Registration of Manuscript Images using Rotation Invariant Features

Markus Diem\textsuperscript{1}, Martin Lettner\textsuperscript{1}, and Robert Sablatnig\textsuperscript{1}

\textsuperscript{1}Pattern Recognition and Image Processing Group, Institute of Computer Aided Automation, Vienna University of Technology, Austria
diem@prip.tuwien.ac.at

Abstract  One medieval Slavonic manuscript is recorded, investigated and analyzed by philologists in collaboration with computer scientists. The aim of the project is to develop algorithms that support the philologists by automatically deriving the description and restoration of the scripts. The parchment partially contains two scripts, where the first script was erased. In addition, the script is degraded due to poor storage conditions. In order to enhance the degraded script, the manuscript pages are imaged in seven bands between 330 and 1000 nm. A registration, aligning the resultant images, is necessary so that further image processing algorithms can combine the information gained by the different spectral bands. Therefore, the images are coarsely aligned using rotationally invariant features and an affine transformation. Afterwards, the similarity of the different images is computed by means of the normalized cross correlation. Finally, the images are accurately mapped to each other by the local weighted mean transformation. The algorithms used for the registration and preliminary results are presented in this paper.

1 Introduction

Since ancient times vellum was used to write on, but the laborious manufacturing made vellum a valuable material. This is why the scripts were erased and the parchment used again. The so-called palimpsests\textsuperscript{1} often comprise vestiges of the original text. During the 19th century scientists used chemical means to read palimpsests that were sometimes very destructive, using tincture of gall or later, ammonium hydroxysulfate. Modern methods of reading palimpsests using multi-spectral image acquisition are not damaging.

The codex, which is analyzed in cooperation with philologists of the University of Vienna, is called Cod. Sin. slav 5N and was found in 1975 at the St. Catherine’s Monastery (Mount Sinai, Egypt). The codex is written in Glagolitsia which is the oldest known Slavic alphabet and consists of 162 pages. Each page was captured eleven times in different spectral bands which results in a total number of 1782 images that have to be enhanced.

In this paper the results of a multi-spectral image acquisition system and a method for a fully automatic image registration are presented. The aim of multi-spectral imaging is to maximize the contrast between the erased and the second script as well as enhancing and making the degraded script visible respectively. Since manual operations such as repositioning the pages or camera changes are performed between the acquisitions, a registration that aligns one spectral image to the other is necessary. If the images are registered, they can be combined, using for instance a principal component analysis.

In a previous approach, manuscript images were registered using solely the cross correlation. In order to get reliable results, the template images, which are details of the so-called reference image, needed to comprise at least one character ($\approx 130 \times 130 \text{px}$). An image to which all other images, called sensed images, are aligned to is referred to as reference image. Since the cross correlation needed to be computed over approximately a quarter of the sensed image, two similar characters could be mistaken. But the main weakness of this approach is the dependency on rotations between the images. Since the manuscript pages are repositioned between both cameras, the algorithm must be able to handle rotations up to $180^\circ$ between the images.

Thus, a modified scale-invariant feature transform, which is rotationally invariant, aligns the images coarsely. Wrong point matches are eliminated by computing the distance to the second nearest neighbor and the RANSAC method. Afterwards the images are aligned by estimating the affine transformation matrix using the Least Squares Solution. Having aligned the images coarsely, the cross correlation is computed between a $16 \times 16 \text{px}$ template image and a $32 \times 32 \text{px}$ search window of the sensed image. A local weighted mean mapping function, which is a local polynomial transformation, is estimated by means of the corresponding control points. Thus, non-rigid local distortions caused by the changing curvature of the pages are corrected.

The paper is organized as follows. The following section characterizes related work and gives an overview in the range of image acquisition. Section 3 and Section 4 describe the image acquisition and the image registration in more detail. Numerical and visual results of the stated methods are given in Section 5. Finally, the last section gives a conclusion and an outlook.

