

张柏洲<sup>1</sup> 李飚<sup>1</sup>

1. 东南大学建筑学院; zhangbaizhou@seu.edu.cn

Zhang Baizhou<sup>1</sup> Li Biao<sup>1</sup>

1. School of Architecture, Southeast University

# 基于大数据特征提取的建筑形态聚类检索方法

## 研究<sup>①</sup>

——以大学校园为例

# Research on Clustering and Retrieval Method of Building Form Based on Feature Extraction of Big Data —Take university Campus as an Example

**摘要:**近年来,大数据技术在建筑及城市设计领域的相关探索与应用在不断更新发展。其中,建筑形态相关的可视化数据如何抽象表征,并运用于设计分析与决策过程,是数据驱动的设计方法中的重要命题之一。本研究以大学校园作为研究对象,通过对地图大数据的筛选与要素提取,收集建筑形态布局相关的量化特征,借助计算几何、神经网络等技术手段,研究建筑空间形态角度的案例匹配与支持设计决策的可行性。研究首先从网络开源地图平台对大学校园数据样本进行采集和筛选;进而通过几何规则算法、图像特征提取等方法从地块形态、建筑形态、道路形态三方面分别对大学校园样本进行形态特征的提取;最后将形态特征整合,根据综合特征数据和聚类结果进行近似案例匹配的实验验证。研究结果表明,该研究方法能够将建筑形态要素进行合理的量化表征,并从形态角度将建筑样本进行分类并探究一般规律,从而进一步助力设计者从大量数据中快速获得较为可靠的近似案例检索以支持设计决策。

**关键词:**特征提取; 大数据; 建筑形态; 大学校园; 聚类; 案例检索

**Abstract:** In recent years, the relevant exploration and application of big data technology in architecture and urban design have been continuously updated and developed. Among them, how to abstract the visual data related to architectural form and apply it in design analysis and decision-making is one of the important propositions in data-driven design methods. This study takes university campus as the research object, through the screening of map big data and element selection, collects the quantitative features related to the architectural layout, and studies the feasibility of case matching and supporting design decisions from the perspective of architectural morphology with the help of computational geometry, neural network and other technical means. Firstly, the data was collected and screened from open source map. Secondly, through the computer geometry and neural network technology, the feature extraction methods were performed from three aspects: site,

<sup>①</sup> 国家自然科学基金面上项目,项目批准号:51978139;江苏省研究生实践创新计划项目,项目批准号:SJCX21\_0025。

road and building. Finally, different features were integrated and an experiment of matching approximate cases according to the features and clustering results was conducted to validate the research methods. The results show that the research method can characterize the architectural morphology elements reasonably and quantitatively and classify architectural samples from the perspective of morphology to explore the general law, thus further helping designers to quickly obtain more reliable approximate case retrieval from big data to support design decisions.

**Keywords:** Feature Extraction; Big Data; Building Morphology; University Campus; Clustering; Case Retrieval

## 1 研究背景

城市形态学 (urban morphology) 作为从形态分析的角度对城市空间及其发展演化进行的研究, 自 19 世纪以来逐步兴起<sup>[1]</sup>。以康泽恩学派、凡尔赛学派、意大利学派等为主的经典城市形态分析方法将建筑、开放空间、街区、用地单元、街道等基础空间要素进行了抽象与分析, 这些方法在当代的进一步发展极大助力了城市形态研究, 并为相关设计实践提供了形态学的分析视角。20 世纪后期到 21 世纪, 随着城市形态量化分析手段的成熟, 新的分析技术与经典方法不断结合, 单一或多要素整合的城市形态量化分析方法不断涌现, 在技术和需求上均有着系统化发展, 例如空间句法、Spacematrix、Place Syntax, 等等<sup>[2]</sup>。

