Pattern Recognition 48 (2015) 3615–3626

Contents lists available at ScienceDirect

Pattern Recognition

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

Automatic ink mismatch detection for forensic document analysis

Zohaib Khan*, Faisal Shafait, Ajmal Mian

School of Computer Science and Software Engineering, University of Western Australia, 35 Stirling Highway, Crawley, WA 6009, Australia

ARTICLE INFO

Article history:
Received 9 December 2013
Received in revised form
22 December 2014
Accepted 3 April 2015
Available online 15 April 2015

Keywords:
Hyperspectral imaging
Multispectral document examination
Band selection

ABSTRACT

A key aspect of handwritten document examination is to investigate whether some portion of the text was modified, altered or forged with a different pen. This paper demonstrates the use of hyperspectral imaging for ink mismatch detection in a handwritten note. We propose a novel joint sparse band selection technique that selects informative bands from hyperspectral images for accurate ink mismatch detection. We have developed an end-to-end camera-based hyperspectral document imaging system and collected a database of handwritten notes which has been made publicly available. Algorithmic solutions are presented to handle specific challenges in camera-based hyperspectral document imaging. Extensive experiments show that the proposed band selection method selects the most informative bands and improves average accuracy up to 15%, compared to using all bands.

1. Introduction

The human eye is sensitive to light in the visible range to distinguish materials based on their color [1]. However, humans are unable to distinguish between two apparently similar colors [2] due to the trichromatic nature of the human visual system. When a document is manipulated with the intention of forgery or fraud, the modifications are often done in such a way that they are hard to catch with a naked eye. The forger not only tries to emulate the handwriting of the original writer, but also uses a pen that has a visually similar ink compared to the rest of the note. Hence, analysis of inks is of critical importance in questioned document examination.

The outcome of ink analysis can potentially lead to the determination of forgery, fraud, backdating and ink age. Of these, one of the most important tasks is to discriminate between different inks which we term as *ink mismatch detection*. There are two main approaches to distinguish inks, destructive and non-destructive examination. Chemical analysis such as thin layer chromatography (TLC) [3] is a destructive test which separates a mixture of inks into its constituents via capillary action. However, TLC compromises the originality of a sample, of a forensic evidence. Furthermore, observing any noticeable differences in a chromatograph incurs considerable time.

An alternative approach is to employ spectral imaging to differentiate apparently similar inks. Spectral imaging captures subtle differences in the inks which is valuable for mismatch detection as shown in Fig. 1. A hyperspectral image (HSI) is a series of discrete narrow-band images in the electro-magnetic spectrum. In contrast to a three channel RGB image, an HSI captures finer detail of a scene in the spectral dimension as shown in Fig. 2. Hyperspectral imaging has recently emerged as an efficient non-destructive tool for detection and identification of forensic traces such as bloodstain analysis and latent print analysis [4]. It is also useful to questioned document examination for ink discrimination, document age estimation and restoration of historical documents [5–8].

Brauns and Dyer [9] developed a hyperspectral imaging system for forgery detection in potentially fraudulent documents in a non-destructive manner. They prepared written documents with blue, black and red inks and later introduced alterations with a different ink of the same color. Using fuzzy clustering, the ink spectra were grouped into one or more cluster indicating their degree of association. They qualitatively showed that the inks can be separated into two different classes. The absence of quantitative results and slow imaging process collectively limit the applicability of their system to practical ink mismatch detection. In contrast, our approach is quantitative instead of qualitative-subjective analysis.

A relatively improved hyperspectral imaging system for the analysis of historical documents in archives was developed by Padaon et al. [10]. Their use of narrowband tunable light source reduced the chances of damage to a document due to excessive heat generated by a strong broadband white light source. However, its extremely slow acquisition time (about 15 min) [11] resulted in prolonged exposure. Therefore, the benefit gained by a tunable light source may be nullified and the productivity of the system in terms of the number of documents captured is reduced. Our proposed system captures hyperspectral images in only a fraction of that time using an *electronically tunable filter* which is fast, precise and has no moving parts.

* Corresponding author.
E-mail address: zohaib.khan@uwa.edu.au (Z. Khan).

http://dx.doi.org/10.1016/j.patcog.2015.04.008
0031-3203/© 2015 Elsevier Ltd. All rights reserved.

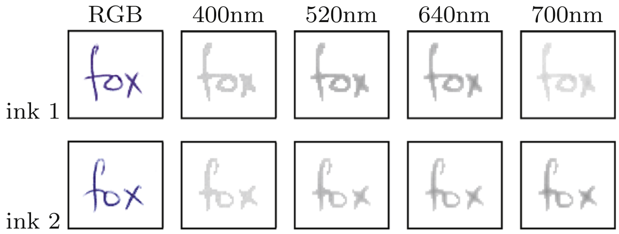

Fig. 1. The images highlight the discrimination of inks at different wavelengths offered by spectral imaging. A word written in two different blue inks is shown in this example. Observe that the two inks appear similar at short wavelength and gradually appear different at longer wavelengths. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

In commercial hyperspectral document imaging systems [4,12] the examiner needs to select a suspected portion of a note for ink mismatch analysis. The examiner has to search through hundreds of combinations of the different wavelengths to visually identify the differences in inks, which is laborious. For instance, to analyze a 33 band HSI in an exhaustive search, the total number of band combinations is of the order of $\approx 10^{10}$, which is not feasible in time critical scenarios. This procedure is not required in our proposed *automatic* document analysis approach. Note that acquisition of all bands is time consuming and limits the number of documents that can be scanned in a given time. Moreover, the resulting data is huge and some bands with low energy contain significant system noise. Therefore, it is desirable to select the most informative subset of bands, thereby reducing the acquisition time, and increase the accuracy by getting rid of the noisy bands. Since, hyperspectral images are densely sampled along the spectral dimension, the neighboring bands are highly correlated. This redundancy makes hyperspectral images a good candidate for sparse representation and band selection [13].

