一个基于扫描串的统一整体矢量化算法*

李 宾 谭建荣 彭群生

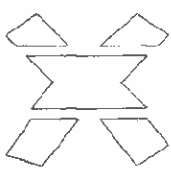
(浙江大学 CAD&CG 国家重点实验室 杭州 310027)

摘要 工程图纸扫描识别和字符笔画提取一直是模式识别中的难点问题.为克服细化算法逐象素处理的局部特性,已提出了一些整体算法.基于图段的整体识别算法突破了细化算法在象素层次做局部操作的局限,力图抓住图象的宏观特征进行整体识别;但它在处理小线素、曲线及字符图象时仍有较大局限性.本文采用扫描串(行程编码)作为基本处理单元,动态扫描整个图象数据,利用预先建立的信息头指导后续跟踪,得到了更精确的矢量化结果.本文提出的算法可适用于各种不同工程图纸的处理,用来提取汉字笔画也得到了满意的结果.

关键词 工程图纸扫描识别,字符笔画提取,整体识别,扫描串.

中图法分类号 TP391

在模式识别的应用方面,由于工程图纸扫描识别(矢量化)与光学字符识别 OCR(optical character recognition)在工程设计与文档自动处理上有重要的实用价值.因此在学术界及开发部门引起了广泛注意.截止到目前,已提出了众多不同的算法,其中使用得最广泛的是细化算法.Suen 等人[1]调查了 100 多种细化算法,并把它们分为两类:基于迭代删除象素的算法和基于非象素的算法.基于迭代删除象素的方法放置一个 $n\times n$ 的窗口到图象上,利用一个查找表来决定是否保留窗口中心的黑象素.它的优点是简单、易于并行实施;但缺点也是显而易见的,计算量大,对噪声太敏感,节点处有畸变,因此需要做预处理.



两条相交直线包含了5个图段

图1

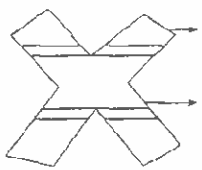


图2

按照 Pavlids[2]细化方法,对噪音的敏感来源于局部判别,将来的目标在于使用整体信息以减少各种可能性.现在已有不少文章[3~6]探讨利用图象的整体信息进行矢量化或特征抽取.文献[4]在国内率先提出了工程图扫描图象的整体识别思路,力图从宏观上进行整体识别.它的核心思想是:关注整个图象的拓扑结构,用尺寸约束、校正图形,最长线素优先识别,实行动态采样,智能剔除坏点.相比于传统的细化算法,整体识别的思路无疑深化了对矢量化问题的认识,是对工程图纸扫描识别的巨大突破.但在进一步的研究实践中发现,识别过程中局部与整体的关系是辨证统一的;仅仅停留在某一局部层次进行图象识别将增加判断的难度与识别的错误率;而上升到整体,利用宏观特征进行判别必然要落实到图象的局部信息.由此看来,整体与局部之间的合理划分并建立相应的图象数据模型将是矢量化中的一个关键问题.文献[4]中图象处理的核心单元为连通图段,它由一串上下相互单连通的行程编码组成,包含了较大范围内的图象信息(如图 1 所示);整体识别算法中直线、圆弧的识别过程就是关于图段的分裂、合并与匹配过程.由于图段的建立完全由按水平方向扫描图象时行程编码的分裂与合并关系所决定,因此有一定的盲目性,在处理复杂图纸过程中会带来相当的麻烦与错误.尽管基于图段的算法在识别长直线与圆弧时有较高的效率,但是图象中还存在大量的字符、小特征及小图素,用图段来描述这一部分图象则显得力不从心.同时,建立图段时我们还没有考虑到含有曲线图纸的处理,例如地图、等高线图和气象图的处理,现在我们也希望将整体识别算法进一步扩展到不同的应用中去.因此,需要在继承整

---

* 本文研究得到国家杰出青年科学基金资助.作者李宾,1972 年生,博士生,主要研究领域为工程图纸扫描识别,字符识别,模式识别.谭建荣,1954 年生,教授,博士导师,主要研究领域为工程图扫描图象整体识别,工程图离线参数化技术,工程曲线、曲面的计算机辅助设计和工程信息可视化.彭群生,1947 年生,教授,博士导师,主要研究领域为工程图纸扫描识别,真实感图形基础算法,三维几何造型,计算机动画,科学计算可视化及虚拟现实.

