed in our work to encode the image quality into the codebook, which is then used to facilitate the quality assessment task. In this framework, the target OCR accuracy is integrated into the training phase of sparse representation, which enhances the discriminant capability of the system.

The remainder of this paper is organized as follows. In Sec. II, we briefly describe the proposed DIQA system. Section III introduces the discriminative sparse representation based feature extraction for document images, which is followed by a quality assessment technique we employed in section IV. Section V provides the details of the experimental set up and results. We conclude our work in Sec. VI.

II. System Overview

We illustrate the flowchart of the proposed DIQA system in Fig. 1. Our DIQA approach is based on the sparse representation for images and a linear regression method for the quality estimation.

In the proposed system, the data are camera-captured document images, which are suffered from varying degrees of degradations. Given a set of document images, the data is separated into a training set and a test set. For both sets, the document images are divided into small patches, where the normalized raw feature of each patch is considered as the basic atom in our system.

For each training image, a set of patches are randomly selected and used to train a codebook, where the sparse representation is applied. Unlike the other codebook training methods, such as K-means clustering used in [8] and latent Dirichlet allocation used in [9], which are unsupervised learning approaches ignoring the disciminability of the trained codebook, we propose to use the target OCR accuracies of the training data to guide the sparse representation's training. Based on the trained codebook, each patch from a training image can be represented by the combination of codewords and the entire document image is vectorized by a pooling approach. Thereafter, the linear regression training is carried out to obtain a model that can be utilized to map images to OCR confidences.

For each test image, the same steps as the training phase are taken, where the trained codebook and sum pooling are employed to generate document image vectors, and regression models are used to predict the OCR scores for the images.

III. Discriminative sparse representation

A. Patch extraction

Sparse representation (or coding) has been successfully applied to natural images for restoration, segmentation, classification, and other tasks [14], [15], [16]. Based on the unlabeled training data, sparse representation learns overcomplete basis sets, in which the high level features can be obtained to ensure the minimum reconstruction error under sparsity constraints.

Normally, prior to the training phase for sparse representation, the basic training atoms are constructed. In our scenario, given a set of M training document images, we divide each document image evenly into small patches, whose size is $B \times B$. To each patch, the raw feature is normalized by subtracting the mean and dividing by the standard deviation of the entire training set. Then the normalized feature of each patch is used as the local descriptor and considered as a single visual word in our system. For each image I, we randomly select N patches within it to represent this image: $\boldsymbol{X}_I = [\vec{\boldsymbol{x}}_1, \vec{\boldsymbol{x}}_2, \cdots, \vec{\boldsymbol{x}}_N]$, where $\vec{\boldsymbol{x}}_n \in R^d$, $d = B \times B$. For the entire training set, we have a bag-of-visual-words (BoV) whose size is $K = M \times N$.

B. Conventional sparse representation training

Based on the collected training visual words, sparse representation attempts to represent each word $\vec{\boldsymbol{x}}$ using basis vectors $\vec{\boldsymbol{c}}_1, \vec{\boldsymbol{c}}_2, \cdots, \vec{\boldsymbol{c}}_n \in R^d$ and a sparse vector of coefficients (sparse code) $\vec{\boldsymbol{\phi}} \in R^n$, which satisfies:

$$\vec{\boldsymbol{x}} \approx \sum_{i=1}^{n} \phi_i \vec{\boldsymbol{c}}_i = [\vec{\boldsymbol{c}}_1, \vec{\boldsymbol{c}}_2, \cdots, \vec{\boldsymbol{c}}_n] \cdot \vec{\boldsymbol{\phi}} = \boldsymbol{C} \cdot \vec{\boldsymbol{\phi}} \tag{1}$$

and ensures only a few codewords from the codebook $\boldsymbol{C} \in R^{d \times n}$ are enough to represent the data. Here, n indicates the codebook size. Normally, we choose $n \geq d$, which means the new basis is overcomplete.

