Occluded Character Restoration Using Active Contour with Shape Priors

Rafi Cohen, Klara Kedem, Itshak Dinstein, Jihad El-Sana
Ben-Gurion University, Beer-Sheva, Israel
{*rafico,el-sana,klara*}*@cs.bgu.ac.il, dinstein@ee.bgu.ac.il*

Abstract

Broken or partially visible characters is common phenomenon in historical documents. It stems from various factors, such as overlaid text or degradation. Restoring such characters is necessary for document analysis applications. This paper presents a new approach for restoring underlaying Hebrew broken characters that were partially occluded by Arabic text in a palimpsest. We apply text recognition on the fragments of the Hebrew letters and select the k candidate letters that match best the fragments of the Hebrew letters. We then complete the broken Hebrew characters using active contour with the k candidates as shape priors. We use a modified geodesic active contour, which we tailored to occluded text restoration. It is initialized on the fragments of the Hebrew text, then it undergoes an expansion phase and a contraction phase via the occluding Arabic text to form the restored Hebrew character. We measure the distance between the completed character and its corresponding priors and choose the shape with the minimal distance as the reconstructed character. Experimental results are presented. On average 68% *of the characters were correctly reconstructed.*

1 Introduction

Parchment was used in the middle ages as one of the main writing materials. Due to its high cost, the writing was often scraped off or washed to be reused. Documents consisting of reused parchments are called *palimpsests*. Many palimpsests are valuable, as they may contain, under the overlaid text, an important historical information. A well known example is the Archimedes palimpsest [14, 20]. The original text in palimpsests is often broken, due to the occlusion caused by the more recent writing.

In this paper we address the problem of restoring broken characters caused by occlusion. More specifically, we address the restoration of Hebrew text that was partially occluded by Arabic text (Figure 1).

The research in restoration of occluded and broken characters could be classified to three main approaches: in-painting, active contours, and rule-based systems. Next, we briefly overview related work in each category.

In-painting is defined as the automatic filling of an unknown image region [5]. Most of in-painting algorithms are for natural images and do not fit well text images. Bertozzi *et al.* [6] suggest an in-painting method for text images that is based on a slightly modified Cahn-Hilliard equation, which models the phase separation in binary fluids. Chang *et al.* [10] address the problem of visible watermark characters, occluded by foreground content. Their algorithm is based on an exemplar-based in-painting which reconstructs the missing areas by referring to similar patches from undamaged areas. Hollaus *et al.* [13] perform in-painting of the Archimedes palimpsest using a high-order Markov random field. These in-painting methods do not use character models for the reconstruction and this causes smearing in the in-painted area, and inability to restore severely broken characters. Allier *et al.* [1] restore broken printed Latin characters using the classical active contour model, with a GVF field and a character model. The model is automatically selected using a bank of Gabor filters, and the completion is done in the degraded area. Droettboom [11] propose a method based on graph combinatorics to merge broken characters of historical documents. The goal of his algorithm is to find an optimal way to join connected components on a given page that would maximize the mean confidence of all characters. Several approaches for restoring broken digits in binary images extend the

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.243

boundaries of the character using masks, morphological operators and region-growing [19, 22, 21].

We use an active contour with shape prior to restore the underlaid text in palimpsests. In this work, we assume that the input consists of segmented Arabic text (the overlaid layer) and partially occluded Hebrew text (the underlaid layer). We apply our text recognition algorithm [17] to find a constant number, k, of candidate letters that match best the fragments of the Hebrew letters. We then complete the broken Hebrew characters using a 2-phase active contour for occluded text (more in Section 3) with the k candidates as shape priors.

We developed a 2-phase active contour for occluded text. We compared the results of our active contour with the k candidate letters using the *shape context* algorithm [4] and chose the shape with the minimal distance as the reconstructed Hebrew character.

The paper is organized as follows, In Section 2 we describe briefly active contours with shape prior, then in Section 3 we describe our reconstruction approach followed by experimental results in Section 4. Finally, we suggest directions for future work in Section 5.

