The 10th International Symposium on Virtual Reality, Archaeology and Cultural Heritage VAST (2009)
K. Debattista, C. Perlingieri, D. Pitzalis, and S. Spina (Editors)

Analysis of Document Snippets as a Basis for Reconstruction

M. Diem, F. Kleber & R. Sablatnig[1]

[1] Institute of Computer Aided Automation, Vienna University of Technology, Austria

Abstract

In Archaeography, Philology, Forensics, and related research areas fragments of documents are very common. These fragments are the basis for the subsequent reconstruction process, where the goal is to make the original information spread over several fragments visible again. The fragments can originate from paper shredders, hand torn pages or in the case of ancient manuscripts this is due to bad storage conditions, or other destroying facts. So we can distinguish between an "on-purpose" destruction because the information contained on the pages should not be readable anymore or a "time-induced" destruction for ancient documents which is unintentional. Nevertheless the reconstruction of document fragments is an interesting research question. This paper shows a preliminary step for the page reconstruction namely the automatic orientation of snippets in order to eliminate the rotation in the later reconstruction (puzzling) process. Furthermore features like paper color and the color of the inks used are analyzed as a pre-classification step to find matching snippets. In the case of "on-purpose" destruction there is no a-priori information on which fragment belongs to which page which makes a reconstruction based on thousands of fragments from unknown sources difficult since the combinatorial effort explodes (NP-hardness). Preliminary results on orientation and color segmentation are presented and show that these pre-processing steps can be performed reliably and can be used for reconstruction and snippet classification.

Categories and Subject Descriptors (according to ACM CCS): I.4.0 [Image Processing and Computer Vision]: General—I.5.4 [Pattern Recognition]: Applications—Text Processing

1. Introduction

The reconstruction of destroyed documents is important in different applications and sciences like forensics [UR05, NS07] and archaeology. The reconstruction of ancient manuscripts that have been fragmented due to bad storage conditions as well as the reconstruction of worth keeping manuscripts or books that have been destroyed by disasters (e.g. [Cur09]) need an automated solving of this task to archive objects of historic and cultural value. Due to the collapse of the historical archive of cologne a total of more than 18 shelve kilometers have been overwhelmed by rubbish. Another example is the reconstruction of the manually torn "Stasi-files" of Germany for historic investigations [NS07]. Within this project approximately 600 million snippets [NS07] from about 45 million pages exist. It is stated that the fields of forensics, tax fraud investigation, customs investigation and the suppression of business crime are in need of automated recovery algorithms.

The reassembling of torn documents is related to the traditional puzzle games like 2D pictorial cardboard puzzles [DD07]. The main difference to canonical jigsaw puzzle games [YS03, Tyb04] is the irregular shape of the fragments and the content (mainly text in documents compared to images in jigsaw puzzles). Although manually torn documents are processed as 2D puzzle, it is possible that paper tears in different layers which causes overlapping parts of fragments.

2D puzzles furthermore can be divided into pictorial and apictorial reconstruction methods [FG64]. In apictorial reconstruction problems only the shape of the fragments can be considered as information to assemble a single fitting structure. Compared to pictorial reconstruction problems the shape as well as the information printed on the fragments (e.g. in terms of printed pictures or text) are accounted to find the correct solution. An instance of an edge matching puzzle [DD07], where all puzzle pieces have the similar shape, are shredded documents (e.g. [UR05]), where all parts of the documents exist of equal stripes. As a result, only the texture information of the pieces can be used to solve the problem.

DOI: 10.2312/VAST/VAST09/101-108

Arising problems are the described mismatch of the outer boundaries, fragments that belong to different pages or even manuscripts/books. An additional problem is that it is not known a priori if there are missing pieces or not [Tyb04]. As a result, heuristics used to solve jigsaw puzzles like assembling the outer boundary and afterwards the interior [BW89] cannot be applied to torn documents. Therefore the aid of pictorial reconstruction methods in combination with shape matching algorithms can solve even large instances of this problem [NS07]. The complexity of puzzle problems (NP-hardness, depending on the type of the puzzle) is presented in [Tyb04, DD07].

In this paper a pre-calculation of document snippets is described to enable a clustering of the provided data as described in [NS07]: according to the printed or handwritten text information the orientation is calculated. Additionally, the color of the used inks and of the paper are analyzed. Further statistical analysis e.g. handwritten vs. printed text, line spacing, paper type (blank, lined, checked) can be analyzed too. After this step the snippets can be further clustered according to their color. As a result, the search space for the shape matching is reduced.