本文通讯联系人:李宾,杭州 310027,浙江大学 CAD&CG 国家重点实验室

本文 1997-04-14 收到原稿,1997-06-12 收到修改稿

体识别核心思想的前提下,探索新的思路与算法.

文献[5]在这方面做了有益的尝试,提出了基于扫描串和连通网络图的图象数据模型,并给出了扫描串关于线素的分类算法(它所提出的扫描串就是通常所称的行程编码).已有不少文献[7~9]利用行程编码来细化处理字符图象.所谓扫描串(行程编码)就是在图象某一行(列)中所有连通黑象素的集合(如图 2 所示),它介于象素与图段之间,因此具有更大的灵活性.总而言之,这类算法以扫描串作为基本处理单元,利用扫描串分裂、合并的关系及长度的变化(有时也引入了先验的规则)来判断它的属性(属于水平笔画、竖直笔画或倾斜笔画),然后进一步做细化处理.文献[7~9]中的扫描串属性判别还只是利用了扫描串自身的特征(或者最邻近扫描串的信息),仅仅是局部算法,容易产生错误的结果.最近的一篇文章[6]对此做了改进,更多地利用整体特征和轮廓信息来细化印刷体汉字,得到了更精确的结果.

在已有工作的基础上,本文采用扫描串来描述图象数据,充分利用扫描串之间的连通关系及轮廓信息来进行矢量化.线素跟踪前,预先搜索满足强约束条件的扫描串集合,建立相应的种子信息头,用来指导后续的跟踪;跟踪时,利用信息头的预测分析,判断新加入扫描串的归属及类别;由于充分利用了扫描串端点的轮廓信息,因此整个跟踪过程更加精确有效,可以有效地抵御噪声的干扰.扫描图象数据时根据扫描串线宽来动态调整扫描方向,这样算法处理非常灵活.相比于文献[4]中算法,我们新提出的算法不仅可以处理更加复杂的图纸,还可以用来提取字符图象的笔画,因此具有更大的优越性.

本文第 1 节从扫描串按线素分类的角度,重新考察了矢量化过程,并分析了噪音的整体特性;第 2 节给出了具体的跟踪算法;第 3 节演示了对两幅工程机械图和一幅汉字扫描图象的处理结果,验证了算法的统一性和有效性;最后是总结.

1　基于扫描串分类的矢量化过程分析

首先我们来回顾一下扫描串[5]的定义:图象扫描数据某一行(列)中所有连通的黑象素的集合.沿水平方向的扫描串用集合可以表示为

$$S_{y,k}=\{(x,y)\mid x_0<x<x_1, p(x,y)=1 \text{ 且 } p(x_0,y)=p(x_1,y)=0\},$$

其中 $p(x,y)$为象素(x,y)的值,取值 0 和 1;沿垂直方向的扫描串定义类似.

矢量化就是要将图象数据转换为图形,并保持相应的拓扑结构.图形包括直线、圆弧、曲线及字符的笔画等.为使算法具有更大的通用性及灵活性,以适用于不同的应用环境,我们并不打算提取直线及圆弧,而致力于将图象转换为短折线段表示.如果矢量化得到的短折线段结果较好,那么进一步拟合为长直线、圆弧及自由曲线就非常简单.本文将以扫描串为基本处理单元,提取满足一定几何约束条件的短折线段.在本文中,我们将这种短折线段称为基元.

从匹配、分类的角度出发,矢量化就是将扫描串按基元分类,把每一个扫描串映射到相应的基元.扫描串可分为简单串和复合串[5],简单串仅属于一条基元,而复合串属于多条基元.简单串具有聚集特性,可以在较强的约束条件下搜索得到它们,建立它们的几何参数,然后利用这些整体信息来指导下一步的跟踪.在基元参数头指导下跟踪扫描图象时,要不断判别新加入扫描串与基元之间的偏差度.这时,扫描串端点信息(图象轮廓信息)得到充分利用,比较基元理想延伸预测到的扫描串和实际扫描图象得到的扫描串,线宽差别、两侧轮廓的偏差等 5 个指标得到评价,于是满足条件的扫描串将被加入.

