Science and Justice 53 (2013) 55–59

Contents lists available at SciVerse ScienceDirect

Science and Justice

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

Objective ink color comparison through image processing and machine learning

Charles E.H. Berger *

Netherlands Forensic Institute, P.O. Box 24044, 2490 AA The Hague, The Netherlands
Leiden University, P.O. Box 9520, 2300 RA Leiden, The Netherlands

ARTICLE INFO

Article history:
Received 30 January 2012
Received in revised form 5 September 2012
Accepted 6 September 2012

Keywords:
Forensic science
Questioned documents
Ink comparison
Color deconvolution
Image analysis
Feature vector

ABSTRACT

Making changes or additions to written entries in a document can be profitable and illegal at the same time. A simple univariate approach is first used in this paper to quantify the evidential value in color measurements for inks on a document coming from a different or the same source. Graphic, qualitative discrimination is then obtained independently by applying color deconvolution image processing to document images, with parameters optionally optimized by support vector machines (SVM), a machine learning method. Discrimination based on qualitative results from image processing is finally compared to the quantitative results of the statistical approach. As color differences increase, optimized color deconvolution achieves qualitative discrimination when the statistical approach indicates evidence for the different source hypothesis.

1. Introduction

This paper deals with the type of questioned document case where the accusation is that entries were changed or added after the signing of some official document. In such a case a comparison of the colors of the ink of the original writing and suspected changes can be important. It should be noted that the actual ink color comparison is just one out of many factors when considering whether an accused made fraudulent changes to a document. Other factors include the comparison of the handwriting, motive, skill, access to the original pen, and time of the addition. Our ink color comparison approach is entirely based on the analysis of images of the document; other non-destructive optical methods for the comparison of inks have been described elsewhere [1–4].

We define the relevant hypotheses for an ink color comparison of samples *A* and *B* (with colors *a* and *b*), as follows:

H_s Samples *A* and *B* come from the *same* specific instance of a blue ballpoint.

H_d Sample *A* comes from some other blue ballpoint than sample *B*.

Our aim is two-fold: First, we will find a quantitative measure for the color difference of the inks on the document, and the value of the evidence in the colors for the hypotheses mentioned above. That can be done by defining a feature vector for the ink color and analyzing within source and between source variation [5]. Color differences and variation are all based on these feature vectors rather than the colors themselves. Second, we will qualitatively discriminate inks by processing the images with a method called color deconvolution [6].

We will present a method to obtain optimized parameters for color deconvolution and with that, optimized qualitative discrimination. We will then be able to compare the qualitative results after the image processing to the quantitative results from the simple statistical analysis. While the benefits of being able to quantify evidential value are clear, one might wonder whether the qualitative discrimination outperforms quantitative discrimination because it employs the human visual system.

2. Methods

2.1. Preparation of the samples

For this study, 262 *blue* ballpoint pens from the collection of the Netherlands Forensic Institute (NFI) were used. Samples to be used for population data were prepared by writing lines with all 262 ballpoint pens on a single sheet of standard white copy paper. To get an impression of the intra-source (or within source) variation for a single ballpoint pen, 25 samples were written with the same ballpoint pen on the same sheet. The imaging was done by scanning all samples in one large, high resolution scan (1270 dpi, or pixels of 20×20 μm), with a high quality scanner (CreoScitex Eversmart Jazz). After acquiring the image, it was sliced into a collection of images of all the separate samples.

Samples to be used for color separation experiments were made by choosing 2 pens and writing the number of the first pen three

* Netherlands Forensic Institute, P.O. Box 24044, 2490 AA The Hague, The Netherlands.
E-mail address: c.berger@nfi.minvenj.nl.

1355-0306/$ – see front matter
http://dx.doi.org/10.1016/j.scijus.2012.09.003

times. The second pen was then used to write its number over the second entry of the number of the first pen, and cross out the third with a spiraling line; and finally to write the number of the second pen on the right side.

2.2. Extracting the colors

All image processing and calculations were carried out with MATLAB® (The Mathworks, Inc., Bioinformatics Toolbox™). Colors of the inks and paper were extracted by segmentation of the image into areas with ink or paper only.

The segmentation was carried out as follows. First, the image was reduced from full color to gray levels, then reduced to black and white with a threshold determined by Otsu's algorithm [7]. After that, a mask was created using binary morphological operations. For the ink mask, the image was reduced to a cleaned up skeleton of the inked entries. This skeleton was dilated to result in a mask that stays within the inked portion of the image. The paper mask was formed by eroding the segmented paper part of the image, to make sure it only covers the paper background area.

Fig. 1 shows an example of the segmentation with the original image, both masks and the combined result. With the masks, the average color of a segment can be determined, but all the individual pixels of the segment can be analyzed as well.

