retrieval of given words in documents can be interpreted as matching of an input graph (keyword) with a large set of graphs (document). More formally, in order to spot a certain keyword w_i, all p graph instances $g_{i1}, \ldots, g_{ip}$ of that word w_i occurring in the training set are matched against all graph words in each text line using our adapted graph matching procedure. That is, for a given word w_i and a specific text line s pairwise distances between all prototypical graphs $g_{i1}, \ldots, g_{ip}$ and the m word graphs $g_{'1}, \ldots, g_{'m}$ from text line s are obtained first. The minimum of these graph distances serves as a distance function $d(w_i,s)$ of the keyword's word class w_i to the text line s. If the distance $d(w_i,s)$ of a keyword to the text line is below a given threshold, the text line s and the word from s having the minimum distance is returned as a positive match to the keyword w_i.

A very important aspect of the whole project is the experimental evaluation of our novel graph based procedure for keyword spotting. We plan to carry out exhaustive experimental evaluations on the Miroslav Gospels. Miroslav Gospels is a 362-page illuminated manuscript Gospel Book on parchment with very rich decorations. It is one of the oldest surviving documents written in Old Church Slavonic and represents one of the most precious and significant documents in cultural heritage of Serbia (see Figure 1 for a detailed view of a page of the Miroslav Gospel).

Figure 1: The Miroslav Gospel (source: http://upload.wikimedia.org/wikipedia/commons/3/36/Miroslavs_Gospel.jpg)

The outlined project has been submitted as a proposal for joint research project to the SCOPES programme (Scientific co-operation between Eastern Europe and Switzerland). The research project will be mainly carried out by the Technical Faculty in Bor at the University of Belgrade (Serbia) and the Institute for Information Systems at the University of Applied Sciences and Arts Northwestern Switzerland. Further information about the SCOPES programme can be found at http://www.snf.ch/E/international/europe/scopes/Pages/default.aspx

Links:
Video lecture on our novel algorithmic framework for approximate graph distances:
http://videolectures.net/gbr07_riesen_bgm
The algorithmic framework for approximate graph distances:
http://www.fhnw.ch/wirtschaft/iwi/gmt
Miroslav Gospels:
http://en.wikipedia.org/wiki/Miroslav_Gospel
SCOPES programme:
http://www.snf.ch/E/international/europe/scopes

References:
[1] T. M. Rath, R. Manmatha: "Word spotting for historical documents", Int. Journal on Document Analysis and Recognition, p. 139-152, 2007
[2] K. Riesen, H. Bunke: "Approximate graph edit distance computation by means of bipartite graph matching", Image and Vision Computing, 27(4):950-959, 2009.
[3] A. Fischer, K. Riesen, H. Bunke: "Graph similarity features for HMM-based handwriting recognition in historical documents", in proc. Int. Conf. on Frontiers in Handwriting Recognition, pages 253–258, 2010.

Please contact:
Kaspar Riesen
University of Applied Science and Arts Northwestern Switzerland, Olten, Switzerland
Tel: +41 79 688 77 19
E-mail: kaspar.riesen@fhnw.ch

Highly Degraded Recto-verso Document Image Processing and Understanding

by Emanuele Salerno and Anna Tonazzini

The ITACA project (Innovative tools for cultural heritage archiving and restoration) is investigating new approaches to treat severe back-to-front interference in digital images of two-sided documents. This work is part of a vast research program on the study and preservation of historical documents, which, since 2004, has been supported in various forms by European funds.

Distinguishing the foreground pattern (eg text or graphics) from interference is a basic step in document understanding. When dealing with back-to-front interference, this is not always an easy task, as the front and rear patterns are mixed nonlinearly and the interference strength varies from point to point. When imaging a physical document, the transparency of the paper combined with the features of the capture device cause the content of each side to appear in the opposite side's image as "show-through interference". A similar effect (bleed-through) is caused by the ink bleeding from one side to the other.

These effects are intrinsically nonlinear. Moreover, the non-uniformity of the paper support, the varying strength of text strokes, and possible uncompensated variations in illumination often make these processes non-stationary. Based on a slight generalization of a model proposed in the literature for moderate show-through, we presented a

nonlinear model that can account for show-through and, partly, for bleed-through [1]. We then augmented the model by introducing a space varying interference strength [2]. This new model is not yet fully non-stationary, as the convolutive effects and the typical non-linearities are considered constant throughout the image, but the experimental results compare very favourably with previous approaches, and the possibilities of discriminating foreground and interference are broadened significantly.

