İSTANBUL BÜYÜKŞEHİR BELEDİYESİ

İstanbul Üniversitesi

KUMID
KÜLTÜREL MİRASIN DOSTLARI DERNEĞİ

Consiglio Nazionale delle Ricerche
Istituto di Metodologie Chimiche

Association Investing in Culture

5th International Congress on

"Science and Technology for the Safeguard of Cultural Heritage in the Mediterranean Basin"

Istanbul, Turkey

22 – 25 November 2011

RESTORATION OF HISTORICAL RGB MANUSCRIPTS VIA CORRELATED COMPONENT ANALYSIS*

Bedini Luigi[1], **Tonazzini Anna**[1]

[1]National Research Council of Italy, ISTI-CNR, Pisa, Italy, name.surname@isti.cnr.it

One of the most common degradations affecting historical documents which are written or printed on both sides of the page is see-through, that is an undesired pattern in the background, caused by the text in the reverse side of the page. Such distortion can significantly degrade the readability of the document or make difficult the automatic analysis of its content.

Several approaches for see-through reduction have been investigated, mainly for greyscale documents, and exploiting the availability of scans of both sides (*recto* and *verso*). Recently, the interest in applying Blind Source Separation (BSS) algorithms for solving this problem has increased noticeably. The appearance of the degraded recto and verso scans is first modelled as a parametric superimposition of the uncorrupted recto and verso images, and then a separation algorithm is used to estimate both the mixing parameters and the ideal front and back side images (*sources*). The assumption of a linear mixing model has led to BSS algorithms such as Independent Component Analysis (ICA) or Non-negative Matrix Factorization (NMF). Some works have also addressed more realistic non-linear and/or convolutional mixing models, and separation algorithms based on adaptive filtering, wavelet transforms, or image regularization techniques.

In spite of the assumption of a linear mixing model, ICA has proven to be very cost-effective and versatile for application to different typologies of data and several instances of document restoration and analysis. For example, it can be easily extended to the analysis of multispectral scans of a single-sided document containing multiple information layers, or when the data are the RGB recto and verso scans of a colour document. This latter case is particularly interesting if the aim is to produce a restored visible document that, while cleansed of the unwanted interferences, maintains its useful features, e.g. the original colour, as much as possible. Nevertheless, ICA assumes independence or at least uncorrelation of the individual sources, that is, it forces an unrealistic uncorrelation between the recto and verso ideal images.

In this paper, we consider the problem of removing interferences from pairs of recto-verso RGB documents, and show that the use of recently proposed Correlated Component Analysis techniques, based on second order statistics, allows to remove the uncorrelation assumption. Our method, working in the Fourier domain, is based on the joint estimation of the mixing parameters and the source spectra and cross-spectra. This estimation is performed by alternating minimization, with respect to the mixing parameters and the spectra, of a suitable cost function. Once the estimates are available, the individual sources can be recovered, at each Fourier mode, either by simply inverting the mixing matrix, or, when noise is present, by Wiener filtering, since the estimated spectra can be effectively exploited. The method is very fast, and can be easily extended to account for blur kernels on the sources. Indeed, blur usually affects the interfering patterns, for effect of light or ink spreading through the support. The experimental results performed show that this method significantly outperforms ICA, by permitting separation to be achieved also when the individual sources are largely correlated. This is especially true when the patterns that interfere from a side to the other of the page are sensibly blurred.

* This work is partially supported by European funds, through the Calabria Region program PIA 2007-2013, project no. 1220000119 AMMIRA - Multispectral acquisition, enhancement, indexing and retrieval of artworks.