The paper is organized as follows: Section 2 reviews the state of the art of 2D reconstruction and skew estimation methods. In Section 3 the pre-calculation of all snippets is described, while in Section 4 the experimental results are discussed. Finally a conclusion is given in Section 5.

2. Related Work

There are several studies for the automated assembly of jigsaw puzzles, e.g. [NDH08, CFF98, BW89]. A puzzle related to archaeological problems is the one published in [PPE*02]. In this paper fragmented wall paintings (1600 b.c.) are reconstructed by matching the fragments' contours. In [Kol08] the reconstruction of the Severan Marble Plan of Rome, or Forma Urbis Romae (more than 1000 fragments, destroyed in the 5th century) is treated. A boundary incision matching method was developed, which uses topographic features in addition to the boundary information [Kol08]. It is stated that a matching algorithm simply based on contour matching did not lead to a solution.

For text documents the reassembly of strip-shredded [PR08, UR05] or cross cut-shredded [PR09] documents is the main research area in the forensic domain. An approach that is dealing with the reconstruction of torn paper is presented in [Ber08] (only the shape is considered, texture information is not used).

Methods proposing algorithm for skew estimation include techniques based on projection profiles [BK97, SZHZ07], the Hough transformation [HFD95, AF00] and methods based on properties of the Fourier transform [PT97]. A summary and a classification of skew estimation algorithm is shown in [Lv04, Hul98]. A drawback of projection profile based techniques is that narrow lines do not produce a significant peak [LSZT07]. In [SK00] the document image is smoothed and then a bounding rectangle of the Connected Components (CC) with a minimal area is calculated. The orientation of the bounding rectangle determines the skew of the document. According to [KFG02] the problems of skew estimation algorithms are the following:

- "*Restriction of detectable angle range*
- *Restriction on type or size of fonts*
- *Dependence on page layout*
- *A specific document resolution is required*
- *High computational cost*
- *Limitation to specific application*
- *Large text areas are required*
- *[...] Furthermore, the proposed algorithms can estimate the dominant skew angle and cannot deal with the cases of handwritten pages where the text lines may not be parallel to each other.*"

The main benefit of the proposed algorithm is the possibility to handle fragments of stamps' size up to a mutual page size. Additionally all points except the need of a large text area and the fact of different orientated text lines are fulfilled.

Algorithms for determining the up/down orientation on a text page are mainly based on statistical analysis of ascenders and descenders, e.g. [Cap00, BKD95]. The main drawback is the dependence on the script as well as the language used within a document. In [Ara05] an algorithm analyzing the openness of characters is published, which can also be applied to different scripts with a dominant directionality. For a detailed description see [Ara05].

3. 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 which are accumulated into an orientation histogram. In addition, a color segmentation is performed and the color of the ink/paper is analyzed.

3.1. Rotational Analysis

The rotation estimation of a given snippet is inspired by Lowe's SIFT [Low04]. Thus, the gradient magnitude $m(x,y)$ and the gradient orientation $\theta(x,y)$ of a snippet are computed:

$$m(x,y) = \sqrt{\begin{array}{l}(L(x+1,y)-L(x-1,y))^2+\ldots \\ +(L(x,y+1)-L(x,y-1))^2\end{array}} \qquad (1)$$

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

where $L(x,y)$ is the observed image. Subsequently an orientation histogram is constructed with bins corresponding to the orientation $\theta(x,y)$. Hence, each pixel is accumulated into the bin corresponding to its orientation and weighted by the gradient magnitude $m(x,y)$. Peaks in the resulting histogram indicate the main orientation of a given snippet.

Since reflected gradient vectors indicate whether a border is black-white or white-black but do not make a statement about the exact local orientation, $\theta(x,y)$ is computed on the interval $[-\pi/2 \;\; \pi/2]$ (see Section 4). Therefore, reflected gradient vectors are accumulated to the same bins of an orientation histogram which leads to improved peaks.

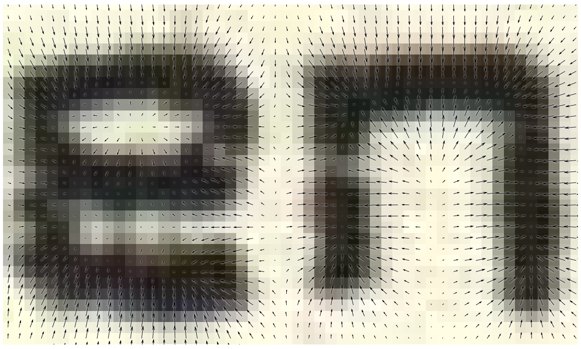

Figure 1: *Gradient vectors of a manuscript image.*

To determine the global orientation, simply the highest histogram bin could be taken into account. However, if a straight torn border is present in a given snippet, the highest peak would correspond to this border, not to text lines in the snippet. In order to solve this problem, the highest peak relative to its neighbors is taken into account rather than the global maximum of the orientation histogram. For this purpose, the histogram is smoothed with a gaussian ($\sigma = 3$) which discards small local maxima. Then the local maxima $l(x_i)$ are detected in the smoothed histogram and again allocated in the original histogram. Finally, the peak is detected by:

$$p = \max(l(x_i) - median[l(x_{i-j}) \; l(x_{i+j})]) \tag{3}$$

where $l(x_i)$ is the i-th local maximum, p is the resulting peak and j determines the interval of the local neighborhood.

So far the accuracy of the determined main orientation depends on the number of bins in the orientation histogram (e.g. 1° if 180 bins are used). In order to increase the accuracy – independent to the number of orientation bins – a 2^{nd} order polynomial is interpolated to the peak and its two neighbors. Hence, the main orientation of a snippet is determined by the maximum of the polynomial with an arbitrary accuracy.

3.2. Quadrant Estimation

The orientation histogram determines the main orientation within a quadrant $[0 \; \pi/2]$. In order to determine the quadrant, the content of a snippet needs to be considered. Since the best fitting ellipses are computed for content analysis, the quadrant assignment is based on these. Therefore, the snippet is thresholded using the Otsu [Ots79] binarization and multiplied with its mask (see Section 3.4) which discards the background. Subsequently the best fitting ellipses are computed for each blob by means of the normalized second moments. A blob generally represents a word which is guaranteed by a previous smoothing of the image.

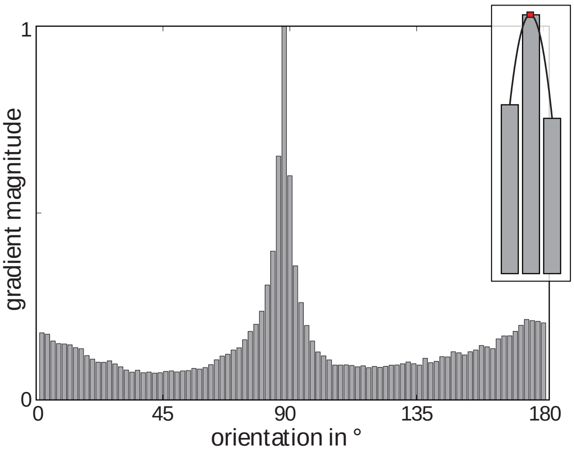

Figure 2: *Orientation histogram with spline interpolation*

In order to determine the quadrant, the ellipses are first rotated relative to the main orientation. Than they are accumulated into an x and y bin depending on their angle. A weight based on the angle, size and aspect ratio is assigned to each ellipse. Thus, ellipses having a relative orientation of 45° have a lower weight than those with 1°. If the resulting y bin is higher than the x bin, the snippet needs additionally to be rotated by 90°.

Since the entropy of snippets can be low (only a few words written), it is necessary to establish a confidence measure c. This measure is defined by:

$$c = \begin{cases} 1 - \frac{\min(x_b, y_b)}{\max(x_b, y_b)} - \frac{1}{n} & c > 0 \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

where (x_b, y_b) are the x and the y bins respectively and n is the number of ellipses. This measure allows for a soft clustering decision in the final matching algorithm.

3.3. Up/Down Orientation

The page up/down orientation determination is based on the work of Caprari [Cap00]. Therefore the decision is based on the frequency of ascenders and descenders of roman letters and arabic numerals. The frequencies of the single characters for English and German are shown in Table 1 [Beu05, Lew05]. It can be seen that the occurrence of ascenders is dominating. Capital letters, which will rise the asymmetry of ascender/descender and therefore enhance the result of the given method, are not considered in this statistics.

Stroke	Letters	Frequency English	Frequency German
Descender	j,p,q,y	4.15%	1.12%
Ascender	b,d,f,h,k,l,t	27.92%	24.19%
Neither	a,c,e,i,m,n,o,r, s,u,v,w,x,z	67.93%	74.69%

Table 1: *Character frequency in English and German [Beu05, Lew05]*

Caprari analyzes the asymmetry of the line histogram based on the ascender and descender frequency. Since the algorithm is sensitive to the correct skew, the entire page is divided into stripes. The dependency of the up/down rotation to the relative rotation error has been tested on the dataset with a varying amount of stripes (1,2,4,6,8). It turned out that the best results are gained when a snippet is divided into 6 stripes (see Section 4).