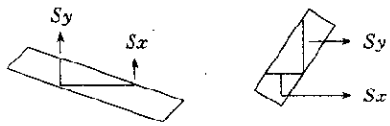
跟踪过程中扫描串的灵活特性在于,一方面由于它是一连串黑象素的集合,可以看作扫描线与图象中线素的交线,因此扫描串具有一定的整体性质;其次,扫描串的端点也就是图象的轮廓信息,因此可充分利用它来获取局部特征;另外,它关于 X 和 Y 方向的对称性,使得动态改变扫描方向成为可能,这样明显改善了矢量化质量,与图段结构相比有更大的优越性.

按照扫描串按基元分类的观点,我们认为,图象中噪声的影响具有整体性.由于 X 和 Y 方向所有扫描串的端点构成了图象的轮廓点,因此扫描串端点对图象信息而言是充分且完备的.噪音对图象信息的影响集中体现在对轮廓点及扫描串端点的影响上.当扫描串端点按基元分类完成后,在没有噪声的理想情况下,应该没有剩余端点;然而由于噪声的介入,必然会剩下多余的无法确定归属的点,噪声就包含在这些点中.所以,在基元跟踪阶段,噪声难以和其他基元的图象信息正确区分,只能在所有基元提取完毕后再利用整体信息予以识别.在我们的算法中,由于利用了整体信息,采用 5 个指标从不同方面来评价新加入扫描串和基元之间的偏差度,因此算法非常稳健,可充分提取整幅图象中性态良好的基元直线段.对于图象信息粘连、模糊不清,难以自动处理部分,我们预备引入知识,进行后续的智能整体识别,而这部分图象并未对基元提取产生明显影响,因此为后续的智能识别奠定了坚实的基础.

2 基元提取跟踪算法

为了提取基元直线段，我们首先搜寻图象，找到一连串满足一定几何约束的相互单连通的扫描串，把它作为种子点，种子点将用来指导后续的跟踪；跟踪的结束由扫描串相互的匹配度来决定，以避免复合线素的产生，保证基元信息的单一性；跟踪同时标记对应的图象，记录提取的象素信息. 下面我们来详细介绍算法的思想和关键步骤.

2.1 动态扫描时扫描方向的确定



$Sy<Sx$,沿Y方向扫描　　$Sx<Sy$,沿X方向扫描

图3

当在图象上搜索扫描串时，我们并不沿固定的 X 或 Y 方向扫描，而是根据实际的线宽信息动态决定沿 X 还是 Y 方向扫描. 沿单一方向扫描的缺点是显而易见的：假如沿水平方向扫描，在处理水平线时，容易丢失附着在水平线上的其他线素信息，不易获得线宽信息，不易分割具有不同线宽的水平线. 因此，跟踪提取某一基元时，我们采用“最短线宽”的原则来确定扫描方向. 具体的过程描述如下(如图 3 所示)：

GetScanDir (POINT *seed*) /* 确定从种子点 *seed* 开始作下一步扫描的方向 */
{
　step1. 分别沿 X 和 Y 方向扫描，得到包含 *seed* 点的扫描串 Sx 和 Sy；
　step2. 计算 Sx 和 Sy 的线宽；
　step3. 若 Sx 和 Sy 的线宽都大于最大线宽，则返回，表明不能从点 *seed* 进行下一步搜索；
　step4. 否则返回具有较小线宽的扫描串代表的方向，表明可以进行下一步搜索；
}

仅从一个点搜寻扫描串，利用其线宽来决定扫描串方向，似乎太随机；但在扫描整幅图象时，没有被标记的每一个扫描串都有机会作为候选种子扫描串，因此从整体上看来，仍具有稳定性.

