基于两方向动态时间规整的无分割手写汉字检测*

黄志敏[1]，姚舜奕[2]，熊玉洁[2]
(1. 公安部第三研究所，上海 200031；2. 华东师范大学 上海市多维度信息处理重点实验室，上海 200241)

摘 要： 中文文本布局复杂、汉字种类多、书写随意性大，因而手写汉字检测是一个很有挑战的问题。针对上述问题，提出了一种无分割的手写中文文档字符检测的方法。该方法用 SIFT 定位文本中候选关键点，然后基于关键点位置和待查询汉字大小来确定候选字符的位置，最后用两个方向动态时间规整(dynamic time warping，DTW)算法来筛选候选字符。实验结果表明，该方法能够在无须将文本分割为字符的情况下准确找到待查询的汉字，并且优于传统的基于 DTW 字符检测方法。

关键词： 手写汉字检测；无分割；SIFT；动态时间规整
中图分类号： TP391.4 **文献标志码：** A **文章编号：** 1001-3695(2016)11-3499-04
doi:10.3969/j.issn.1001-3695.2016.11.066

Two-directional dynamic time warping based Chinese handwritten segmentation-free word spotting

Huang Zhimin[1], Yao Shunyi[2], Xiong Yujie[2]
(1. *The Third Research Institute of Ministry of Public Security, Shanghai* 200031, *China*; 2. *Shanghai Key Laboratory of Multidimensional Information Processing, East China Normal University, Shanghai* 200241, *China*)

Abstract: Large variety of Chinese characters and handwriting styles and the complexity of Chinese handwritten documents layout lead a huge challenging for the Chinese handwriting word spotting. This paper proposed a segmentation-free word spotting method for Chinese handwritten documents. Firstly, the method used the SIFT keypoint detector to locate the candidate keypoints in document images. Then it determined the candidate character regions by the keypoints' locations and the size of query word image. At last, it applied the two-directional dynamic time warping (DTW) to refine the candidate regions. The experimental results show that the proposed method can detect the query word in the document images with high mean average precision and the two-directional DTW outperforms the traditional DTW.

Key words: Chinese handwritten word spotting; segmentation-free; SIFT; dynamic time warping

0 引言

手写汉字检测旨在手写文档中寻找到需要查询的某个字，即判定某一候选区域是否含有要检测的汉字。当今社会存在着大量有价值的手写文档，许多都作为图像的形式进行保存，然而对于手写文档的关键字查询仍是一个极具挑战性的问题。由于中文文本不仅布局复杂、书写随意性大，使得传统的单字 OCR 识别变得不可行。而字符检测不用识别出每一个汉字，可以利用图像检索中的技术来解决字符检测问题，因而字符检测引起越来越多的学者关注。

Manmatha 等人[1]首先提出了字符检测的概念，将一个传统 OCR 识别问题转变为一个验证性的问题。这种理念弱化了对文档检索前进行准确字符的分割，甚至不需要对文档进行任何切割的预处理。基于文档被分割的层次，字符检测方法可以分为基于字(字符)的、基于文本行的和无须分割的三个类别。

基于字分割字符检测技术意味着首先将文本图像切分为独立的字符或字，然后将这些字和待查询的字进行匹配并判定它们是否为相同的字符。Rath 等人[2]研究了特征的表示方法，他们发现使用上/下轮廓信息、笔画像素/背景转换数目和投影信息的组合作为特征结合 DTW 可以获得最好的结果。在文献[3]中，Zhang 等人使用基于轮廓特征的两维 DTW 方法进行中文书法字符检索。

因为在同一行中两个字符间的空间很小，难以准确分割，所以基于行分割字检测方法被提出。基于行的字检测方法首先将文本图像分割为文本行，然后将待查询的字符与文本行进行匹配判定文本行中是否含有该字符。基于文本行分割的字符检测方法避免字符分割的困难，而文本行分割相对字符分割要容易些。文献[4]提出一种隐马尔可夫模型(HMM)的方法进行字符检测；文献[5]提出了基于多层反馈网络(RNN)的字符检测方法；Huang 等人在文献[6]中提出了一种使用汉字上下文字符模型进行中文字符检测的方法，进而提高字符检测的判别能力。