3.4. 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 [MTG05, HYT*04]).

To segment the snippets into background and printed information, color spaces (RGB, CIE $L^*a^*b^*$, HSV, XYZ) [TT03] have been tested. It turned out, that for the segmentation of the background the luminance channel of the CIE $L^*a^*b^*$ color space is the best choice (see Section 4). Since all snippets are scanned using high-end devices the recording conditions (e.g. illumination) remain stable for all digitized images. Hence, the color temperature is not significantly varying which allows for directly comparing captured colors.

To determine a threshold the 1-D histogram of the luminance channel is calculated and smoothed. It is assumed that the background (paper, parchment,...) has a higher luminance value than the foreground (writings, images) and the luminance histogram therefore can be represented by two gaussian distributions. As a result the two gaussian distributions representing the background p_b and the foreground p_f are fitted into the histogram using Expectation Maximization (EM) [CJST07, HKM08] to approximate a Gaussian Mixture Model [CBGM02]:

$$p(x \mid \theta) = \sum_{i \in \{b,f\}} \alpha_i p_i(x \mid \theta_i) \tag{5}$$

where α_i represent the mixing weights and θ represents the parameters of the gaussian distributions p_i [CBGM02]:

$$p_i(x \mid \theta_i) = \frac{1}{\sigma_i \sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i}} \tag{6}$$

The threshold th for segmenting the background is set to

$$th = \mu_{p_b} - 2\sigma_{p_b} \tag{7}$$

whereas μ represents the mean value, and σ is the standard deviation of the gaussian determined by the EM algorithm. According to the definition of a normal distribution this threshold includes approx. 95.4% of all background pixels. Figure 6 shows a snippets luminance histogram and the fitted gaussians to determine the threshold. The background color is calculated as the mean value of all RGB values in the determined background region. Since edges of the writing or snippets have a continous color gradient (e.g. a blue character fades out to the background color) the pixel values are weighted with the normalized gradient magnitude values: $1 - \|m(x,y)\|$ (see Equation 1). Note that the calculations are done only in the snippet region, since a mask is used to eliminate the region outside the snippet. Due to the scanning process the snippets have a uniform illumination and a defined background color. Subject to these limitations the threshold to determine the mask image can be calculated using a global threshold approach (e.g. Otsu [Ots79]).

The threshold for the text is determined in the Saturation channel of the HSV color space and black/gray colors are determined in the V channel. Each blob is analyzed separately to determine the color of the writing by calculating the mean values of the blob's pixels for the R, G and B channel. To reduce effects like the fading out of the ink at the borders the pixel values are weighted in the same way as the border pixels of the background.

After calculating the mean color value for each blob (black/gray and color segmentation is processed separately) the number of colors determined is approximately in the range of 100 up to 200. Figure 3 shows two degraded characters to illustrate the problem of the determination of a single color for a character and the importance of the weighting with the gradient magnitude. Additionally it explains the occurrence of slight differences in the mean color of 2 characters written with the same color. It can also be seen that using K-means for color segmentation in this application results in an oversegmentation.

To reduce the quantity of color values a 3D RGB Histogram is calculated. This is done by determining the local maxima of the color histogram which is also known as hill climbing segmentation [OAM03, AEWS08].

4. Results

In order to generate the groundtruth, a tool was developed which allows for a manual assignment of the text (print/manuscript), the paper (void/lined/checked) and the

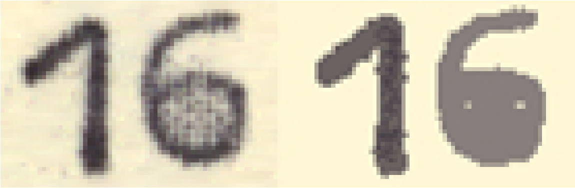

Figure 3: *Original degraded characters (a) segmented characters with determined color value*

rotation angle. The annotated test set consists of 690 images containing torn documents of all classes. The snippets' sizes range from $3.1cm^2$ to $378cm^2$ at a resolution of $300dpi$. Thus, the document analysis system must be able to handle snippets of a stamp's size with hardly any content up to half a DIN A4 page.

A cross validation of the manually tagged groundtruth was performed on 150 images so that the variation of different operators can be analyzed. The resulting median error of the rotational angle – which was annotated by dragging a line in a given image – is 0.14° ($q_{0.75}$: 0.29°). The maximal error of 2.86° can be traced back to the fact that some snippets do not have an obvious main direction (e.g. if handwritten lines are not parallel to each other).

