基于广义 Hough 变换的手写汉字文档关键词提取

陈 睿,唐 雁
(西南大学 计算机与信息科学学院,重庆 400715)

摘 要:提出了一种基于广义 Hough 变换的手写汉字文档关键词提取技术。对于待提取的手写文档图像,采用字符像素逐点匹配和投票的方式进行广义 Hough 变换,在参数空间中定位出手写关键词图像的位置。本技术对传统的广义 Hough 变换进行了修改,突破了形状匹配需要完整轮廓信息的局限,简化了局部特征的计算,对手写汉字文档图像中具有局部形变、部分旋转和缩放的手写关键词能够有效提取。对于提取的相同关键词建立训练集,用签名识别的方法对书写者建模,能够达到书写者身份鉴别的目的。

关键词:广义 Hough 变换;参考表;匹配与投票;笔迹鉴别
中图分类号:TP391.1;TP391.4 **文献标识码**:A **文章编号**:1674-7720(2013)06-0075-04

Keyword extraction from handwritten Chinese document image using improved generalized Hough transform

Chen Rui, Tang Yan
(School of Computer and Information Science, Southwest University, Chongqing 400715, China)

Abstract: In this paper, we present a keyword extraction methodology from handwritten Chinese document image using improved generalized Hough transform. In the voting phase, the features of each character pixel of handwritten character image are compared with the ones in the reference table. A vote is made in the parameter space when the features match each other and the location of the keyword is found by the cluster of the votes. In our method, the generalized Hough transform is modified so that the complete contour information of the shape is not necessary, and the computation of the local features is simplified. The handwritten keywords with local shape variation, slight rotation and zooming can be extracted from handwritten Chinese document image effectively. The same keywords from a same document can be used as training samples, which are used to model writers based on the signature verification methods to accomplish the goal of writer identification.

Key words: generalized Hough transform; reference table; match and vote; writer identification

从手写汉字文档图像中提取关键词是一项重要的任务,它能够作为文本相关笔迹鉴别的预处理步骤,其实质是从数字图像中识别和定位目标物体。目标物体识别一般通过将目标物体模型的特征与数字图像中检测到的实体的特征进行匹配的方式实现。学术界提出了大量基于整体和局部方法的能够抗旋转和位移的目标物体识别和定位技术[1-3]。整体方法基于整体特征,如边界和区域。这些方法包括不变矩、Fourier 描述子和互相关。局部方法使用局部特征,包括关键点(Dominant Point)、局部最大曲率和多边形近似。

Hough 变换是检测直线、圆和其他解析曲线的有效方法,对它的研究一直非常活跃[4-5]。Hough 变换在目标识别中有效的原因就在于其把目标的识别转化为对在参数空间投票多少的判定。最初的 Hough 变换只能用来检测形状有解析表达式的目标。为了检测形状任意的、没有解析表达式的目标,BALLARD[5]提出了广义 Hough 变换(GHT)算法。GHT 的实质也是让轮廓边界点进行投票,只是投票地点不是由表达式的参数确定,而是定义一个参考点和一套投票机制,通过投票的集中程度来判定目标是否存在。GHT 解决了任意形状边界目标的识别,但它需要目标物体的完整轮廓边界点信息。

对于手写汉字文档图像中的关键词提取问题,也可以看作是一类特殊的目标识别问题,其特殊之处在于手写汉字字符和单词没有一个完整的轮廓,即使找到汉字

的最小外包轮廓也不能描述汉字的大量内部形变结构特性。然而，基于点对点匹配投票思想的广义 Hough 变换同样适用于手写汉字的目标识别问题。这就需要对传统的基于轮廓边界点的投票机制进行改进，改为基于每个字符像素的投票机制，同时局部特征和匹配策略也要做相应的修改。

参考文献[6]提出了一种改进的 GHT 变换算法，能有效地对具有旋转和部分遮挡的物体进行识别。它采用一个两阶段的 GHT 策略，将三维参数空间(旋转角度 θ、横轴坐标 x 和纵轴坐标 y)转化为一个一维参数空间(θ 空间)和一个二维参数空间($x-y$ 空间)进行投票，将 θ 空间中投票的结果用作 $x-y$ 空间投票的条件，降低了运算量，提高了识别精度。但是该算法有很大的局限性：(1)它需要目标的完整轮廓，这对汉字字符和单词不适用；(2)它假定场景图像中只包含一个待识别的目标物体，否则 $x-y$ 空间的投票无法进行，这就无法实现对场景中包含的多个目标物体进行匹配定位。

