Journal of Cultural Heritage xxx (2017) xxx–xxx

Available online at
ScienceDirect
www.sciencedirect.com

Elsevier Masson France
EM|consulte
www.em-consulte.com/en

Original article

Subjective and objective quality assessment of degraded document images

Atena Shahkolaei [a,b,*], Hossein Ziaei Nafchi [a,b], Somaya Al-Maadeed [a,b], Mohamed Cheriet [a,b]

[a] Synchromedia Laboratory for Multimedia Communication in Telepresence, École de technologie supérieure, Montreal, QC H3C 1K3, Canada
[b] Department of Computer Science & Engineering, Qatar University, Doha, Qatar

ARTICLE INFO

Article history:
Received 10 May 2017
Accepted 4 October 2017
Available online xxx

Keywords:
Document image quality assessment
Degraded document images
Physical noises
MSCN coefficients
Human visual system
log-Gabor filter

ABSTRACT

The huge amount of degraded documents stored in libraries and archives around the world needs automatic procedures of enhancement, classification, transliteration, etc. While high-quality images of these documents are in general easy to be captured, the amount of damage these documents contain before imaging is unknown. It is highly desirable to measure the severity of degradation that each document image contains. The degradation assessment can be used in tuning parameters of processing algorithms, selecting the proper algorithm, finding damaged or exceptional documents, among other applications. In this paper, the first dataset of degraded document images along with the human opinion scores for each document image is introduced in order to evaluate the image quality assessment metrics on historical document images. In this research, human judgments on the overall quality of the document image are used instead of the previously used OCR performance. Also, we propose an objective no reference quality metric based on the statistics of the mean subtracted contrast normalized (MSCN) coefficients computed from segmented layers of each document image. The segmentation into four layers of foreground and background is done on the basis of an analysis of the log-Gabor filters. This segmentation is based on the assumption that the sensitivity of the human visual system (HVS) is different at the locations of text and non-text. Experimental results show that the proposed metric has comparable or better performance than the state-of-the-art metrics, while it has a moderate complexity. The developed dataset as well as the Matlab source code of the proposed metric is available at http://www.synchromedia.ca/system/files/VDIQA.zip.

1. Research aims

In the literature, several objective metrics were proposed to evaluate the quality of natural images, but these metrics may not be applied to historical document images, which contain physical degradations. In this paper, a new blind objective metric is proposed that can assess the quality of historical document images remarkably well. The proposed metric is tested on VDIQA dataset, which is also developed in this paper based on the human judgments.

2. Introduction

Heritage is our legacy from the past, what we live with today, and what we pass on to future generations [1]. Old manuscripts and documents constitute an important part of the heritage. They are the memory of human cultures, their history, their achievements, their lifestyle and their individual and social behaviors [2]. Recent years have seen increasing efforts in archiving and digitizing these valuable documents in order to safeguard them against deterioration and to process them by signal processing techniques [2–6]. Having a large amount of the documents archived, it is very interesting to use automatic techniques to extract and use meaningful information from their images. Because of aging, improper handing and environmental factors, these documents suffer from low to high degrees of degradation. These physical degradations appear on the digitized document images as well. According to the taxonomy of Lins [7], these types of degradation refer to the "physical noises" in document images. The following is a list of the physical noises [7]: folding marks, paper translucency, paper aging, paper tex-

* Corresponding author. Synchromedia Laboratory for Multimedia Communication in Telepresence, École de technologie supérieure, Montreal, QC H3C 1K3, Canada.
E-mail address: atena.shahkolaei.1@ens.etsmtl.ca (A. Shahkolaei).

https://doi.org/10.1016/j.culher.2017.10.001
1296-2074/

ture, paper punching, stains, torn-off regions, worm holes, readers' annotations, physical blur, carbon copy effect, scratches and cracks, sunburn, and inadequate printing. Physical noises can greatly affect the performance and accuracy of the document processing algorithms.

In order to reduce the effect of the physical noises on the document processing algorithms, it is common to automatically enhance document images. The performance of the current enhancement techniques is not reliable, and the reason can be explained by two arguments. Firstly, document images might be extremely degraded such that their enhancement is beyond the ability of the current enhancement methods. In addition, the enhancement techniques may over/under-estimate the severity of degradations, which leads to an inaccurate enhancement. We note that other imaging techniques like multi-spectral may help [8,9], but it is beyond the scope of the current study.

More specifically, we noted that the performance of the current enhancement methods is far beyond the human visual system's (HVS) ability to distinguish type and severity of degradations. Computational models of HVS that are able to assess common distortions appearing in natural images, synthetic images, etc. are available and evolving [10–14]. Several objective image quality assessment (IQA) metrics based on these models have been proposed. Similarly, document images can be assessed and benefit from the quality metrics. Unfortunately, less attention has been paid to the quality assessment of physically degraded document images.

Given the ability of the HVS to recognize degradations in document images, it is very interesting to develop degradation assessment metrics that mimic the HVS. In turn, enhancement methods or other processing algorithms can be developed or tuned in relation to these metrics. Visual document image quality assessment (VDIQA), if not used by the enhancement methods, can have other applications. It helps in the classification of the document images on the basis of their visual quality. For example, an extremely degraded document can be picked and treated manually in order to extract maximum information or to prevent further document deterioration. The performance expectation of the document processing algorithms can be estimated through VDIQA metrics.

