引文格式：铁占琦. 利用改进的Hausdorff距离匹配多尺度线要素[J]. 地理空间信息, 2024, 22(5): 62-65. doi:10. 3969/j.issn. 1672-4623.2024.05.012

利用改进的Hausdorff距离匹配多尺度线要素

铁占琦[1]
（1. 河南省有色金属地质矿产局，河南 郑州 450000）

摘　要：传统Hausdorff距离对噪声较敏感，在进行多尺度线要素匹配时，易导致漏匹配和误匹配。为提高匹配正确率，提出了一种改进的Hausdorff距离算法。针对部分匹配问题，采用曲线分割算法，以短曲线分割长曲线；针对点位分布差异问题，采用曲线加密算法，以匹配曲线的中间节点进行双向加密；针对曲线点集自身的噪声问题，以距离集合的中位数作为相似性指标，判断匹配要素是否为同名要素。选取不同尺度的行政区划界线和道路网数据进行匹配，以验证该算法。结果表明，该算法具有较好的匹配效果。

关键词：多尺度线要素；要素匹配；Hausdorff距离；相似性指标
中图分类号：P283　　**文献标志码：**B　　**文章编号：**1672-4623（2024）05-0062-04

Multi-scale Linear Feature Matching Using Improved Hausdorff Distance

TIE Zhanqi[1]
(1. Henan Nonferrous Metals Geology and Mineral Resources Bureau, Zhengzhou 450000, China)

Abstract: In the matching process of multi-scale linear feature, traditional Hausdorff distance is prone to be affected by noisy data. In order to improve the robustness of algorithm, we proposed an improved Hausdorff distance algorithm. Considering the problem of partial matching, we used the curve segmentation algorithm to segment the long curve with short curve. Aiming at the difference problem of point distribution, we used the curve encryption algorithm to gather the curve with the center node. With respect to the noise magnitude of curve points set, taking the median of distance set as the similarity index, we judged whether the matching features were the same. We selected different scales of administrative boundaries and road network data to validate the performance of proposed algorithm. The result indicates that the algorithm can effectively solve the matching problem of multi-scale linear feature.

Key words: multi-scale linear feature, feature matching, Hausdorff distance, similarity index

多源空间数据表达的同一地物通常存在差异，这种差异主要源于：①数据生产标准和要求不同；②数据获取途径和时间不同[1-2]。数据的差异性阻碍了空间数据在集成、分析和共享等方面的应用[3]，解决该问题的主要手段是进行数据匹配。同名实体匹配是依据空间实体的几何、拓扑和语义等特征，计算同一地物在多源空间数据中的差异度或相似度，进而识别出同名实体的过程[2]。美国人口调查局最早开展同名实体匹配研究，开发了地图自动合并系统，采用迭代匹配算法，通过建立空间数据节点间的链接实现同名实体匹配，缺点是只能处理一对一匹配关系[4-5]。随着研究的不断深入，新的匹配算法层出不穷，匹配准确率、效率和适用性等方面得到极大提高。匹配算法大致可归纳为3类：①几何匹配[6-13]，以空间实体的几何特征相似性为匹配依据；②拓扑匹配[14-15]，以空间实体的空间关系相似性为匹配依据；③语义匹配[16]，以空间实体的属性数据相似性为匹配依据。不同尺度的空间实体经过概括、抽象、简化等一系列复杂变换后，其拓扑关系难以保持一致；而不同来源的属性数据在逻辑结构、数据类型等方面存在较大差异，转换处理过程中也可能遭到破坏，或者空间实体本身就不包含属性数据。相较于几何匹配，拓扑匹配和语义匹配在多尺度要素匹配中适用性较差。几何相似性度量是线要素匹配的关键，度量指标包括距离[6-9]、长度[10]、方向[11]、形状[12-13]等。其中，距离指标更能反映线要素的相似性，常采用L2距离、Hausdorff距离和Fréchet距离作为度量指标。Hausdorff 距离又称最大最小距离，是度量点集合相似性的有力工具。传统Hausdorff距离采用最大距离表征不匹配程度，对噪声十分敏感。Huttenlocher D P[17]等以最小距离升序集合的第 n 个值

