文章编号：1672-5328(2026)01-0123-04　中图分类号：U491　文献标志码：D　DOI:10.13813/j.cn11-5141/u.2026.0104

短视频驱动的多模态城市意象感知方法研究动态

李毅喆[1,2]
(1. 同济大学交通学院，上海201804；2. 同济大学道路与交通工程教育部重点实验室，上海201804)

摘要：选取来自国际学术期刊的论文，以概述形式对城市交通理论方法、实证分析等学术研究成果进行总结性介绍，旨在增强城市交通业界和学界对国际学术动向和研究热点的关注，促进学术交流。《利用海量短视频对上海城市形象进行多模态感知》一文以上海市为例，从空间、景观、社会与情感4个维度，构建多模态的城市意象感知框架，探讨短视频大数据在城市意象分析中的应用。研究发现，上海市城市意象呈现“核心集聚-外围扩散”的空间结构；景观意象表现为现代都市、传统建筑、自然景观与消费空间等多样化主题；情感分析显示，上海市总体情感偏正，但在新型冠状病毒感染疫情期间波动显著。这一研究为城市规划、城市品牌建设及社会情绪管理提供了新的数据支持和理论依据。

关键词：城市意象；短视频；多模态；深度学习；情感分析；上海市

Academic Dynamics on Urban Image Perception Based on Multi-Perspective Short Videos

Li Yizhe[1,2]
(1. College of Transportation Engineering, Tongji University, Shanghai 201804, China; 2. The Key Laboratory of Road and Traffic Engineering, Ministry of Education, Shanghai 201804, China)

Abstract: A review of selected papers from international academic journals is presented to summarize research findings, theoretical approaches, and empirical analyses of urban transportation. The aim is to enhance the communication between industrial and academic fields in urban transportation, highlight international research focuses, and promote academic exchange. Using Shanghai as a case, the article "Utilizing Massive Short Video Data for Multi-Perspective Perception of Shanghai's Urban Image" builds a multi-perspective framework from four dimensions: space, landscape, society, and emotion. It examines the application of short-video big data in urban image analysis. This study finds that Shanghai's urban image shows a spatial pattern of “core concentration and outward diffusion”. The landscape imagery include diverse themes, such as the modern metropolis, traditional buildings, natural scenery, and consumer spaces. Sentiment analysis shows that Shanghai generally has a positive emotional tone. However, the sentiment fluctuated significantly during the COVID-19 pandemic. This study provides new data support and a theoretical basis for urban planning, city branding, and social emotion management.

Keywords: urban imagery; short videos; multi-perspective; deep learning; sentiment analysis; Shanghai

收稿日期：2025-10-16

作者简介：李毅喆(2002—)，男，安徽寿县人，硕士研究生，研究方向为交通规划、设计与运行管理，电子邮箱213576172@qq.com。

引用格式：李毅喆. 基于多模态短视频的城市意象感知研究动态[J]. 城市交通，2026，24(1)：123-126.

Li Yizhe. Academic dynamics on urban image perception based on multi- perspective short videos[J]. Urban transport of China, 2026, 24(1): 123-126.

研究背景

城市意象是城市认知与品牌建构的重要组成部分，直接关系到城市形象传播、空间认同与可持续发展。自凯文·林奇(Kevin Lynch)提出《城市意象》(《The Image of the City》)以来，相关研究逐渐从空间可读性与形态结构，扩展至社会文化、情感心理等多重维度。传统研究方法，例如问卷调查、深度访谈与认知地图绘制，虽然揭示了个体对城市空间的认知机制，但存在数据量有限、主观偏差大、更新周期长等问题，难以全面捕捉公众对城市的真实感知与情绪反馈。