Thus, given all training words $\boldsymbol{X}_{d \times K} = [\vec{\boldsymbol{x}}_1, \vec{\boldsymbol{x}}_2, \cdots, \vec{\boldsymbol{x}}_K]$, the sparse representation finds an optimal solution to maximize the posterior estimation of the codebook $\boldsymbol{C}$ and the sparse codes $\boldsymbol{\Phi}_{n \times K} = [\vec{\boldsymbol{\phi}}_1, \vec{\boldsymbol{\phi}}_2, \cdots, \vec{\boldsymbol{\phi}}_K]$ for all training words, or minimize the reconstruction error:

$$< \hat{\boldsymbol{C}}, \hat{\boldsymbol{\Phi}} >= \underset{\boldsymbol{C}, \boldsymbol{\Phi}}{\arg\min} \|\boldsymbol{X} - \boldsymbol{C}\boldsymbol{\Phi}\|_F^2 + \lambda \sum_{i=1}^{K} \|\vec{\boldsymbol{\phi}}_i\|_1, \quad \text{s.t. } \|\vec{\boldsymbol{c}}_i\|^2 \leq t, \forall i = 1, \cdots, n, \tag{2}$$

where the first term in Eq.2 is the data fitting term and the second term is the sparsity regularization term, and $\|\cdot\|_1$ means the ℓ_1 norm.

For the optimization problem stated in Eq. 2, the object function is convergence for $\boldsymbol{C}$ when $\boldsymbol{\Phi}$ is fixed, and it is convergence for $\boldsymbol{\Phi}$ when $\boldsymbol{C}$ is fixed. But the object function is not convergence for both $\boldsymbol{C}$ and $\boldsymbol{\Phi}$ simultaneously. Thus, the typical approach to obtain the optimization solution for Eq. 2 is to reduce the original problem to two sub-optimization problems: alternatively performs the codebook learning and sparse codes encoding steps. In [17], Aharon *et al.* proposed a K-SVD algorithm that calculated the sparse codes of the samples based on the current dictionary and updated the dictionary atoms to better fit the data. In [18], Lee *et al.* derived Lagrangian-dual approach to solve the problem for dictionary learning, and the sparse codes $\boldsymbol{\Phi}$ were obtained by using a modified least angle regression (LARS) method. This efficient sparse coding algorithm was employed in our work.

C. Discriminative sparse representation training

However, despite competitive performance achieved by conventional sparse representation based codebook learning, the lack of discriminability for the learned codebook limited its applicability for the classification tasks. Therefore, several discriminative sparse representation training approaches were proposed. Zheng and Tao trained an optimal classifier which applied an overcomplete discriminative dictionary, where the small within-class scatter and big between-class scatter constraint was maintained [19]. Based on the same constraint, Liu *et al.* presented a discriminant sparse coding method to fulfill the image classification task, where the Lagrangian-dual algorithm was carried out to obtain the optimal codebook [20].

To facilitate the DIQA task, we also propose a discriminative sparse representation training method in this work. Unlike previous discriminative training approaches, which were designed for classification purpose, in the proposed DIQA system we train a codebook suitable for the regression task which maps images to continuous quality scores.

Given an initial trained codebook, for each training document image I, we assume $\boldsymbol{\Phi}_I$ represents the sparse codes for the training patches from I. Let f_p denote the pooling operation (we use sum pooling in this work), then the image vectorization step can be represented by:

$$\vec{\boldsymbol{z}}_I = f_p(\boldsymbol{\Phi}_I) = \frac{1}{N} \sum_{i=1}^{N} \vec{\boldsymbol{\phi}}_i, \tag{3}$$

where $\vec{\boldsymbol{z}}_I \in R^n$ is the feature vector for image I.

To map the image feature to a quality score, the following formula can be applied:

$$y_I \approx \vec{\boldsymbol{\theta}}^{\mathsf{T}} \cdot \vec{\boldsymbol{z}}_I, \tag{4}$$

where y_I is the target quality score of image I and $\vec{\boldsymbol{\theta}}$ is the coefficients that force the regression system to best fit the training data, which satisfies:

$$\hat{\vec{\boldsymbol{\theta}}} = \underset{\vec{\boldsymbol{\theta}}}{\arg\min} \sum_{i=1}^{M} \|y_i - \vec{\boldsymbol{\theta}}^{\mathsf{T}} \cdot \vec{\boldsymbol{z}}_i\|^2. \tag{5}$$