收稿日期：2022-12-07。
项目来源：2021年度河南省财政地质勘查资助项目（2021-6）。
作者简介：铁占琦（1978—），本科，高级工程师，研究方向为空间数据融合与测绘工程，E-mail：25570549@qq.com。

作为相似性指标，提出部分Hausdorff距离（PHD）；邓敏[18]等以最小距离升序集合的中位数作为相似性指标，提出中位数Hausdorff距离。上述方法在一定程度上减弱了噪声影响，但应用于多尺度线要素匹配仍存在一些问题。多尺度线要素匹配常出现一条曲线的部分与另一条曲线匹配的情况，非匹配部分要素会转化成“噪声”；受制图综合影响，同名要素可能会出现节点数量差异大、点位分布不对称等情况，上述情况会导致Hausdorff的最小距离集合整体偏大，此时PHD算法的 n 值难以确定，中位数也难以反映曲线的相似度。鉴于此，本文提出了改进的Hausdorff距离算法。该算法能够解决多尺度线要素匹配中常见的部分匹配和点位差异问题，从而提高匹配运算的正确率。

1　Hausdorff距离算法

1.1　Hausdorff距离

给定两条曲线 L_1、L_2，其离散点集合分别为 $A\{N_{11},N_{12},\cdots,N_{1m}\}$、$B\{N_{21},N_{22},\cdots,N_{2n}\}$，存在两个单向Hausdorff距离，即

$$\begin{cases} h(A,B)=\max\limits_{N_{1p}\in A}\left\{\min\limits_{N_{2q}\in B}\left\|N_{1p}-N_{2q}\right\|\right\} \\ h(B,A)=\max\limits_{N_{2q}\in B}\left\{\min\limits_{N_{1p}\in A}\left\|N_{1p}-N_{2q}\right\|\right\} \end{cases} \tag{1}$$

式中，$\left\|N_{1p}-N_{2q}\right\|$ 为点 N_{1p} 到 N_{2q} 的欧氏距离；$\min\limits_{N_{2q}\in B}\left\|N_{1p}-N_{2q}\right\|$ 为集合 A 中每一点到集合 B 中最近点的距离集合；$\min\limits_{N_{1p}\in A}\left\|N_{1p}-N_{2q}\right\|$ 同理；$h(A,B)$ 为距离集合中的最大值；$h(B,A)$ 同理。

双向Hausdorff距离是单向Hausdorff距离 $h(A,B)$、$h(B,A)$ 中的较大者，也是Hausdorff距离最基本的形式，即

$$H(A,B)=\max\{h(A,B),h(B,A)\} \tag{2}$$

Hausdorff距离能度量两个点集间的最大不匹配程度，算法简单、易实现，但对噪声十分敏感。噪声影响示意见图1，点 N_{25} 偏离其真实位置（方形空心点），为噪声点。受其影响，两条曲线的Hausdorff距离为 $\left\|N_{25}-N_{15}\right\|$，该值大于无噪声时曲线的Hausdorff距离，曲线的不匹配程度增大、相似度降低。

1.2　中位数Hausdorff距离

对于给定的两条曲线 L_1、L_2，中位数Hausdorff距离的定义为：

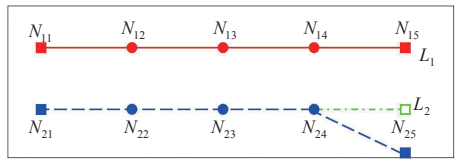


图1　噪声影响示意图

$$\begin{cases} h_m(A,B)=\text{median}\left\{\min\limits_{N_{2q}\in B}\left\|N_{1p}-N_{2q}\right\|\right\} \\ h_m(B,A)=\text{median}\left\{\min\limits_{N_{1p}\in A}\left\|N_{1p}-N_{2q}\right\|\right\} \\ H_m(A,B)=\max\{h_m(A,B),h_m(B,A)\} \end{cases} \tag{3}$$

式中，$h_m(A,B)$ 为最短距离集合顺序排序后的中间值；$h_m(B,A)$ 同理；$H_m(A,B)$ 为二者中的较大者。

中位数Hausdorff距离能够描述点集间距离分布的中心趋势，不受局部特征影响。采用中位数Hausdorff距离求解图1中两条曲线的相似度能避开噪声点 N_{25} 的影响，具有一定的稳健性。