2 Active contours with shape prior

Active contours have been used successfully for detection and segmentation in various image processing applications [7, 18, 16, 9]. The geodesic active contour (GAC) model [8] is an *edge based* active contour, in which a curve is evolving to minimize the following energy term:

$$E_{GAC} = \int g(|\nabla I(C(s))|)ds, \qquad (1)$$

where g is an inverse edge indicator function (e.g., $g(I) = 1/(1+|\nabla I|^2)$), and C is the evolving curve. The steady state of the active contour is reached by evolving each contour point according to Eq. (2), where κ is the contour curvature, $\vec{N}$ is the unit normal to the curve, and ν is a real constant, which controls the inflation and shrinking of the contour; e.g. negative value of ν causes inflation.

$$\frac{\partial C}{\partial t} = g(I)(\nu + \kappa)\vec{N} - (\nabla g \cdot \vec{N})\vec{N}, \qquad (2)$$

The implementation of the geodesic active contour is based on the level set framework [15]. The main idea of the level set method is to embed the evolving contour C as the zero level set of an implicit function ϕ, defined in a higher dimension $C(t) = \{(x, y)|\phi(x, y, t) = 0\}$.

In order to integrate a shape as a prior knowledge to the segmentation process, it is necessary to represent the shape as a signed distance function (SDF) [16, 18]. In the level-set framework, integrating the shape becomes natural. Let ϕ be the SDF for the segmentation, and ψ be the SDF of the embedding shape. Then, their shape difference can be defined as in equation (3), where $H(x)$ is the Heaviside step function [9].

$$E_{shape} = \int (H(\phi) - H(\psi))^2 dx, \qquad (3)$$

3 Our approach

Our restoration method works at character level and aims to restore Hebrew characters occluded by Arabic text in palimpsests. The algorithm consists of four steps which are discussed next.

3.1 Character fragment extraction

In the first step of the algorithm we compute the boundaries of the connected components which are fragments of the Hebrew characters. To segment these fragments, we initialize an inflating geodesic active contour on the fragments, and set the pixels of the foreground Arabic letters as barriers. We allow the active contour to evolve until it converges to a stable solution at the boundary of the fragments of characters (Figure 3(c)). Once the character fragments are detected, we manually assign a bounding box for each character to group the fragments of the occluded character.

3.2 Preliminary recognition

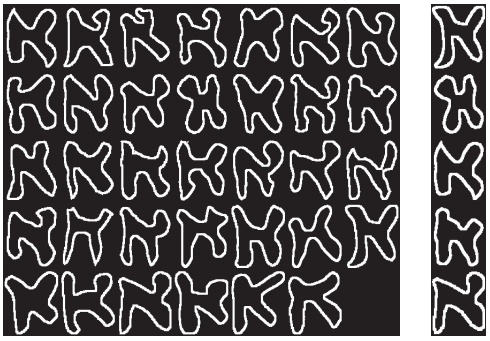
For each document we create a training set of models consisting of some completely visible representatives of characters in the text document (not all the characters are occluded). The recognition of a partially occluded character, c, is performed by computing the distance between each of the models and c. The closest k models are selected as potential candidates for c. The score of matching the partially occluded character, c, and a candidate model, m, is computed by the GSC method in [12, 17],which employs normalized features and flexible window sizes.

(a) (b)

Figure 1. (a) A Patch from the Palimpsest "L 120 SUP c. 7", and (b) its corresponding segmented Hebrew And Arabic texts in white and green texts, respectively.

3.3 Model set selection for a character

To reduce the computational cost of the reconstruction using shape-prior, we use a method similar to the one suggested by Bar-Yosef *et al.* [3], with a different way for picking the representative models. For each character c we manually build a set of reference models that represent the variations among the different shapes of c in the document. We rearrange the signed distance function (SDF) of the training characters as column vectors, and apply principal component analysis (PCA) to reduce dimensionality [3]. In this paper, we use the first N principal components and cluster the N-dimensional vectors into k clusters using k-means. For each cluster, we choose the model closest to the mean as the representative shape model (of the cluster). Figure 2(a) shows a set of 34 characters from which we choose five models (see Figure 2(b)).