近年来, 随着计算机技术在建筑及城市设计领域的不断介入, 以及城市空间形态信息的数据化与标准化, 城市形态与数据信息结合的相关探索在不断发展, 该研究领域也试图不断突破统计式的“分析”, 进而借助大数据和计算机技术, 试图搭建从形态分析到设计决策辅助的桥梁, 以推动数据驱动的城市形态研究方法的变革。Dillenburger<sup>[3]</sup> 的研究基于计算机几何和数据库原理, 将城市地块的形状属性、环境属性、特定信息属性等量化数据整合, 建立了一种城市地块信息的快速检索系统, 使用者通过输入目标地块, 即可快速从大量数据集中检索出量化指标相近的结果, 简化了案例研究的过程。Moosavi<sup>[4]</sup> 的研究从城镇尺度出发, 收集了遍布全球的一百多万个城市、乡镇、村庄的路网形态数据, 通过预训练的 Auto-Encoder 神经网络, 自动学习了城市形态结构并用向量数据进行表示, 由此可对空间结构、方向、图形、密度、局部形变等城市形态要素进行比较, 并进一步分析了全球城市形态的主

要模式与分布。上述两个研究对于城市形态问题进行了技术策略上的探索, 但主要针对特定尺度的城市形态与数据, 忽略了建筑类型与功能的影响因素。蔡陈翼等<sup>[5]</sup> 聚焦于住宅区这一特定类型, 研究提取 4172 个南京市住宅区地图数据, 并将其处理为图像, 使用 GoogLeNet 开源深度神经网络模型将平面图像提取为多维向量数据, 继而可以通过降维数据对住宅区形态相似程度 (特征向量距离) 进行计算和比较。该方法对特定类型建筑的形态提取以及相似案例检索起到了较好的效果, 但受限于样本数量和所在地 (均来自南京), 未能完全发挥数据驱动分析方法的优势。

因此, 以特定建筑类型为目标, 提取大量城市数据的形态特征, 建立建筑及城市的案例检索机制, 在当下成为可能。本研究将基于大数据采集、计算机几何、神经网络等技术手段, 以大学校园为例, 通过对采集的校园平面形态数据进行筛选与量化特征提取, 研究空间形态角度的案例检索与支持设计决策的可行性。

## 2 数据样本采集

作为数据驱动的技术方法研究, 充足且可靠的数据样本是本研究的必要前提。在网络资源愈加丰富的当下, 以谷歌地图、OpenStreetMap、高德地图等为代表的各类网络地图平台逐渐成为城市形态相关数据的理想来源。本次课题研究所需要的数据样本均来自 OpenStreetMap<sup>①</sup> (以下简称 OSM) 平台, 使用 Python 程序进行自动获取 (图 1)。

<sup>①</sup> OpenStreetMap (<https://www.openstreetmap.org>) 是一个开源地图数据平台, 允许用户自由下载获取。



图 1 OpenStreetMap

(图片来源: <https://www.openstreetmap.org>)

## 2.1 数据获取

从原始的地理空间数据中提取建筑与城市形态分析的有效信息十分重要，并且需要专业性保证的研究环节<sup>[6]</sup>。对于本研究的研究对象，数据采集的第一步是在全球范围内获取尽可能多的大学校园POI数据以及有效矢量数据。

OSM 中数据存储的结构为树状，其图元类型分为三类：节点 (node)，路径 (way)，关系 (relation)。在三类数据之下，又进一步存储了经纬度坐标、名称、地址、功能等属性标签 (tags)，每个标签均由键 (key) 与值 (value) 进行一一映射。其中，大学校园作为一种特定建筑类型，其在地图数据中以特定的封闭路径存储，通过查询 OpenStreetMap Wiki<sup>①</sup> 的 Map Features，可以得知大学校园所对应的标签为：key = “amenity”，value = “university”。根据此标签，可以从地图中提取全部大学校园这一特定类型数据。

由于全球范围的地图数据量过于庞大，且不同国家地区的地图数据质量参差不齐，本研究抽样选择了各大洲具有一定代表性，数据量充足、完整的 30 个国家和地区进行原始数据下载，其中亚洲 6 个，大洋洲 2 个，欧洲 17 个，北美洲 3 个，南美洲 1 个，非洲 1 个。通过 Python 程序自动筛选上述的“大学”标签，所有的大学校园数据均可从这 30 个

原始数据集分离并提取出来。提取后，共有 38126 个大学校园数据样本可作为原始数据集。样本信息包括 OSMID、名称、经纬度、国家等 POI 信息，以及地块、道路、建筑物等矢量信息（图 2）。

