

Extraction of Specified Objects from Binary Images Using Object Based Erosion Transform: Application to Hebrew Calligraphic Manuscripts

Itay Bar-Yosef, Isaac Beckman and Klara Kedem
Ben-Gurion University
Computer science department.
{itaybar,beckmani,klara}@cs.bgu.ac.il

Itshak Dinstein
Ben-Gurion University
Electrical Engineering department
dinstein@ee.bgu.ac.il

Abstract

This paper presents a method for automatic extraction of certain pre-specified objects from binary images. The method is applied to letters from historic Hebrew manuscripts and is based on object based erosion transform. The extraction method deals very well with common problems of historical documents, such as broken or merged characters, and can be applied to different writing styles. An average correct extraction rate of 96% is achieved.

1. Introduction

Paleography is the study of ancient handwritten manuscripts. Among other things, it deals with dating and localizing ancient and medieval scripts, and with studying the development of letter shapes. In order to perform these tasks automatically, one should be able to automatically extract pre-specified letters. The Hebrew alphabet consists of 22 letters, five of them have special forms when appearing at the end of a word. Although each character possesses a distinct shape, some characters have similar shapes to other characters. Some examples are given in Figure.1. The Hebrew *STAM* handwriting used in our research, is influenced both by time and place – different geographical regions over different periods of historical time use different scripts of the same alphabet.

There are several papers dealing with retrieval of complicated characters or extraction of pre-defined symbols. A system for retrieval of chinese calligraphic characters is reported in [1], where characters are represented by an approximated point context. In Saykol et. al [2], features based on angular and distance span of shapes are used for symbol extraction. The symbols are maintained in a codebook



(a)



(b)

(c)

Figure 1. (a) Letters with unique shapes – Tzadik, Aleph, Shin and Lamed. Letters with high resemblance: (b) The letters Vav and Zain. (c) The letters Kaf and Beit.

for the purpose of content-based image retrieval of Ottoman documents. A segmentation-free approach for recognition of arabic text is presented in [3]. Text primitives are extracted using mathematical morphology in order to recognize words. They report promising results for symbol extraction and word recognition.

Historical documents often suffer from degradations which cause deformed, broken or merged characters. A good extraction method should be able to handle this artifact. We present a segmentation-free approach for extraction of pre-specified letters from his-

torical Hebrew manuscripts. Our work is related to the work presented in [3]: Instead of extracting text primitives for recognition of words, we use mathematical morphology to extract the pre-defined letters. We apply our method on Hebrew calligraphic manuscripts and demonstrate it on the Hebrew letter Aleph. Results show that our method performs very well in extraction of selected letters, deals very good with broken or merged characters and that our method can be applied to a large variations of writing styles.

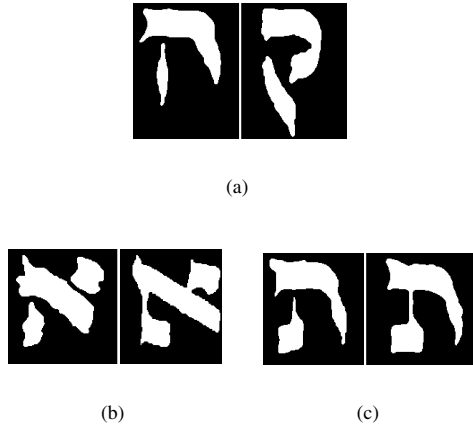


Figure 2. (a) Some of the Hebrew letters are composed of more than one connected component (CC), for example the letters Kuf and Hei (b,c) Although most of the Hebrew letters are composed by one CC, their strokes are often disjoint and appear as several CC's.

The extraction method is based on the well known erosion transform. For another use of the erosion transform, see Haralick et.al [4]. The extraction process is composed of several stages: structuring element generation, character extraction, character validation and structuring element adaptation. Our paper is organized as follows: Section 2 describes the *object based erosion transform*. Section 3 describes the structuring element generation, and Section 4 describes the extraction process. Experimental results are presented in Section 5, and Section 6 summarizes and discusses future work.