\textsuperscript{1}Greek: \textit{palimpsestos} - scraped again
2 Related Work

There have been efforts in image analysis of historical documents [1, 2, 15]. In general, differences between image analysis of ancient versus modern documents result particularly from the aging process of the documents. Some related studies in image analysis of historical documents are covered in this section.

Multi- and hyper-spectral imaging has been used in a wide range of scientific and industrial fields including space exploration like remote sensing for environmental mapping, geological search, medical diagnosis or food quality evaluation. Recently, the technique is getting applied in order to investigate old manuscripts [1]. Two prominent representatives are the Archimedes Palimpsest [2] and Tischendorf’s Codex Sinaiticus [3]. Easton et al. were the first to capture and enhance the erased writing of the famous Archimedes palimpsest by multi-spectral methods [2]. The system they propose is modeled on the VASARI illumination system developed at the National Gallery of London [13]. In that project it turned out that the adoption of spectral imaging produces higher and better readability of the texts than conventional thresholding methods. Balas et al. developed a computer controllable hyper-spectral imaging apparatus, capable of acquiring spectral images of 5nm bandwidth and with 3nm tuning step in the spectral range between 380-1000 nm [1]. This device was selected as the instrument of choice for the Codex Sinaiticus Digitization Project conducted by the British Library in London [3].

Spectral images of palimpsests and other ‘latent’ texts have also been enhanced by the Italian company Fotoscientifica Re.co.r.d.® [6] which provided for instance the pictures for the EC project Rinascimento virtuale, devoted to the decipherment of Greek palimpsest manuscripts [6]. The EC project IsyReaDeT developed a system for a virtual restoration and archiving of damaged manuscripts, using a multispectral imaging camera, advanced image enhancement and document management software [17].

Two cameras, in contrast to the mentioned imaging systems, are used in this approach. A grayscale camera with an automatic filter wheel takes seven images in different spectral bands. Additionally, color images and UV fluorescence images are taken with a second camera. By aligning the images from both cameras to each other up to eleven channels per pixel are available for further processing steps.

3 Image Acquisition

Since photographic techniques in the visible range have proven to be insufficient with degraded manuscript pages, spectral imaging is applied [1, 15]. Images in different wavelengths provide information that is invisible to the human eye [15]. Generally, there are narrow spectral bands at which the maximum difference in the reflectance characteristics of each ink exists. The aim of multi-spectral imaging is to provide spectral image cubes, where the third dimension contains spectral information for each pixel. The degraded script is enhanced by combining the spectral information.

For the acquisition of the manuscripts a Hamamatsu C9300-124 camera is used. It records images with a resolution of 4000 × 2672px and a spectral response between 330 and 1000 nm. A lighting system provides the required IR, VIS and UV illumination. In order to speed-up the acquisition process software was developed which controls the Hamamatsu camera and the automatic filter wheel that is fixed on its object lens. Thus, the user can specify which optical filters to use and camera parameters such as exposure time. Having specified all parameters, the software takes the spectral images and stores them on the harddisk. Low-pass, band-pass and high-pass filters are used to select specific spectral ranges. The near UV (320 nm - 440 nm) excites, in conjunction with specific inorganic and organic substances, visible fluorescence light [12]. Up to now the historical manuscripts are recorded with UV fluorescence and UV reflectography. In principle, grabbing the visible fluorescence of objects is possible with every camera. UV reflectography is used to visualize retouching, damages and changes through e.g. luminescence. Therefore the visible range of light has to be excluded in order to concentrate on the long wave UV light. This is achieved by applying low-pass filters and using exclusively UV light sources.

Figure 1: Illustration of the acquisition system, with the Hamamatsu and the Nikon camera.

