2009 10th International Conference on Document Analysis and Recognition

Restoration and segmentation of highly degraded characters using a shape-independent level set approach and multi-level classifiers

Reza Farrahi Moghaddam, David Rivest-Hénault, and Mohamed Cheriet
Synchromedia Laboratory for Multimedia Communication in Telepresence,
École de Technologie Supérieure, Montréal (QC), H3C 1K3, Canada
{reza.farrahi-moghaddam, david.rivest-henault}.1@ens.etsmtl.ca, mohamed.cheriet@etsmtl.ca

Abstract

Segmentation of ancient documents is challenging. In the worst cases, text characters become fragmented as the results of strong degradation processes. New active contour methods allow to handle difficult cases in a spatially coherent fashion. However, most of those method use a restrictive, a priori shape information that limit their application. In this work, we propose to address this issue by combining two complementary approaches. First, multi-level classifiers, which take advantage of the stroke width a priori information, allow to locate candidate character pixels. Second, a level set active contour scheme is used to identify the boundary of a character. Tests have been conducted on a set of ancient degraded Hebraic character images. Numerical results are promising.

1. Introduction

Cultural studies and heritage preservation are subjects of growing interests that received a large amount of attention in recent years [2]. However, extraction and analysis of information from historical documents requires costly manual labor. Automatic processing and study of documents can be done, but there are many obstacles in this way. On a regular basis, historical documents suffer from severe degradations which have been introduced along time and can be of very different nature. Bleed-through effect, ink fading, deterioration of paper material and cellulose structure are a few examples of such degradation types. Also, many physical parameters such as the type of paper and ink that were used are involved. Many methods for specific types of degradation have been developed [7, 8, 14]. These methods aim to restore the document image by removing the degradation and enhancing the strokes.

The case of very degraded texts where many of the stroke pixels have been destroyed is one of most challenging problems [1, 3, 5, 9, 19]. Usually, the goal is the improvement of some OCR output. However, in cases where legibility of the text is the main target, an accurate reconstruction of the characters is needed. Different approaches, such as morphological methods [19] and active contours methods using a shape prior [1, 3], have been used for the restoration of broken characters. For example, Bar-Yosef *et al*. [3] take profit in an adaptive prior shape.

Shape-independent methods tend to act on a per-pixel basis and, indeed, produce results that are not spatially coherent and create unacceptable artificial links [17]. On the other hand, the performance of shape-based methods is intimately linked to the availability of very similar characters in the the training set. Arguably, those methods are integrated restoration and recognition processes, which are not desirable in many situation. There is, therefore, a need for restoration methods which are robust and, at the same time, uses shape-independent a priori information. It could then be possible to use theses methods to provide enhanced and uniform images to the succeeding stages, such as recognition/understanding processes.

In this work, we use the level set framework and content-level classifiers to introduce a segmentation method which use a small set of parameters to represent *a priori* information about the stroke width, its regularity and its intensity distribution. While our method can also be perceived as a prior-based method, it is worth noting that we only use local shape priors. That is, we do not use any global, i.e. at character or word level, shape information of the handwritten text. Our method is then independent of any recognition process. The goal of this work is not to provide a method superior to the shape-based methods, but it is to eliminate the need for global shapes *a priori* knowledge. In fact, as an application, our method can be used as a preprocessing step for shape-based methods. Besides, the level set framework is actually a general, open and multi-scale framework which can host several different methods and concepts in an easy, fast and stable way at the same time. Also, it provides

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.107

continuous boundaries which is very important in the case of degraded documents enhancement. As such, this framework can be used as an interesting integration framework for characters segmentation.

The paper is organized as follows. In section 2, a brief description of the problem and also a priori information is provided. Then, the multi-level classifiers are discussed in section 3. The computation of the probability density function is presented in section 4. In section 5, the main level set formalism is provided. It follows by the experimental results and evaluation in section 6. Finally, the summary of work is presented in the conclusion.

2. Problem statement and a priori information

A grayscale image of a degraded character on an old document is given by $u(r)$ where $r = (r_x, r_y)^T \in \Omega$. The domain Ω is a rectangle with H and W dimensions. The character image is very degraded and many of stroke pixels have been erased. Also, some parts of adjoining characters are presented on the image. The goal is to reconstruct the original character as accurately as possible. It is assumed that the average stroke width, w, is available as a priori information.