2. Object Based Erosion Transform

Let I be a set of the foreground pixels in a binary image. Each object in I is represented by one or more con-

nected components. See, for example Figure 2, in which sets of connected components represent some Hebrew Calligraphic characters. Denote by $\Omega = \{\omega_1, \dots, \omega_N\}$ the set of object classes, each class representing a letter. For each object class $\omega_i, i = \{1, \dots, N\}$, we generate a structuring element S_i , such that the number of translations in which S_i is contained in an object class ω_i is maximal, and the number of translations in which S_j is contained in an object class ω_i when $j \neq i$ is minimal. The erosion operator \ominus , causes objects to shrink. The amount and the way that they shrink depends upon the choice of the structuring element. We show that when a suitable structuring element is used, the connected components of substantial area in the binary image $D_n = I \ominus S_n$, are associated with objects belonging to class ω_n . Consider the following definition of the erosion operation:

$$D_n = I \ominus S_n = \{ (r, c) | (S_n)_{(r,c)} \subseteq I \}$$

For each foreground pixel (r, c) in D_n , the structuring element S_n translated by (r, c) is contained in I . Denote the set of all foreground connected components of I by $C = \{C_1, C_2, \dots, C_M\}$, and the set of all foreground connected components in image D_n by $CD^n = \{CD_1^n, CD_2^n, \dots, CD_L^n\}$ (see Figure 5(b)).

Claim 1: If the structuring element S_n is connected, then for each connected component CD_i^n in the eroded image D_n , there exists a connected component C_k such that $CD_i^n = C_k \ominus S_n$.

Proof: For each pair of connected (neighboring) pixels $(p, q) \in S_n$ and $(r, s) \in S_n$, the set $(S_n)_{(p,q)} \cup (S_n)_{(r,s)}$ is also connected.

Claim 2: If S_n is a union of K connected components, $S_n = \cup_{k=1}^K S_{n,k}$, then for each connected component CD_i^n , there are at most K connected components in C , such that

$$CD_i^n = \{ \cup_{j=1}^{K_i} C_{i,j}^n \} \ominus S_n \equiv C_i^n \ominus S_n, 1 \leq K_i \leq K,$$

where $C_{i,j}^n$ is the j^{th} component among the set of connected components representing the i^{th} object of class ω_n .

Proof: $S_n = \cup_{k=1}^K S_{n,k}$, where $S_{n,k}$ is a connected component. According to *Claim 1*, there is one connected component, say $C_{i,k}^n$, such that $CD_{i,k}^n = C_{i,k}^n \ominus S_{n,k}$. This is true for $k=\{1, 2, \dots, K\}$. $CD_i^n = \cap_{k=1}^K CD_{i,k}^n$ is the connected component containing all translations for which $(S_n)_{(r,c)} \subseteq C_i^n$. Since some of the K components $C_{i,k}^n$ can be interconnected, C_i^n contains at most K connected components. Figure 3 illustrates a letter "Aleph" represented by two connected components (in gray color). Superimposed on the letter (in white) is the structuring element dilated by the respective CD_i^n component.

The erosion transform for class ω_n is defined as $E_i^n = \# \{ CD_i^n \}$. It is the number of all possible translations

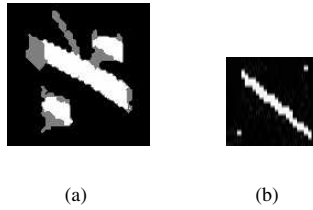


Figure 3. (a) The letter Aleph with the superimposed dilated structuring element. (b) The structuring element.

of S_n such that it is included in C_i^n . A high value of E_i^n indicated that the component C_i^n may represent an element belonging to ω_n .

3. Structuring element generation

The structuring element S_n for class ω_n is generated in the following manner. Let $C_i^n, i = \{1, \dots, T_n\}$ be T_n sets of connected components, representing a training set of T_n elements of class ω_n . We calculate the maximum intersection (under translation) of these sets, and denote it by CS_n . $CS_n = \max(\cap_{i=1}^{T_n} C_i^n)$. The set of pixels belonging to CS_n is contained in each one of the training set elements. The structuring element S_n is a pseudo medial axis of CS_n . Figure 4 illustrates the process of generating a structuring element for the letter Aleph.

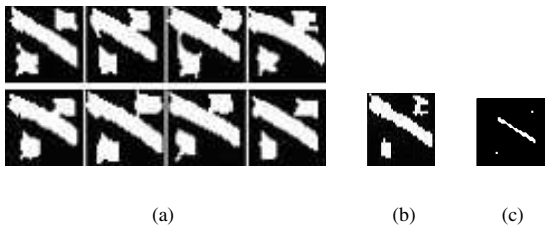


Figure 4. (a) A set of eight Alephs. (b) Their intersection. (c) The structuring element.

4. The Letter Extraction process

The extraction process is composed of several stages. In the first stage, we use the erosion transform to extract the candidate letters. As can be seen in Figure 6, there are some cases where the structuring element S_n is contained in a combination of several characters. Therefore, in the second stage a validation procedure is invoked in order to decide for each extracted character, whether it belongs to class ω_n or not. In order to make the extraction process robust and insensitive to different writing styles, an adaptation process of the structuring element S_n is applied in the last stage.