基于字符和基于文本行的字符检测方法都依赖于对文本图像的预处理。预处理中的错误会很容易导致字符检测的失败，因此字符检测的趋势是采用基于无分割的方法。近年来一些基于无分割的字符检测方法被提出。基于无分割的字符检

收稿日期：2015-09-24 修回日期：2015-11-12 基金项目：国家科技支撑计划资助项目(2011BAK05B04)；上海市科委资助项目(14DZ2260800)

作者简介：黄志敏(1960-)，男，广东梅州人，副研究员，主要研究方向为图像处理及模式识别(mouse902@163.com)；姚舜奕(1990-)，男，江苏盐城人，硕士研究生，主要研究方向为图像处理及模式识别；熊玉洁(1989-)，男，湖南湘乡人，博士研究生，主要研究方向为模式识别与智能系统.

测方法直接用待查询的字符与整个文本图像进行匹配而无须将文本图像分割为行或者字符; 匹配待查询的字符和文本图像中特定的候选区域,进而判定这个候选区域是否为待查询的字符。在文献 [7] 中,作者提出基于滑动窗口的词袋模型来进行字符检测。在文献 [8] 中,Rothacker 等人采用 HMM 对字符的空间结构进行约束,以克服词袋模型空间信息丢失的缺陷。对每个候选区域特征,以列为单位编码,得到一个列序列特征,最后采用维特比算法进行解码并得到相似性。在文献 [9,10] 中 Zhang 等人应用热核描述符(heat kernel signature,HKS) 进行字符局部特征描述,以适应字符的非刚体变换。他们用 SIFT 检测器进行关键点定位,然后用 HKS 描述子来描述这些关键点。考虑到汉字分割的困难,本文方法采用了基于关键点定位的无分割汉字检测技术,并用字符轮廓和结构信息作为特征,与两方向的 DTW 算法结合进行相似度度量。首先用 SIFT 来获取候选关键点,本方法采用文献 [4] 中使用的局部特征作为每个候选区域的特征。尽管使用候选关键点缩减搜索空间,但是仍然有许多候选区域并非对应真正的待查询的字符,因此采用两方向的 DTW 对该候选区域进行进一步的精确匹配,得到每个候选区域是待查询字符的可能性。

1 候选区域定位

为了避免使用滑动窗口来遍历整个文本图像,本方法采用关键点定位的方法以减少候选区域数目。本文基于 SIFT 算法进行候选关键点的提取,进而确定需要精确匹配的区域。Lowe 于 1999 年提出 SIFT,并在 2004 年总结了不变性的特征检测方法,完善了基于尺度空间的特征匹配算法 SIFT。SIFT 特征点在图像旋转、尺度变换、仿射变换和光照变化条件下都有良好的不变性,因此使用 SIFT 可以检测出字符中具有不变性的稳定点,以便能够进行进一步的字符间的精确匹配。尽管 SIFT 可以检测出图像中的稳定点,但是这些点中仍有不少背景点,因此可以通过点出的特征点在其领域内的相对灰度值去除关键点中的背景点。设点 kp 为通过 SIFT 算法提取的关键点,其坐标为(kp_x,kp_y) ,若其灰度值低于 th_0 则为背景点。th_0 定义为

$$\frac{1}{2}\times\frac{\sum_{i=-knr}^{i=knr}\sum_{j=-knr}^{j=knr}F(kp_x+i,kp_y+j)}{(2knr+1)\times(2knr+1)}$$

