Rekonstruktion zerrissener Manuskripte/Dokumente: Vorbereitende Berechnung von Eigenschaften einzelner Dokumentschnipsel

Reconstruction of Torn Manuscripts/Notes: Preliminary Determination of Snippet Features

Florian Kleber, Markus Diem and Robert Sablatnig
Insitut für Rechnergestützte Automation
Technische Universität Wien
Favoritenstr. 9/1832, 1040 Wien
Tel.: +43-1-58801-18361, Fax: +43-1-58801-18392
E-mail: {kleber,diem, sab}@prip.tuwien.ac.at, Internet: http://www.prip.tuwien.ac.at

Zusammenfassung:
In Wissenschaften wie Forensik oder Archäographie spielt die Wiederherstellung zerstörter Dokumente oder Manuskripte eine bedeutende Rolle. Auch wenn die Motivation aus Bereichen wie der Bekämpfung von Kriminalismus durch Forensik oder wissenschaftliche Erforschung historischer Manuskripte in der Philologie stammt, will man allgemein zerstörte Information durch ein Zusammenfügen von Fragmenten wiederherstellen. Die einzelnen Fragmente können entweder durch einen Aktenvernichter („Shredder"), manuelles („händisches") Zerreissen oder bei altertümlichen Dokumenten durch Umwelteinflüsse (schlechte Lagerbedingungen) entstanden sein. Für die Rekonstruktion einer großen Anzahl (1000 und mehr) von Fragmenten müssen zusätzliche Eigenschaften ausser der Form berücksichtig werden. Dieser Beitrag zeigt die vorbereitende Bestimmung von zusätzlichen Eigenschaften von Fragmenten: die Bestimmung der Orientierung und der Schriftfarben bzw. Papierfarben. Erste Ergebnisse haben gezeigt, dass diese Eigenschaften verlässlich bestimmt werden können.

Abstract:
Sciences like Forensics, Archaeography and also related research areas need to reconstruct disrupted documents or manuscripts. Although the intention comes from different motivations, e.g. fight against business crime in forensics or scientific research on ancient manuscripts, the main goal is to make the original information spread over several fragments visible again. Therefore the fragments can originate from paper shredders, hand torn pages, or environmental effects (bad storage conditions). The reconstruction of document fragments is an interesting research question, especially if the amount of fragments is in the range of 1000 and higher. This paper will show a preliminary step for the reconstruction of documents, namely the automatic determination of the snippets' orientation. Additional calculated features are the writing and the paper color. These features will support the matching algorithm. Preliminary results show that these pre-processing steps can be performed reliably and can be used for snippet classification.

Introduction

The combinatorial problem of recomposing paper fragments into a single structure can be compared with the traditional 2D pictorial cardboard puzzles, also known as jigsaw puzzles. Freeman and Garder (Freeman and Garder 1964) have been one of the first who dealt with the automatic assembling of jigsaw puzzles. The automated solving can be divided into shape based matching techniques (apictorial) or techniques that analyze the visual content of the fragments (pictorial) (Freeman and Garder 1964). In apictorial reconstruction problems only the shape of the fragments can be considered as information to assemble a single fitting structure. Compared to

pictorial puzzles where the shape as well as the texture information of fragments are accounted to find the correct solution.