## 2.2 数据筛选

数据样本的完整性与可靠性是本研究深入推进的关键。网络地图数据的存在信息重复、信息缺失的情况，因此需要进行筛选处理以保证样本质量。当前 38126 个样本出现的问题集中在重复数据和无效数据两方面。

重复数据主要指原始数据集中存在重复或未整合的样本。例如，部分校园内的建筑物数据被单独记录在地图信息内，导致同一个大学校园的名称在原始数据集中重复出现。此类数据需要对原始数据的属性标签进行二次筛选，剔除 “building” 等属性的样本。

无效数据主要指地图信息的缺失的样本。例如，校园区域内缺失建筑物信息或缺失道路信息。此类数据需要对原始数据集的图元信息进行检测，若校园区域内建筑物几何图元或道路几何图元数量为 0，则剔除该样本。

<sup>①</sup> OpenStreetMap Wiki (<https://wiki.openstreetmap.org/>) 是 OSM 的指南文档。

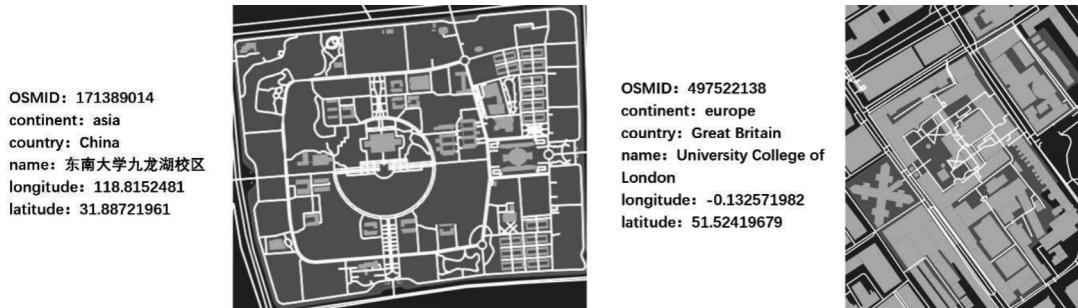


图 2 POI 数据与可视化的矢量数据

(图片来源：作者编写程序生成后自绘)

经过两次筛选清洗后，共计剩余 11124 个样本  
可作为有效数据集（表 1）。

续表

表 1 数据筛选统计表

大洲	国家和 地区	原始数据量/ 个	冗余数据 筛选剩余/ 个	无效数据 筛选剩余/ 个
欧洲	英国	1489	703	653
	法国	1584	605	541
	意大利	1496	406	345
	德国	4820	654	615
	荷兰	748	65	65
	西班牙	1497	410	377
	瑞士	187	58	56
	波兰	1459	496	472
	俄罗斯	5327	912	810
	芬兰	138	56	56
	瑞典	264	71	68
	奥地利	457	65	63
	丹麦	185	44	32
	比利时	513	112	95
	挪威	114	30	28
	捷克	297	57	56
	乌克兰	1120	349	303
亚洲	中国	2446	1579	1024
	日本	1654	998	891
	韩国	508	371	293
	印度	651	557	357

大洲	国家和 地区	原始数据量/ 个	冗余数据 筛选剩余/ 个	无效数据 筛选剩余/ 个
亚洲	马来 西亚+ 新加坡+ 文莱	246	154	126
	美国			
北美洲	加拿大	427	162	155
	墨西哥	1286	1045	573
	澳大利亚	521	209	178
大洋洲	新西兰			
	巴西	2962	1944	1040
非洲	南非	158	102	72
	总计	38126	14303	11124

### 3 特征提取与聚类检索实验

在采集到的有效数据集中，样本矢量数据包含了地块形态、建筑形态、道路形态这三类对平面形态分析影响最大的要素，后续的研究将从这三类形态数据出发，通过 Python 编程，分别进行特征提取，以便定量地对样本间的异同性进行分析，进而实现案例匹配辅助决策的目标。

本研究针对大学校园空间形态的特征提取工作主要采用几何计算与深度卷积神经网络（deep convolutional neural networks, DCNN）两种方法来进行