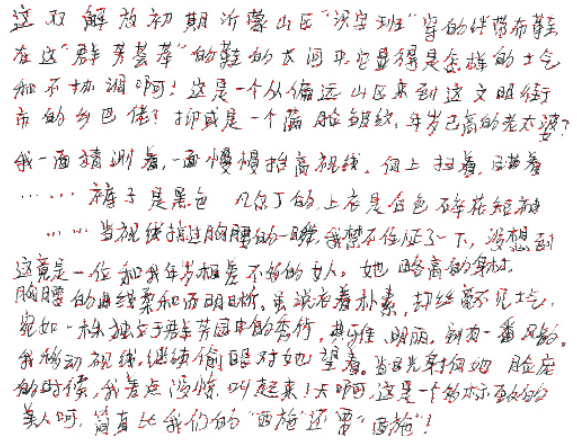
其中: $\pm 2knr$ 为 kp 的邻域,F 为像素点的灰度值函数。

去除在 SIFT 算法定位出的关键点中背景点中的关键点,这样既可以降低误匹配点数对,还能够提高精确匹配的运算速度。

图 1 展示了关键点定位的一个示例。图 1(a) 是要进行查询的汉字“的”,红色的点(以点为中心画出的横线,见电子版) 显示了由该方法得到的“的”的关键点; 图 1(b) 则表示要检索的文档中去除背景点的关键点的位置。图 1(a) 中红色的点位是由 SIFT 检测出的关键点,图 1(b) 中红色点是由 SIFT 检测出的至少与(a) 中一个点距离小于某一阈值的点。

对于文本图像中每一个关键点,本方法为其构造一个其与待查询字符位置相应的候选区域。设定待查询字符图像是 Q,其宽为 q_w,高为 q_h,要匹配的文本图像为 D。图 2 展示了一个关键点匹配对。图中线段显示了文本图像中的关键点 b 与待查询的字符中的关键点 a 是一对匹配点,并用矩形框框出了通过该匹配对确定的候选区域。设 D 中要匹配的点为 b,其在文本图像中的坐标为(d_{bx},d_{by}) ,在 Q 中选择与点 b 距离最小的点 a 作为其匹配的点,两者构成一组匹配对; 设点 a 的坐标为(q_{ax},q_{ay}) ,以待查询字符的大小作为候选区域的大小,则依据点 a 在 Q 中的相对位置和点 b 的位置,可以求出候选区域左上角坐标为($d_{bx}-q_{ax}$,$d_{by}-q_{ay}$) 。实验中,选择待查询的字符的宽和高分别作为候选区域的宽和高,则候选域的高度为 q_h,宽度为 q_w。

(a)待查询的“的”



(b)待匹配的文本

图 1 关键点定位示例

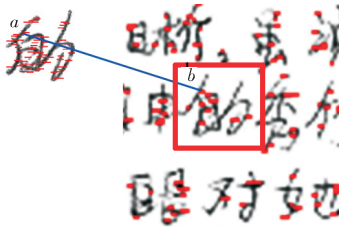


图 2 关键点匹配对

2 字符特征提取

a) 将候选区域的每一列像素视为一个单位,并对每一列像素进行特征提取,使得原始字符图像表示为一个垂直方向的特征序列; b) 将候选区域每一行像素视为一个单位,并对每一行像素进行特征提取,用一个水平方向的特征序列对原始图像进行再次表示。这样就可以将两维的图像数据转换为两个一维的特征数据序列,并通过两个方向的特征描述来保持字符的垂直和水平属性。