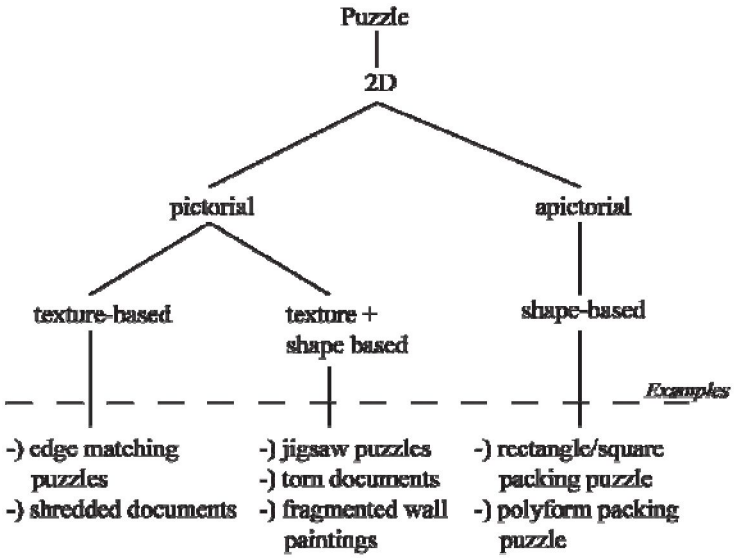


Figure 1: Classification of 2D puzzles

Figure 1 illustrates the classification of different puzzle types. Edge matching puzzles exist of pieces having the same shape, whereas only the texture information of the pieces is used to assemble the correct solution. An example of a shape matching puzzle in the document reconstruction area are pages that have been destroyed by a paper shredder (stripes of the same size).

Although puzzles are well known popular games, the automated solving is of great interest in different scientific areas like archaeology or forensics (Ukovich and Ramponi 2005, da Gama Leitao and Stolfi 2002). In this area artefacts of broken pottery e.g. (Kampel and Sablatnig 2004) have to be reconstructed as well as shredded or torn documents in the e.g. forensics (Ukovich and Ramponi 2005). Desasters like the collapse of the historical archive of cologne (Curry 2009) where a total of more than 18 shelve kilometers have been overwhelmed by rubbish need an automated solving of this task to save objects of historic and cultural value. Another example is the reconstruction of the manually torn “Stasi-files” of Germany for historic investigations (Nickolay and Schneider 2007).

Arising problems by using only shape matching techniques are described in the following section. To solve reconstruction problems with a number of fragments of 1000 and higher pictorial reconstruction methods in combination with shape matching algorithms have to be applied (Nickolay and Schneider 2007). In this paper a pre-calculation of document snippets is described that can be used in addition to shape features of fragments. The calculated characteristics are the rotation of a snippet as well as the writing and the paper color. In the Outlook Section additional features currently under development are described.

This paper is organized as follows: Section 2 shows the problems of torn documents compared to jigsaw puzzles. In Section 3 the features to cluster fragments are discussed which is followed by the Results and a Conclusion.

Torn Documents vs. Jigsaw Puzzles

The definition of the given problem by Freeman and Garder (Freeman and Garder 1964) is stated as an “*arrangement of a set of given pieces into a single, well-fitting structure with no gaps left between adjacent pieces*”. On the basis of this definition we can define the main problems by dealing with fragments from manuscripts or different documents compared to canonical jigsaw puzzles.

In realistic scenarios fragments from different pages or even books/manuscripts are possible. As a result *multiple* single fitting structures (pages) have to be assembled. A brute-force approach by comparing each snippet with every other snippet will not be feasible due to the computational effort. By analyzing the printed or written information of the snippet it is possible to cluster the fragments to pieces of possible pages. This can be done by analyzing the paper type (blank, checked, lined), the type of the writing (handwritten vs. printed) and layout analysis such as the line spacing or the font size (or image vs. text). Figure 2 shows the possible features that can be used to cluster fragments from different pages.

