ARTICLE IN PRESS

Pattern Recognition ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Junction detection in handwritten documents and its application to writer identification

Sheng He*, Marco Wiering, Lambert Schomaker

Institute of Artificial Intelligence and Cognitive Engineering, University of Groningen, PO Box 407, 9700 AK Groningen, The Netherlands

ARTICLE INFO

Article history:
Received 12 December 2014
Received in revised form
21 April 2015
Accepted 26 May 2015

Keywords:
Handwriting recognition
Junction detection
Cross-script
Writer identification
Junclets

ABSTRACT

In this paper, we propose a novel junction detection method in handwritten images, which uses the stroke-length distribution in every direction around a reference point inside the ink of texts. Our proposed junction detection method is simple and efficient, and yields a junction feature in a natural manner, which can be considered as a local descriptor. We apply our proposed junction detector to writer identification by *Junclets* which is a codebook-based representation trained from the detected junctions. A new challenging data set which contains multiple scripts (English and Chinese) written by the same writers is introduced to evaluate the performance of the proposed junctions for cross-script writer identification. Furthermore, two other common data sets are used to evaluate our junction-based descriptor. Experimental results show that our proposed junction detector is stable under rotation and scale changes, and the performance of writer identification indicates that junctions are important atomic elements to characterize the writing styles. The proposed junction detector is applicable to both historical documents and modern handwritings, and can be used as well for junction retrieval.

1. Introduction

Singular structural features [1] are informative elements in visual patterns. Especially, where curvilinear lines form a cross, there exist small, informative areas. Such crossing regions, or junctions in this paper, are of primary importance for character perception and recognition. The junctions can be categorized into different types such as L-, T-(or Y-) and X-junctions [2] according to the number of edges they connect, or the number of branches they have. Fig. 1 shows several artificial junctions. Given a combination of them, people can easily recognize the corresponding character. For example, given the junction set {(a), (b), (c)} in Fig. 1, the character '**A**' will pop up in our brain. Similarly, the combination of {(d), (e)} results in the character '**F**', putting the set {(d), (e), (f)} together will form the character '**E**', {(e), (g)} will form character '**H**', {(a), (h), (a)} will be character '**M**', and the different arrangement {(h), (a), (h)} will be character '**W**'. From this example we can conclude that junctions are important atomic elements for some English characters, and such atomic elements are shared between different characters. For instance, the junction (e) in Fig. 1 is shared between '**H**','**E**' and '**F**'.

Junctions are also prevalent in handwritten scripts for languages that use the Roman alphabets, some of which have inherent junctions. Since Chinese characters are composed of line-drawing strokes, they naturally contain many junction points [3]. Characters in other scripts probably also contain junctions, such as Arabic characters [4]. Junctions are often the consequence of overwritten curved traces of handwriting, or are the consequence of connecting strokes between characters, as illustrated in Fig. 2. Junctions reflect the local geometrical and structural features around the singular, salient points in handwritten texts. Hence, it is natural to use junctions in handwritten document analysis. Liu et al. [5] have shown the efficiency of using fork points on the skeletons for Chinese character recognition. It has also been used to extract features for Arabic handwriting recognition (see the survey [6]).

In this study, we take the assumption that junction shapes are not guaranteed to be identical for different writers. Furthermore, even the same characters written in different historical periods contain different junction shapes. Generally, the differences are from three aspects: first, the length of branches of the junctions are variant. Second, the angles between each branch are also different between different writers or in different periods. Third, the type of junctions might be changed. We believe that those differences are caused by individual writing habits which can be considered as one type of biometric feature. Such features can be used for writer identification and historical document dating [7].

Based on the observations that junctions are prevalent in handwritten documents and they are different when generated by different writers as mentioned above, we propose an approach to detect junctions in handwritten documents and evaluate the

* Corresponding author. Tel.:+31 503637410.
E-mail addresses: heshengxgd@gmail.com (S. He), M.A.Wiering@rug.n (M. Wiering), L.Schomaker@ai.rug.nl (L. Schomaker).

http://dx.doi.org/10.1016/j.patcog.2015.05.022
0031-3203/

Please cite this article as: S. He, et al., Junction detection in handwritten documents and its application to writer identification, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.05.022

