Pattern Recognition 42 (2009) 3355 – 3364

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

RSLDI: Restoration of single-sided low-quality document images

Reza Farrahi Moghaddam*, Mohamed Cheriet

École de Technologie Supérieure, Synchromedia Laboratory for Multimedia Communication in Telepresence, Montréal (QC), Canada H3C 1K3

ARTICLE INFO

Article history:
Received 7 August 2008
Accepted 15 October 2008

Keywords:
PDE-based image processing
Classifiers
Degradation modeling
Nonlinear model
Document enhancement and restoration
ICA
Bayes modeling

ABSTRACT

This paper addresses the problem of enhancing and restoring single-sided low-quality single-sided document images. Initially, a series of multi-level classifiers is introduced covering several levels, including the regional and content levels. These classifiers can then be integrated into any enhancement or restoration method to generalize or improve them. Based on these multi-level classifiers, we first propose a novel PDE-based method for the restoration of the degradations in single-sided document images. To reduce the local nature of PDE-based methods, we empower our method with two flow fields to play the role of regional classifiers and help in preserving meaningful pixels. Also, the new method further diffuses the background information by using a content classifier, which provides an efficient and accurate restoration of the degraded backgrounds. The performance of the method is tested on both real samples, from the Google Book Search dataset, UNESCO's Memory of the World Programme, and the Juma Al Majid (Dubai) datasets, and synthesized samples provided by our degradation model. The results are promising. The method-independent nature of the classifiers is illustrated by modifying the ICA method to make it applicable to single-sided documents, and also by providing a Bayesian binarization model.

1. Introduction

Preprocessing is the key to processing and understanding historical and very old documents. If a document image is preprocessed well, many of the common tools used for processing normal documents, or modified versions of them, can be applied to historical document images. However, as these documents are usually severely degraded, very unstructured and variable in content and appearance, the preprocessing itself becomes a difficult problem.

The problem of restoring a document image suffering from bleed-through degradation, which is a major task in the analysis of very old documents, has been studied from several points of view [1–9]. The methods can be categorized in two groups: methods which work on double-sided document images, and methods which work on single-sided document images. Double-sided images contain valuable information, but are very difficult to create because of the need for an accurate registration process. This is an especially challenging problem in the case of very old documents, as the paper is most often very deformed. The ICA method, for example, can be applied to restore double-sided document images [3]. It is reasonably fast, but suffers from its global characteristics. An example from the second group of methods is entropy-based binarization [4], in which a statistical adjustment, based on entropy, is performed between the input gray-scale and output binary versions of an image. Methods belonging to the second group are applicable to single-sided datasets, which are more readily available. However, these methods are usually restricted to binarization or edge detection techniques.

If we look at the degradation problems in documents from a physical point of view, it appears that many of them are the result of some kind of diffusion process which occurs over time [10–18]. In these cases, the appropriate use of an inverse diffusion model can result in very good and direct restoration of these types of physical diffusion. Moreover, the denoising nature of diffusion models provides clear output, without the need for extra denoising postprocessing [19]. The main point about diffusion methods is their great similarity with the physical processes of degradation [20].

However, there are some barriers to the application of diffusion-based methods to the problem of restoring very old documents. The first is the diffusion of ink, which causes a loss of sharpness of the strokes. This side-effect can be prevented by using small values for the diffusion coefficients. However, the diffusion coefficients must be large enough to denoise and remove the background of the document. The second is the local characteristics of diffusion models (in the sense of spatial coordinates). In other words, in diffusion models, only the data from the neighbors of a pixel can alter the information of that pixel. If a pixel is surrounded by, for example, dark regions, there is no way to insert bright or white information into that pixel.

* Corresponding author.
E-mail addresses: reza.farrahi-moghaddam.1@ens.etsmtl.ca, imriss@yahoo.com (R. Farrahi Moghaddam), mohamed.cheriet@etsmtl.ca (M. Cheriet).

0031-3203/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.10.021

This high spatial dependency of the diffusion models limits their ability in information layer (source) separation and removal.

In this work, by introducing multi-level classifiers, we modify the diffusion-based methods to enhance single-sided document images which suffer from bleed-through or similar effects. These multi-level classifiers extend from the local level to the global level, and provide simple and direct access to physical and meaningful information.

The structure of the paper is as follows: in Sections 2 and 3, multi-level classifiers are discussed. Then, in Section 4, a diffusion-based method is modified using two regional classifiers, the flow fields, and one content classifier, the estimated background. Through this modification, the new model will successfully remove the interference from degraded document images. The results of applying the modified method on real and synthetic samples are presented and shown to be promising. In Section 5, a modified version of the ICA method, which, by definition, is not applicable to single-sided document images, is applied to enhance these images. In Section 6, a simple Bayesian model for the enhancement of single-sided document images based on multi-level classifiers is given. Finally, in Section 7, we provide an analysis of the computational cost of the various methods.