行。几何计算方法针对信息较单一的形态信息（例如地块）有比较便捷的处理方法；深度卷积神经网络方法则通过图像特征提取的方式，将较复杂的形

态信息（道路、建筑）处理为特征向量，进而便可通过计算特征向量之间的欧几里得距离来完成匹配推荐检索的实验验证（图 3）。

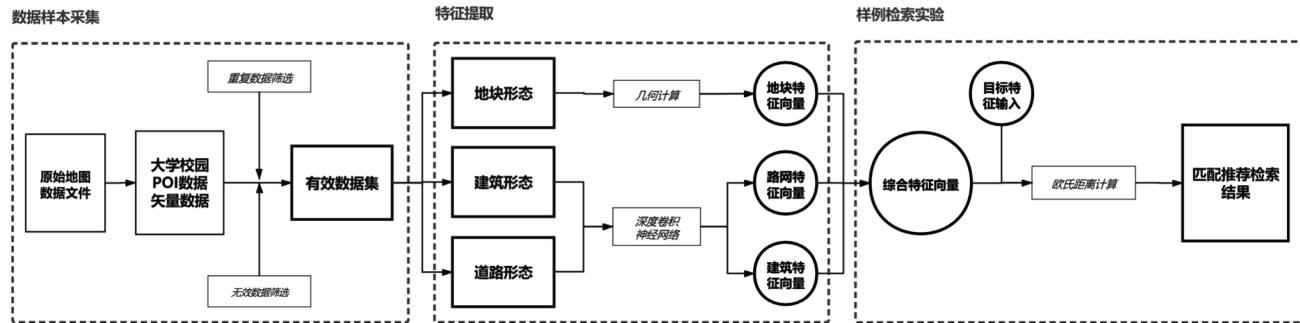


图 3 研究方法流程

(图片来源：作者自绘)

### 3.1 特征提取

#### 3.1.1 地块形态

地块作为城市中的重要边界要素，体现了一个建筑单体或建筑群在城市中的空间边界和功能覆盖区域。地块轮廓的形态信息可以通过样本中的多边形兴趣面数据（area of interest, AOI, 表示一个区域状的地理实体）进行提取与分析。通常情况下，地块形态的特征描述常以定量的基础指标（面积、长度、宽度等）和定性的类型区分来加以定义（方形、三角形、L 形等），但上述方法对地块形态几何属性的描述精确度不足，不足以精细对比每个样本间的差异，因此需要对任意地块多边形的形态进行更精准和更泛用的特征描述方法。

Boyce-Clark Index 是 1964 年被提出的一种计算

平面几何图元形状指数的方法<sup>[7]</sup>，该方法通过几何图形中心向外均匀发射的若干条密集射线与外轮廓的交点来综合计算形状指数，以表示几何形状的“异形”程度。本研究则将地块的多边形轮廓等距剖分为  $n$  个等分点，通过计算全部等分点与地块中心点的连线长度  $l$ ，并根据极坐标进行起点与排列顺序的统一后，便可得到一个  $n$  维向量  $(l_1, l_2, l_3, \dots)$  来表征任意形状、任意大小地块的几何形态特征。向量数值的整体大小体现了地块规模的区分，数值的变化趋势则体现了地块形态的凹凸变化（图 4）。这一算法规避了原 Boyce-Clark 方法对复杂形状适应性差的情况，并可用较高维的向量（ $n$  维）代替形状指数的低维度（1 维）特征表示，使地块形状特征的信息储存更加精确。