3. Content-level classifiers: Stroke cavity map

The first step of our broken characters restoration method is to translate the stroke width information into a form that is usable by the level-set method. In this section, a field, called Stroke Cavity Map (SCM), will be developed in order to represents the possible stroke pixels. The details of SCM computation are provided in subsection 3.2. Before computing SCM, we need to extract valuable stroke pixels which are presented on the document image. It will be done using Stroke Map classifier [8] in the next subsection.

3.1. Stroke map

Stroke Map (SM) was first introduced in [8]. SM finds possible stroke pixels based on a kernel method [18] and use the stroke width w as an input parameter. In the case of very degraded characters, the strokes are not continuous and are thinned. Therefore, a reduced value of the parameter w must be used in order to capture true stroke pixels and we use $w/2$ to compute SM. An example of a degraded character and its corresponding SM is provided in Fig. 1.a) and 1.b).

3.2. Stroke cavity map

Stroke cavity map (SCM) contains all pixels which are probable to be on a stroke. In addition to SM pixels, all pixels which are between two SM pixels of less than the stroke width distance will be added. This is done using a binary kernel $K_{r_1,r_2}(r)$:

$$K_{r_1,r_2}(r) = \begin{cases} 1 & ||r_1 - r_2|| \leq w \quad \& \quad r \in R(r_1, r_2, t) \\ 0 & \text{otherwise} \end{cases}$$

where $R(r_1, r_2, t)$ is a rectangle from pixel r_1 to pixel r_2 with height t which represents the pen thickness. SCM can simply be initialized:

$$SCM(r) = \begin{cases} 1 & r \in SM \\ s & \exists r_1, r_2 \in SM \quad s.t. \quad K_{r_1,r_2}(r) = 1 \\ 0 & \text{otherwise} \end{cases}$$

where s is a value between 0 to 1. Fig. 1.c) shows the SCM corresponding to Fig. 1.a). Finally, in order to use a priori information that the character is in the center of image, the SCM is modified using a spatial decay transform as follows:

$$\begin{aligned} SCM(r) = SCM(r) \\ \times \left(1 + \tanh\left(2\frac{-|r_x - W/2| + d_{scm}}{d_{scm}}\right)\right) \\ \times \left(1 + \tanh\left(2\frac{-|r_y - H/2| + s_{scm}d_{scm}}{d_{scm}}\right)\right), \end{aligned}$$

where d_{scm} stands for the average character width of the text. As the spacing between lines is usually more that character spacing, d_{scm} is multiplied by a factor s_{scm} in the vertical direction. The final SCM of Fig. 1.a) is shown in Fig. 1.d).

4. Estimation of intensity *pdfs*

A small set of characters manually segmented by an expert is used in order to estimate the pixel intensity Probability Density Functions (*pdf*) associated with the pixel classes *text* and *background*. Let x_j^k be the number of pixel with intensity $j \in \{0 \dots 255\}$ labeled as class $k \in \{text, bg\}$ and n_k the total number of pixel labeled k. In this work, the *pdfs* are estimated by using a normalized smoothed histograms:

$$pdf_k(i) = \frac{1}{n_k} \sum_{j=1}^{n_k} \frac{x_j^k}{h\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{i-j}{h}\right)^2\right\},$$

where, h is a bandwidth parameter. In turn, the estimated probability that a certain pixel intensity i occurs given a pixel class is $P(i|k) = pdf_k(i)$.

5. Level set formulation

Within the level set framework, a set of closed, but possibly disjoint, contours C on $\Omega \subset \mathbb{R}^2$ is represented by the

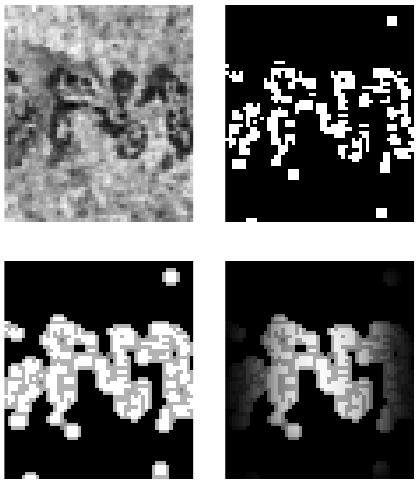

Figure 1. a) An image of a degraded character [3], b) corresponding SM obtained using $w/2 = 2$, c) Initial SCM after applying the binary kernel k, and d) SCM of (a) after applying the decay transform.