随着移动互联网与社交媒体的兴起，人们通过影像记录与分享城市生活场景，形成了丰富的数字足迹。特别是短视频平台(如抖音)的爆发式增长，使公众在自发表达城市体验的同时，也生成了庞大的可视化数据，为城市意象研究提供了新的数据来源与研究视角。与传统媒体相比，短视频具有即时性、沉浸性与多模态(multimodal)融合性等特征，能够更真实地呈现城市的空间肌理、社会活动与情感氛围，体现出一种“群众化影像建城”的新趋势。

在此背景下，《利用海量短视频对上海城市形象进行多模态感知》一文提出以短视频大数据驱动的多模态城市意象感知方法。该研究以上海市为例，整合地理信息、深度学习与情感分析(Sentiment analysis)方法，从空间、景观、社会与情感4个维度系统刻画城市意象的构成与空间分布，展示了新媒体时代下城市形象研究的数字化、自动化和多模态化发展方向。

研究方法

基于“数据采集—语义识别—聚类分析—情感挖掘”的流程，该研究构建了多模态感知的城市意象分析框架。具体分为4个阶段：数据预处理、景观语义分割、模型分析与意象解构，实现从短视频原始数据到多模态城市意象的自动识别与量化。

1）数据采集与预处理。

研究选取“上海”“城市印象”“旅游景点”“现代都市”等关键词，从抖音平台批量爬取视频，共收集4 013条样本。经人工筛选与内容去重后，保留2 180条高质量短视频。每条视频均提取标题、标签、点赞数、评论数、位置信息等元数据(Metadata)。为解决部分视频未包含地理位置标签的问题，借助GazPNE2地名识别模型与Tongyi Tingwu平台的语音转写技术，从视频文本和音频中自动提取地名，并通过百度地图地理编码生成地理坐标。

在视频帧处理阶段，研究采用哈明距离(Hamming Distance)方法对连续帧进行相似性检测：

$$hamming\ distance(H_1,H_2)=\frac{1}{n}\sum_{i=1}^{n}\left[H_1(i)\oplus H_2(i)\right],$$

式中：$hamming\ distance(H_1,H_2)$ 为 H_1 和 H_2 之间的哈明距离，即两个等长字符串之间对应位置上不同字符的个数；n 为哈明值 H_1 和 H_2 的长度，即各自包含的位数(bit)，假设 H_1 和 H_2 是长度为 n 的二进制字符串，那么 n 就是其长度；$H_1(i)$ 和 $H_2(i)$ 分别为哈明值 H_1 和 H_2 的第 i 位的值，这里假设 H_1 和 H_2 是二进制字符串，每一位的值可以是0或1；$\oplus$ 为异或运算(Exclusive OR, *XOR*)符号，对于每一位 $H_1(i)$ 和 $H_2(i)$ 来说，如果值不同(即一个为0，另一个为1)，异或运算结果为1，如果值相同(都是0或都是1)，异或运算结果为0。

当两帧的哈明值差距小于设定阈值时，即视为重复帧并予以剔除。最终得到9 255张代表性图像，构成帧级数据集，为后续语义识别奠定了基础。

2）城市要素识别与景观语义分割。

在视觉分析阶段，采用U-Net深度学习模型对视频帧进行语义分割，识别城市中的9类关键要素，包括建筑、交通、植被、水体、天空、公共设施和商业空间等。U-Net的编码-解码结构能够在保持图像细节的同时实现像素级分类。模型在Cityscapes与MIT Indoor Scenes数据集上训练，其评估指标包括像素准确率与交并比。

$$PA=\frac{\sum_{t=0}^{k}p_{tt}}{\sum_{t=0}^{k}\sum_{f=0}^{k}p_{tf}},$$

式中：PA 为像素准确率，即模型正确分类的像素占总像素的比例/%；k 为分类任务中的类别总数/个；p_{tt} 为第 t 类别中被正确分类的像素数量，具体来说，p_{tt} 为真实标签为类别 t 且预测也为类别 t 的像素数量/个；p_{tf} 为真实标签为类别 t 但被预测为类别 f 的像素数量，这里 t 和 f 可以是同一个类别，也可以是不同的类别。

