Number 111 October 2017

ERCIM NEWS

www.ercim.eu

Special theme:

Digital Humanities

Also in this issue:

Research and Innovation:

Trend Analysis of Underground Marketplaces

2) the manner in which data that are not digitised or shared become "hidden" from aggregation systems;
3) the fact that data is human created, and lacks the objectivity often ascribed to the term;
4) the subtle ways in which data that are complex almost always become simplified before they can be aggregated.

We approach these questions with a humanities research perspective and remain committed to humanities methodologies and forms of knowledge, but we make use of social science research tools to look at both the humanistic and computer science approaches to the term "data" and its many possible meanings and implications. Our core shared discourse of the digital humanities allows us to use these methods and knowledge in a contextualised, socially relevant manner, a strength of our consortium that is further enhanced by our inclusion of both ethnographic/anthropological and industrial perspectives.

Led by Trinity College Dublin, the KPLEX team spans four countries, taking in Freie Universität Berlin (Germany), DANS-KNAW (The Hague) and TILDE (Latvia). Each of the K-PLEX project partners addresses an integrated set of research questions and challenges. The research teams have been assembled to pursue a set of questions that are humanist-led, but broadly interdisciplinary, including humanities and digital humanities, data management, anthropology and computer science, but also including stakeholders from outside of academic research able to inform the project's evidence gathering and analysis of the challenges, including participation from both a technology SME (TILDE) and a major national ICT research centre (ADAPT, Ireland). In addition, KPLEX takes in the experiences of a large number of major European digital research infrastructure projects federating cultural heritage data for use by researchers, through the contributions by TCD (Dublin) and KNAW-DANS (The Hague). These projects (including CENDARI, EHRI, DARIAH-EU, DASISH, PARTHENOS, ARIADNE and HaS) have all faced and progressed the issues surrounding the federation and sharing of cultural heritage data. In addition, two further projects that deal with non-scientific aspects of researcher epistemics are also engaged, namely the "Scholarly Primitives and Renewed Knowledge Led Exchanges" project (SPARKLE, based at TCD) and the "Affekte der Forscher" (based at FUB). These give the KPLEX team and project a firm baseline of knowledge for dealing with the question of how epistemics creates and marks data.

The KPLEX project kicked off in January 2017, and will conclude in March 2018, presenting its results via a composite white paper that unites the findings of each research team, with each research team also producing a peer reviewed academic paper on their findings. Over the coming months the project will be represented at DH conferences in Liverpool ("Ways of Being in the Digital Age"), Austria ("Data First!? Austrian DH Conference"), Manchester ("Researching Digital Cultural Heritage International Conference") and Tallin ("Metadata and Semantics Research Conference").

Links:
[L1] https://kplex-project.com/,
Twitter: @KPLEXProject,
Facebook: KPLEXProject

References:

[1] T Presner: "The Ethics of the Algorithm", in Probing the Ethics of Holocaust Culture.
[2] J Edmond: "Will Historians Ever Have Big Data?" In Computational History and Data-Driven Humanities. doi:10.1007/978-3-319-46224-0_9.
[3] L Gitelman, ed., "'Raw Data' is an Oxymoron".

Please contact:

Jennifer Edmond, Georgina Nugent Folan, Trinity College Dublin, Ireland
edmondj@tcd.ie, nugentfg@tcd.ie

Restoration of Ancient Documents Using Sparse Image Representation

by Muhammad Hanif and Anna Tonazzini (ISTI-CNR)

Archival, ancient manuscripts constitute a primary carrier of information about our history and civilisation process. In the recent past they have been the object of intensive digitisation campaigns, aimed at their preservation, accessibility and analysis. At ISTI-CNR, the availability of the diverse information contained in the multispectral, multisensory and multiview digital acquisitions of these documents has been exploited to develop several dedicated image processing algorithms. The aim of these algorithms is to enhance the quality and reveal the obscured contents of the manuscripts, while preserving their best original appearance according to the concept of "virtual restoration". Following this research line, within an ERCIM "Alain Bensoussan" Fellowship, we are now studying sparse image representation and dictionary learning methods to restore the natural appearance of ancient manuscripts affected by spurious patterns due to various ageing degradations.

The collection of ancient manuscripts serves as history's own closet, carrying stories of enigmatic, unknown places or incredible events that took place in the distant past, many of which are yet to be revealed. These manuscripts are of great interest and importance for historians to study people of the past, their culture, civilisation and way of life. Most of the ancient classic documents have had a very narrow escape from total annihilation. Thus, digital preservation of our documental heritage has been one of the first focusses of the massive archive and library digitisation campaigns performed in the recent years. This, in turn, has contributed to the birth of digital humanities as a science. In addition to preservation, computing technologies applied to the digital images of these documents have quickly become a powerful and versatile tool to simplify their study and retrieval, and to facilitate new insights into the documents' contents.