2　改进的Hausdorff距离算法

2.1　分割Hausdorff距离

对于如图2a所示的同名线要素，L_1 的子集 $\{N_{11},N_{12}\}$ 与 L_2 匹配，即 L_1 与 L_2 部分匹配，常见于多尺度线要素匹配中。若采用中位数Hausdorff距离计算二者的相似度，因受 L_1 非匹配子集 $\{N_{13},N_{14},N_{15}\}$ 的影响，两条曲线的 $H_m(A,B)$ 为 $\left\|N_{13}-N_{22}\right\|$，该值大于部分匹配曲线的Hausdorff距离，无法反映二者的真实相似度，且曲线长度差异越大，相似度越低。

为此，本文提出了一种分割Hausdorff距离算法（SHD）[16]，利用短曲线的起点和终点对长曲线进行分割，再计算分割后曲线的中位数Hausdorff距离。具体步骤为：①获取短曲线 L_2 的起点和终点；②遍历长曲线 L_1 上所有的点，查找与 L_2 起点和终点距离最短的点 N_{1i} 和 N_{1j}；③以点 N_{1i}、N_{1j} 将 L_1 分割为3个部分，选取起点和终点为 N_{1i} 和 N_{1j} 的曲线作为新曲线 L_1'；④计算 L_1' 和 L_2 的中位数Hausdorff距离。采用该算法求解的曲线相似度能避开非匹配部分的影响，仅计算 L_1' 和 L_2 的中位数Hausdorff距离，有效解决了部分匹配问题。

2.2　加密Hausdorff距离

线要素的相似度除受噪声和部分匹配影响外，还与曲线上的点位分布有关。如图2b所示，曲线 L_1 的离散点集合为 $A\{N_{11},N_{12}\}$，两条曲线存在节点数量差

异大、点位分布不对称的情况。若采用中位数Hausdorff距离计算二者的相似度，两曲线的 $H_m(A,B)$ 为 $\|N_{11}-N_{22}\|$ 或 $\|N_{12}-N_{24}\|$，该值无法反映二者的真实相似度，且曲线节点数量差异越大、点位分布对称性越差，相似度越低。为此，本文提出了一种加密Hausdorff距离算法（DHD），即利用两条曲线的中间节点进行双向加密，再计算加密后曲线的中位数Hausdorff距离。算法具体步骤为：

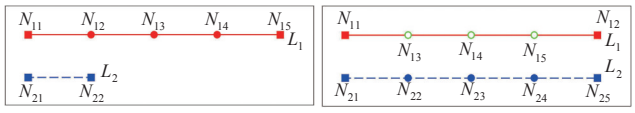
1）从起点出发赋予曲线每一点“里程”属性 M。

2）对于曲线 L_1，其加密点个数为 L_2 的中间节点数为 $n-2$ 个。第 i 个加密点的坐标计算步骤为：①遍历曲线 L_1 上所有的点，找到 N_{1p} 点，使得 $M(N_{1p})\leqslant M(N_{1(i+m)})\leqslant M(N_{1(p+1)})$，其中 $M(N_{1(i+m)})=M(N_{2(i+1)})\times\frac{M(N_{1m})}{M(N_{2n})}$；②利用点 N_{1p} 和 $N_{1(p+1)}$ 计算坐标方位角 α；③利用点 N_{1p} 和 $N_{1(i+m)}$ 计算水平距离 D；④利用 α 和 D 计算点 N_{1p} 和 $N_{1(i+m)}$ 的坐标增量 $\Delta x=D\cos\alpha$、$\Delta y=D\sin\alpha$，进而得到第 i 个加密点坐标。

3）采用同样算法对曲线 L_2 进行加密。

4）计算加密曲线的中位数Hausdorff距离。

图2b中空心圆点为算法所得的加密点，再计算加密曲线的中位数Hausdorff距离，即可有效消除点位差异影响。



a 曲线分割　　b 曲线加密

图2 改进的Hausdorff距离算法

3 应用实例分析

本文选取郑州市某区域的线状交通要素进行道路网匹配应用研究。图3中红色实线为1∶100万道路数据，共21个要素；蓝色虚线为1∶25万道路数据，共523个要素。经人工比对，两类道路共115个匹配对。分别利用中位数Hausdorff距离算法、SHD和DHD算法进行道路网匹配。首先对原始数据进行预处理（投影、坐标变换、拓扑错误检查等）；再以1∶100万道路网为参考数据，1∶25万道路网为待匹配数据，通过上述算法计算要素相似度；最后利用预先设置的阈值判断匹配要素是否为同名要素。