1 改进的广义 Hough 变换算法

针对手写汉字的特点，对广义 Hough 变换做了改进。首先将模板图像和文档图像进行二值化和骨架化，对模板图像中的每个字符像素提取 5 个局部特征作为参考表的内容，分别是分叉数、曲率、法线方向、重心夹角和重心距离；再用文档图像中的每个字符像素与参考表中的每一项进行匹配投票。下面依次介绍这 5 种局部特征。

1.1 分叉数

分叉数是与字符像素相连接的分支数目。传统的分叉数计算方法是对字符像素的 8 个邻接点计算顺时针(或逆时针)方向上像素值由 1 到 0 的跳变数目。由于手写汉字骨架图中形变大、分支多，断裂和冗余连接普遍存在，采用传统的分叉数计算方法不能充分体现字符的分叉程度。针对这个问题，提出了一种改进的计算分叉数的方法。

设经过二值化和骨架化的模板图像为 R，$R=\{r_{ij}|i=1,\cdots,m;j=1,\cdots,n\}$，其中 m 和 n 分别为模板图像的行数和列数，字符像素定义为 R 中值为 1 的点。对于当前考虑的字符像素 r_{ij}，其半径为 t 的邻域定义为：$A(r_{ij})=\{r_{xy}|x=i-t,i-t+1,\cdots,i+t;y=j-t,j-t+1,\cdots,j+t\}$。下面给出字符像素 r_{ij} 在半径为 t 的邻域中的分叉数的计算。

算法 1：$y=\text{ForkNum}(R,i,j,t)$

(1)置 $y=0$，$A(r_{ij})=\{r_{xy}|x=i-t,i-t+1,\cdots,i+t;y=j-t,j-t+1,\cdots,j+t\}$；

(2)从 $A(r_{ij})$ 的左上角开始沿顺时针方向将 $A(r_{ij})$ 的边界点依次放入点集 P，$P(r_{ij})=A_{1,1\cdots 2t+1}\cup A_{2\cdots 2t+1,2t+1}\cup A_{2t+1,2t\cdots 1}\cup A_{2t\cdots 2,1}=\{p_i|i=1,\cdots,8t\}$；

(3)依次从 P 中读取每个像点 $p_i(i=1,\cdots,8t)$ 的值，当 $p_i=1$ 且 $p_{i+1}=0$ 时，置 $y=y+1$；当 $i=8t$ 时，将 p_1 作为 p_{8t+1} 考虑；

(4)输出分叉数 y，算法结束。

如图 1 所示，黑色方块表示字符像素，其中包含白色圆点的黑色方块表示当前考虑的字符像素，方形区域代表当前考虑的二维邻域，其大小为 11×11(半径为 5)。图 1(a)~图 1(f)分别代表分叉数从 1~6 的情况。如果采用传统的分叉数计算方法，这 6 种情况下的分叉数都为 2。从图中可以看出，采用本文方法得到的分叉数能够较准确地反映出字符的局部分叉程度。邻域大小的选择很重要，当邻域大小为 3×3 时，本方法等价于传统的分叉数计算方法。当邻域过大时，由于笔画断裂的原因，得到的分叉数不能准确反映字符的局部分叉程度。本实验中，采用字符大小的 20%左右的邻域计算效果最好。例如，对于平均字符大小为 25×25 的汉字文档图像，采用 11×11 的邻域进行计算。

1.2 曲率

曲率是表示曲线弯曲程度的量。平面曲线的曲率就是针对曲线上某个点的切线方向角对弧长的转动率，通过微分来定义，表明曲线偏离直线的程度。曲率越大，表示曲线的弯曲程度越大。对于汉字字符图像的每个像素，其曲率反映了笔画的弯曲特性。对于给定大小的邻域，采用边界点到中心点连线的向量夹角来表示曲率，其范围是[0，180°)。