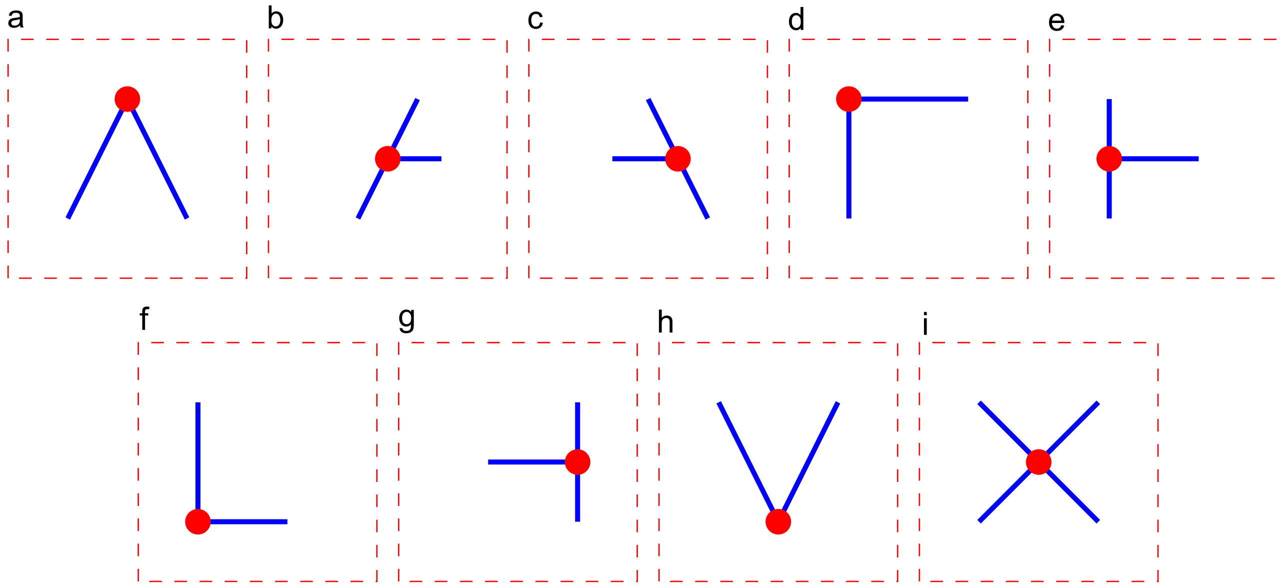

Fig. 1. Artificial junctions. The red dots are the center points of the junction, and the blue lines are the branches. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

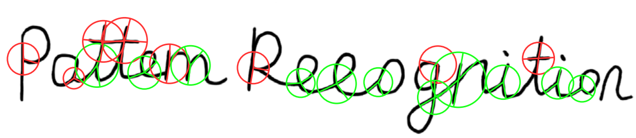

Fig. 2. Junctions in handwritten characters. The junctions marked in light (red) are within a character itself, and the green ones are from stroke touching. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

performances of using these detected junctions for writer identification. The contributions of this paper are summarized as follows: (1) we propose a simple yet effective method for junction detection in handwritten documents. (2) Our junction detector yields a junction feature, which can be considered as a mid-level feature representation. Furthermore, a new representation of handwritten documents is proposed based on the detected junctions, termed as *Junclets*, which are the primitive junctions of the document. The main advantage of the proposed method compared to junction detection in line-drawing images in [8] is that our proposed junction method can yield a junction feature in a natural manner. In addition, the benefit of the proposed *Junclets* compared to COnnected-COmponent COntour (CO^3) [9] and Fraglets [8] is that it does not rely on any segmentation or line detection which are challenging problems in document images, especially in historical documents where a connected component may span several lines due to touching ascenders and descenders.

This paper is organized as follows: we review the related work for junction detection and writer identification in Section 2. The third section presents the details of the proposed method. The writer identification method based on Junclets is described in Section 4. The experimental results are given in Section 5. We give the conclusion and future work in the last section.

2. Related work

2.1. Junction detection

In natural images, junctions are often detected based on template matching, contours, or gradient distributions. The template-based method for junction detection has been proposed in [2], in which the junction detection problem is formulated as one of finding the parameter values of the junctions that yield a junction which best approximates the template data by minimizing an energy function. The energy function has two parts: scale and location of junctions in images and the junction parameters, which are the number of wedges, wedge angles and wedge intensities. In [10], a novel junction detector is proposed by fitting the neighborhood around a point to a junction model, which segments the neighborhood into wedges by determining a set of radial edges. Two energy functions are used for radial segmentation, and junctions with the most energy are selected as junction candidates, followed by junction refinement to suppress the junctions on the straight edges. The contour-based approach [11] considers junctions as points at which two or more distinct contours intersect, and junctions are localized based on the combination of local and global contours using the *global probability of boundary* (gPb) [12]. Finally, a probability of junction operator is designed to compare the keypoints found by junctions to those detected by the Harris operator. Recently, Xia et al. [13] introduced a novel meaningful junction detection method based on the *a contrario* detection theory, called *a contrario* junction detection (ACJ). The strength of a junction is defined as the minimum of the branch strengths which is a measurement of the consistency of the gradient directions in an angular sector. Junctions are detected whose strength is greater than a threshold which is estimated by the *a contrario* approach. Compared to other methods, this approach requires fewer parameters, and is able to inhibit junctions in textured areas.

In [8], junctions are computed by searching for optimal meeting points of median lines in line-drawing images. There are three main steps in this method: (1) region of support determination by the linear least squares for 2-junctions, and crossing-points in skeleton lines for n-junctions, where $n=3, 4$. (2) Distorted zone construction by a circle centered at candidate junction points whose diameter is equal to the local line thickness. (3) Extracting the local topology which is a set of skeleton segments linked with a connected component distorted zone and junction optimization.

Su et al. [14] propose a method for junction detection in 2D images with linear structures. The Hessian information and correlation matrix measurements are combined to select the candidate junction points. The potential junction branches of candidate junctions are found, based on the idea that the linear

Please cite this article as: S. He, et al., Junction detection in handwritten documents and its application to writer identification, Pattern Recognition (2015), http://dx.doi.org/10.1016/j.patcog.2015.05.022