4.1. Performance of the Rotational Analysis

Based on this annotated test set, the rotational analysis described in Section 3.1 is evaluated. Additionally, parameters needed are determined so that a good generalization performance of the methods is given. The relative orientation error is computed by:

$$e_1 = \|mod(\alpha_c,\ 90) - mod(\alpha_m,\ 90)\| \tag{8}$$

$$e = \min(e_1,\ 90 - e_1) \tag{9}$$

where α_m is the manually determined rotation angle, α_c is the computed angle and e is the final error. This error computation considers solely the difference of angles within one quadrant $[0\ \pi/2]$ since the quadrant estimation does not consider if a snippet is upside down or not.

The statistical moments of the relative angle error are given in Table 2. In this evaluation 678 images are considered even though the test set consists of 690 images as mentioned before. This arises from the fact that some snippets exist (e.g. no content, no straight border) where the main orientation cannot be assigned to.

Number of images:	678	
Mean error:	1.95°	$\sigma \pm 6.13°$
Median error:	0.37°	$q_{0.25}$ 0.16° – $q_{0.75}$ 0.82°

Table 2: *Relative rotational errors.*

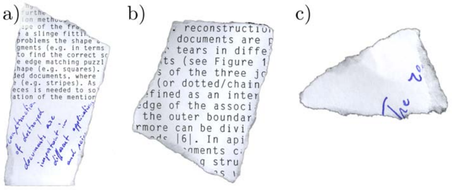

Figure 4: *Different examples of rotated snippets. The first snippet (a) has a relative orientation error of* 0.11°*, where (b) has an error being* 0.98° *and (c)* 18.58°.

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°. Figure 4 shows three differently rotated snippets. In (a) and (b) snippets are shown that are rotated correctly (Error < 1°). The snippet shown in (c) is one of the 32 outliers. Since hardly any text is visible in this example, the main orientation cannot be assigned automatically.

In addition to the tests made concerning the relative orientation, a test was performed on the quadrant estimation. There, 47 (7.63%) images could not be rotated correctly as a consequence of false binarization and errors of the relative orientation estimation. In addition to the full testset, a set consisting of 164 snippet images was created which contain at least 5 partially visible text lines. On this test set 5 (3.05%) snippets were not rotated correctly. If 5 images having a relative error above 5° are not regarded, the quadrant estimation fails in 3 cases (1.89%).

The results gained with the method for determining the up/down orientation are given in Table 3 when varying the number of stripes. On the whole testset 9.04% were not correctly rotated even though images are present which contain solely capital letters or less than a half textline.

# Stripes	Total	Relative Error < 5°
1	10.98% (18/164)	5.37% (8/149)
2	8.54% (14/164)	3.36% (5/149)
4	7.32% (12/164)	1.34% (2/149)
6	6.71% (11/164)	1.34% (2/149)
8	6.71% (11/164)	2.01% (3/149)

Table 3: *Results of the up/down orientation determination if the number of stripes is changed.*

4.2. Parameter Evaluation

In the presented method, three crucial parameters need to be determined (θ, n_b, σ). In order to improve the generalization performance of the method, they are evaluated on the given test set.

First θ is evaluated, which is the angle of the gradient vectors. In Figure 5 four boxplots are shown. The first three of which show the relative error of θ when it is set to $(2\pi, \pi, \pi/2)$. The last boxplot shows the before mentioned cross validation so as to visualize the statistical error of the groundtruth. Each rectangle of a boxplot is between the $q_{0.25}$ and $q_{0.75}$ quantile of the error distribution. The red line within the rectangle shows the median. The whiskers correspond to the most extreme values below 1.5 times of the interquartile range. Red crosses indicate statistical outliers.

Computing θ between $[-\pi/2\ \pi/2]$ (second boxplot) results in the best performance, as a consequence of reasons mentioned in Section 3. If θ is computed between $[-\pi/4\ \pi/4]$ orthogonal gradients are accumulated into the same orientation bins of the histogram. This results in more outliers if italic manuscripts are to be considered.

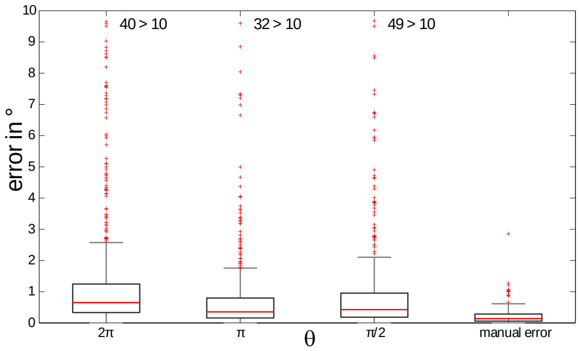

Figure 5: *Evaluation of* θ.