By considering Eq. 3 and constructing a target score vector $\vec{\boldsymbol{y}}_i \in R^N$ for each document image, whose elements are duplication of score y_i, the optimization of Eq. 5 can be converted to an equivalent problem:

$$\hat{\vec{\boldsymbol{\theta}}} = \arg\min_{\vec{\boldsymbol{\theta}}} \sum_{i=1}^{M} \|\vec{\boldsymbol{y}}_i - \vec{\boldsymbol{\theta}}^\intercal \cdot \boldsymbol{\Phi}_i\|^2 \quad (6)$$

$$= \arg\min_{\vec{\boldsymbol{\theta}}} \|\boldsymbol{Y} - \vec{\boldsymbol{\theta}}^\intercal \cdot \boldsymbol{\Phi}\|^2, \quad (7)$$

where $\boldsymbol{Y}_{1\times K} = [\vec{\boldsymbol{y}}_0^\intercal, \vec{\boldsymbol{y}}_1^\intercal, \cdots, \vec{\boldsymbol{y}}_M^\intercal]$ is a vector that concatenates target score vectors from all training images and $\boldsymbol{\Phi}$ is the sparse codes for all training patches.

Then, the sparse representation learning phase can be formulated by adding the Eq. 7 into the original Eq. 2 to form an updated object function:

$$\begin{aligned} < \hat{\boldsymbol{C}}, \hat{\boldsymbol{\Phi}}, \hat{\vec{\boldsymbol{\theta}}} >= &\arg\min_{\boldsymbol{C},\boldsymbol{\Phi},\vec{\boldsymbol{\theta}}} \|\boldsymbol{X} - \boldsymbol{C}\boldsymbol{\Phi}\|_F^2 \\ &+ \alpha\|\boldsymbol{Y} - \vec{\boldsymbol{\theta}}^\intercal \cdot \boldsymbol{\Phi}\|^2 + \lambda \sum_{i=1}^{K} \|\vec{\boldsymbol{\phi}}_i\|_1, \quad (8) \\ &\text{s.t. } \|\vec{\boldsymbol{c}}_i\|^2 \le t, \forall i = 1, \cdots, n; \|\vec{\boldsymbol{\theta}}\|^2 \le t, \end{aligned}$$

By taking a closer look into Eq. 8, we can observe that the second term of Eq. 8 has the same mathematical expression as the data fitting term when we relax the constraints by considering both $\vec{\boldsymbol{\theta}}$ and $\boldsymbol{\Phi}$ as variables. Then we can have the final optimization function:

$$\begin{aligned} < \hat{\tilde{\boldsymbol{C}}}, \hat{\boldsymbol{\Phi}} >= &\arg\min_{\tilde{\boldsymbol{C}},\boldsymbol{\Phi}} \left\| \boldsymbol{A} \left[\tilde{\boldsymbol{X}} - \tilde{\boldsymbol{C}}\boldsymbol{\Phi} \right] \right\|_F^2 + \lambda \sum_{i=1}^{K} \|\vec{\boldsymbol{\phi}}_i\|_1, \quad (9) \\ &\text{s.t. } \|\vec{\boldsymbol{c}}_i\|^2 \le t, \forall i = 1, \cdots, n+1, \end{aligned}$$

where $\tilde{\boldsymbol{X}} = \begin{pmatrix} \boldsymbol{X} \\ \boldsymbol{Y} \end{pmatrix}$ is the augmented feature matrix, $\tilde{\boldsymbol{C}} = \begin{pmatrix} \boldsymbol{C} \\ \vec{\boldsymbol{\theta}}^\intercal \end{pmatrix}$ is the augmented codebook and $\boldsymbol{A}_{(d+1)\times(d+1)} = \text{diag}(1, 1, \cdots, 1, \alpha)$ is a diagonal matrix which controls the weight of each feature element.

For Eq. 9, we can see that it has the same formulation as Eq. 2 except a linear transform introduced by $\boldsymbol{A}$, which means the existing fast sparse coding training algorithms can be easily applied on Eq. 9 to calculate the optimized discriminative codebook and sparse codes for each patch.