(a) (b)

Figure 2. (a) A training set of 34 aleph characters; (b) the 5 representatives chosen by PCA and clustering.

3.4 Opening barriers for a character

The development of the modified geodesic active contour (GAC) for character reconstruction is guided by a shape-prior energy (see Eq. (4)). The energy of the prior's shape, E_{Shape}, stays constant during the development of the active contour, while the E_{GAC} is changing according to the phase of the algorithm.

$$E = E_{GAC} + E_{Shape} \tag{4}$$

To complete the reconstruction of the Hebrew

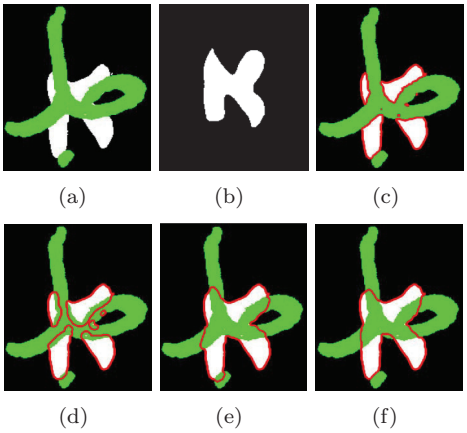


Figure 3. Illustration of the steps in our algorithm. (a) input; (b) the chosen shape-prior; (c) the edges of the seen Hebrew letters, as computed by the geodesic active contour, with the Arabic letters as barriers; (d) the active contour breaching the barriers into the arabic component. (e) the maximal inflation of the active contour; and (f) the computed borders of the Hebrew text are shown.

letters we remove the barriers caused by the Arabic components (the foreground content), by discarding the external forces on the contour inside the Arabic characters. This enables the Hebrew fragments to expand and merge with each other through the occluding Arabic text. This is done by setting the input of the inverse indicator function, g, from Eq. (2) to $g(I_A + I_H)$, where I_A is defined as:

$$I_A(x,y) = \begin{cases} 255, & (x,y) \text{ is an Arabic pixel,} \\ 0, & \text{otherwise ,} \end{cases} \quad (5)$$

and I_H is defined in a similar way. We also set the value of ν (Eq. (2)), which controls the inflationary behaviour of the active contour, to a negative value.

Once the number of connected components generated by the active contour equals the number of connected components of the character model, the algorithm switches phase. In the second phase it discards the inflating force and the contour starts to shrink to its final shape (the prior). This is done by setting the value of ν to zero in pixels that belong to the Arabic components. Figures 3(a) and 3(b) show the input dataset and the selected prior, respectively. Figures 3(c)-(f) illustrate the development of the active contours.

3.5 Final recognition

In our implementation we use k models and N clusters. Hence, the reconstruction using shape-prior is performed kN times for each broken letter. The distance between the reconstructed shape, and the corresponding prior is measured using shape context [4]. We choose the shape with the minimal distance as the reconstructed character.

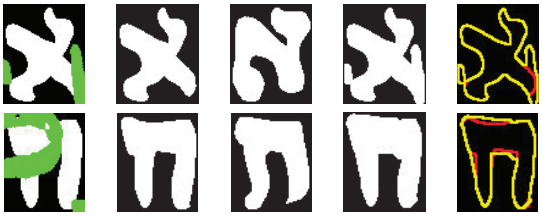
4 Experimental results

We experimented with our approach using various palimpsests from the Arabic manuscripts of the Ambrosiana from Milan library. We generated the ground truth (GT) data by using Bar-Yosef's binarization method [2] to segment the Arabic characters. The segmentation of the Hebrew character fragments was done manually and was verified by experts in Jewish history.