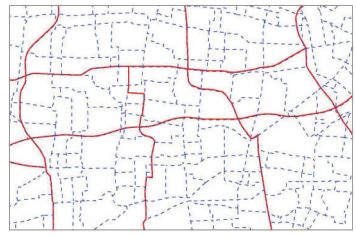


图3 道路网匹配

匹配阈值的选择非常重要，阈值过小会出现较多的漏匹配，阈值过大则会出现较多的误匹配。本文选取5组阈值进行匹配结果比较，其单位与地图实际坐标单位一致。匹配阈值、正确匹配数 C、漏匹配数 L、错误匹配数 M 的计算结果见表1，可以看出，对于同一算法，匹配阈值越大，正确匹配数越多，误匹配数也越多，误匹配主要来自于非匹配要素，其距离参考要素较近，当二者相似度小于阈值时，造成误匹配；对于同一匹配阈值，SHD和DHD算法的正确匹配数明显增多，DHD算法在正确匹配数和误匹配数方面都比SHD算法有一定的改善。

表1 道路网匹配结果

匹配阈值	中位数Hausdorff距离			SHD			DHD		
	C	L	M	C	L	M	C	L	M
200	6	109	0	20	95	4	59	56	0
400	9	106	0	73	42	7	78	37	3
600	15	100	0	83	32	13	89	26	5
800	21	94	6	94	21	17	97	18	9
1 000	27	88	12	101	13	22	104	10	14

当匹配阈值为1 000时，中位数Hausdorff距离算法和DHD算法的匹配结果见图4，红色实线为正确匹配要素，绿色虚线为漏匹配要素，蓝色点画线为错误匹配要素。要素匹配正确率从23.5%升至90.4%，效果显著；且匹配增速明显，中位数Hausdorff距离算法所需时间为57.8 s，SHD算法为12.4 s，DHD算法为15.2 s（需进行曲线加密），运算效率得到较大提高。然而，存在两个方面的缺陷：①漏匹配，一些要素未匹配上；②误匹配，引入一定数量的错误匹配要素。



a 中位数Hausdorff距离算法　　b DHD算法

图4 两种算法阈值为1 000的匹配结果

4 结 语

多尺度要素匹配是空间数据融合领域的一大难

题。本文在研究传统Hausdorff距离的基础上，提出了SHD和DHD算法。理论和实验证明，该算法适用于多尺度要素匹配情形，且计算效率较高。不足之处在于，随着匹配阈值的增大，正确匹配率提高的同时，错误匹配要素数量也在增加，匹配结果仍需人工干预，未能实现自动化处理。下一步研究工作的重心是进一步降低误匹配数量，最终实现自动化匹配。