IV. Document image quality assessment

Based on the learned codebook $\tilde{\boldsymbol{C}}$, a final encoding step is taken to calculate the sparse codes of each patch for the training set, where the last row of $\tilde{\boldsymbol{C}}$ is ignored during this step. To assess the quality of document image I, the feature vector $\vec{\boldsymbol{z}}_I$ of this image is initially computed according to the sum pooling described in Eq. 3 and the obtained sparse codes $\boldsymbol{\Phi}_I$ for patches within this image.

With the feature vectors of training images, a Ridge regression model is trained to predict the scores. Ridge regression is a linear model whose loss function is the linear least squares function and its regularization term is given by the ℓ_2 norm.

Given the image features and their corresponding target quality scores $\{(\vec{\boldsymbol{z}}_1, y_1), \cdots, (\vec{\boldsymbol{z}}_M, y_M)\}$, the Ridge coefficients $\vec{\boldsymbol{\omega}} \in R^n$ minimize the object function:

$$\hat{\vec{\boldsymbol{\omega}}} = \arg\min_{\vec{\boldsymbol{\omega}}} \sum_{i=1}^{M} \|\vec{\boldsymbol{\omega}}^\intercal \cdot \vec{\boldsymbol{z}}_i - y_i\|_2^2 + \gamma\|\vec{\boldsymbol{\omega}}\|_2^2 \quad (10)$$

where M is the number of training images and γ is the tuning parameter that controls the strength of the regularization term.

The optimal estimation of $\vec{\boldsymbol{\omega}}$ for Eq. 10 can be solved by applying the data augmentation approach:

$$\hat{\vec{\boldsymbol{\omega}}} = (\boldsymbol{Z}^\intercal \boldsymbol{Z} + \gamma I)^{-1} \boldsymbol{Z}^\intercal \epsilon \quad (11)$$

where $\boldsymbol{Z}_{M\times n} = [\vec{\boldsymbol{z}}_1, \cdots, \vec{\boldsymbol{z}}_M]^\intercal$ and $\epsilon = [y_1, \cdots, y_M]^\intercal \in R^M$.

To each test image, the sparse codes of patches and feature vector for the image can also be calculated with the same procedure as training images. Then, by relying on the trained Ridge regression model, its quality score is computed by:

$$\hat{y}_i = \hat{\vec{\boldsymbol{\omega}}}^\intercal \vec{\boldsymbol{z}}_i. \quad (12)$$

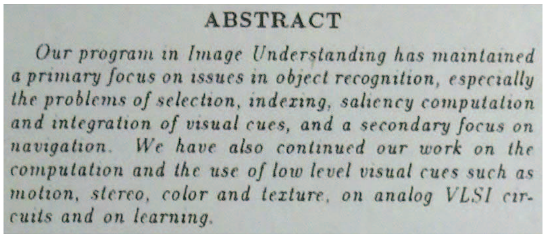
V. Experimental results

We followed the experimental setup used in [9] to evaluate the proposed discriminative sparse representation based DIQA method, where the experiments were carried out on a public Sharpness-OCR-Correlation (SOC) dataset [21]. In this dataset, a total of 175 high resolution document images were captured from 25 "ideally clean" documents with different focus lengths. The entire dataset was split into 25 subsets, where each subset contained $6 \sim 8$ images from the same document but with different degradation degrees. For each image, OCR scores were provided by three differen OCR engines. In our experiments, we used the average OCR score for the training and evaluation purpose. To perform our experiments, 102 images from 15 subsets were randomly selected as the training set. The remaining 73 images from 10 subsets were used as our evaluation set.

In Fig. 2, we show three document images with different degradation degrees from the test set. Fig. 2(a) is a high quality image which can be easily read and recognized by OCR system. Fig. 2(b) is a document image with light degradations but still can be recognized. And Fig. 2(c) was an image with severe blurs which lost its readability.

Prior to the training and test phase, each image in the dataset was divided into small patches without overlaps, and 3000 patches were randomly selected from each document image. Thus, we created a training set with 306000 visual words and a test set with 219000 visual words.