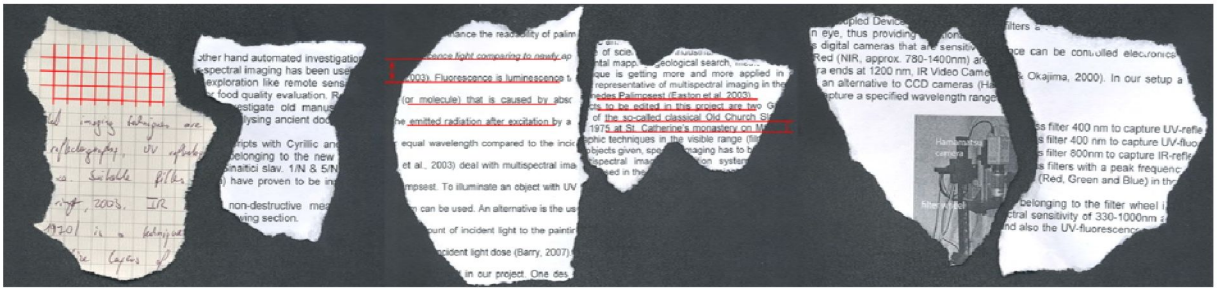


Figure 2: Possible snippet features for page-clustering: paper type, line spacing, text vs. image

Figure 3 shows correct aligned snippets of a page. It can be seen that paper tears in different layers, which causes overlapping parts. In this case the shape of the outer boundaries will match only partially and overlapping regions have to be treated. In addition gaps occur if borders are demolished or even small fragments are lost.

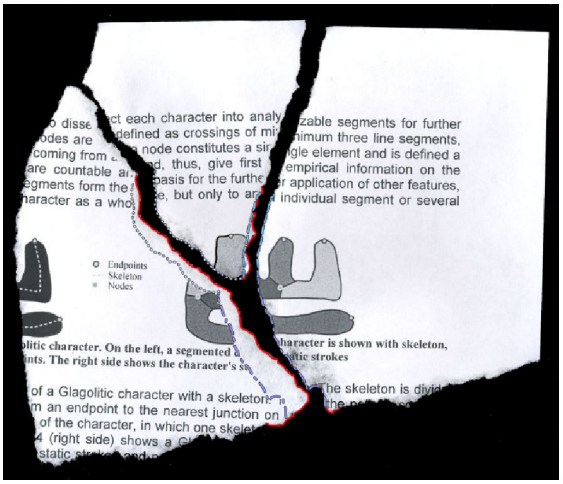


Figure 3: Torn paper with marked matching boundaries/edges

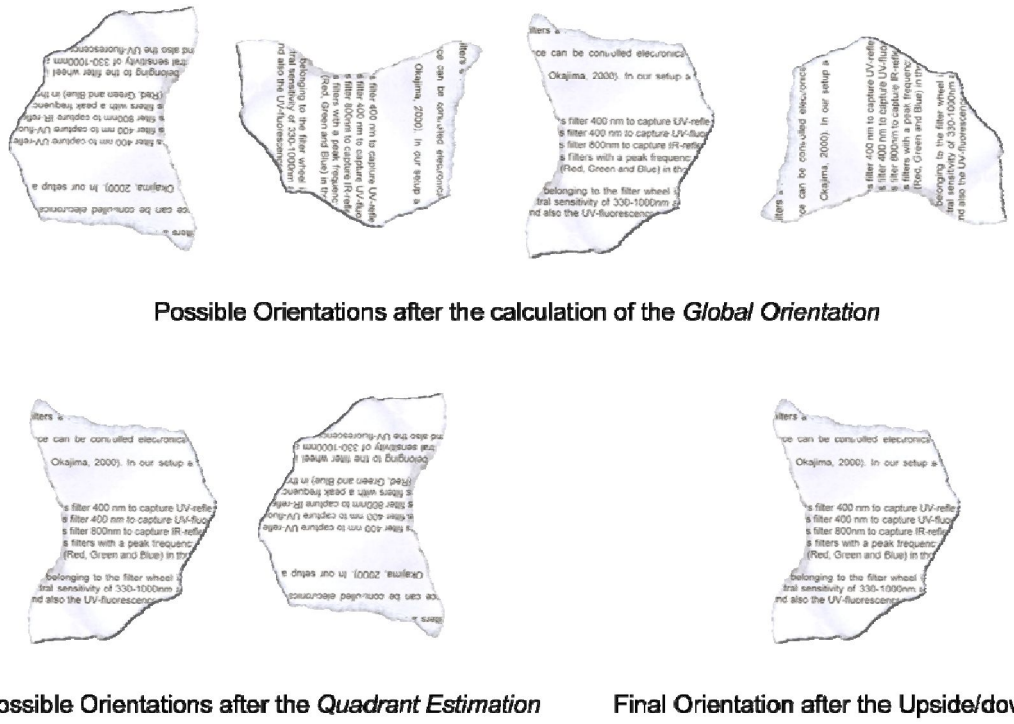
Additional differences of document fragments compared to canonical jigsaw puzzles is the irregular shape of the fragments and the content (mainly text compared to images in jigsaw puzzles). Common human heuristics like assembling the outer boundary and afterwards the interior (Burdea and Wolfson 1989) cannot be applied if border pieces are missing or are unidentifiable due to corrupt borders.
In this paper additional features (rotation, paper color, text color) are calculated to overcome the above listed problems.