A total number of 765 occluded characters were restored using the proposed algorithm. In our implementation we set $N = k = 5$, and measured the restoration accuracy using the Normalized Symmetric Difference (NSD) metric (see Eq. 6, where L is the ground-truth character shape, and $\widehat{L}$ the restored character). We used $|S|$ to denotes the number of pixels contained in the shape S. Note that NDS is a dissimilarity measure given in percentage, i.e., lower values correspond to higher restoration accuracy.

$$NSD = \frac{|L \backslash \widehat{L}| + |\widehat{L} \backslash L|}{|L|} \times 100 \quad (6)$$

Figure 6 reports the average recognition rates and NSD for Hebrew characters from different datasets. The overall percentage of occluded characters which we correctly recognized is 67.58%, and the average NSD is 8.28%. This low NSD value indicates that the restored character has high similarity with the ground-truth character. In computing the statistics, we omitted the Hebrew letters ו,ן,י which are almost indistinguishable in the Hebrew handwriting. In the future we plan to use a dictionary to improve the recognition rate of these letters. An examination of the unrecognized characters, reveals that 68.26% of the unrecognized characters failed to be recognized in the preliminary recognition [17]. That is, none of the 5 candidate characters suggested was the ground truth character. For an analysis of the possible reasons for that we refer the reader to [17]. Figure 5 illustrates the results



(a) (b) (c) (d) (e)

Figure 4. Normalized Symmetric Difference: (a) the occluded character; (b) the ground-truth; (c) the prior; (d) the restoration result; and (e) finally the ground-truth (Red) and the restored character (Yellow) overlaid together. The NSD values for the upper and lower rows are 15% and 8%, respectively.

(a) Input

(b) Hebrew ground truth

(c) Result of proposed method

Figure 5. Results for "L 120 SUP c. 7" palimpsest.

of our algorithm on a text line, and Figure 4 depicts the computation of the NSD on two different Hebrew characters.

5 Conclusions and future work

We have presented an approach restoring broken Hebrew characters, using active contours with shape-prior. We presented a simple scheme for choosing the correct shape prior.

Our future work will focus on improving some aspects of the algorithm. In our current implementation, the bounding box of character fragments (that belong to the same letter) is manually chosen. We intend to automate this process by first dividing the document into lines, and then to apply a sliding window over each line. The size of the window is determined by the range of character dimensions in the document. A classifier will be used to determine whether the fragments in the windows belong to one character or not. We also intend to experiment with different classifiers.

Acknowledgment

This research was supported in part by the Israel Science Foundation grant no. 1266/09, DFG-Trilateral Grant no. 8716, the Lynn and William Frankel Center for Computer Sciences, and the Paul Ivanier Center for Robotics and Production Management at Ben-Gurion University, Israel. We would like to thank Prof. Uri Ehrlich and Uri Safrai from the Goldstein-Goren department of Jewish thought, Ben-Gurion University of the Negev, for their assistance in generating the ground truth. The authors also thank Irina Rabaev for providing us the source code of her algorithm.

References

[1] B. Allier, N. Bali, and H. Emptoz. Automatic accurate broken character restoration for patrimonial documents. *International Journal on Document Analysis and Recognition*, 8(4):246–261, 2006.

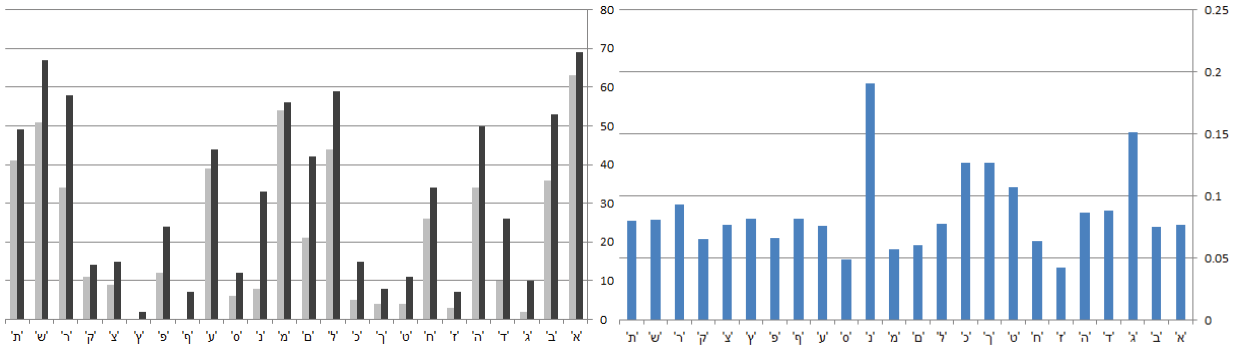
[2] I. Bar-Yosef, I. Beckman, K. Kedem, and I. Dinstein. Binarization, character extraction, and writer identification of historical hebrew calligraphy documents. *International Journal on Document Analysis and Recognition*, 9(2):89–99, 2007.