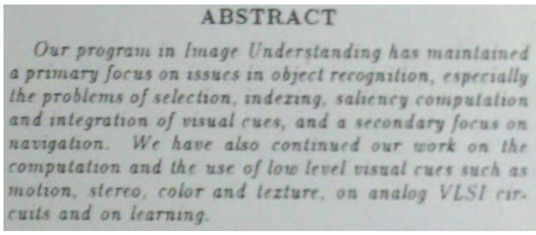
In order to find the optimal parameters for the sparse representation, we initially explored different sets of patch size d and codebook size n for the conventional sparse representation training. Based on these initially trained codebooks, each training image's feature was created by using sum pooling and max pooling, where the sparse codes of patches within this image were used. Then the Ridge regression models were trained on the top of these image features. The same procedure was carried out for the test set to extract image features and predict the quality scores. Thereafter, Spearman



ABSTRACT

Our program in Image Understanding has maintained a primary focus on issues in object recognition, especially the problems of selection, indexing, saliency computation and integration of visual cues, and a secondary focus on navigation. We have also continued our work on the computation and the use of low level visual cues such as motion, stereo, color and texture, on analog VLSI circuits and on learning.

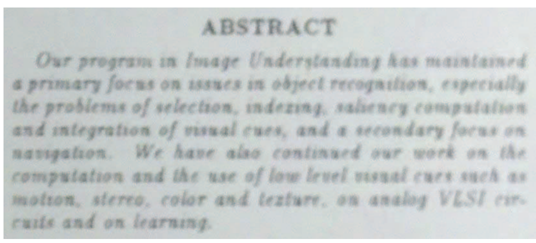
(a) The high quality document image.



ABSTRACT

Our program in Image Understanding has maintained a primary focus on issues in object recognition, especially the problems of selection, indexing, saliency computation and integration of visual cues, and a secondary focus on navigation. We have also continued our work on the computation and the use of low level visual cues such as motion, stereo, color and texture, on analog VLSI circuits and on learning.

(b) The degraded document image.



ABSTRACT

Our program in Image Understanding has maintained a primary focus on issues in object recognition, especially the problems of selection, indexing, saliency computation and integration of visual cues, and a secondary focus on navigation. We have also continued our work on the computation and the use of low level visual cues such as motion, stereo, color and texture, on analog VLSI circuits and on learning.

(c) Severe degraded document image.

Fig. 2: Sample document images with different degradation degrees in our corpus.

rank correlation was used to evaluate the performance of these parameters which is illustrated in Fig. 3.

From Fig. 3, we can discover that the overall performance of sum pooling was superior to max pooling based DIQA. It can also be observed that with the same patch size, the system with bigger codebook size n normally produced more reliable results than the one with smaller codebook size. This phenomenon proved the effectiveness of the overcomplete basis provided by the sparse representation. By analyzing all parameter combinations, we found that the system with patch size of 10×10 and the codebook size of 150 obtained the highest Spearman correlation score, which is 0.923. Therefore, we selected patch size $d = 10 \times 10$ and codebook size $n = 150$ for our discriminative sparse representation training, and the sum pooling was taken for the image feature vectorization.

To perform the discriminative sparse representation training, the original feature of each visual word was expanded one element by adding the corresponding target quality score, which was coming from its document's OCR accuracy. The value of the last element for the new feature vector was normalized into the same range as the other elements. Prior to the training, we empirically set the weight α for linear constraint in Eq. 8 as 0.1 in our experiment. Then the modified conventional training method described in [18] was taken to obtain the discriminative codebook. In Fig. 4, we show the obtained codebook, where the first 100 elements of each

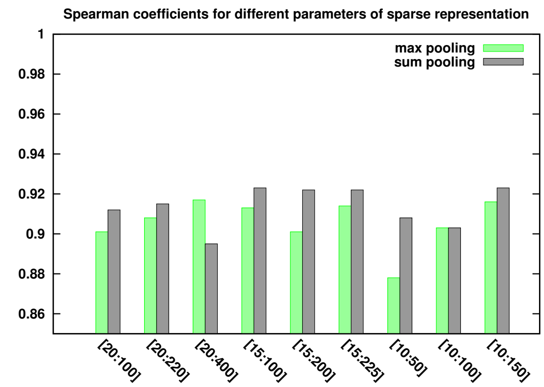


Fig. 3: The Spearman rank coefficients for different parameters of sparse representation. The first element of each parameter pair in x-axis means the edge size B for patches and the second element is the codebook size n.