By evaluating the second parameter it turned out, that the performance increases as the number of bins is decreased from initially 180 to 90. This is because the polynomial interpolation is more robust as the bin values are more stable if n_b is reduced. Another aspect is, that especially the main orientation of small snippets is more likely to be mistaken if more histogram bins are used. When decreasing n_b further, the orientation histogram becomes too coarse ($\tilde{x} = 0.35°$ when $n_b = 90$, $\tilde{x} = 0.62°$ when $n_b = 45$) which leads to a low performance on large snippets (e.g. half a DIN A4).

Varying σ of the gaussian filter kernel is especially crucial because different features of a snippet are regarded if the gaussian kernel is changed. Hence, noise reduces the performance when a small σ (e.g. 0.5) is chosen. But with increasing σ the scale of the image is decreased resulting in a coarser representation of the image which is especially crucial for snippets with low content. The proposed method performs best on the given test set if σ is chosen to be 3.

4.3. Results of the Color Segmentation

The segmentation and determination of the background color and the writing's color was tested on the same set of snippets used for rotational analysis (690 images).

Figure 6 shows the luminance histogram of the snippet presented in Figure 7(a). The fitted gaussian distributions representing the background and the foreground (writing) are also shown in Figure 6. It can be seen, that the luminance channel can be used to find a threshold value to segment the background. The result of the background as well as the foreground segmentation is shown in Figure 7. It can be seen, that the method described in Section 3.4 clearly differentiates color from black/gray text components. A different example of a snippet is shown in Figure 8 presenting the segmented image regions and the final mean colors calculated. 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.

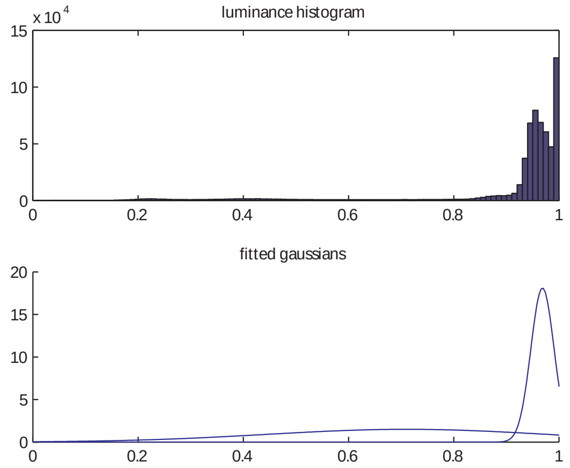

Figure 6: *Fitted gaussians to determine the threshold for the background*

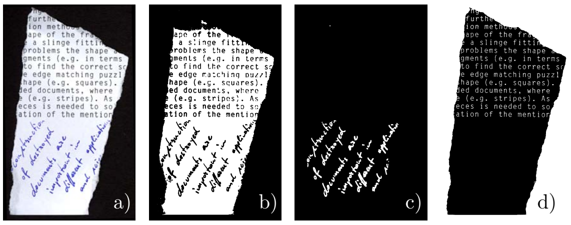

Figure 7: *Segmentation result of a snippet: (a) original image, (b) segmented background, (c) segmented colors and (d) segmented black/gray ink.*

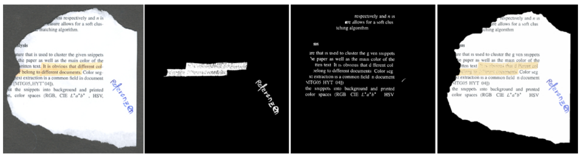

Figure 8: *Segmentation result of a snippet: (a) original image, (b) segmented colors, (c) segmented black/gray and (d)final image colors*

5. 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. The tearing paper problem deals with the re-assembling of torn documents.

To minimize the search space for matching the following calculations are performed on each snippet: a rotational analysis to determine the alignment and the color of the ink/paper is distinguished. Without this information only the shape information can be used, which cannot solve large instances of the problem.

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. Furthermore a matching heuristic that uses the information provided by the introduced methods will be developed.

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

[AEWS08] ACHANTA R., ESTRADA F., WILS P., SÜSSTRUNK S.: Salient Region Detection and Segmentation. In *Int. Conf. on Computer Vision Systems (ICVS '08)* (2008), vol. 5008, Springer Lecture Notes in Computer Science, Springer Berlin / Heidelberg, pp. 66–75.

[AF00] AMIN A., FISCHER S.: A document skew detection method using the hough transform. *Pattern Analysis and Applications 3*, 3 2000 (2000), 243–253.