2.3. Defining the feature vector and univariate analysis

Rather than comparing colors directly, we will compare derived feature vectors instead. Our feature vector should take into account that the color of the ink in the document image will not only depend on the source, but also on the thickness of the deposited layer of the ink, and the background color of the paper. Fig. 2 shows the three-dimensional histogram of the red, green and blue (RGB) color components in a document image with 2 inks (black and blue) on white paper. The within variation of the RGB colors results in elongated clusters for both inks, that extend from the spherical paper background color cluster. This is due to differences in ink coverage in the pixels in and along the edge of the ink line. Colors of pixels of areas with low ink coverage do not differ much from the color of the paper background, while pixels of areas with high ink coverage have the color of the pure ink. The colors of all other pixels of inked areas vary between these two extremes. Even for a black and a blue ink, the variation in RGB colors from one ink is often larger than that between the inks. Using the RGB colors directly for discrimination would therefore not give good results.

For the purpose of discrimination, we want to minimize within variation while maximizing between variation. We therefore choose the two average spherical angles (x,y) of the elongated clusters with the average background color as the origin to define the feature vector (see Fig. 2). This feature vector captures the direction of the elongated clusters of ink colors and minimizes within source variation

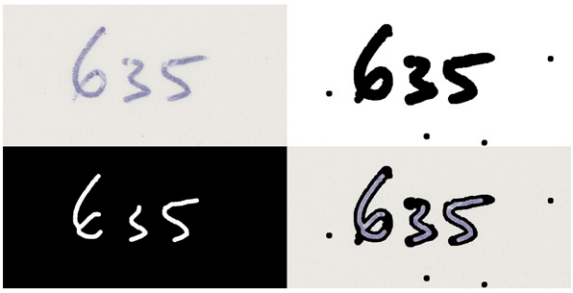

Fig. 1. Segmentation of writing samples, with from upper left to bottom right: original image; background segment (white); ink segment (white); composite of background and ink segment from original, with remaining pixels in black.

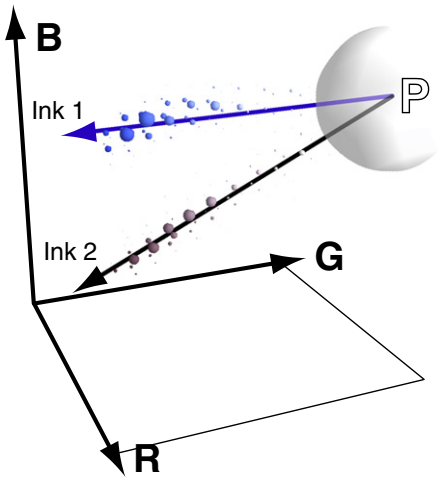

Fig. 2. Three-dimensional histogram of the red, green and blue (RGB) color components in a document image with 2 inks (black and blue) on white paper.

while maximizing between source variation, thus optimizing discrimination. With two spherical angles our feature vectors are defined in a 2D feature space. Our comparison is simply defined as the Euclidian distance between two feature vectors in feature space, resulting in a color distance or difference.

To study the within and between source variation for the blue ballpoint pen collection, the feature vectors of the average ink colors of all samples were extracted. We will carry out a simple univariate analysis for the estimation of the evidential values for the hypotheses of same or different source, based on the color differences of multiple ink traces from a single pen (within source variation) and single traces of multiple pens (between source variation). More advanced bivariate methods were described in an earlier paper (see Ref. [5]).

2.4. Color deconvolution

Based on the extracted colors and feature vectors we can use color deconvolution image processing to also obtain qualitative color discrimination (Ref. [6]). The essence of the color deconvolution algorithm is based on a coordinate transformation. The vectors from the background color (P) to the respective ink colors (Ink 1, Ink 2) can be used as a basis in 3D RGB color space (see Fig. 2). When a third vector perpendicular to these two is defined, every color can be defined as a combination of these 3 vectors rather than as a combination of RGB components. The advantage of having colors expressed in those alternative unit vectors is that we can now easily remove a color component, or display components that are to be discriminated as green and red on a white background. For the purpose of this paper the components to be discriminated will be shown as a lighter and darker one on a neutral gray background.

The simplest way to determine the parameters for color deconvolution is to determine the average colors that are to be discriminated with the same masking method as before. These average colors give the aforementioned vectors that define the new basis in 3D RGB color space.

2.5. Support vector machines

In an alternative approach we do not average the colors from the segmented areas, but we determine the feature vectors for separate pixels within the segmented inked areas. This will lead to two (possibly overlapping) clusters in feature space, related to the Ink 1 and Ink 2 segment pixels. The extent to which two clusters can be seen in