The quality of the digital records, however, depends on the current status of the original manuscripts, which in most cases are affected by several types of degradation, such as spots, ink fading, or ink seeping from the reverse side, due to bad storage conditions and the fragile nature of the materials (e.g., humidity, mould, ink chemical properties, paper porosity). In particular, the phenomenon of ink seeping is perhaps the most frequent one in ancient manuscripts written on both sides of the sheet. This effect, termed as bleed-through, becomes visible as an unpleasant and disturbing degradation pattern, severely impairing legibility, aesthetics, and interpretation of the source document.

In general, bleed-through removal is addressed as a classification problem, where image pixels are labelled as either background (paper texture), bleed-through (seeped ink), or foreground (original text). This classification problem is very difficult, since the intensity of both foreground text and bleed-through pattern can be so highly variable as to make it extremely hard or even impossible to distinguish them. In addition, when the aim is to obtain a very accurate and plausible restoration, another very critical issue is to replace the identified bleed-through pixels with appropriate replacement colour values, which do not alter the original look of the manuscript.

Recently, we proposed a two-step method to address bleed-through document restoration from a pre-registered pair of recto and verso images of the manuscript. First, the bleed-through pixels are identified on both sides [1]; then, a sparse representation based image inpainting technique is applied to fill-in the bleed-through pixels, by taking into account their propensity to aggregation, and in accordance with the natural texture of the surrounding background.

Sparse representation methods are reported with state-of-the-art results in different image processing applications. These methods process the whole image by operating on a patch-by-patch level. For this specific application, our aim is to reproduce the background texture to maintain the original look of the document. In the sparse representation setup, an over-complete dictionary is learned using a set of training patches from the recto and verso image pair. In the training set we only select patches with no bleed-though pixels. This choice speeds up the training process since it excludes non-informative image regions. For each patch to be inpainted, we first search for its mutual similar patches in a small bounded neighbourhood window. We used a block matching technique with Euclidean distance metric as similarity criterion. The similar patches are grouped together in a matrix. A group-based sparse representation method [2] is exploited to find the befitting fill-in for the bleed-through strokes. The use of similar patch groups incorporates local information that helps to preserve the natural colour/texture continuation property of the physical manuscript.

An original degraded manuscript and its restored version are presented in Figure 1. It is worth noting that our algorithm can be directly applied to inpaint any other possible interference pattern detected in the paper support (e.g., stains). We are also studying the extension of the method to the restoration of broken or faded foreground characters.

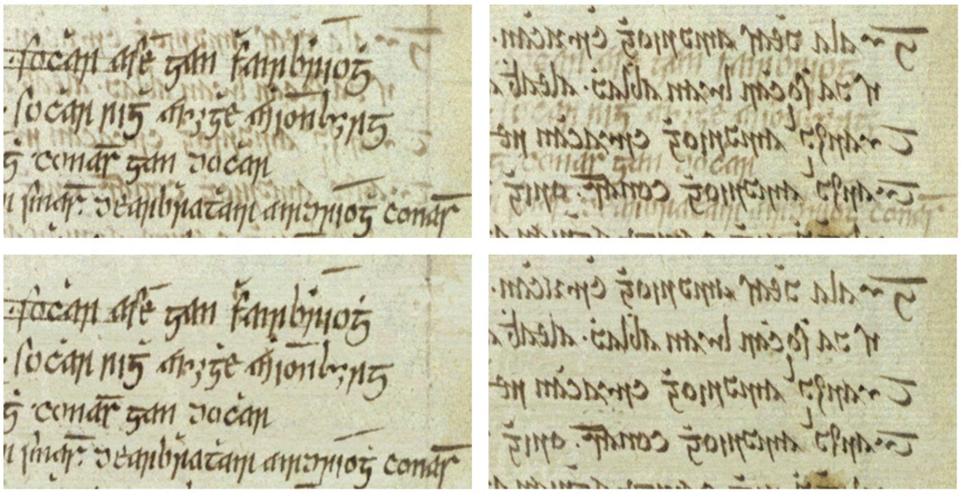

Figure 1: A visual comparison of an original ancient manuscript effected by bleed-through degradation and its restored version using our method. The first row shows the degraded recto and verso pair, and the restored images are presented in the second row.

References:

[1] A. Tonazzini, P. Savino, and E. Salerno: "A nonstationary density model to separate overlapped texts in degraded documents," Signal, Image and Video Processing, vol. 9, pp. 155–164, 2015.

[2] J. Zhang and D. Zhaocand W. Gao: "Group-based sparse representation for image restoration," IEEE Trans. Image Process., vol. 32, pp. 1307–1314, 2016.

Please contact:

Anna Tonazzini, ISTI-CNR, Pisa
+39 3483972150
anna.tonazzini@isti.cnr.it