Seven filters allow the recording of the document pages in seven different bands ranging from 330-1000 nm. Additionally, a RGB color image and a UV fluorescence image of each manuscript page are taken using a Nikon D2Xs camera. Due to the automatic image acquisition system the registration of the images is solely needed for the correction of the differing distortions caused by the filter changes. Therefore, a simple correlation based approach and a consecutive local transformation could be applied. Since the grayscale images, taken in varying bands with the Hamamatsu camera...
system, shall be aligned to the color images taken with the Nikon camera, a more extensive registration method needs to be implemented. The image acquisition system is illustrated in Figure 1. The manuscript pages are first captured with the Hamamatsu camera and then moved in order to image them with the Nikon camera.

4 Image Registration

Following the acquisition of the manuscripts, the images have to be registered. Image registration is a fundamental task in image processing used to match two or more pictures taken under different conditions.

As stated above image registration is the process of estimating the ideal transformation between two different images of the same scene taken at different times and/or different viewpoints. It geometrically aligns images so that they can be overlaid. There is a wide variety of different methods (especially in remote sensing and medical imaging applications) like the use of corresponding structures or mapping functions [16] which can be adapted for this application. An overview of image registration methods is given by Zitová and Flusser [18].

4.1 Previous Work

An automatic image registration has been implemented for a previous project. Thereby the control points are localized using an Otsu thresholding approach [14]. In order to reduce low frequency image effects caused by the aging of the manuscripts the images are convolved with a homomorphic filter. Having localized the control points in one image and enhanced both images the correspondence between the images is computed using a normalized cross correlation. The cross correlation needs to be computed between a template image and a search image gradient magnitude $m(x, y)$ and the orientation $\theta(x, y)$ are computed for each pixel of the smoothed image $L(x, y)$.

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + \ldots + (L(x, y+1) - L(x, y-1))^2}$$ (1)

$$\theta(x, y) = \tan^{-1} \left( \frac{L(x, y+1) - L(x, y-1)}{L(x+1, y) - L(x-1, y)} \right)$$ (2)

An orientation histogram with 36 bins corresponding to 360° is created. Each sample added to the histogram is weighted by its gradient magnitude and a Gaussian weight. Afterwards, the histogram is smoothed with a Gaussian kernel. The maximum of the histogram indicates the dominant direction of local gradients. Figure 3 shows a test image where the black arrows indicate the dominant orientation
for each control point. In contrast to Lowe’s example images all arrows have the same lengths since they all have the same scale.

In order to compute a local descriptor that characterizes each control point the image gradients $m(x,y)$ and the orientations $\theta(x,y)$ in a $16 \times 16px$ window around each control point are considered. The coordinates of the descriptor and the gradient orientations are rotated relative to the control point orientation so that the features are rotationally invariant. Each gradient is weighted by a Gaussian window of $\sigma = 8$ so that the descriptor does not change significantly with small changes in the position of the window. The control point descriptor consists of eight $4 \times 4$ planes where each plane represents the spatial distribution of the gradients for different directions. The location of a gradient in the local descriptor depends on the rotated coordinates and the orientation. Each gradient is interpolated to its eight neighbors of the control point descriptor.

After the features are computed for both images, they are matched using the nearest-neighbor algorithm. The Euclidean distance between the descriptor of each control point of the reference and the sensed image is computed. The correspondence of two control points is indicated by the minimal Euclidean distance. Since a control point may exist solely in one of the two images, corresponding control points are rejected if their distance to the nearest-neighbor is less than $0.8$ times the distance to the second-nearest neighbor. Control points which have more than one correspondence are discarded too. Having discarded the control points according to this scheme $\approx 200$ corresponding control points are left for an image with $391 \times 493px$.

Since false matches can exist after discarding the previously mentioned control points and one outlier changes the transformation estimation of the Least Squares Solution dramatically, the RANSAC method is used to discard all remaining outliers. RANSAC was introduced by Fischler and Bolles [4]. This approach computes the affine transformation using three randomly selected matching points. Having tested all remaining control point pairs, the model is reestimated from the entire set of hypothetical inliers. These steps are repeated until the distances between points and the model meet a given threshold. This method discards in our approach $\approx 8.3\%$ of the matching control points.