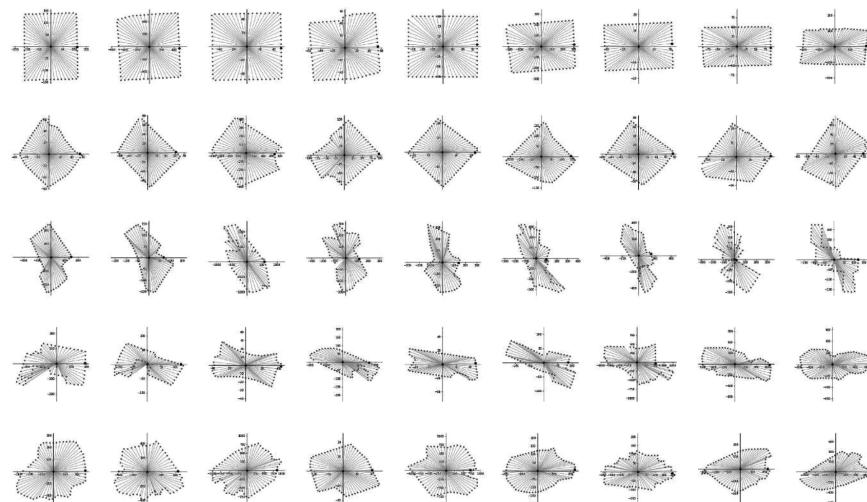


图 4 地块形态特征提取结果

(图片来源：作者编写程序生成)

### 3.1.2 建筑与道路形态

建筑形态与道路形态是大学校园与城市形态关系的要素，二者密切相关且相互影响，在地图数据中均以矢量形式的节点和多段线进行储存。现有的关于建筑与道路形态特征取研究经历了从定性到定量的转变，探讨了量化指标与空间认知的关系<sup>[8]</sup>，例如密度、整合度等，更加适合宏观到中观的类型区分与解析，但若从大量数据中精确到案例与案例的比较，则有所欠缺。

本研究对建筑和道路形态特征的提取采用了神经网络计算的方法，以获得可被更精确抽象描述的

量化特征。首先，将建筑形态和道路形态数据预先分层（图 5），分别提取并处理为像素图片，继而利用预训练的 DCNN 模型对形态图像进行特征提取，以保证每个样本的特征高度精确且各异。此处使用了 PixPlot<sup>①</sup> 所提供的深度卷积神经网络模型，对建筑形态图像和道路形态图像进行特征提取，该模型将神经网络输出层倒数第二层的 2048 维向量数据作为每个输入图像的特征向量，表示图像特征。经过处理，11124 个道路形态图像样本和 11124 个建筑形态图像样本被处理为等长且可比较的 2048 维向量。

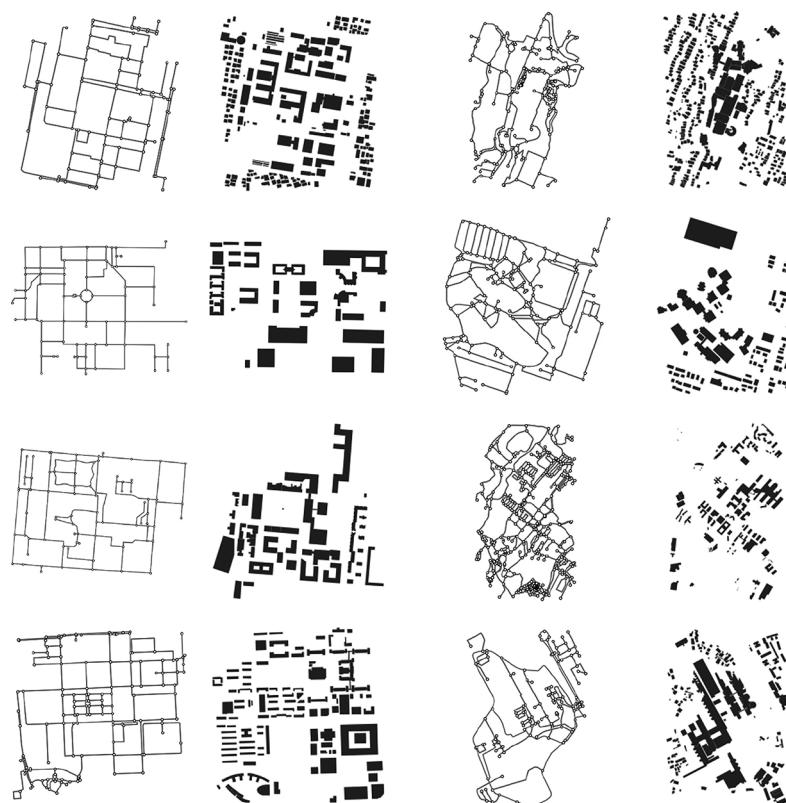


图 5 道路形态与建筑形态数据分层

（图片来源：作者编写程序生成）

## 3.2 案例检索实验

### 3.2.1 聚类可视化

经过全部形态要素的特征提取的工作后，每个大学校园样本的形态特征均被整合在维度等长的特征向量中，通过计算机比较向量每一维的数值，可以自动比较出样本间形态特征的相似与否。随后经过数据归一化（data normalization）和数据降维（data reduction）处理，即可获得形态数据样本聚类的可视化结果。如图 6 所示，可见具有相似形态特