intersection of a surface $\phi(r) : \Omega \rightarrow \mathbb{R}$ and of the zero level $z = 0$. By convention, $\phi(r)$, called Level Set Function (*lsf*), is approximatively as a signed distance function of C. If $\phi(r_i) \geq 0$ the pixel at r_i is labeled as *text* else, if $\phi(r_i) < 0$, the pixel is labeled as *background*.

Starting from an initial and arbitrary guess, $\phi(r)$ evolves according to the following governing equation:

$$\frac{\partial \phi(r)}{\partial t} = \delta_\phi \left(F_{pdf} + \nu F_R + \mu F_{scm} \right), \qquad (1)$$

where ν and μ are free parameters and

$$\delta_\phi = \frac{1}{\pi} \frac{1}{(1 + \phi^2)} \qquad (2)$$

is a regularized Dirac function that limit the evolution of ϕ around the zero level [4].

The first term, F_{pdf}, is the data force. It tends to attract inside the contours pixels that are more likely to be part of the text and to repulse pixels more likely to be part of the background. It is expressed as follows [11]:

$$F_{pdf}(r) = \log \left(pdf_{text}(u(r)) \right) - \log \left(pdf_{bg}(u(r)) \right)$$

The second term, F_R is the common minimal boundary length regularization force [12, 13]:

$$F_R = \nabla \cdot (\nabla \phi / |\nabla \phi|)$$

Finally, he last term is the SCM force:

$$F_{scm}(r) = \begin{cases} SCM(r) - 1 & \phi(r) \geq 0 \\ SCM(r) & \phi(r) < 0 \end{cases}$$

This force attracts inside the contours plausible, but possibly erased, candidate stroke pixels.

The evolution of (1) by a numerical scheme results in the displacement of the contour toward the character boundary, resulting in a segmentation. For details on the numerical implementation of a related level set method, we refer the reader to [16]. In section 6, our segmentation method of is tested on a set of degraded characters.

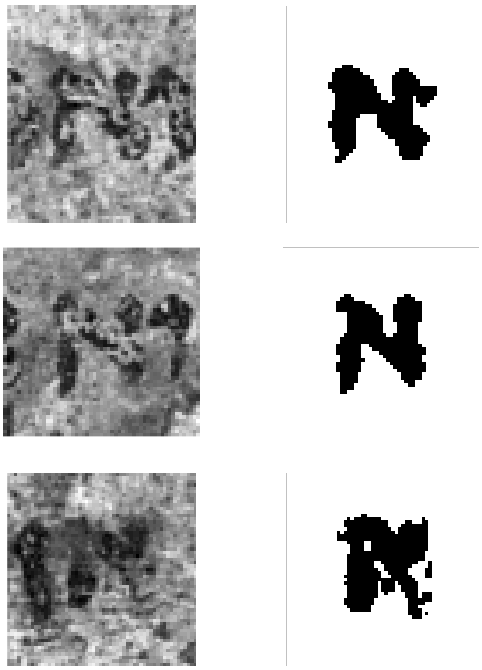

Figure 2. Left column: Input images of degraded characters from [3]. Right column: Segmentation obtained using the level set method presented in this work.

6. Experiments

6.1 Experimental setup

Our method has been applied to a small set of very degraded character images. For this experiment, we set the various parameters of our method as follows. Three parameters of our method can be set using a priori information concerning the document. Those have been fixed as follows: $w = 4$, $d_{scm} = W/4$, $s_{scm} = 1.5$. The bandwidth parameter for the smoothing of the histogram has been set to $h = 3$. Finally, the three remaining parameter had to be hand picked, we used $s = 0.7$, $\nu = 2.5$ and $\mu = 3.0$.