本方法依据文献 [2] 中的特征组合对字符进行特征提取,对每一列(行) 提取四维特征。因为要将这四个特征组合成一个特征序列,所以要对这四个特征进行归一化。下面以列像素为例:

a) 前景像素个数占该列总像素个数比重为 f_0;

b) 上边界轮廓距离 f_1 表示最高的前景像素点到区域上边界的距离与该列总像素的比重;

c) 下边界轮廓距离 f_2 表示最低的前景像素点到区域下边界的距离与该列总像素的比重;

d) 前景背景像素转换频率 f_3 表示从前景像素转换为背景像素和背景像素转换为前景像素的次数占该列总像素的比重。

通过这四个特征可以描绘出字符垂直方向的轮廓和结构信息。因为汉字的书写是从左往右、从上往下,所以,对字符进

行水平方向的轮廓和结构信息同样具有判别信息,因此对水平方向也进行同样的特征提取。最终每个待查询字符或字符候选区域得到两个方向的特征序列。

3　关键字匹配

本文使用两方向的 DTW 进行字符和候选域相似度的匹配,以两个方向的加权距离作为最终的距离。给定一个宽度为 M 的候选区域图像 C 和宽度为 N 的待查询字符图像 Q,由于候选区域的大小和待查询的字符大小一致,所以 $M=N$。以垂直方向特征序列为例,在第 2 章中已经为每一列提取 f_0、f_1、f_2、f_3 四个维度的特征,则 C 中第 i 列像素 C_i 与 Q 中第 j 列像素 Q_j 间的距离采用欧氏距离,定义为

$$d(c_i,q_j)=\sum_{k=0}^{3}(c_{if_k}-q_{jf_k})^2 \tag{1}$$

其中: c_{if_k} 是 C 中第 i 列的第 k 维特征, q_{jf_k} 是 Q 中第 j 列的第 k 维特征。

考虑到书写具有有序性和连续性,即匹配的两个序列之间不出现交叉匹配的两列和每一列都能找到与之匹配的列,本方法和文献 [2] 一样,采用连续性限制确保图像 C 的第 i 列和图像 Q 的第 j 列累加距离 $D(c_i,q_j)$ 由且仅由 $D(c_i,q_{j-1})$、$D(c_{i-1},q_j)$、$D(c_{i-1},q_{j-1})$ 和 $d(c_i,q_j)$ 共同决定,即

$$D(c_i,q_j)=\min\left\{\begin{matrix}D(c_i,q_{j-1})\\D(c_{i-1},q_j)\\D(c_{i-1},q_{j-1})\end{matrix}\right\}+d(c_i,q_j) \tag{2}$$

其中: i 和 j 都大于 1。

尽管与相同的字符之间存在书写差异,但是这种差异应该保持在局部的小范围内,因此要对全局路径进行约束,以保持局部结构不变性,同时能够加速问题的求解。限制匹配上的两列 C_i 和 Q_j 之间的空间距离要不大于 r 个像素。

$$\begin{aligned}&\|i-j\|\leqslant r\\&r=\lceil k\times \mathrm{seqLength}\rceil\end{aligned} \tag{3}$$

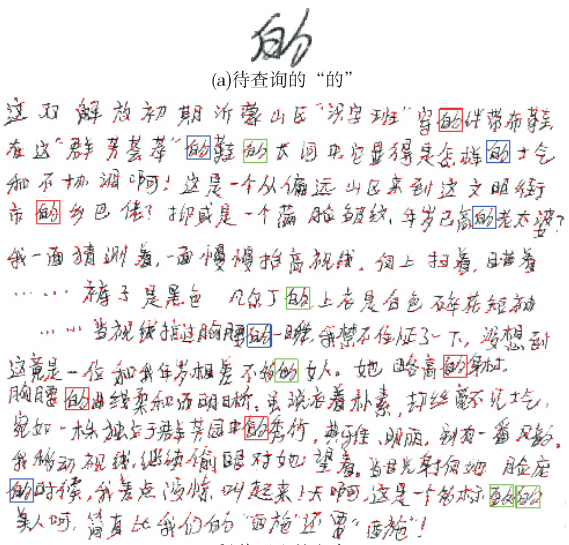
其中: k 是一个常量系数,seqLength 是特征序列的长度。

采用动态规划的方法可以加速求解水平方向的最终距离 $D(c_M,q_N)$。由于待查询的字符 Q 的特征序列和候选域的特征序列长度一致,所以无须对累加距离 $D(c_M,q_N)$ 按序列长度进行归一化。考虑到字符的长宽不一致,对最终的贡献也不一致,因此本文采用加权的两方向的 DTW 算法计算两个字符的最终距离,权重为序列本身的长度:

$$\mathrm{dist}(C,Q)=D(c_N,q_N)\times N+D(c_M,q_M)\times M \tag{4}$$

最后按照每个候选域 C 和待查询字符 Q 距离 $\mathrm{dist}(C,Q)$ 升序排列,得到候选区域列表。由于每个字符都会有若干的关键点,所以需要对重叠的候选域进行消除。按照得到的候选域列表,对于列表中的每一个候选域,若在列表中存在排序在其之前,且重叠面积比率($\mathrm{overlapRatio}=\dfrac{\mathrm{area}Q\cap\mathrm{area}C}{qw\times qh}$, $\mathrm{area}Q\cap\mathrm{area}C$ 为 C 和 Q 的重叠面积,显然 $\mathrm{area}Q$ 和 $\mathrm{area}C$ 面积均为 $qw\times qh$) 超过阈值 overlap1,则予以消除。最后得到列表即为最终检测结果。

图 3 显示了用本文方法进行字符检测结果的一个样例。图中,(a) 是待查询的字符“的”; (b) 是待查询的文档图像的可视化检测结果。红色的点是找到的关键点,红色矩形框内是返回列表中的前 5 个,蓝色矩形框是返回列表中的第 6 ~ 10 个,绿色框中的是列表中的第 11 ~ 15 个(见电子版)。



(a)待查询的“的”

(b)待匹配的文本

图 3　本文方法字符检测结果的样例

4　实验分析

本文实验在 CASIA-HWDB 2.1 中文测试集上进行。该数据含有 60 个书写者抄写的 300 页文本,每个书写者抄写 5 页;所有的图像以 300 DPI 的分辨率扫描,以 8 位灰度级图像存储。该数据库含有每一个汉字的位置、大小信息。对每一个文本页面,使用出现 5 次及以上的汉字作为待查询的字符,并在该字符出现的页面进行检索。对于基于字符分割的方法,使用真值信息对文档进行分割,用分割好的字符作为检测对象;对于无分割的方法,直接以整个文档作为检测对象。

本文采用主流评价指标,即平均正确率均值[13]。正确率指的是返回的结果中与待查询字符相同字符所占比。假设有 N_q 个待查询的字符,当第 i 个待查询字符共有 Rel_i 个相关字符,则若检索出 i_k 个字符时,有 Rel_{ik} 个字符是正确的,则此时的正确率为

$$p(i_k)=\mathrm{Rel}_{ik}/i_k \tag{5}$$

计算平均正确率的时候要先求出每个位置上的正确率,若该位置返回的结果相关,计算该位置的正确率,若不相关,正确率置为 0;然后对所有的相关字符位置的正确率再求平均。因此,平均正确率定义为

$$AP_i=\frac{\sum_{k=1}^{\mathrm{Reti}}(p(i_k)\times\mathrm{rel}(i_k))}{\mathrm{rel}_i} \tag{6}$$

其中: $\mathrm{rel}(i_k)=\begin{cases}0 & \text{召回的第 } k \text{ 个字符不是待查询的字符}\\1 & \text{召回的第 } k \text{ 个字符是待查询的字符}\end{cases}$

Reti 为第 i 个待检索字符返回的候选域个数。平均正确率均值是将每个待查询的字符的平均正确率求均值即

$$\mathrm{mAP}=\frac{\sum_{i=1}^{N_q}AP_i}{N_q} \tag{7}$$

为了在无分割场景下采用该评价指标,需要定义如何判定检索到的区域是有效的。在实验中,对于给定的待查询的字符图像,一个返回的区域如果与待查询的汉字相同字符的真值重叠面积(定义同 $\mathrm{overlap}_1$) 超过阈值 $\mathrm{overlap}_2$,则认为该返回区域是正确的。直到文档中所有的与待查询的汉字相同的汉字被检索到为止。