Methodology

To cluster the data and to support the matching algorithm the orientation and the color of the inks/paper is calculated. The orientation assignment is based on the gradient orientation of each pixel. To determine the color a foreground/background segmentation is performed on the snippet.

Skew Estimation of a Snippet

The alignment is calculated in 3 steps, called *Global Orientation*, *Quadrant Estimation* and *Up/Down Orientation*. The possible alignment of the fragment after these 3 steps is illustrated in Figure 4.



Possible Orientations after the calculation of the *Global Orientation*

Possible Orientations after the *Quadrant Estimation*

Final Orientation after the Upside/down decision

Figure 4: Orientations of a fragment after the main processing steps

After the Global Orientation step the fragment is aligned either horizontal or vertical in relation to the writing direction. Figure 5 shows two characters and the computed gradients. It can be seen that for printed as well as handwritten text, the main orientation of the gradients is either in writing direction or orthogonal to the writing direction. By accumulating the gradients into an orientation histogram the Global Orientation is determined by a peak in the histogram.

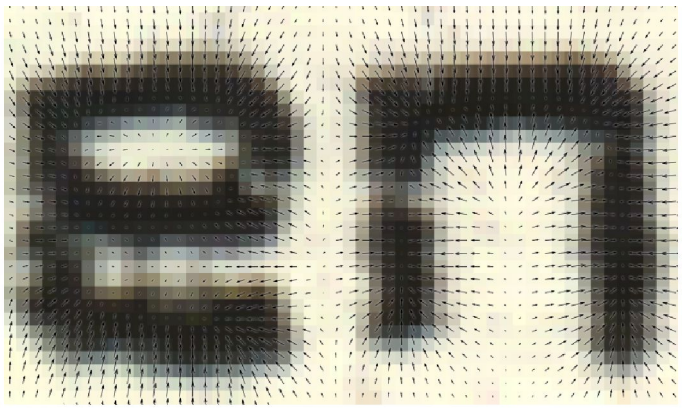


Figure 5: Gradient vectors of a manuscript image

For a detailed description for the calculation of the orientation histogram and the local statistics applied see (Diem, Kleber and Sablatnig 2009).
To determine the correct quadrant (see Figure 4), the snippet is binarized and a local projection profile is calculated. For each blob the minimum area rectangle is calculated. According to the

length, the aspect ratio and the orientation each blob is assigned a weight. Figure 6 shows the calculated minimum area rectangles for a snippet.

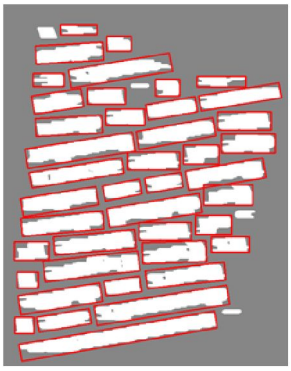


Figure 6: Minimum area rectangles

The *Quadrant Estimation* is based on the orientation of the minimum area rectangles. The upside/down orientation is mainly based on the work of Caprari (Caprari 2000). Therefore the decision is based on the frequency of ascenders and descenders of roman and arabic letters for German and English. For a detailed description see (Kleber, Diem and Sablatnig 2009).