codeword were illustrated and the last element was discarded because it was ignored during the regression training and test phase. From this figure, it can be seen that the codewords captured different type of edge/corner representations with various degradation degrees from our corpus. As revealed in [3], [4] and [5] that the sharpness of characters and edges has strong correlations with document images' qualities, the codebook learned from our experiments had the good potential to represent different type of degradations to estimate image's quality.

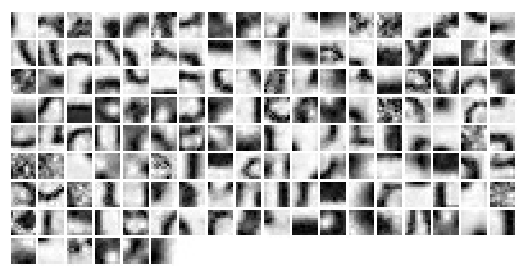


Fig. 4: The codebook trained by the proposed discriminative sparse representation approach.

To train the Ridge regression model, the method described in Sec. IV was implemented, where the reduced codebook from $\tilde{\boldsymbol{C}}$ was used to generate sparse codes for each training patch, and the the sum pooling was applied for image vectorization. Based on the obtained codebook $\tilde{\boldsymbol{C}}$ and regression model, we estimated the OCR scores for all test document images.

In order to evaluate the performance of the proposed method, Spearman rank coefficient ρ_s and Pearson coefficient ρ_p were used to measure the correlation between predicted OCR scores and the ground truth. In this experiment, we obtained $\rho_s = 0.928$ and $\rho_p = 0.935$ for the entire test set.

Furthermore, we compared the discriminative sparse representation based DIQA with the other three state-of-the-art DIQA approaches: Edge gradient based DIQA [4], K-Means plus max pooling based DIQA [8] and LDA based DIQA

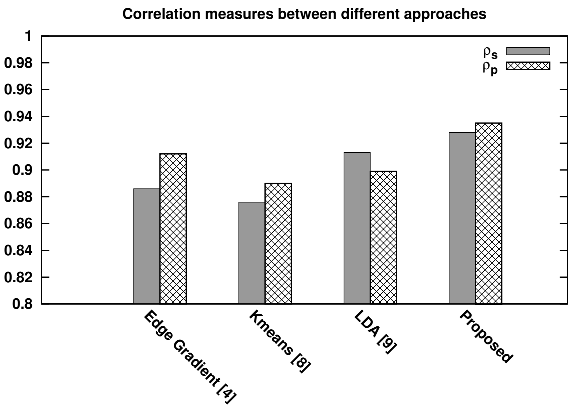


Fig. 5: The Spearman (ρ_s) and Pearson (ρ_p) correlations of different methods.

[9]. Among these methods, edge gradient based DIQA used hand-crafted features, whilst the other two approaches used unsupervised features. In Fig. 5, we illustrated the ρ_s and ρ_p of four DIQA algorithms, from which we can see both Spearman coefficient and Pearson coefficient of the proposed method were higher than the coefficients of the other approaches. Also from this figure, we can conclude that the codebook obtained from a semi-supervised sparse representation training has more discriminability, which benefits our DIQA task.

VI. Conclusion

In this paper, we present a DIQA approach which is based on discriminative sparse representation. Unlike the conventional DIQA methods where the features are empirically designed for specific degradations or the codebook are trained using unsupervised technique without discriminative capability, we introduce the target quality score into the training phase to guide the codebook and sparse codes' learning. The experimental results show that the trained codebook contains plenty of edge representations with different degree of degradations and obtains the discriminability, which outperforms traditional hand-crafted and unsupervised codebook based DIQA methods.

In the future, other semi-supervised feature learning approaches will be studied. We also plan to explore other image feature vectorization approaches to boost DIQA accuracy.

References

[1] P. Ye and D. Doermann, "Document image quality assessment: A brief survey," in *12th International Conference on Document Analysis and Recognition (ICDAR)*, 2013, pp. 723–727.