实验中参数设置如下: $\mathrm{knr}=5$, $\mathrm{overlap}_1=0.2$, $\mathrm{overlap}_2=0.5$。

表 1 的前两行显示了使用的两方向 DTW 和文献 [2] 中 DTW 在基于分割情况下在 CASIA_2.1 数据库中的检测结果。

在基于分割的两个实验中，待查询字符也是作为候选字符的。使用传统的 DTW 方法的 mAP 是 66.61%（待检索字符中不含原待查询字符时为 60.41%），而两方向 DTW 的 mAP 为 72.40%（待检索字符中不含原待查询字符时为 65.31%）。从表中可以看出，使用两个方向对字符进行描述使得 mAP 提升了 5.79%，说明同时对字符进行水平和垂直方向的特征提取并用加权的 DTW 算法进行相似度度量是有效的。表 1 中第三行显示采用改进的 SIFT 算法进行关键点定位和两方向 DTW 相结合在 CASIA_2.1 进行字符检测中的结果。由于采用无分割，实验结果必然低于基于分割的两方向 DTW 的实验结果，但是实验结果仍然高于原始的基于分割的 DTW [2]，说明了该方法的有效性。

表 1 三种算法在 CASIA 库上的实验结果

方法	是否将文档分割为字符	mAP
DTW [2]	是	66.61%（60.41%）
两方向 DTW	是	72.40%（65.31%）
SIFT + 两方向 DTW	否	67.29%