Color Analysis

An additional feature that is used to cluster the given snippets is the color of the paper as well as the main color of the printed or handwritten text. It is obvious that different colors of inks/paper belong to different documents. Color segmentation for text extraction is a common field in document analysis (see (Mancas-Thillou and Gosselin 2005) (Hase, et al. 2004)).
For the segmentation of the foreground, it is assumed that two Gaussians represents the background and the foreground. To determine a threshold a Gaussian Mixture Model is applied. After the segmentation, the mean color for each blob is calculated. To avoid the influence of degraded characters (see Figure 7), the pixel values are weighted with $1-||m(x,y)||$, where m is the gradient magnitude of the pixel at the coordinates x,y.

Figure 7: Degraded characters and their mean color value (without weighting with the gradient magnitude)

The mean color for each blob is accumulated in a 3D RGB color histogram. The local maxima determine the existing colors of a snippet. This is done to reduce the amount of colors. For a detailed description see (Diem, Kleber and Sablatnig 2009).

Results

The proposed skew estimation was tested on a dataset of 678 images. The results are summarized in Table 1. The difference of 1,58° between the mean error and the median error can be traced back to the fact that 32 outliers exist where the rotational analysis completely fails. These outliers push the mean error while snippets that are consistent to the requirements have an error below 1°.

Number of Images	678	
Mean Error	1,95°	σ±6,13°
Median Error	0,37°	$q_{0,75}$:0,82°
Wrong Quadrants	6,64%	error < 5°: 1,89%
Wrong Up/Down	6,71%	error < 5°: 1,34%

Table 1: Result of the skew estimation

The segmentation and determination of the background color and the writing color was tested on the same set of snippets used for rotational analysis (678 images). The result of the background as well as the foreground segmentation is shown in Figure 8.

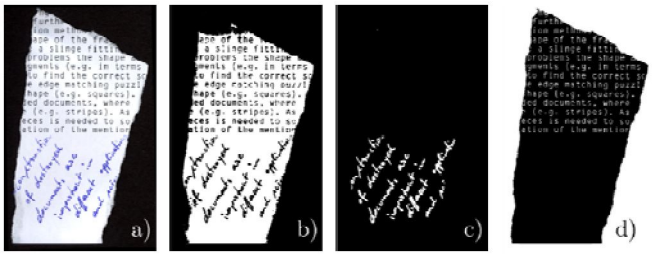


Figure 8: Segmentation result of a snippet: (a) original snippet (b) segmented snippet (c) segmented colors (d) segmented black/gray

Up to now, the evaluation of all images was based on visual criteria. However, for an improved evaluation, groundtruth data will be provided by manually annotating colored text.

Conclusion

In this paper a prerequisite, namely the calculation of characteristics of snippets, for a combined shape and pictorial approach that solves the tearing paper problem is presented. As future work additional methods to determine characteristics like the type of the writing (handwritten vs. printed), the line spacing and the paper type (blank, checked, ruled) will be developed. In addition the document layout will be analysed. Figure 9 shows preliminary results of the analysis of the writing and the layout.

