Pattern Recognition 47 (2014) 248–259

Contents lists available at ScienceDirect

Pattern Recognition

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

Virtual unrolling and information recovery from scanned scrolled historical documents

Oksana Samko *, Yu-Kun Lai, David Marshall, Paul L. Rosin

School of Computer Science & Informatics, Cardiff University, Queens Buildings, 5 The Parade, Cardiff CF24 3AA, UK

ARTICLE INFO

Article history:
Received 16 November 2012
Received in revised form
6 June 2013
Accepted 11 June 2013
Available online 3 July 2013

Keywords:
Parchment restoration
Digital unwrapping
Document processing
Text retrieval
Volumetric scanning

ABSTRACT

The objective of our work is to enable the reading of fragile scrolled historical parchments without the need to physically unravel them, thus providing valuable information to a wide range of scholarly disciplines. This problem has not been investigated by the computer vision community properly yet due to the need for parchment scanning technology: standard X-ray equipment is not sufficient as there is a requirement to extract out parchment ink in addition to the parchment's underlying structure. Effective data recovery is also compromised as content from historical scrolled documents is inaccessible due to the deterioration of the parchment. We create a 3D volumetric model of a scrolled parchment's underlying geometry and perform digital unwrapping of the parchment, producing a readable image of the text as an output. The proposed recovery framework consists of structure preserving anisotropic filtering in combination with robust segmentation, surface modelling and ink projection. We demonstrate with real examples how our algorithm is able to recover the underlying text and to solve the major challenge for scrolled parchment analysis, namely segmentation of connected layers and processing the data without user interaction.

1. Introduction

Much of the history of the western world is written on parchment, a dry, treated, skin-derived writing medium [25]. The material was primarily designed as a writing medium that was smooth and flat; durability over millennia was probably not a prime consideration. Now, the information content of this complex medium is sometimes impossible to access without causing considerable damage or permanently altering the object to an unacceptable level. In some cases, their physical deterioration is at such an advanced state that any attempt to unravel the document manually would cause catastrophic fragmentation, destroying the internal information. Use of X-ray microtomography, a new direction in digital document analysis [18], provides a digital copy of a scrolled parchment as a 3D volumetric object, see Fig. 1. We utilise this 3D representation to recreate a virtual parchment model as an input for a subsequent information recovery framework.

Digital document restoration has been an extremely active area of research in recent years [6,8,12,15,23,28,32,35]. Current efforts have provided a new level of accessibility to many valuable literary works. However, not much attention has been paid to the analysis of scrolled parchments. Traditionally document restoration approaches concentrate on regular photographic images and non-scrolled surfaces [20], which are easier to process.

Brown and Seales [6] proposed a general de-skewing algorithm for arbitrary warped documents based on 3D shape. Doncescu et al. [12] reported a similar method, where a laser projector is used to project a 2D light network on the document surface to capture 3D shape, and then 2D distortions of the surface are corrected with a two-pass mesh de-warping algorithm. Cao et al. [8] presented an algorithm to rectify the warping of a bound document image: they built a general cylindrical model, and then used the skeleton of horizontal text lines in the image to estimate the model parameters. Pilu [23] introduced a method for distorted document restoration which is based on physical modelling of paper deformation by an applicable surface. Yamashita et al. [32] introduced a shape reconstruction method using a two-camera stereo vision system. Except for Cao's work [8] and a few others [34,35], most of the current approaches require special setup (equipment, illumination) to assist in 3D shape discovery. Moreover, they can only handle smooth distortion of the image surfaces.

The most related work was undertaken by the EDUCE project [17], which attempted to read a scrolled document from a 3D scan. However, very few results on document unrolling have been reported [21,26]. The results were only shown on small contrived samples and would not scale up to real parchment with many layers which are frequently compacted together. The segmentation

* Corresponding author. Tel.: +44(0)29 2087 0389.
E-mail addresses: O.Samko@cs.cf.ac.uk (O. Samko), Yukun.Lai@cs.cf.ac.uk (Y.-K. Lai), Dave.Marshall@cs.cf.ac.uk (D. Marshall), Paul.Rosin@cs.cf.ac.uk (P.L. Rosin).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.06.015