参考文献

[1] 李小雨,王艳东,吴胜. 众源地理数据质量评价系统设计与实现[J]. 地理空间信息,2020,18(3):45-47

[2] 孙群. 多源矢量空间数据融合处理技术研究进展[J]. 测绘学报,2017,46(10):1 627-1 636

[3] 曹文涛,郭明武,贾小霞,等. 多尺度地理空间要素非一致性检测技术研究[J]. 地理空间信息,2020,18(8):94-97

[4] 胡菡,张亮,石婷婷. 多源地理要素变化识别研究[J]. 地理空间信息,2020,18(6):53-55

[5] 陈玉敏,龚健雅,史文中. 多尺度道路网的距离匹配算法研究[J]. 测绘学报,2007,36(1):84-90

[6] 秦育罗,宋伟东,张在岩,等. 顾及几何特征和拓扑连续性的道路网匹配方法[J]. 测绘通报,2021(8):55-60

[7] 邵世维,刘辉,肖立霞,等. 一种基于Fréchet距离的复杂线状要素匹配方法[J]. 武汉大学学报(信息科学版),2018,43(4):516-521

[8] 郝燕玲,唐文静,赵玉新,等. 基于空间相似性的面实体匹配算法研究[J]. 测绘学报,2008,37(4):501-506

[9] 安晓亚,刘平芝,杨云,等. 一种线状要素几何相似性度量方法及其应用[J]. 武汉大学学报(信息科学版),2015,40(9):1 225-1 229

[10] 钟东. 线要素质量评价与多尺度水系匹配方法研究[D]. 武汉:武汉大学,2018:24-28

[11] 刘涛,杨树文,李轶鲲,等. 空间线群目标方向相似度计算模型[J]. 测绘科学,2013,38(1):156-159

[12] 丁险峰,吴洪,张宏江,等. 形状匹配综述[J]. 自动化学报,2001,27(5):678-694

[13] 秦育罗,郭冰,孙小荣. 改进Hausdorff距离及其在多尺度道路网匹配中的应用[J]. 测绘科学技术学报,2020,37(3):313-318

[14] Xiong D, Sperling J. Semiautomated Matching for Network Database Integration[J]. ISPRS Journal of Photogrammetry and Remote Sensing,2004,59(1/2):35-46

[15] 许俊奎,武芳,钱海忠,等. 一种空间关系相似性约束的居民地匹配算法[J]. 武汉大学学报(信息科学版),2013,38(4):484-488

[16] 程钢,卢小平. 顾及通名语义的汉语地名相似度匹配算法[J]. 测绘学报,2014,43(4):404-410

[17] Huttenlocher D P, Klanderman G A, Rucklidge W J. Comparing Images Using the Hausdorff Distance[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1993,15(9):850-863

[18] 邓敏,钮沭联,李志林. GIS空间目标的广义Hausdorff距离模型[J]. 武汉大学学报(信息科学版),2007,32(7):641-645

（上接第56页）

[5] 李安平,翟仁健,殷吉崇,等. 顾及空间结构关系的居民地自动合并方法[J/OL]. (2022-07-19)[2022-12-25]. https://kns.cnki.net/kcms2/article/abstract?v=RsYUK1W_Gsgq9WkJxi8OgUVapdgPIKZsdhalu9BP1twH2Q35tKfeU4a9yu16_6JmD4t0num1sas_3c9c-q96_KZlTRH2K9tnpbuljKZO-nP7TPFAYQXU3BbxfouoE6Ex7fugIhtAuqo=&uniplatform=NZKPT&language=CHS

[6] 马京振,周炤,张付兵. 顾及多特征约束的建筑物组合化简方法[J]. 地球信息科学学报,2023,25(2):277-287

[7] 蔡娇楠. 顾及格式塔原则的街区建筑物MST聚类[D]. 西安:长安大学,2018

[8] 孙前虎. 基于多约束的建筑群聚类方法研究[D]. 长沙:中南大学,2011

[9] 金澄,安晓亚,陈占龙,等. 矢量居民地多边形多级图聚类划分方法研究[J]. 武汉大学学报(信息科学版),2021,46(1):19-29

[10] Li Z. Algorithmic Foundation of Multi-scale Spatial Representation[M]. CRC Press,2007

[11] 何海威,钱海忠,刘闯,等. 采用决策树算法进行居民地自动综合[J]. 测绘科学技术学报,2016,33(6):623-628

[12] Wagemans J, Feldman J, Gepshtein S, et al. A Century of Gestalt Psychology in Visual Perception: II. Conceptual and Theoretical Foundations[J]. Psychological Bulletin,2012,138(6):1 218

[13] Yan H, Weibel R, Yang B. A Multi-parameter Approach to Automated Building Grouping and Generalization[J]. Geoinformatica,2008,12(1):73-89

[14] 艾廷华,郭仁忠. 基于格式塔识别原则挖掘空间分布模式[J]. 测绘学报,2007,36(3):302-308

[15] Zahn C T. Graph-theoretical Methods for Detecting and Describing Gestalt Clusters[J]. IEEE Transactions on Computers,2006,C-20(1):68-86

[16] Zhang X, Ai T H, Stoter J. Characterization and Detection of Building Patterns in Cartographic Data: Two Algorithms[M]. ISPRS,2012