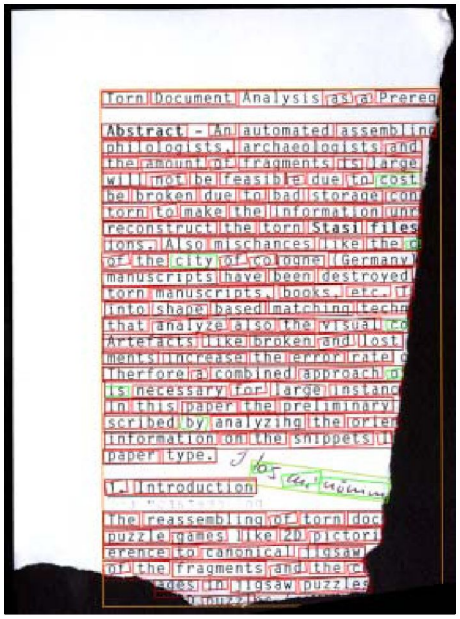


Figure 9: Snippet layout and text classification

Acknowledgement

This work was supported by the Austrian Science Fund under grant P19608-G12 and by the Fraunhofer-Institute for Production Systems and Design Technology (IPK), Berlin.

References

Burdea, B G, and H J Wolfson. "Solving jigsaw puzzles by a robot." *Robotics and Automation, IEEE Transactions on* 5 (Dec 1989): 752-764.

Caprari, , R. „Algorithm for text page up/down orientation determination.“ *Pattern Recogn. Lett.* (Elsevier Science Inc.) 21 (2000): 311--317.

Curry, Andrew. *Archive Collapse Disaster for Historians, Spiegel Online International.* 2009. http://www.spiegel.de/international/germany/0,1518,611311,00.html (accessed March 2009).

da Gama Leitao, H.C., and J Stolfi. "A Multiscale Method for the Reassembly of Two-Dimensional Fragmented Objects." *IEEE Trans. Pattern Anal. Mach. Intell.* (IEEE Computer Society) 24 (2002): 1239--1251.

Diem, Markus, Florian Kleber, and Robert Sablatnig. "Analysis of Document Snippets as a Basis for Reconstruction." *VAST 2009.* St. Julians, Malta: Eurographics, 2009. 101-108.

Freeman, H, and L Garder. "Apictorial Jigsaw Puzzles: The Computer Solution of a Problem in Pattern Recognition." *Electronic Computers, IEEE Trans. on* EC-13 (April 1964): 118-127.

Hase, H, M Yoneda, S Tokai, J Kato, and Y Suen. "Color segmentation for text extraction." *Int. J. Doc. Anal. Recognit.* (Springer-Verlag) 6 (2004): 271--284.

Kampel, M, and R Sablatnig. "On 3D Mosaicing of Rotationally Symmetric Ceramic Fragments." *17th International Conference on Pattern Recognition.* Cambridge, UK: IEEE Computer Society, 2004. 265--268.

Kleber, Florian, Markus Diem, and Robert Sablatnig. "Torn Document Analysis as a Prerequisite for Reconstruction." *15th Int. Conference on Virtual Systems and Multimedia, VSMM.* Vienna, Austria: IEEE, 2009. 143-148.

Mancas-Thillou, C, and B Gosselin. "Color Text Extraction from Camera-based Images the Impact of the Choice of the Clustering Distance." *8th Int. Conference on Document Analysis and Recognition (ICDAR).* Washington D.C., USA: IEEE Computer Society, 2005. 312--316.

Nickolay, B., and J. Schneider. *Virtuelle Rekonstruktion vorvernichteter Stasi-Unterlagen. Technologische Machbarkeit und Finanzierbarkeit - Folgerungen für Wissenschaft, Kriminaltechnik und Publizistik.* Edited by J. Weberling and G. Spitzer. Vol. 21. Berlin: Schriftenreihe des Berliner Landesbeauftragten für die Unterlagen des Staatssicherheitsdienstes der ehemaligen DDR (German), 2007.

Nielsen, Ture R, Peter Drewsen, and Klaus Hansen. "Solving jigsaw puzzles using image features." *Pattern Recogn. Lett.* (Elsevier Science Inc.) 29 (2008): 1924--1933.

Ukovich, A, and G Ramponi. "Features for the reconstruction of shredded notebook paper." *Image Processing (ICIP 2005). IEEE Int. Conf. on.* 2005. III-93-6.