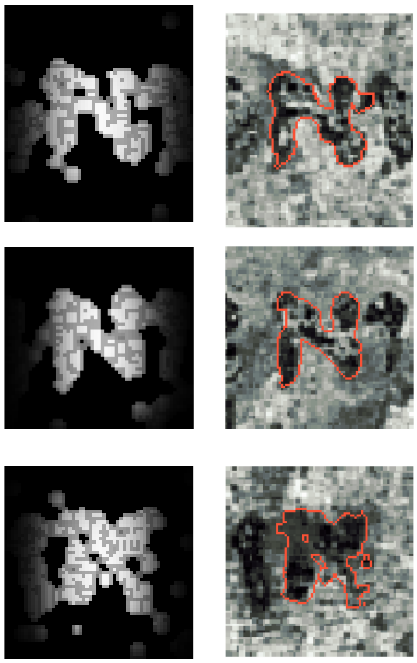

Figure 3. Left column: SCM of input images of Figure 2. Right column: The final contour is shown in color over the input images.

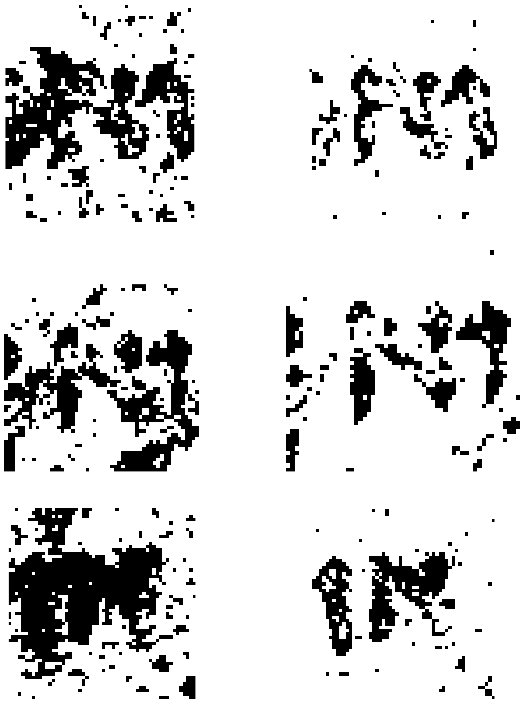

Figure 4. Left column: Binarized version of input images of Figure 2 obtained using Otsu's method. Right column: Binarized version of input images of Figure 2 obtained using Niblack's method. In the Niblack's method, following set of parameters is used(using notation of [15]): $k = -1.7$ **and neighborhood size**$= 30$**.**

Three example input characters are shown on Fig. 2. The input images which have lost many of their stroke pixels are shown on the left column of the figure. The recovered strokes are continuous and smooth. It is worth noting that these results are obtained without using a priori shape information. The SCM and the contour of each case is shown in Fig. 3. For the sake of comparison, the results of application of global and local thresholding to the input images are presented in Fig. 4. As it can be seen from the figure, both thresholding methods are incapable of preserving stroke pixels of light intensity. On the other hand, without using any shape dependent prior, the proposed method is able to recover many erased stroke pixel and produce interesting, spatially coherent results.

6.2 Evaluation

In order to evaluate the performance of the proposed method, Euclidean distance to the manual segmentation is computed for different methods. For each image, an error mask of the manual segmentation is computed using Euclidean distance [6] (using *bwdist* function of Matlab [10]). Then, the output of a method is mapped on the error mask. In Fig. 5, Euclidean distances between the outputs of the proposed method and the manual segmentations are presented. The distances between the two images is finally computed by summing up all distances and normalizing the result to the area of the manual image. The distance is summarized by the following formula:

$$d(u, u_{manual}) = \frac{||u \times EDT(u_{manual})||}{||u_{manual}||}$$

where EDT represents Euclidean distance transform. Figure 6 shows the mean and variation of the distance d for different method in the logarithmic scale. It can easily be seen that the method 1 which is our proposed method provides a good result even comparing to a shape-based method [3].

7. Conclusion

A novel segmentation method for highly degraded character images is presented. The method is implemented within the level set framework. Using a set of parameters

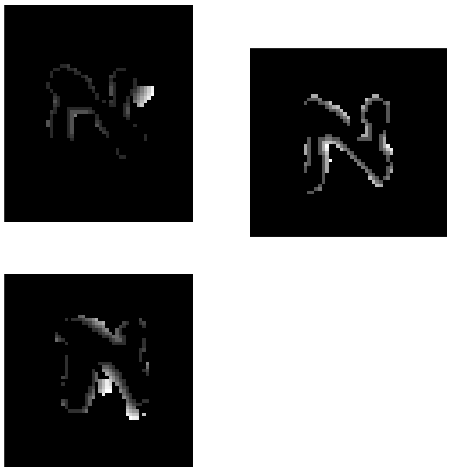

Figure 5. Computed Euclidean distances of the results of our proposed method to the manual segmentations. The distances are marginal and small.