1.1. Image processing paradigms

Images, and especially document images, are very complex and contain a large amount of data and information. Most of this information is in the form of (two-dimensional) relations between colors or intensities. Analysis and processing of document images is performed using various paradigms, such as statistical methods [3], variational models [21], PDE-based models [19], non-local filters [22], and spectral models (wavelet transforms) [23], among others. In some of these paradigms, such as the statistical models, the spatial relations of the information are mostly ignored. In contrast, there are models, such as non-local models, which only consider the restrictive spatial similarities. Certainly, the most powerful solution to document image analysis will use all the spatial relations. One straightforward possibility is a huge neural network which accepts all the document image information. Of course, such a model, or similar models from other paradigms, is not feasible because of computational limitations and the high dimensionality of the model. Also, designing it and training it will be a difficult, if not impossible, task, because its huge size makes it very difficult to understand, especially for concepts such as convergence, uniqueness, and existence. To add to the complexity of the problem, the nature and characteristics of document images vary dramatically. These variations are very noticeable in historical documents, where cultural aspects not only influence the content and structure of the text, but also the type of paper and ink. Even the manner in which the documents are kept is culture-dependent and varies noticeably from one culture to another. Finally, in many historical documents, even on the same page, there are extreme variations in content, color, and structure. Also, because of the age of these documents, they are most often highly degraded: many of the strokes have cuts, the color is not uniform across or along the strokes, and the edges are extremely weak.

In Fig. 1, we provide a general model for preprocessing such difficult images. In this model, the classifiers, which will be discussed in the next section, are used to provide proper exchange of information between sites on the image. Although in many modern documents there is consistency between the information content and the structure of the text in different parts or sites on a page, the intensity and structure of the text changes dramatically from site to site and location to location, even on a single page in very old documents, magazines, and newspapers. The classifiers will act as connecting bridges for the exchange of information over the whole the document

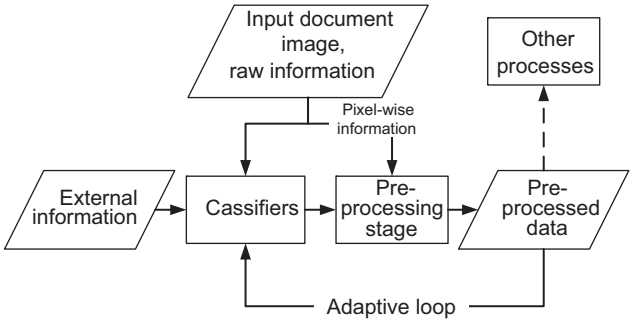

Fig. 1. A schematic diagram of the preprocessing model.

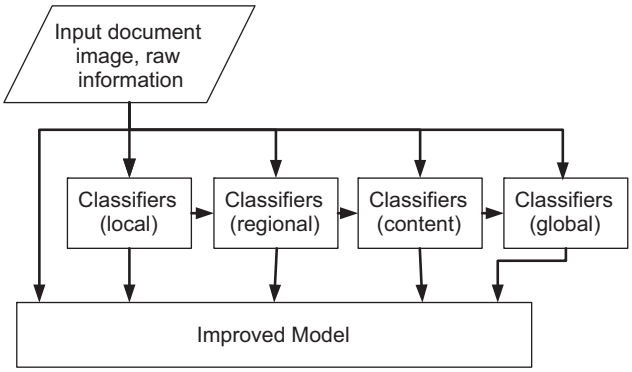

Fig. 2. Multi-level classifiers mediate between the raw image data and the model.

image. In this way, while keeping the computational window small, useful information from different levels is available for modeling (see Fig. 2). The model is not restricted to a specific type. Actually, all preprocessing methods can be modified and improved by incorporating the new information (classifiers). Each paradigm has its own benefits and advantages. However, using the proper classifiers, all the paradigms will provide a satisfactory enhanced result. Also, preprocessed information can be used as new input for the computation of the classifiers. Note that the classifiers here are pixel-wise. In other words, the information at each pixel is enumerated by adding valuable classifiers, and all this information is used to enhance the methods.

The main issue is the development and selection of the classifiers. Actually, any function or procedure can be assumed to be a classifier. Below, the process of introducing classifiers will be discussed in a level-based approach. Although blind data mining and the attempt to discover good classifiers from the set of unlimited possible classifiers is usually used and seems interesting, a basic and physical understanding of the problem and its characteristics can speed up the identification of the proper classifiers. These basic rules and guides must be so general and adaptive as to provide enough room for the variable nature of images otherwise the resulting classifiers will be very restricted and not perform well.

2. Multi-level classification of very old document images

The enhancement of the document images is usually achieved by applying some evolving method to the data. However, it has been observed that global and data-independent methods are not the best choice, and so data-dependent adaptive models and methods are usually used. Proper labeling of the input data will reduce the complexity of the models, and multi-level classifiers can provide a robust and adaptive labeling method. We divide the levels of