4.1. Letter extraction

The extraction process of objects belonging to class ω_n , is as follows. Given a gray scale image, we first binarize it using the algorithm presented in [5]. This results in a binary image I . Following the binarization, we normalize I 's height such that lines height in I will be equal on all documents. This is done in order to make the structuring elements $S_n, n = \{1, \dots, N\}$ insensitive to different object sizes. Then, we apply the erosion transform on I , using the structure element S_n of the training set. The eroded image D_n , contains a set of connected components, CD_i^n , each representing a match between S_n and the corresponding component in I , C_i^n (see Figure 5b).

4.2. Validation process

For each $C_i^n, i = \{1, \dots, N_x\}$, where N_x is the number of letters extracted with S_n , we compute a measure V_i^n for validation as follows.

$$V_i^n = \frac{\# \max(C_i^n \cap CS_n)}{\# CS_n},$$

where CS_n is the maximum intersection (under translation) between the training objects of class ω_n ($0 \leq V_i^n \leq 1$). $C_i^n \notin \omega_n$ if $V_i^n \geq THS$. We used in our experiments $THS = 0.9$.

4.3. Adaptation of the structuring element

When dealing with writing styles different from the style of the training set, the extraction process often yields poor results. In order to adapt to the new style, we generate a new structuring element based on the extracted letters. After extraction and validation of C_i^n (according to sections 4.1,4.2), we use the extracted characters as a training set for generating a new structuring element S_n as described in Section 3.

We summarize the overall letter extraction algorithm as follows:

1. Structuring element generation.

- Let $C_i^n, i = \{1, 2, \dots, T_n\}$ be the characters of the training set of a certain letter ω_n .
- Compute the maximum intersection (under translation):

$$CS_n = \max(\cap_{i=1}^{T_n} C_i^n)$$
- Compute the pseudo medial axis of CS_n

$$S_n = \text{Medial}(CS_n)$$
- S_n is the structuring element for the letter ω_n .

2. Character Extraction.

- Compute $D_n = \{I \ominus S_n\}$
- Compute $E_i^n = \# \{CD_i^n\}$, for $i=1$ to the number of connected components in D_n .
- C_i^n are the corresponding extracted characters from I .

3. Validation measure.

- Compute $V_i^n = \frac{\# \max(C_i^n \cap CS_n)}{\# CS_n}$.

4. Structuring element adaptation.

- Define the new training set T . $C_i^n \in T$, if $E_i^n \geq THS$.
- Repeat steps 1-3 to extract the new C_i^n .
- $C_i^n \in \omega_n$ if $V_i^n \geq THS$ and $E_i^n \geq E_i^m$ for all $m \neq n$.

5. Experimental results

At this stage all the experiments were performed on the letter Aleph. Four experiments demonstrate the performance of the character extraction algorithm. The first three experiments were issued on letters of similar style without the validation and adaptation. The fourth experiment was on letters of different style and included both the validation and adaptation methods.

Experiment 1: Nine documents were processed in this experiment. A structuring element was generated for each of the documents using four Alephs manually selected from each document. The first experiment is summarized in Table. 1.

Experiment 2: Eight documents were processed. A structuring element was generated using eight manually selected Alephs, one from each document. The results are summarized in Table. 2.

Experiment 3: Eight documents were processed. The structuring element generated for Experiment 2 was applied to these documents, which are not the ones used in the previous experiment. The results of Experiment 3 are presented in Table. 3.