[Ara05] ARADHYE H. B.: A generic method for determining the up/down orientation of text in roman and non-roman scripts. In *Proc. of the 8th Int. Conf. on Document Analysis and Recognition (ICDAR'05)* (Washington, DC, USA, 2005), IEEE Computer Society, pp. 187–191.

[Ber08] BERGER F.: *Ein hybrides Verfahren zur automatischen Rekonstruktion von handzerrissenen Dokumentenseiten mittels geometrischer Informationen*. Master's thesis, Vienna University of Technology, Inst. of Computer Graphics and Algorithms, Austria, 2008.

[Beu05] BEUTELSPACHER A.: *Kryptologie*. vieweg, 2005.

[BK97] BAGDANOV A. D., KANAI J.: Projection profile based skew estimation algorithm for jbig compressed images. In *Proc. of the 4th Int. Conf. on Document Analysis and Recognition (ICDAR '97)* (Washington, DC, USA, 1997), IEEE Computer Society, pp. 401–406.

[BKD95] BLOOMBERG D. S., KOPEC G. E., DASARI L.: Measuring document image skew and orientation. *Document Recognition II 2422*, 1 (1995), 302–316.

[BW89] BURDEA B., WOLFSON H.: Solving jigsaw puzzles by a robot. *Robotics and Automation, IEEE Transactions on 5*, 6 (Dec 1989), 752–764.

[Cap00] CAPRARI R. S.: Algorithm for text page up/down orientation determination. *Pattern Recogn. Lett. 21*, 4 (2000), 311–317.

[CBGM02] CARSON C., BELONGIE S., GREENSPAN H., MALIK J.: Blobworld: image segmentation using expectation-maximization and its application to image querying. *Pattern Analysis and Machine Intelligence, IEEE Transactions on 24*, 8 (Aug 2002), 1026–1038.

[CFF98] CHUNG M. G., FLECK M., FORSYTH D.: Jigsaw puzzle solver using shape and color. *Signal Processing Proceedings (ICSP '98). 1998 4th Int. Conf. on 2* (1998), 877–880.

[CJST07] CHENG D.-C., JIANG X., SCHMIDT-TRUCKSAESS A.: Image segmentation using histogram fitting and spatial information. *Advances in Mass Data Analysis of Signals and Images in Medicine, Biotechnology and Chemistry 4826* (2007), 47–57.

[Cur09] CURRY A.: Archive collapse disaster for historians. *Spiegel online international* (04th march 2009). http://www.spiegel.de/international/germany/0,1518,611311,00.html.

[DD07] DEMAINE E. D., DEMAINE M. L.: Jigsaw puzzles, edge matching, and polyomino packing: Connections and complexity. *Graphs and Combinatorics 23*, 1 (2007), 195–208.

[FG64] FREEMAN H., GARDER L.: Apictorial jigsaw puzzles: The computer solution of a problem in pattern recognition. *Computers, IEEE Trans. on EC-13*, 2 (April 1964), 118–127.

[HFD95] HINDS S. C., FISHER J. L., D'AMATO D. P.: A document skew detection method using run-length encoding and the hough transform. *Document image analysis 1* (1995), 209–213.

[HKM08] HENDERSON N., KING R., MIDDLETON R. H.: An application of gaussian mixtures: Colour segmenting for the four legged league using hsi colour space. In *RoboCup 2007: Robot Soccer World Cup XI* (Berlin, Heidelberg, 2008), Springer-Verlag, pp. 254–261.

[Hul98] HULL J. J.: Document image skew detection: Survey and annotated bibiliography. In *Document Analysis System II, World Scientific* (1998), Hull J. J., Taylor S. L., (Eds.), pp. 40–64.

[HYT*04] HASE H., YONEDA M., TOKAI S., KATO J., SUEN Y.: Color segmentation for text extraction. *Int. J. Doc. Anal. Recognit. 6*, 4 (2004), 271–284.

[KFG02] KAVALLIERATOU E., FAKOTAKIS N., G. K.: Skew angle estimation for printed and handwritten documents using the wigner-ville distribution. *Image and Vision Computing 20* (2002), 813–824.

[Kol08] KOLLER D. R.: Virtual Archaeology and Computer-Aided Reconstruction of the Severan Marple Plan. In *Beyond Illustration: 2D and 3D Digital Technologies as Tools for Discovery in Archaeology, British Archaeological Reports International Series* (2008), Frischer B., Dakouri-Hild A., (Eds.), Archaeopress, pp. 125–134.

[Lew05] LEWAND R. E.: *Cryptological Mathematics*. The Mathematical Association of America, 2005.