2.2 种子基元的建立

图象的信息体现在轮廓点即扫描串的端点，从构成线素的若干连通扫描串来看，信息表现在轮廓边缘处及线宽的变化. 信息完备的基元，我们认为它的轮廓边缘变化均匀，且具有大致恒定的线宽，因此拟合为直线段时，有较高的正确性，基本无噪声影响. 通过观察，我们发现，轮廓边缘的变化具有连续性，局部范围内的变化并不明显，仅从有限的少数几条扫描串，难以判断轮廓的变化是否正常；而扫描串线宽的变化则十分明显，不论遇到噪声，还是上下相邻的扫描串产生多连通对应，线宽都会发生相对明显的变化. 因此，搜寻判断种子点时，我们采用“线宽均匀”的判断准则. 下面是具体的过程描述：

VerifySeed (POINT *seed*) /* 假设扫描方向已被确定，验证是否可以从 *seed* 点建立基元 */
{
　step1. 沿扫描方向搜寻包含 *seed* 点的扫描串 s，放入扫描串 *buffer*；
　step2. 记数器置为 0，跟踪方向定为向上(左)；
　step3. 沿扫描方向定向搜寻和 s 单连通的扫描串 t；
　step4. 若记数器值小于 4、t 存在且与 s 相互单连通、t 与 s 之间的线宽差别小于阈值且 t 的端点象素值至少有一个未做任何标记，则把 t 送入到 *buffer*，记数器值加 1，否则跳转到 step6；
　step5. $s=t$，返回 step3 继续循环；
　step6. 记数器重新置为 0，跟踪方向定为向下(右)，s 置为 *buffer* 中的第 1 条扫描串；
　step7. 循环执行 step3～5；
　step8. 若 *buffer* 中存放的扫描串数目小于 5，则返回，表明验证失败；
　step9. 计算 *buffer* 中存放扫描串线宽的平均长度及线宽变化的方差；
　step10. 若方差值大于阈值或平均线宽大于最大线宽，则返回，表明验证失败；
　step11. 建立基元的参数头，用来指导后续跟踪，同时将 *buffer* 中所有扫描串覆盖的象素点标记为 3，然后成功返回.
}

2.3 基于基元参数头引导的跟踪

跟踪的目的在于确定属于当前基元的扫描串，因此可以利用已建立的基元参数头来引导后续的跟踪. 我们利用延伸原则，先计算基元延伸到当前行时理想的扫描串 t，然后计算实际跟踪得到的扫描串 s 以及 s 相对与 t 的偏差度 bias(如图 4 所示). 若 s 完全属于当前基元，我们采样 s 的所有象素点，并将其标记为 3；若 s 并不完全属于当

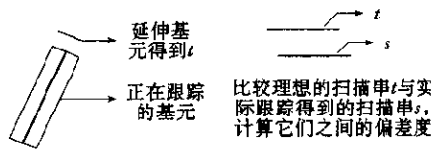


图4

前基元,则只处理 s 的具有较小偏差度的单侧轮廓点,并进行估计采样;若 s 相对与 t 的偏差度较大超过阈值或发生多连通,则跟踪结束. bias 有 5 个方面的属性:线宽差别、左轮廓点的差别、右轮廓点的差别、s 与 t 的匹配度及 s 总的偏差度. 设 s 为三元组 (s_0,s_1,y),t 为 (t_0,t_1,y),l 为区间 $[s_0,s_1]$ 与 $[t_0,t_1]$ 交集的长度,lineWidth 为基元参数头中存放的线宽,则 bias 的计算公式为

```
bias[0]=|s1-s0+1-lineWidth|/(lineWidth+3);        /*线宽差别*/
bias[1]=|s0-t0|/(lineWidth+3);                    /*左轮廓点差别*/
bias[2]=|s1-t1|/(lineWidth+3);                    /*右轮廓点差别*/
bias[3]=l/max(s1-s0+2,lineWidth+1);               /*s 与 t 的匹配度*/
bias[4]=0.5bias[0]+0.25*bias[1]+0.25*bias[2];     /*s 总的偏差度*/
```