Figure 1 shows our data model: two coupled nonlinear equations, written in terms of the front and rear optical densities. As is known, optical density is related logarithmically to the commonly used reflectance value. The appearance of each side of the document is a nonlinear combination of the foreground pattern and a blurred and attenuated version of the pattern on the opposite side. Given the front and rear-side appearances, inverting this model means estimating the show-through point spread functions (PSF's), the show-through gains, and the front and rear-side pure patterns. Strictly speaking, a non-stationary model

$$D_r^{obs} = D_r + q_v[h_v * (1 - e^{-D_v})]$$

$$D_v^{obs} = D_v + q_r[h_r * (1 - e^{-D_r})]$$

$D_{r,v}^{obs}$ = *recto* and *verso* observed optical densities

$D_{r,v}$ = *recto* and *verso* pure density patterns

$q_{r,v}$ = *recto* and *verso* (space variant) show-through gains

$h_{r,v}$ = *recto* and *verso* (space invariant) show-through PSF's

Figure 1: Non-linear and non-stationary model for back-to-front interference. The asterisk means convolution.

should include two space-variant kernels rather than two PSF's, and the show-through gains should depend on space. As a first attempt to treat non-stationary back-to-front interference, we use fixed PSF's all over the image pair while allowing the gains to depend on space. Approximated PSF's can simply be evaluated beforehand; once this is done, a straightforward formula allows us to estimate the gains in all the pixels where interference is present. An easy and fast constrained maximum likelihood scheme is then used to estimate the pure patterns. The results obtained are very promising, even though a general strategy is still to be found to cope with the saturation effects occurring where the foreground pattern occludes the show-through.

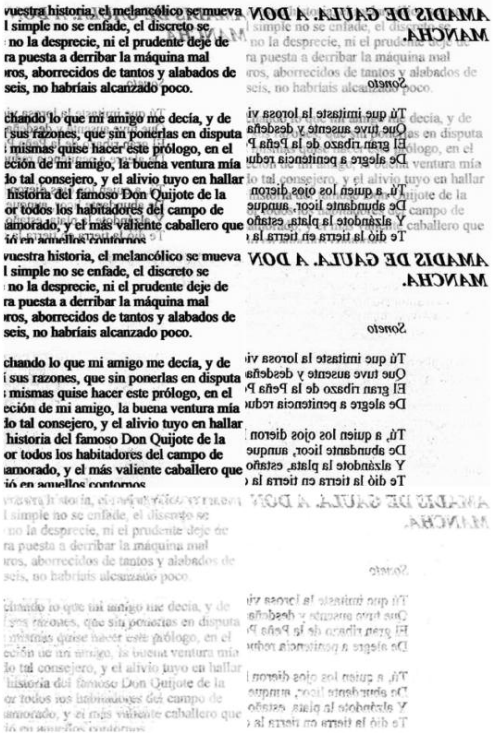

Figure 2: Example of non-stationary show-through removal. Top: observed image pair (appearance). Middle: restored image pair (estimated pure patterns). Bottom: show-through gain maps (linear grayscale, black=maximum).

Figure 2 shows a typical result obtained in an apparently non-stationary case: the removal of the back-to-front interferences is very effective, and the gain maps show us that the distortion is significantly non-stationary. A comparison between the new results and those obtained through stationary models confirms the potential advantages achievable by accounting for non-stationarity. The type of fixed non-linearity we adopt is the same as that introduced for weak show-through in the paper that inspired our work [3]. It is likely that a stronger show-through or other phenomena, such as bleed-through, would need different non-linearities. Finding a suitable model for documents showing different kinds of interference might even prove to be impossible. In such cases, introducing non-stationarity could produce an additional advantage: providing an accurate and comprehensive data model could become less important - we could even return to a linear model. This is the focus of our current research, and our first results are corroborating our conjecture.

The ITACA project (POR-FESR Calabria 2007-2013) is led by TEA Sas, Catanzaro, Italy, a firm that provides consulting and cultural heritage digitization, processing, and management. The activity described here is being conducted at ISTI-CNR, Pisa. The research teams at ISTI and TEA are both members of ERCIM's MUSCLE working group, on Multimedia Understanding through Semantics, Computation, and Learning.

Links:

http://www.isti.cnr.it/research/unit.php?unit=SI

http://www.teacz.com

References:

[1] E. Salerno et al: "Nonlinear model identification and see-through cancellation from recto-verso documents", Int. J. Docum. Anal. Recogn., 2013, DOI 10.1007/s10032-012-0183-y

[2] A. Tonazzini, et al: "Removal of non-stationary see-through interferences from recto-verso documents", in Machine Learning and Data Mining in Pattern Recognition, 2013, ISBN 978-3-942952-22-4, pp. 151-158),

[3] G Sharma: "Show-through cancellation in scans of duplex printed documents", IEEE Trans. Im. Proc., 2001, DOI 10.1109/83.918567)

Please contact:

Emanuele Salerno
ISTI-CNR, Italy
E-mail: emanuele.salerno@isti.cnr.it