Experiment 4: Six documents were processed in the

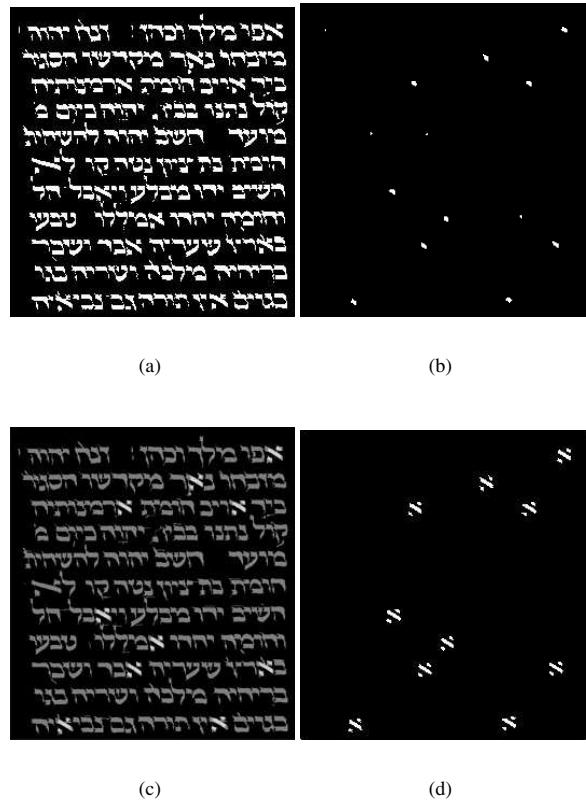


Figure 5. (a) The input image I (b) I eroded by the structuring element of Aleph. (c) The image in b dilated by the structuring element superimposed on the input image. (d) The extracted Alephs.



Figure 6. An example of false detection

fourth experiment. This experiment evaluates the adaptation of the structuring element to different style. The structuring element generated for Experiment 2 was applied to these documents. The results of Experiment 4 are presented in Table. 4.

Number of Documents	9
Number of Characters	6775
Total Number of Alephs	541
correct classification as Alephs	524
% Correct classification As Alephs	96.9%
False classification as Alephs	7

Table 1. Experimental results, Experiment 1.

Number of Documents	8
Number of Characters	5950
Total Number of Alephs	475
correct classification as Alephs	461
% Correct classification As Alephs	97.0%
False classification as Alephs	7

Table 2. Experimental results, Experiment 2.

Number of Documents	8
Number of Characters	5507
Total Number of Alephs	449
correct classification as Alephs	430
% Correct classification As Alephs	95.7%
False classification as Alephs	3

Table 3. Experimental results, Experiment 3.

Number of Documents	6
Number of Characters	5300
False classification as Alephs	8
Total Number of Alephs	380
correct classification as Alephs (stage 1)	198
correct classification as Alephs (stage 2)	348
% Correct classification (stage 1)	52%
% Correct classification (stage 2)	91%

Table 4. Experimental results, Experiment 4.

6. Summary and discussion

We present a new method for extracting pre-specified characters from historical Hebrew manuscripts. Our approach is segmentation-free and does not follow the classical recognition scheme, i.e. segmentation, feature extraction and classification. We demonstrate our method on the Hebrew letter Aleph. The extraction method is composed of four stages: Structuring element generation, character ex-

traction, character validation and adaptation of the structuring element. Our method deals very well the artifacts of historical documents such as broken or merged characters. The first three experiments show very promising results. 96% correct extraction of the letter is achieved. The last experiment demonstrates the effectiveness of the adaptation to different writing styles. An increase from 52% correct extraction in the first extraction stage to 91% is achieved after the adaptation process. Experiments with extraction of different letters with distinct shape show very similar results as reported here for the letter Aleph. When trying to extract letters with high resemblance such as those seen in Figure 1(b), we believe that we should use a hierarchical approach. Similar letters should be extracted as sets of letters and then further processing is required for recognition of the particular letters. We further plan to apply our method on a larger set of letters written in different writing styles. Further research is needed in the adaptation stage and in the extraction of letters with high resemblance.

7. Acknowledgment

We thank Professor Malachi Beit-Arie and Dr. Edna Engel of the Hebrew University for their collaboration in this project. We thank Professor R. M. Haralick of the Graduate Center of CUNY for the help and discussions while Professor I. Dinstein visited his laboratory in New York. Our work was supported in part by the Paul Ivanier Center for Robotics and Production Management, Ben-Gurion University, Israel.

References

- [1] Y. Zhuang, X. Zhang, J. Wu, X. Lu, "Retrieval of Chinese Calligraphic Character Image". 5th Pacific Rim Conference on Multimedia, Tokyo, Japan. Part I, pp. 17-24, 2004.
- [2] E. Saykol, A.K. Sinop, U. Gudukbay, O. Ulusoy, A.E Cetin, "Content-Based Retrieval of Historical Ottoman Documents Stored as Textual Images". IEEE trans. on image processing, vol. 13, no. 3, march 2004. pp. 314-325.
- [3] B. Al-Badr, R. M. Haralick. "A segmentation-free approach to text recognition with application to Arabic text." IJDAR 1 (3). pp. 147-166 (1998)
- [4] M. Schauf, S. Akoy, and R.M. Haralick, "Model-based shape recognition using recursive mathematical morphology", Fourteenth International Conference on Pattern Recognition, (1998), pp. 202-204
- [5] Itay Bar-Yosef, "Input sensitive thresholding for ancient Hebrew manuscript", Pattern Recognition Letters Vol.26 pp.1168-1173. June 2005