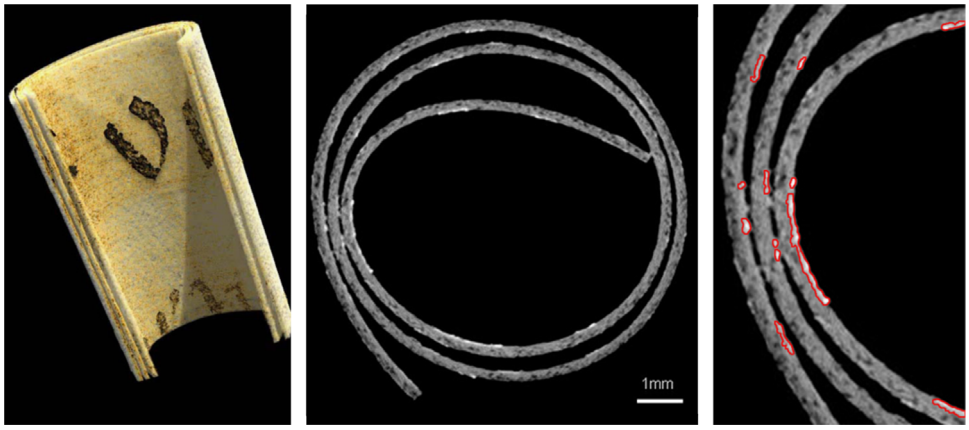

Fig. 1. A small cut sample from a historical parchment scanned with the high definition XMT scanner. Left: volume rendered cutaway view with pseudo-colouring. Middle: tomographic slice with ink on the surface of the parchment (bright pixels). Right: close up of the slice with possible ink locations highlighted (red regions). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

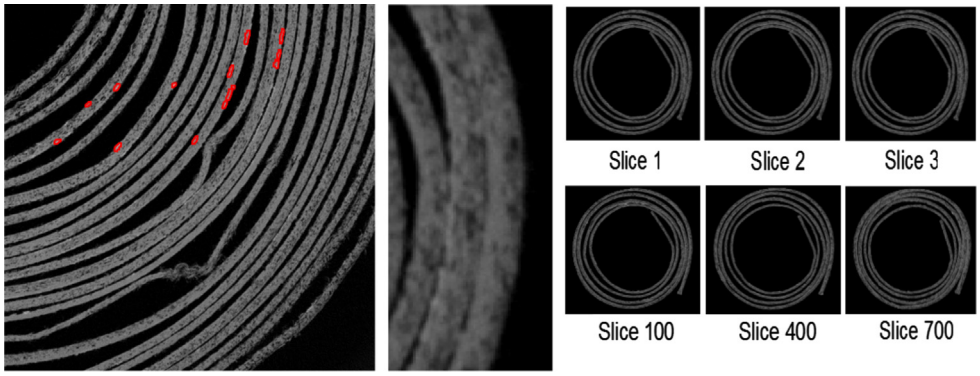

Fig. 2. An example of parchment data. On the left is a section which contains ink on both (inner and outer) sides of the parchment; ink appearances are partially indicated by the red regions. A close up is shown in the middle demonstrating the weak boundaries between layers. On the right are shown several slices of the same scroll. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

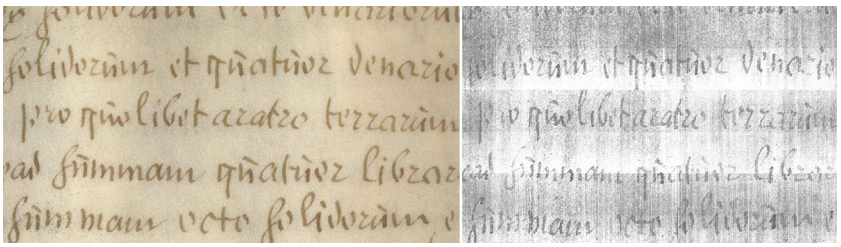

Fig. 3. Left: example of an unrolled photographed parchment. Right: reconstruction of its scanned scrolled version by the proposed framework.

stage of that work was performed semi-manually [21]. Apart from that, no other results on virtual parchment unrolling have been reported. X-ray scanning technology that is typically deployed for medical data analysis [7,14,33] does not meet a key requirement of our application: precise recovery of the ink from a parchment's weak boundaries.

The parchment shape – a tightly scrolled 3D object – makes its processing more challenging than the traditional document information recovery models. The separation of parchment layers is a major problem for parchment analysis. Parchment is essentially animal skin and has an irregular sponge-like structure; also its thickness may vary across a document surface. As a result of degradation over time, parchment may convert to its entropic form, gelatin, making the boundary between its layers difficult to observe even with the human eye. Image noise, low contrast and scanning artifacts may lead to even more indistinct parchment structure boundaries. As can be seen in Fig. 2, it is difficult to handle the parchment segmentation task satisfactorily. A general algorithm can destroy damaged areas because of parchment's latent texture (oversegment), and fail to split tightly connected layers with zero gradient (undersegment) at the same time.

The shape of the parchment smoothly changes from slice to slice, but can differ significantly across the whole scroll. The parchment ink thickness is only a few voxels deep (represented by the light pixels close to the parchment boundary), thus it is very important to carefully process the boundary to avoid losing important information due to incorrect segmentation. Poor contrast between ink and the parchment itself makes this task even more difficult. Also, often the ink remains inside the parchment layer but is lost from its surface due to natural decay of the parchment ink elements. Because of these parchment ink properties, traditional mapping techniques are inapplicable for parchment visualisation. Other challenging parchment characteristics are arbitrary wrapping shape, multiple page parchments, and