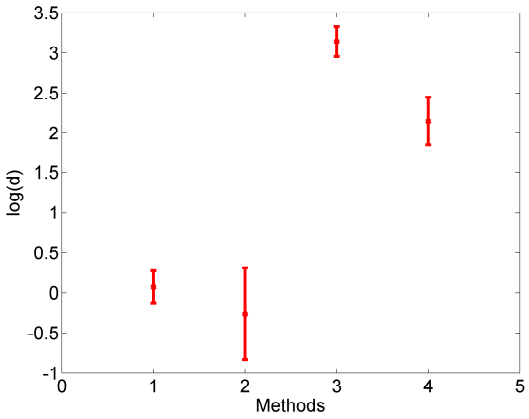

Figure 6. Mean and variation of the distance d for different methods. Method 1 is our proposed method. Method 2 is adaptive shape-based level set method [3]. Method 3 is Otsu's method, and Method 4 is Niblack's method.

representing a priori knowledge, a content-level classifier is introduced which controls the evolution of level set function in order to preserve the erased pixels. The most important a priori information used is the average stroke width. The segmentation force of the method is based on the *pdf* distributions of the stroke and background regions.

In the future works, other types of segmentation forces will be considered. Also, using larger datasets, performance of the content-level classifiers, especially SCM, will be improved. Introduction of higher level features, such as the flow field, for covering very extreme cases is another direction.

References

[1] B. Allier, N. Bali, and H. Emptoz. Automatic accurate broken character restoration for patrimonial documents. *Int. J. Doc. Anal. Recognit.*, 8(4):246–261, 2006.

[2] A. Antonacopoulos and A. Downton. Special issue on the analysis of historical documents. *IJDAR*, 9(2):75–77, Apr. 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*, In Press, Corrected Proof:–.

[4] T. Chan and L. Vese. Active contours without edges. *Image Processing, IEEE Transactions on*, 10(2):266–277, 2001.

[5] M. Droettboom. Correcting broken characters in the recognition of historical printed documents. In *3rd ACM/IEEE-CS*, pages 364–366, Houston, Texas, 2003.

[6] R. Fabbri, L. D. F. Costa, J. C. Torelli, and O. M. Bruno. 2d euclidean distance transform algorithms: A comparative survey. *ACM Comput. Surv.*, 40(1):1–44, 2008.

[7] R. Farrahi Moghaddam and M. Cheriet. Low quality document image modeling and enhancement. *IJDAR*, to be published, 2009.

[8] R. Farrahi Moghaddam and M. Cheriet. RSLDI: Restoration of single-sided low-quality document images. *Pattern Recognition*, to appear in Special Issue on Handwriting Recognition, 2009.

[9] J. D. Hobby and T. K. Ho. In *ICDAR'97*, pages 394–400. IEEE Computer Society, 1997.

[10] The Mathworks Inc., Natick, MA. *MATLAB Version 7.5.0*.

[11] N. Paragios and R. Deriche. *International Journal of Computer Vision*, 46:223–247, 2002.

[12] D. Rivest-Hénault and M. Cheriet. Image segmentation using level set and local linear approximations. *Image Analysis and Recognition*, 4633/2007:234–245, 2007.

[13] J. A. Sethian. *Level Set Methods and Fast Marching Methods*. Cambridge University Press, 1999.

[14] A. Tonazzini, E. Salerno, and L. Bedini. *IJDAR*, 10(1):17–25, June 2007.

[15] O. Trier and A. Jain. *Pattern Analysis and Machine Intelligence, IEEE Trans. on*, 17(12):1191–1201, 1995.

[16] L. A. Vese and T. F. Chan. *Inter. J. Computer Vision*, 50(2):271–293, 2002.

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

[18] X. Ye, M. Cheriet, and C. Suen. *Image Processing, IEEE Trans. on*, 10(8):1152–1161, 2001.

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