征的大学校园平面可被聚合为聚类组团，分布在二维空间上的不同位置，相似程度越高的大学校园平面相距越近，反之相距越远。

### 3.2.2 案例检索

根据整合的特征向量以及聚类结果即可输入目标案例，从有效数据集中进行案例匹配推荐的测试。

<sup>①</sup> PixPlot (<https://dhlab.yale.edu/projects/pixplot>) 是 Yale Digital Humanities Lab Team 开发的开源项目，该项目利用预训练的卷积神经网络对图像进行特征提取。

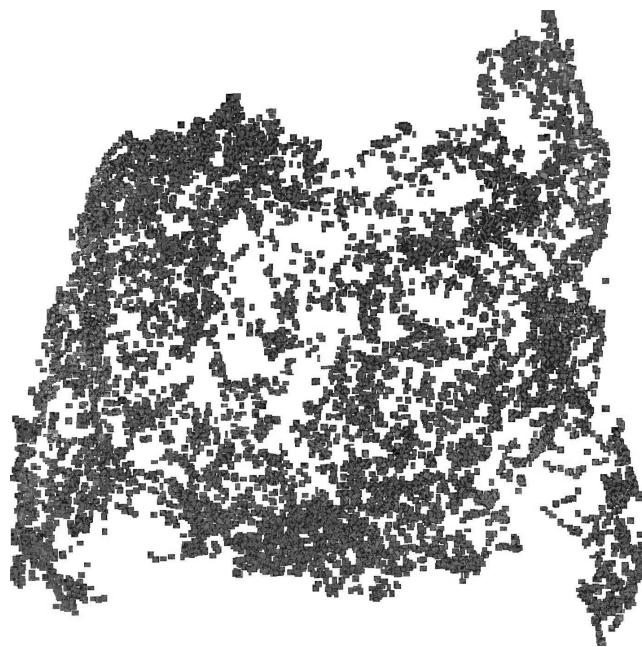


图 6 数据样本聚类可视化结果

(图片来源：作者编写程序生成)

此步测试的机制为：将输入样本按照相同的特征提取方法得到表征其形态的特征向量，继而根据有效数据集中的样本特征向量依次计算欧氏距离并排序，距离最近的若干个样本即为匹配推荐结果。

图 7 展示了三种不同类型的大学校园平面形态测试数据所匹配推荐的前 5 个最相似样本。测试数据 1 的结果显示，地块轮廓相对方正，尺度小；地理位置多位于密集街区中，周边环境紧凑；校园规模很小，仅有几个位数的建筑物数量。测试数据 2 的

结果显示，基地轮廓方正，正南正北坐落于城市地块中；道路与建筑布局呈一定轴线性，有明显的中心；检索结果的区位几乎均位于中国境内，显现出一定的地域化特征。测试数据 3 的结果显示，地块轮廓在正交轴网上，但呈现一定的锯齿状边界；校园多位于正交路网的市区内，尺度紧凑细密，且多出现在北美境内。三类测试数据匹配的案例结果均有较高的形态特征相似度。