[2] T. Kanungo and Q. Zheng, "Estimating degradation model parameters using neighborhood pattern distributions: An optimization approach," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 26, no. 4, pp. 520–524, 2004.

[3] N. Nayef and J.-M. Ogier, "Metric-based no-reference quality assessment of heterogeneous document images," in *Proc. SPIE, Document Recognition and Retrieval XXII*, vol. 9402, 2015, pp. 94 020L–94 020L–12.

[4] X. Peng, H. Cao, K. Subramanian, R. Prasad, and P. Natarajan, "Automated image quality assessment for camera-captured OCR," in *18th IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 2621–2624.

[5] V. Kieu, M. Visani, N. Journet, J. Domenger, and R. Mullot, "A character degradation model for grayscale ancient document images," in *21st International Conference on Pattern Recognition (ICPR)*, 2012, pp. 685–688.

[6] J. Kumar, F. Chen, and D. Doermann, "Sharpness estimation for document and scene images," in *21st International Conference on Pattern Recognition (ICPR)*, 2012, pp. 3292–3295.

[7] C. Hale and E. Barney Smith, "Human image preference and document degradation models," in *Ninth International Conference on Document Analysis and Recognition*, vol. 1, 2007, pp. 257–261.

[8] P. Ye and D. Doermann, "Learning features for predicting OCR accuracy," in *21st International Conference on Pattern Recognition (ICPR)*, 2012, pp. 3204–3207.

[9] X. Peng, H. Cao, and P. Natarajan, "Document image OCR accuracy prediction via latent dirichlet allocation," in *13th International Conference on Document Analysis and Recognition*, 2015.

[10] L. Kang, P. Ye, Y. Li, and D. Doermann, "A deep learning approach to document image quality assessment," in *IEEE International Conference on Image Processing (ICIP),*, 2014, pp. 2570–2574.

[11] T. Bhowmik, T. Paquet, and N. Ragot, "Ocr performance prediction using a bag of allographs and support vector regression," in *Document Analysis Systems (DAS), 2014 11th IAPR International Workshop on*, 2014, pp. 202–206.

[12] A. Ben Salah, N. Ragot, and T. Paquet, "Adaptive detection of missed text areas in OCR outputs: application to the automatic assessment of OCR quality in mass digitization projects," in *Proc. SPIE, Document Recognition and Retrieval XX*, vol. 8658, 2013, pp. 865 816–865 816–12.

[13] M. Rusiñol, J. Chazalon, and J.-M. Ogier, "Combining focus measure operators to predict ocr accuracy in mobile-captured document images," in *11th IAPR International Workshop on Document Analysis Systems (DAS)*, 2014, pp. 181–185.

[14] J. Mairal, M. Elad, and G. Sapiro, "Sparse representation for color image restoration," *Image Processing, IEEE Transactions on*, vol. 17, no. 1, pp. 53–69, Jan 2008.

[15] Y.-C. Liu and H.-T. Chen, "Unsupervised scene segmentation using sparse coding context," *Machine Vision and Applications*, vol. 24, no. 2, pp. 243–254, 2013.

[16] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, June 2009, pp. 1794–1801.

[17] M. Aharon, M. Elad, and A. Bruckstein, "k -svd: An algorithm for designing overcomplete dictionaries for sparse representation," *Signal Processing, IEEE Transactions on*, vol. 54, no. 11, pp. 4311–4322, Nov 2006.

[18] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," in *Advances in Neural Information Processing Systems 19*, 2007, pp. 801–808.

[19] H. Zheng and D. Tao, "Discriminative dictionary learning via fisher discrimination k-svd algorithm," *Neurocomputing*, vol. 162, pp. 9 – 15, 2015.

[20] B.-D. Liu, Y.-X. Wang, Y.-J. Zhang, and Y. Zheng, "Discriminant sparse coding for image classification," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, 2012, pp. 2193–2196.

[21] J. Kumar, P. Ye, and D. Doermann, "A dataset for quality assessment of camera captured document images," in *Camera-Based Document Analysis and Recognition*, 2014, pp. 113–125.