The applications of VDIQA and the hypothesis that a document image can be assessed even though it cannot be enhanced, motivated us to develop the first VDIQA dataset based on HVS judgments. There are few document image quality assessment (DIQA) datasets available in the literature, and most of them are evaluated based on the OCR performance. In what follows, we mention why human judgments are used in the introduced dataset instead of OCR accuracy. Firstly, the OCR engines are not perfect, specially for some of the languages, old writing styles and fonts. The second reason is that higher OCR accuracy does not necessarily mean that a document image is of high quality, rather it may mean that the text region is not degraded. This avoids using DIQA for applications like automatic age estimation of documents. In this paper, we use the term VDIQA when human judgments are used and DIQA when OCR accuracy is used. Image quality assessment (IQA) of natural images, mostly focuses on the assessment of distortions caused in the acquisition setup, compression, data transfer, etc. However, we suppose that high-quality images of documents are available in our study. In other words, document images are free from the distortions caused in the acquisition stage. Of course, other distortions, such as blur effect can be assessed; but the complexity of the problem will increase if other distortions are taken into account.

IQA datasets of natural images [11,15], synthetic images [14], photo-retouched images [16], and screen content images [17] are publicly available in a fast growing field of research. However, there has been little effort to develop datasets and metrics for quality assessment of the document images. The majority of the datasets that have been introduced are either not available or not available to the public [18–21]. The very few quality metrics for DIQA are either not available or not available to the public [18,20–24]. The authors of [25] surveyed DIQA/VDIQA metrics and datasets. The DIQA datasets can have different characteristics. Images in a dataset might be in the form of color, gray-level, or binary. Each image may show a character, a word, a sentence, or a full page. OCR accuracy and mean opinion score (MOS) are two data types available in these datasets for the evaluation purposes. The majority of the VDIQA datasets is in the form of binary images [18,19,21]. The dataset introduced in [26] consists of 135 gray-level images with blur distortion. The MOS values for this dataset are collected and computed by crowd sourcing. To the best of our knowledge, there is no dataset of old and degraded document images consisting of full page color image with associated MOS values. The lack of such a dataset motivated us to develop the first of its kind.

Quality of images can be assessed subjectively and objectively [27]. Subjective evaluation is more accurate and ecologically valid. However, it involves human participation; therefore, it is time-consuming and expensive. To overcome these limitations, objective metrics have been proposed in the literature. Objective IQAs are mathematical models that approximate the results of subjective IQA. The main goal of objective IQA is to supply quality metrics that can automatically predict perceived image quality. According to the availability of non-distorted images, objective image quality assessment can be classified into three categories: full-reference (FR) [10,28–31], reduced-reference (RR) [32–34], and no-reference (NR) IQA models [12,35–38]. In NR-IQA, only distorted images are available. This paper focuses on historical document images, for which the reference image is not available.

NR-IQA metrics perform according to the statistical regularities of natural images in spatial and transformed domains. The deviation between statistical regularities of distortion-free and distorted images is considered in the design of the NR-IQA models. The so-called NR-IQA metric DIIVINE [35], first classifies distortion types. Then, subband coefficients of discrete wavelet transform (DWT) are fitted by generalized Gaussian distribution (GGD). The statistics of GGD determine the severity of distortions and quality scores are thus estimated by regression. BLIINDS-II [36] is a non-distortion specific NR-IQA metric based on the statistics of the discrete cosine transform (DCT) coefficients. The popular NR-IQA metric BRISQUE [12] uses the statistics of natural images in the spatial domain. The distribution of mean subtracted contrast normalized (MSCN) coefficients in two image scales is fitted by symmetric GGD and asymmetric GGD. MSCN coefficients are widely used by NR-IQA models. Similar to the metric DIVIINE, CurveletQA [37] is also a two-stage distortion classification and distortion severity estimation NR-IQA model. It performs according to the statistics of the curvelet coefficients extracted from the images after applying the curvelet transform. BQMS [38] is a NR-IQA metric specifically proposed for the quality assessment of screen content images. It performs by using the principles of free energy theory. Recently, an effective NR-IQA metric (FRIQUEE [13]) was proposed; it is based on a bag of features approach. The metric extracts a large number of features in spatial and frequency domains, and considers color features extracted from different color spaces.

None of these metrics are designed for quality assessment of document images. Basically, they do not access the distortions appearing on documents because of aging and physical condition. Several DIQA metrics have been proposed to assess specific distortions, such as blur [26,39]. The first non-distortion specific DIQA metric was proposed on the basis of unsupervised feature learning [40]. The non-uniform patches corresponding to the text regions are used to construct a visual codebook. The extracted unsupervised features are then mapped to the OCR accuracy by support

Please cite this article in press as: A. Shahkolaei, et al., Subjective and objective quality assessment of degraded document images, Journal of Cultural Heritage (2017), https://doi.org/10.1016/j.culher.2017.10.001