[Low04] LOWE D. G.: Distinctive Image Features from Scale-Invariant Keypoints. *Int. Journal of Computer Vision 60*, 2 (2004), 91–110.

[LSZT07] LIKFORMAN-SULEM L., ZAHOUR A., TACONET B.: Text line segmentation of historical documents: a survey. *Int. Journal Doc. Anal. Rec. 9*, 2 (2007), 123–138.

[Lv04] LINS R. D., ÁVILA B. T.: A new algorithm for skew detection in images of documents. In *ICIAR (2)* (2004), Campilho A. C., Kamel M. S., (Eds.), vol. 3212 of *Lecture Notes in Computer Science*, Springer, pp. 234–240.

[MTG05] MANCAS-THILLOU C., GOSSELIN B.: Color text extraction from camera-based images the impact of the choice of the clustering distance. In *Proc. of the 8th Int. Conf. on Document Analysis and Recognition (ICDAR'05)* (Washington, DC, USA, 2005), IEEE Computer Society, pp. 312–316.

[NDH08] NIELSEN T. R., DREWSEN P., HANSEN K.: Solving jigsaw puzzles using image features. *Pattern Recogn. Lett. 29*, 14 (2008), 1924–1933.

[NS07] NICKOLAY B., SCHNEIDER J.: Automatische virtuelle Rekonstruktion "vorvernichteter" Stasi-Unterlagen - Machbarkeit, Systemlösung, Potenziale. In *Schriftenreihe des Berliner Landesbeauftragten für die Unterlagen des Staatssicherheitsdienstes der ehemaligen DDR (German)* (Berlin, 2007), Weberling J., Spitzer G., (Eds.), vol. 21, pp. 11–28.

[OAM03] OHASHI T., AGHBARI Z., MAKINOUCHI A.: Hill-climbing algorithm for efficient color-based image segmentation. In *IASTED Int. Conf. on Signal Processing, Pattern Recognition and Applications (SPRA'03)* (2003).

[Ots79] OTSU N.: A threshold selection method from grey level histograms. *IEEE Trans. on Systems, Man, and Cybernetics 9* (1979), 62–66.

[PPE*02] PAPAODYSSEUS C., PANAGOPOULOS T., EXARHOS M., TRIANTAFILLOU C., FRAGOULIS D., DOUMAS C.: Contour-shape based reconstruction of fragmented, 1600 bc wall paintings. *Signal Processing, IEEE Transactions on 50*, 6 (Jun 2002), 1277–1288.

[PR08] PRANDTSTETTER M., RAIDL G. R.: Combining forces to reconstruct strip shredded text documents. In *HM '08: Proc. of the 5th Int. Workshop on Hybrid Metaheuristics* (Berlin, Heidelberg, 2008), Springer-Verlag, pp. 175–189.

[PR09] PRANDTSTETTER M., RAIDL G. R.: Meta-heuristics for reconstructing cross cut shredded text documents. In *ACM: to appear in Proceedings of the Genetic and Evolutionary Computation Conf. (GECCO'09)* (2009).

[PT97] PEAKE G., TAN T.: A general algorithm for document skew angle estimation. In *ICIP97* (1997), pp. 230–233.

[SK00] SAFABAKHSH R., KHADIVI S.: Document skew detection using minimum-area bounding rectangle. In *ITCC '00: Proc. of the Int. Conf. on Information Technology: Coding and Computing (ITCC'00)* (Washington, DC, USA, 2000), IEEE Computer Society, p. 253.

[SZHZ07] SU T.-H., ZHANG T.-W., HUANG H.-J., ZHOU Y.: Skew detection for chinese handwriting by horizontal stroke histogram. In *Proc. of the 9th Int. Conf. on Document Analysis and Recognition (ICDAR '07)* (Washington, DC, USA, 2007), IEEE Computer Society, pp. 899–903.

[TT03] TKALCIC M., TASIC J.: Colour spaces: perceptual, historical and applicational background. In *EUROCON 2003. Computer as a Tool. The IEEE Region 8* (Sept. 2003), vol. 1, pp. 304–308 vol.1.

[Tyb04] TYBON R.: *Generating Solutions to the Jigsaw Puzzle Problem*. PhD thesis, Griffith University, Australia, 2004.

[UR05] UKOVICH A., RAMPONI G.: Features for the reconstruction of shredded notebook paper. *Image Processing (ICIP 2005). IEEE Int. Conf. on 3* (Sept. 2005), III–93–6.

[YS03] YAO F.-H., SHAO G.-F.: A shape and image merging technique to solve jigsaw puzzles. *Pattern Recogn. Lett. 24*, 12 (2003), 1819–1835.