[3] I. Bar-Yosef, A. Mokeichev, K. Kedem, I. Dinstein, and U. Ehrlich. Adaptive shape prior for recognition and variational segmentation of degraded historical characters. *Pattern Recognition*, 42(12):3348–3354, 2009.

[4] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(4):509–522, 2002.

[5] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424. ACM Press/Addison-Wesley Publishing Co., 2000.

[6] A. Bertozzi, S. Esedoglu, and A. Gillette. Inpainting of binary images using the cahn–hilliard equation. *Image Processing, IEEE Transactions on*, 16(1):285–291, 2007.



(a) Occurrences (b) NSD

Figure 6. Statistical results for the palimpsests. (a) dark gray represents the number of occurrences of the Hebrew character on the x-axis and light gray the number of successful recognitions; (b) the NSD for the Hebrew letters on the x-axis.

[7] A. Blake and M. Isard. *Active Contours: The Application of Techniques from Graphics,Vision,Control Theory and Statistics to Visual Tracking of Shapes in Motion*. Springer, 1998.

[8] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *International journal of computer vision*, 22(1):61–79, 1997.

[9] T. Chan and W. Zhu. Level set based shape prior segmentation. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 1164–1170. IEEE, 2005.

[10] L. Chang, J. Sun, M. Suwa, H. Takebe, Y. He, and S. Naoi. Occluded text restoration and recognition. In *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, pages 151–158. ACM, 2010.

[11] M. Droettboom. Correcting broken characters in the recognition of historical printed documents. In *Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, pages 364–366. IEEE Computer Society, 2003.

[12] J. Favata and G. Srikantan. A mutiple feature/resolution approach to handprinted character/digit recognition. *International Journal of imaging Systems and technology 7*, pages 304–311, 1996.

[13] F. Hollaus and R. Sablatnig. Inpainting of occluded regions in handwritings. OAGM/AAPR Workshop, Austrian Computer Society, 2011.

[14] R. Netz and W. Noel. *The Archimedes Codex*. Weidenfeld & Nicolson, 2007.

[15] S. Osher and J. Sethian. Fronts propagating with curvature-dependent speed: algorithms based on hamilton-jacobi formulations. *Journal of computational physics*, 79(1):12–49, 1988.

[16] N. Paragios, M. Rousson, and V. Ramesh. Matching distance functions: a shape-to-area variational approach for global-to-local registration. *European Conference on Computer Vision (ECCV)*, pages 813–815, 2002.

[17] I. Rabaev, O. Biller, J. El-Sana, K. Kedem, and I. Dinstein. Features for hebrew character searching. Technical Report 12-07, Ben-Gurion University of the Negev, May 2012.

[18] M. Rousson and N. Paragios. Shape priors for level set representations. *European Conference on Computer Vision (ECCV)*, pages 416–418, 2002.

[19] Z. Shi and V. Govindaraju. Character image enhancement by selective region-growing. *Pattern recognition letters*, 17(5):523–527, 1996.

[20] The Archimedes Palimpsest Project. `http://archimedespalimpsest.org/`.

[21] A. Whichello and H. Yan. Linking broken character borders with variable sized masks to improve recognition. *Pattern Recognition*, 29(8):1429–1435, 1996.

[22] D. Yu and H. Yan. Reconstruction of broken handwritten digits based on structural morphological features. *Pattern Recognition*, 34(2):235–254, 2001.