下面是算法过程的具体描述.

```
TracePrim(start,跟踪方向 scanFlag)
/*start 为跟踪起始扫描串,scanFlag 为跟踪方向,值为向上(左)或向下(右)*/
{
    step1. 当前扫描串 s 赋值为 start;
    step2. 沿扫描方向按跟踪方向,寻找下一条与 s 单连通的扫描串 t,若在跟踪方向上没有与 s 连通或发生多连通,则跟踪结束返回;
    step3. 那么 t 存在,若 t 的两个端点的象素值全为 3,跟踪结束返回;
    step4. 按上面的公式计算 t 的偏差度 bias;
    step5. 若 bias[3]小于阈值 MATCH_TORRENCE,则跟踪结束返回;
    step6. 否则若 bias[4]小于阈值 BIAS_TORRENCE,则标记 t 覆盖的所有象素点为 3,并记录 t 的中点到基元的数据结构中去;
    step7. 否则执行以下步骤:
      step7.1. 若 bias[1]小于阈值 BIAS_TORRENCE,则根据 lineWidth 及 s0 计算得到一个估计的中点,加入到基元的数据结构中去,并标记左端点象素为 3;否则仅标记左端点象素为 2;
      step7.2. 若 bias[2]小于阈值 BIAS_TORRENCE,则根据 lineWidth 及 s1 计算得到一个估计的中点,加入到基元的数据结构中去,并标记右端点象素为 3;否则仅标记右端点象素为 2;
    step8. s=t,返回 step2,循环执行.
}
```

3　实例与讨论

对于本文提出的算法,我们对不同的扫描图象进行了检验. 图 5 是一幅质量较差的图纸局部,有很多污损和粘连的圆弧及直线. 经过基元提取分割后,从图 6 可以看出,图象清楚、质量较好的部分(包括数字、字符),都正确提取了基元;而粘连模糊的部分,几乎未作任何处理,且对其他基元的提取未产生明显影响. 图 7 是另一幅质量好一点的机械图纸局部,扫描分辨率为 300 dpi. 对于这类图纸,算法处理得比较成功(如图 8 所示),全部图象得到处理,得到了整幅图象的短折线段基元表示,可以直接进行后续的直线、圆弧、符号及机械特征等高层语义的识别与理解. 同时,我们也对印刷体汉字的笔画提取做了试验. 图 9 是清华大学开发的 THOCR 软件 TWReader 5.0 中附带的一个样品图象,原文是一段广告词. 图 10 是经过基元分割提取后的结果,绝大部分笔画都被正确提取;少数笔画被其它笔画分割后,信息比较单薄,但提取基元后标记的图象信息已被保存,将用来结合基元笔画做进一步的处理.