算法 2：$y=\text{Curvature}(R,i,j,t)$

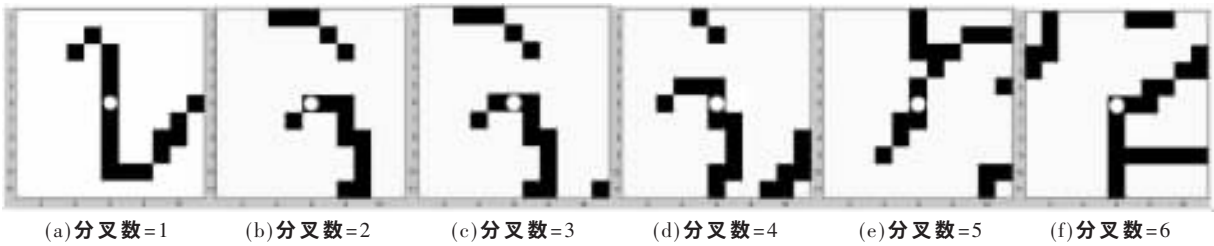
(1)置 $y=0$，用算法 1 的方法计算 $A(r_{ij})$ 和 $P(r_{ij})$；

(2)依次从 P 中读取每个像点 $p_i(i=1,\cdots,4t)$ 的值，将其中连续值为 1 的点划分到集合 $S_j(j=1,\cdots z,z\leqslant 2t)$ 中；如果 $p_1=1$ 且 $p_{4t}=1$，则置 $S_1=S_1\cup S_z$，$z=z-1$；

(3)依次计算 $S_j(j=1,\cdots z,z\leqslant 2t)$ 中每个点集的重心位置 c_j，并将其加入到集合 C 中，得到 $C=\{c_k|k=1,\cdots,z\}$；

(4)依次取出 c_k 和 c_{k+1}，以 c_k 为起点、r_{ij} 为终点建立向量 v_1，以 r_{ij} 为起点、c_{k+1} 为终点建立向量 v_2，计算 v_1 和 v_2 的夹角 θ，置 $y=y+\theta$；当 $k=z$ 时，将 c_1 作为 c_{z+1} 考虑；

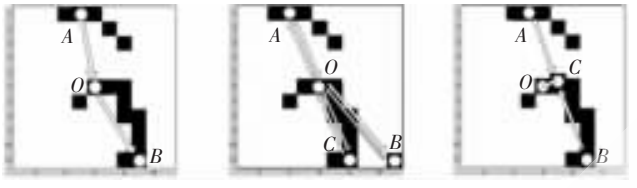
(5)置 $y=y/z$，算法结束。



(a)分叉数=1 (b)分叉数=2 (c)分叉数=3 (d)分叉数=4 (e)分叉数=5 (f)分叉数=6

图 1 各种情况下的分叉数

如图 2(a)所示，白色圆点 O 代表当前考虑的字符像素，白色圆点 A 和 B 代表两个边界点集的中心点，像点 O 的曲率定义为向量 AO 和 OB 的夹角。对于图 2(a)中的像点 O，其曲率为 15.26。对于分叉数大于 2 的字符像素，其曲率为相邻边界点集的中心点两两组合后与字符像素构成向量的夹角的平均值。如图 2(b)所示，白色圆点 O 代表当前考虑的字符像素，白色圆点 A、B 和 C 代表三个边界点集的中心点，像点 O 的曲率定义为向量 AO 和 OB 的夹角、BO 和 OC 的夹角及 CO 和 OA 的夹角的平均值，即 23.20°、151.70°和 5.10°的平均值 117.83°。



(a)双边界点时的曲率 (b)三边界点时的曲率 (c)法线方向的计算

图 2 曲率和法线方向的计算

1.3 法线方向

曲线的法线是垂直于曲线上一点的切线的直线。对于具有相同曲率的笔画段，若其旋转角度不一样，则表明两字符具有较大差异。这个差异可以通过字符像素的法线方向来体现。对于给定大小的邻域，法线方向为边界点连线中点与当前考虑像素的连线构成的向量的方向角，即与横轴按逆时针方向所成夹角，范围是[0，360°)。

算法 3：y=NormalAngle(R，i，j，t)

(1)置 y=0，用算法 1 和算法 2 计算 $A(r_{ij})$、$P(r_{ij})$、集合 S 和集合 $C=\{c_k|k=1,\cdots,z\}$；

(2)依次取出 c_k 和 c_{k+1}，计算 c_k 和 c_{k+1} 的连线中点 m_k，以 m_k 为起点、r_{ij} 为终点建立向量 v，计算 v 与横轴夹角 θ，$\theta\in[0,180°)$；若 v 在三、四象限，置 $\theta=360°-\theta$。置 $y=y+\theta$；当 $k=z$ 时，将 c_1 作为 c_{z+1} 考虑；