Hedjam and Cheriet [14] proposed a hyperspectral band selection algorithm based on graph clustering. They created a band adjacency graph in which the nodes represent the bands and the edges represent the similarity weights between bands. Markov clustering was then used to form distinct clusters of highly correlated bands. In contrast, Martinez et al. [15] proposed a hierarchical clustering structure which minimizes certain distance measures (KL-divergence, MI, etc.) to reduce redundancy in adjacent bands. In both approaches, clustering results in groups of highly correlated bands, and a new challenge of selecting a representative band which accounts for maximum information within each cluster arises. Unlike clustering based methods, sparse representation explicitly indicates a selection of bands which is representative of the maximum information in the complete hyperspectral data. One such approach is to optimize a subspace projection while simultaneously selecting features in a supervised manner [16].

We propose joint sparse PCA (JSPCA) that computes a PCA basis by explicitly removing the non-informative bands in an unsupervised manner. The joint sparsity ensures that all basis vectors share the same sparsity structure whereas the complete hyperspectral data can be represented by a sparse linear combination of the bands. Our work is most similar to Xiaoshuang et al. [17], who solved a linear system for sparse PCA with an $\ell_{2,1}$ constraint for joint sparsity. However, a major drawback of their closed form regression is the instability of solution at high joint sparsity constraint. In contrast, our proposed algorithm produces jointly sparse solutions due to a robust iterative optimization scheme. We demonstrate the joint sparse band selection (JSBS) algorithm for hyperspectral ink mismatch detection and show that it selects fewer bands with higher accuracy compared to the state-of-the-art.

This paper is a significant extension of our preliminary work on ink mismatch detection [18]. The main contributions of the work are:

- A novel *joint sparse PCA* (*JSPCA*) algorithm for dimensionality reduction and feature selection.
- A *joint sparse band selection* (*JSBS*) technique for automatic ink mismatch detection.
- An *efficient hyperspectral document imaging system* and collection of a *new public database* of writing ink hyperspectral images.[1]

The rest of this paper is organized as follows. In Section 2, we present the proposed ink mismatch detection methodology. In Section 3, we describe the database specifications, acquisition and normalization. Section 4 provides details of the experimental setup, evaluation protocol, and analysis of the results. The paper is concluded with a discussion in Section 5.

2. Ink mismatch detection

Ink mismatch detection is based on the fact that the same inks exhibit similar spectral responses whereas different inks are spectrally dissimilar [18]. We assume that the spectral responses of the inks are independent of the writing styles of different subjects (which is a spatial characteristic). Thus, unlike works that identify hand writings by the ink-deposition traces [19], our work solely focuses on the spectral responses for ink discrimination. In the proposed ink mismatch detection framework, the initial objective is to segment handwritten text from the paper. The next task is to select features (bands) from the ink spectra by the proposed band selection technique. Finally, the class membership of each ink pixel is determined by clustering of the ink spectral responses using selected features.

Notations: In the following text, a scalar is denoted as lowercase letter (i), a vector as a lowercase letter in bold font ($\mathbf{x}$), and a matrix as an uppercase letter in bold font ($\mathbf{X}$). X_{ij} is the (i,j)th element of a matrix. $\mathbf{x}^i$ is the ith row and $\mathbf{x}_j$ is the jth column of a matrix. $\mathbf{X}_{\cdot\mathcal{S}}$ is the submatrix obtained by indexing the columns of $\mathbf{X}$ by the index set $\mathcal{S}$ ($\mathbf{X}^{\cdot\mathcal{S}}$ indexes the rows). All vectors are treated as column vectors.

2.1. Handwritten text segmentation

Consider a three dimensional HSI $\mathbf{I} \in \mathbb{R}^{x \times y \times p}$, where (x,y) are the number of pixels in spatial dimension and p is the number of bands in the spectral dimension. The objective is to compute a binary mask $\mathbf{M} \in \mathbb{R}^{x \times y}$ which associates each pixel to the foreground or background. The ink pixels (text) make up the foreground and the blank area of the page is the background. A global image thresholding method, such as the Otsu [20] is ineffective because of the non-uniform illumination over the document (Fig. 3(a)). A local image thresholding method such as Sauvola and Pietikäinen [21] with an efficient integral image based implementation [22] more effectively deals with such illumination variations. The Sauvola's method generates a binary mask according to

$$M_{ij} = \begin{cases} 1 & \text{if } I_{ij} > \mu_{ij}\left(1 + \kappa\left(\frac{\sigma_{ij}}{r-1}\right)\right) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

[1] UWA Writing Ink Hyperspectral Image Database http://www.csse.uwa.edu.au/~ajmal/databases.html.