图 4 显示了两方向 DTW 在无分割和分割情况下的实验结果。

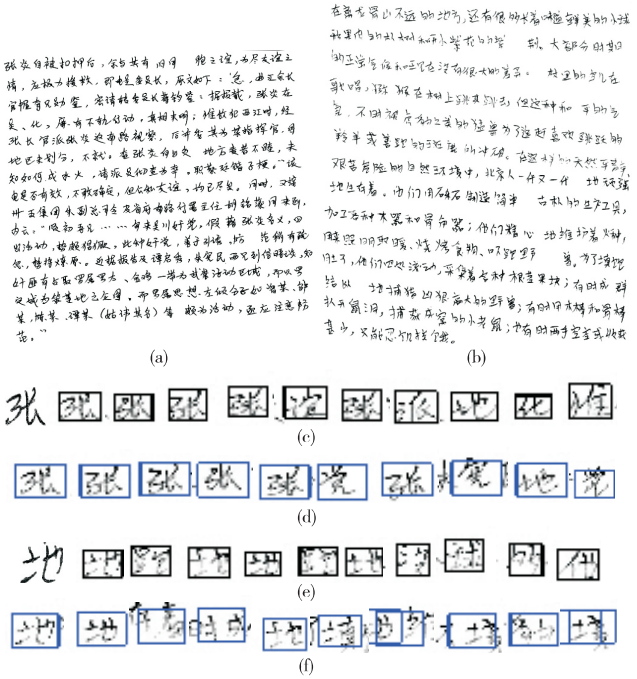


图 4 两方向 DTW 无分割和分割情况实验结果

图 4 中，(a) 和 (b) 是两张文本图像。图 4(c) 左边是一个在 (a) 中待查询的汉字“张”，在 (a) 中“张”出现了 6 次；(c) 右边是用两方向 DTW 算法在分割好的情况下检索出的前 10 个结果，仅有 5 个“张”被检索出。图 4(d) 是用本文提出的无分割的两方向 DTW 的检索结果，在前 7 个候选域中就检索出全部的 6 个“张”字。在图 4(a) 中有 27 个待查询的关键字，基于分割的平均正确率均值为 76.36%，而无分割的平均正确率均值为 88.30%。这说明，通过 SIFT 关键点对候选域进行初步筛选可以排除一些不能与待查询字符相应位置关键点匹配的字符候选域。尽管用分割好的字符真值去匹配待查询的字符，不存在分割差错的问题，但是存在着一些用轮廓特征描述相似却并不是待查询字符的字符，而第一步的 SIFT 关键点的匹配就能很好地解决这个问题；(e) 左边是图 4(b) 中待查询的汉字“地”，“地”在图中出现了 5 次；(e) 右边是用两方向 DTW 算法在分割好的情况下检索出前 10 个结果，有 4 个“地”被检索出。图 4(f) 是用本文提出的无分割的两方向 DTW 的检索结果，在返回结果的前 10 个中仅有 3 个“地”。在图 4(b) 中有 41 个“地”字，基于分割的两方向 DTW 的 mAP 的值为 78.22%，高于无分割的 63.36%。结合表 1 的第 2 行和第 3 行，说明尽管无分割的方法在某些方面可以弥补基于分割的不足，但总体上依旧没有取得基于分割的方法的效果。

5 结束语

本文提出了一种 SIFT 特征点定位与两方向 DTW 相结合的手写中文字符检测方法。该方法通过 SIFT 关键点的粗定位和两方向 DTW 的精确匹配，避免了整页中文字符检索时精确分割的困难，在无分割状态下也取得了较好的 mAP 值。但是该方法对于字符大小变化不具有鲁棒性，在候选区域大小确定方面仍然需要进一步研究。

参考文献：

[1] Manmatha R, Han C, Riseman E M. Word spotting: a new approach to indexing handwriting [C] // Proc of IEEE Conference on Computer Vision and Pattern Recognition. 1996: 631-637.

[2] Rath T M, Manmatha R. Features for word spotting in historical manuscripts [C] // Proc of IEEE Conference on Document Analysis and Recognition. 2003: 218-222.

[3] Zhang Xiafen, Zhuang Yueting. Dynamic time warping for Chinese calligraphic character matching and recognizing [J]. Pattern Recognition Letters, 2012, 33(16): 2262-2269.

[4] Fischer A, Keller A, Frinken V, *et al*. HMM-based word spotting in handwritten documents using subword models [C] // Proc of the 20th International Conference on Pattern Recognition. Washington DC: IEEE Computer Society, 2010: 3416-3419.

[5] Frinken V, Fischer A, Manmatha R, *et al*. A novel word spotting method based on recurrent neural networks [J]. IEEE Trans on Pattern Analysis and Machine Intelligence, 2012, 34(2): 211-224.

[6] Huang Liang, Yin Fei, Chen Qinghu, *et al*. Keyword spotting in unconstrained handwritten Chinese documents using contextual word model [J]. Image and Vision Computing, 2013, 31(12): 958-968.

[7] Rusinol M, Aldavert D, Toledo R, *et al*. Browsing heterogeneous document collections by a segmentation-free word spotting method [C] // Proc of IEEE Conference on Document Analysis and Recognition. 2011: 63-67.

[8] Rothacker L, Rusinol M, Fink G A. Bag-of-features HMMs for segmentation-free word spotting in handwritten documents [C] // Proc of IEEE Conference on Document Analysis and Recognition. 2013: 1305-1309.

[9] Zhang Xi, Tan C L. Handwritten word image matching based on heat kernel signature [C] // Computer Analysis of Images and Patterns. Berlin: Springer, 2013: 42-49.

[10] Zhang Xi, Tan C L. Segmentation-free keyword spotting for handwritten documents based on heat kernel signature [C] // Proc of IEEE Conference on Document Analysis and Recognition. 2013: 827-831.

[11] Lowe D G. Distinctive image features from scale-invariant keypoints [J]. International Journal of Computer Vision, 2004, 60(2): 91-110.

[12] Llados J, Rusinol M, Fornes A, *et al*. On the influence of word representations for handwritten word spotting in historical document [J]. International Journal of Pattern Recognition and Artificial Intelligence, 2012, 26(5): 53-61.