(3)置 $y=y/z$，算法结束。

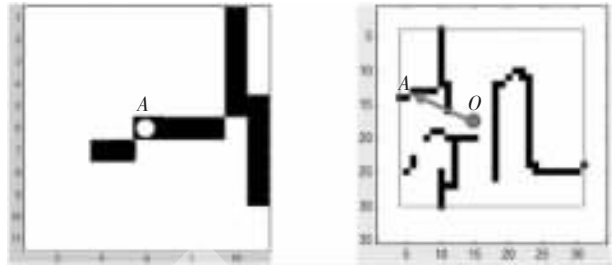
如图 2(c)所示，像点 O 的法线方向定义为线段 AB 的中点 C 与 O 所构成的向量 CO 的方向角，其值为 198.94°。对于图 2(b)中像点 O 的法线方向，由线段 AB、BC、CA 的中点为起点，O 为终点的 3 个向量的方向角的平均值构成，即 213.40°、120.85°和 19.25°的平均值为 117.83°。

1.4 重心夹角

对于模板图像中的每个字符像素，它与模板图像重心连线构成的向量的方向角定义为重心夹角，值域为[0，360°)，记为 GravityAngle(r_{ij})。如图 3 所示，字符像素 A 的重心夹角定义为模板图像重心 O 到 A 构成的向量 OA 的方向角，其值为 151.09°。

1.5 重心距离

重心距离定义为字符像素到模板重心的连线的长度，记为 GravityDist(r_{ij})。如图 3 所示，字符像素 A 的重心距离即 OA 的长度，其值为 11.16。



(a)当前字符像素 (b)重心夹角和距离

图 3 重心夹角和重心距离

1.6 参考表

对于给定模板图像的每个字符像素，计算上述 5 个局部特征，它们构成了参考表 RTable 中的一行。在后续的匹配过程中，对文档图像中的每个字符像素也提取这 5 个局部特征，然后与参考表中的每一行进行匹配，如果匹配度大于某个阈值，则在参数空间中进行投票，最终在参数空间中形成局部峰值，即定位到的模板图像的位置。

对于图 4(a)中的模板图像“计算机”的每个字符像素，计算它的 5 个局部特征，得到图 4(b)中的参考表，其中每一行对应一个字符像素的 5 个局部特征。

(a)模板图像

(b)参考表

图 4 参考表示例

1.7 投票算法

设经过二值化和骨架化的模板图像和待检索的文档图像分别为 R 和 I，其中 $R=\{r_{ij}|i=1,\cdots,m;j=1,\cdots,n\}$，$I=\{i_{xy}|x=1,\cdots,k;y=1,\cdots,l\}$，$k>m$，$l>n$；参考表为 RTable($rt$)$=\{rt_s|s=1,\cdots,z\}$；邻域半径为 t；角度差阈值为 ta；参数空间图像为 S，$S=\{s_{xy}|x=1,\cdots,k;y=1,\cdots,l\}$，$k>m$，$l>n$；匹配度阈值为 tm；匹配度定义为 s_{xy} 与 R 中字符像素个数的比值，投票算法如下：

算法 4：Vote(R，I，Rtable，ta，tm)

(1)依次读取 I 中的每个字符像素 i_{xy}($i_{xy}\neq 0$)，在给定邻域 t 中计算它的 5 个局部特征 ForkNum(i_{xy})、Curvature(i_{xy})、NormalAngle(i_{xy})、GravityAngle(i_{xy})和 GravityDist(i_{xy})；

(2)对于当前 i_{xy}，依次读取参考表的每一行的 5 个特征：RTable(rt_s，1)、RTable(rt_s，2)… RTable(rt_s，5)。如果 RTable(rt_s，1)≠ForkNum(i_{xy})，取下一个 i_{xy}，重复步骤(2)；若 abs(RTable(rt_s，2)−Curvature(i_{xy}))>ta||abs(RTable(rt_s，3)−NormalAngle(i_{xy}))>ta，取下一个 i_{xy}，重复步骤(2)；

(3)用 RTable(rt_s，4)和 RTable(rt_s，5)计算模板图像中以字符像素 rt_s 为起点、重心点 G 为终点的向量 v，做向量运算 $i_{xy}+v$ 得到参考图像 S 中的投票点 s_{xy}，置 $s_{xy}=s_{xy}+1$。取下一个 i_{xy}，转步骤(2)；