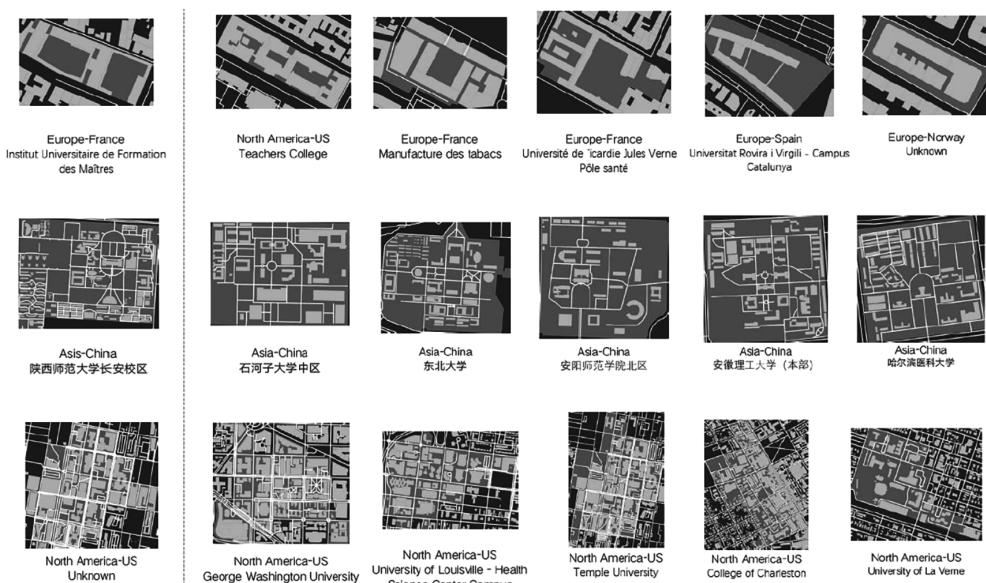


图 7 大学校园平面形态匹配检索测试结果

(图片来源：作者编写程序生成后自绘)

## 4 结语与展望

本研究在大数据的背景下，通过计算机几何、神经网络技术的介入，探索了基于大数据特征提取的建筑形态聚类与检索方法，并以大学校园为例，从地块形态、道路形态、建筑形态三个方面，分别进行特征提取，最终进行了数据聚类与案例匹配推荐的验证。该研究方法的主要成果体现在以下方面：①将建筑与城市形态要素进行了合理的量化表征；②从形态角度将建筑形态样本进行分类，并探究了一般规律；③能够进一步助力设计者从大量数据中快速获得较为可靠的近似案例检索以支持设计决策。

长久以来，针对城市形态问题的特征量化提取研究始终存在，这些研究对城市形态规律的挖掘做出了很大贡献，也为设计实践提供了理性的分析依据与参考。本研究的技术探索表明，大数据特征提取的方法可以对数据进行深入解析，能够为某一特定类型的建筑建立起形态匹配检索的工作链。此外可以预期的是，在数据量扩充以及特征维度增加后，本研究在样本广度与数据深度两方面均有着良好的发展空间，在建筑学问题特征化、类型化研究和案例查找、设计生成方面能够发挥更大作用。

### 参考文献

[1] 段进, 邱国潮. 国外城市形态学研究的兴起与发展 [J]. 城市规划学刊, 2008 (5): 34-42.

[2] 叶宇, 庄宇. 城市形态学中量化分析方法的涌现 [J]. 城市设计, 2016 (4): 56-65.

[3] DILLENBURGER B. Space Index: A retrieval-system for building-plots [J]. 2010.

[4] VAHID MOOSAVI. Urban morphology meets deep learning: Exploring urban forms in one million cities, town and villages across the planet [J]. arXiv preprint arXiv: 1709.02939, 2017.

[5] 蔡陈翼, 李飚, 卢德格尔·霍夫施塔特. 神经网络导向的形态分析与设计决策支持方法探索 [J]. 建筑学报, 2020 (10): 102-107.

[6] MO Y, LI B, WU J, et al. Archibase: A City-Scale Spatial Database for Architectural Research [C]. Proceedings of the 26th CAADRIA Conference-Volume 2, The Chinese University of Hong Kong and Online, Hong Kong, 29 March-1 April 2021, pp. 519-528, 2021.

[7] BOYCE R R, CLARK W A V. The concept of shape in geography [J]. Geographical review, 1964, 54 (4): 561-572.

[8] 甘草, 孙沛. 基于锚点理论的校园路网形态对空间认知的影响研究——以北京大学和清华大学为例 [J]. 西部人居环境学刊, 2020, 35 (4): 88-96.