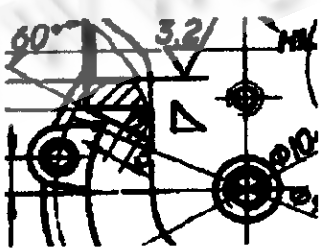


图 5

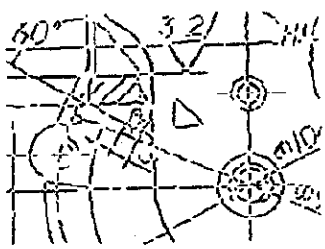


图 6

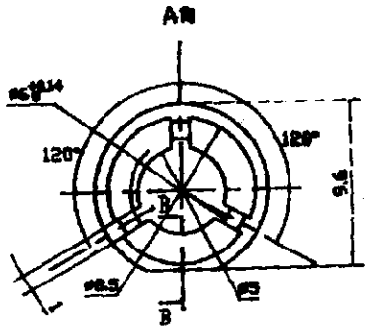


图7

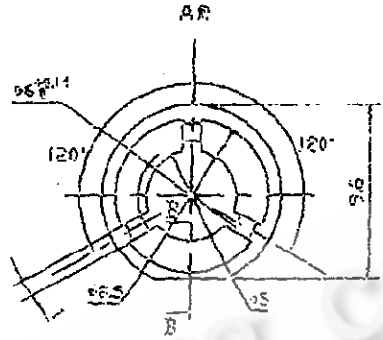


图8

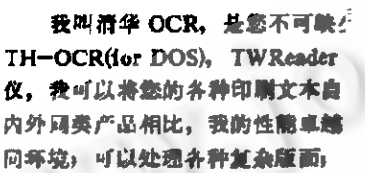


图9

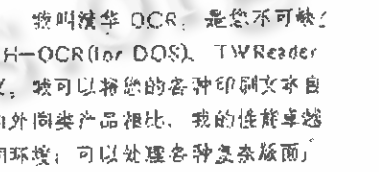


图10

4 结 论

为实现复杂工程图扫描图象的矢量化和文档的自动处理，本文从整体识别的角度出发，提出了一个基于扫描串的统一整体矢量化算法.它具有以下特点：①扫描时根据"最短线宽"原则动态决定扫描方向；②跟踪时首先根据"信息完备性"原则搜寻种子基元，然后利用它来指导后续跟踪；③跟踪每一基元时，标记已被当前基元匹配的扫描串端点，同时保留未被处理的轮廓信息；④图象质量较差的部分，不对基元提取产生明显影响，因此整个算法非常稳定；⑤算法实现简单，不占用大量内存，本质上是一个并行算法；⑥算法具有统一性，能同时用于工程图扫描图象识别和OCR中的笔画提取.

致谢 本文作者特别感谢中国科学院软件研究所的戴国忠研究员和张高博士.戴老师热情的鼓励以及与张高博士对矢量化坦率真诚的讨论直接启发了作者的灵感.

参考文献

1 Louisa Lam, Seong-Whan Lee, Ching Y Suen. Thinning methodologies——a comprehensive survey. IEEE Transactions on Pattern Analysis Machine Intelligence, 1992,14(9):869～885

2 Theo Pavlidis. A vectorizer and feature extractor for document recognition. Computer Vision Graphics and Image Process, 1986,35(1):111～127

3 Yung-Sheng Chen. Segmentation and association among lines and junctions for a line image. Pattern Recognition, 1994,27(9):1135～1157

4 谭建荣.工程图纸扫描图象的整体识别及图形重建的研究[博士论文].浙江大学,1992
(Tan Jian-rong. An approach to global recognition and construction of scanned image of engineering drawings [Ph. D. thesis]. Zhejiang University of China, 1992)

5 李宾,谭建荣,彭群生.基于逐步求精图象数据模型的矢量化方法研究.计算机学报,1996,19(增刊):224～231
(Li Bin, Tan Jian-rong, Peng Qun-sheng. A vectorization method based on PRSIDM. Chinese Journal of Computers, 1996,19 (supplement):224～231)

6 Jenn-Yih Lin, Zen Chen. A Chinese-character thinning algorithm based on global features and contour information. Pattern Recognition, 1995,28(4):493～512

7 Ling-Hwei Chen. A new approach for handwritten character stroke extraction. Computer Process, Chinese Oriental Lang, 1992,6(1):1～17

8 Tseng L Y, Chung C T. An efficient knowledge-based stroke extraction method for multi-font Chinese characters. Pattern Recognition, 1992,25(12):1445~1458

9 Hu G, Li Z N. An X-crossing preserving skeletonization algorithm. International Journal on Pattern Recognition and Artificial Intelligence, 1993,7(5):1031~1053

A Unified Global Vectorization Algorithm Based on Scan Strip

LI Bin TAN Jian-rong PENG Qun-sheng

(*State Key Laboratory of CAD&CG Zhejiang University Hangzhou* 310027)

Abstract The recognition of scanned image of engineering drawings and the extraction of character strokes are very difficult in pattern recognition. To overcome the shortcoming of thinning algorithm, many global algorithms are devised. Among them the global algorithm based on graphical segment directly recognizes the image through the macro-features of image. But it can't process curve and character image. The new algorithm takes scan strip as the basic image process unit and scans the whole image dynamically. A parameter head having been set up is used to supervise the succeeding trace procedure, so more precise vectorization result is obtained. The satisfying result of processing many different kinds of image, including Chinese character image, indicates that the new algorithm is powerful and adaptive.

Key words Recognition of scanned image of engineering drawing, character stroke extraction, global recognition, scan strip.