(4)统计 S 中所有 s_{xy}，s_{xy}>tm，将其作为匹配点画出外包矩形。

2 实验

本文对 96 名学生建立了一个手写文档数据库，每名学生都书写了一段相同的文字。如图 5 所示。图 5(a)中的关键词取自图 5(b)中左上角的“计算机”，它们都经过了二值化和骨架化处理。在图 5(b)中，标记投票数超过给定阈值的点，并对以该点为重心的关键词还原其大小，用方框标出。在这个例子中，角度差阈值取为 30，邻域半径取为 5，匹配度阈值取为 0.07。从图中可以看出，匹配点和方框集中的地方就是关键词出现的地方，算法对关键词“计算机”的提取非常准确。

(a)关键词模板　(b)图(a)对应的文档图像　(c)关键词模板

图 5　用改进的广义 Hough 变换进行关键词定位

图 6 为对图 5(c)中模板图像进行匹配定位的结果，其中图 5(c)的关键词取自图 6 中右下角的“计算机”。在这个例子中，角度差阈值取为 30，邻域半径取为 5，匹配度阈值取为 0.055。从图中可以看出，算法找出了全部 9 个“计算机”的准确位置，但是在图像的右上部分出现了一个误匹配，将“处理图”和“计算机”匹配，其原因为算法在计算字符像素的局部特征时只考虑了该像素邻域中的字符形状特征，不能描述模板图像整体的特征，加上手写汉字骨架图中形变大、分支多，断裂和冗余连接大量存在，影响了算法定位的准确性。针对这个问题，对匹配定位的结果，即提取出的关键词图像，与原模板图像进行相关运算，去掉相关度低于给定阈值的匹配结果。经过实验统计，给定合适的阈值(其范围是[0.5，1.2])，在 96 篇手写文档中“计算机”整词匹配的正确率能够达到 85%。

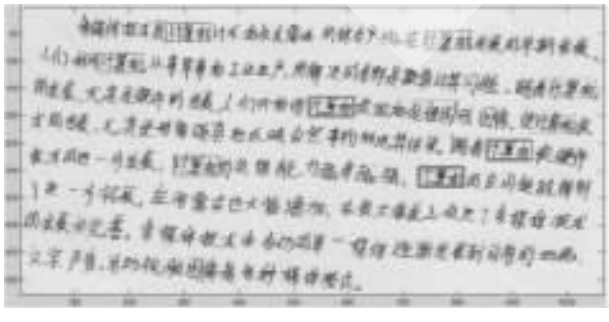


图 6　图 5(c)对应的文档图像

本文提出了一种基于广义 Hough 变换的手写汉字文档关键词提取技术。本技术使用具有形变的手写关键词图像作模板，对该模板的每个字符像素抽取局部特征建立参考表。对于待提取的手写文档图像，采用字符像素逐点匹配和投票的方式进行广义 Hough 变换，在参数空间中定位出手写关键词图像的位置。本算法对手写汉字文档图像中具有局部形变、部分旋转和缩放的手写关键词能够有效提取。后续工作如下：对文档图像进行字符分割，对单个字符进行匹配定位。在此基础上，对超过一定数目且位置相邻的字符进行组合得到整词，再进行整词匹配提取关键词，建立训练集以进行笔迹鉴别。

参考文献

[1] LENG W Y，SHAMSUDDIN S M.Writer identification for Chinese handwriting[J].Int.J.Advance.Soft Comput.Appl，2011，2(2)：160−173.

[2] Tan Jun，Lai Jianhuang，Wang Changdong.A stroke shape and structure based approach for off−line Chinese handwriting identification[J].I.J.Intelligent Systems and Applications，2011，3(2)：1−8.

[3] MOKHTARIAN F，MACKWORTH A K.Scale−based description and recognition of planar curves and two−dimensional shapes[J].IEEE Trans Pattern Anal Mach Intell，1986，8(1)：34−43.

[4] HAN M H，JANG D.The use of maximum curvature points for the recognition of partially occluded objects[J].Pattern Recognition Pattern Recognition，1990，23(1−2)：21−33.

[5] BALLARD D H.Generalizing the Hough transform to detect arbitrary shapes[J].Pattern Recognition，1981，13(2)：111−122.

[6] TSAI D M.An improved generalized Hough transform overlapping objects[J].Image and Vision Computing，1997，15(12)：877−888.

(收稿日期：2012−11−16)

作者简介：

陈睿，男，1979 年生，博士研究生，讲师，主要研究方向：图像处理，模式识别。

唐雁，女，1965 年生，博士，教授，主要研究方向：图像处理，模式识别。