Afterwards, an affine transformation matrix is computed using the Least Squares Solution and all remaining corresponding control points.

### 4.3 Cross Correlation

Having aligned the two images coarsely using modified SIFT Features and a global affine mapping function, a normalized cross correlation is computed at the locations of the previously found control points. The aim of the cross correlation and the subsequent local mapping function is to correct non-rigid distortions. The features detected in the images can be matched by means of the image intensity values in their close neighborhood, the feature spatial distribution, or the feature symbolic description [18]. Cross correlation is an area-based method which does not need features of images which have to be registered. The location of the control points are detected exclusively in the reference image in order to avoid false correspondence. Since the images are coarsely aligned, the search windows in the sensed image are set to the same locations as the template images.

The cross correlation calculates the difference of two image details by means of a modified Euclidean distance. The size of the template image is $16 \times 16px$. Since the images are coarsely aligned by an affine transformation, the search window needs not to be larger than twice the template image. Having defined the image details which need to be compared, the template image is shifted over the entire detail of the sensed image. For each shift the correlation between the template and the search window is computed:

$$c(m, n) = \sum_x \sum_y f(x,y) \ t(x - m, y - n)$$

(3)

where $f(x,y)$ denotes the gray values of a detail of the sensed image and $t(x,y)$ the template image. Varying $m$ and $n$ shifts the template over the search window. The resultant function $c(m, n)$ indicates the strongest correspondence of the template image and the search window by the absolute maximum. Hence the control point of the sensed image is placed at the coordinates of the absolute maximum.

The cross correlation is variant to changes in the image amplitude caused, e.g. by changing lighting conditions. Consequently, the correlation coefficient normalizing the template as well as the search window of the sensed image is computed. The dynamic range of the normalized cross correlation $\gamma(m, n)$ moves, independently to changes in the image amplitude, between $-1$ and $1$.

According to the templates magnitude and the proportion of the template and the search window, the computation performs better in the frequency domain than in the spatial domain.

### 4.4 Local Transformation

Having determined the control points, the parameters of the mapping function are computed. Images which possess only
global distortions (e.g. rotation) may be registered with a global mapping function. Likar and Pernuš mentioned that the global rigid, affine and projective transformations are most frequently used [8]. As a consequence of non-rigid distortions such as the changing lenses or curvature of a single page, the images have to be registered using a curved transformation.

A global mapping function is practicable if a low number of parameters are needed to define the transformation (e.g. rigid or affine transformations). Transformations using polynomials of order $n$ are defined by at least $n + 1$ parameters, which results in a complex similarity functional that has many local optima. To overcome this problem a local mapping function is applied. The local weighted mean method [5] is a local sensitive interpolation method. It requires at least 6 control points which should be spread uniformly over the entire image. Polynomials are computed by means of the control points. Thus, the transformation of an arbitrary point is computed by the weighted mean of all passing polynomials. Besides, a weighting function is defined which guarantees that solely polynomials near an arbitrary point influence its transformation.

5 Results

Having discussed the implemented methods, their results are presented in this section. Both, the modified scale-invariant features and the normalized cross correlation; have been tested on real manuscript images. In addition to tests with real image data, the invariance against rotation of the coarse registration was computed. Therefore two image details of each spectral band and different pages and six RGB image details captured with the Nikon camera were taken into account (see Figure 4). Each test image was rotated 30 times ($0^\circ, 12^\circ, 24^\circ...$) which results in a total of 660 tests. Thus, the rotation between the test image pairs is known. Afterwards the affine transformation model is estimated using the modified SIFT approach. Hence, the rotational error $\varphi_{err}$ of the coarse registration is computed according to:

$$\varphi_{err} = \|\sin^{-1}(\alpha) - \sin^{-1}(-\tilde{\alpha})\|$$

where $\alpha$ is the specified rotation of the test image and $\tilde{\alpha}$ is the estimated rotation to register the rotated test image with the initial test image. The resultant maximal error is $0.87^\circ$ and the minimal error is $\approx 0^\circ$ (exactly: $1.44\times10^{-15}$). The mean error of all 660 tests is $0.06^\circ$. Additionally the mean which is statistically more robust against outliers is mentioned in order to show the trend of the mean error: $0.02^\circ$. Figure 5 shows the minimum, maximum and mean error for each sampled angle. The two peaks in all lines are near 90 and $270^\circ$. This attributes to the maximal interpolation error which increases before it is zero at $0^\circ, 90^\circ, 180^\circ$ and $270^\circ$. Near $0^\circ$ and $180^\circ$ are smaller peaks because the closest sample is $\pm 12^\circ$ whereas it is $\pm 6^\circ$ at $90^\circ$ and $270^\circ$.

Resulting images of the methods are presented in Figure 5 - 9. Figure 6 shows the corresponding control points of the SIFT features before wrong correspondences are discarded. The final set of control points which is used for the estimation of the affine transformation can be seen in Figure 7. After the computation of RANSAC, no false correspondences are left.

Figure 4: Two examples of test images used. Image (a) shows a page detail with degraded characters. Image (b) shows another page captured in a different spectral band with well-preserved characters.

Figure 6: The corresponding control points between two manuscript images before the consistency check. As can be seen, there are still wrong correspondences. White boxes represent control points without a corresponding one.

The affine transformation, estimated with the modified SIFT approach, and the local weighted mean method are compared to each other using spectral and RGB images (see Figure 9). Since the images possess local non-linear distortions, only certain parts of the registered images correspond if an affine mapping function is applied. Hence, the farther the points are away from the corresponding area, the more they differ. That is why a local sensitive mapping function is applied to compute the transformation after the images are coarsely aligned with an affine transformation. Erroneous pixels still exist in the final image. These attribute to the changing shadows caused by the changing page curvature and illumination due to the shifting of the book between the takings.
**Figure 5:** The maximal, minimal and mean error for each sampled angle. The two peaks attribute to the maximal interpolation error near 90° and 270°.

**Figure 8:** A control point at the center of the template image (a), which is 16 × 16px. The search window of the sensed image with the corresponding control point (b). The normalized cross correlation where the strongest peak indicates the location with the strongest correlation (c).

**Figure 7:** The corresponding control points after discarding points with the RANSAC algorithm. The sensed image was rotated about 45°. In contrast to Figure 6, this image contains no wrong correspondences. After the consistency check, 87.5% of the corresponding control points remain.

**Figure 9:** Difference image of two registered images using an affine transformation (a). The errors in the middle of the image result from the changing page curvature. Difference image of the same sample images using the local weighted mean method (b). The pages are registered correctly, remaining errors result from changing illuminations.
6 Conclusion and Outlook

This paper introduces a multi-spectral image acquisition system for ancient manuscripts. Multi-spectral imaging allows philologists to analyze ancient manuscripts contactless. In addition, supplementary information is gained, visualizing characters of the degraded script that cannot be seen by the human eye.

Furthermore a fully automatic registration, aligning two different images with each other, was depicted. The described approach was compared to a previous registration method. Besides discussing the proposed registration approach, the methods were tested on real images. Additionally, numerical results of the coarse registrations accuracy were given in Section 5.

The registration method proposed is planned to be compared to some others (e.g. [8]) and evaluated in more detail by applying it to synthetic images. The aim of the current project is a combination of the acquired images by means of a principal component analysis or comparable methods, in order to enhance the degraded script.

Acknowledgement

This work was supported by the Austrian Science Foundation (FWF) under grant P19608-G12.

References

[3] And the word was made flash. The Economist, March 23rd 2005.