$$IoU=\frac{p_{ii}}{\sum_{j=0}^{k}p_{ij}+\sum_{j=0}^{k}p_{ij}-p_{ii}},$$

式中：IoU 为交并比，用于衡量预测区域与真实区域的重叠程度；p_{ii} 为第 i 类别中被正确分类的像素数量/个，具体来说，p_{ii} 为真实标签为类别 i 且预测也为类别 i 的像素数量(即混淆矩阵的对角线元素)；p_{ij} 为真实标签为类别 i 但被预测为类别 j 的像素数量，这里 i 和 j 可以是同一个类别，也可以是不同的类别；p_{ji} 为真实标签为类别 j 但被预测为类别 i 的像素数量，这里 j 和 i 可以是同一个类别，也可以是不同的类别。

模型整体表现稳定，平均 IoU 为0.69，PA 为0.76，其中“天空”“水体”等类别识别最准确($IoU\approx0.82\sim0.83$)，体现了模型在

城市场景解析中的可靠性。

随后，计算各要素在不同视频帧中的平均比例与标准差，并将这些特征向量输入K-means聚类算法，自动识别景观意象主题。聚类数量通过轮廓系数确定，结果被解释为4类城市景观意象：现代都市、传统建筑、自然景观与消费空间。

为验证聚类效果，研究采用人工标注样本与多模态模型Gemini 1.5的分类结果进行对比，并以Precision，Recall与F1-score衡量一致性。

$$Precision=\frac{TP}{TP+FP}$$，

式中：$Precision$为精确率，衡量模型预测为正类的样本中实际为正类的比例；$TP(True\ Positives)$为真正例，指模型正确地预测为正类的样本数量/个；$FP(False\ Positives)$为假正例，指模型错误地预测为正类的样本数量/个，但实际上这些样本属于其他类别。

$$Recall=\frac{TP}{TP+FN}$$，

式中：$Recall$为召回率，衡量所有实际为正类的样本中，被模型正确预测为正类的比例；$FN(False\ Negatives)$为假负例，指模型错误地预测为负类的样本数量/个，但实际上这些样本属于正类。

$$F_1-score=\frac{2\times Precision\times Recall}{Precision+Recall}$$，

式中：$F_1-score$为综合评估分类模型性能的指标，是一个调和平均值，用于平衡精确率和召回率。

模型预测结果表明，$F_1-score$平均达0.94，显示聚类结果具有较高的稳定性和语义可解释性。

3）社会意象与活动空间识别。

社会意象强调人类活动对城市形象的动态塑造。通过对视频中行为场景的视觉特征与语义标签进行聚类，该研究提炼出4类社会活动主题：餐饮活动、休闲娱乐、户外旅游与公共服务。随后利用核密度估计法绘制空间分布图，揭示不同活动在上海市的空间聚集特征：

① 餐饮活动集中于南京路、淮海中路及武康路一带，反映上海市丰富的都市美食文化；

② 休闲娱乐活动形成以外滩—人民广场为核心的高密集区；

③ 户外旅游活动集中于滨江公共空间与公园体系，例如世纪公园、黄浦江两岸；

④ 公共服务活动集中在交通枢纽与医疗设施周边，这凸显了它们在城市功能中的核心地位。

这些结果揭示了城市生活的多元节奏与空间组织逻辑，也体现出公众认知中“主题化区域”替代行政边界的特征。

4）情感意象分析。

在情感维度上，该研究采用SnowNLP中文情感分析库，利用朴素贝叶斯分类模型对视频、音频转写及评论文本进行情感极性判断。

$$P(c_i|T)=\frac{P(T|c_i)\cdot P(c_i)}{P(T)}$$，

式中：$P(c_i|T)$为在情感类别c_i情况下观察到文本特征向量T的概率；T为文本特征向量；c_i为情感类别(正向或负向)，模型输出的情感值介于0~1之间。通过对2019—2022年长期时间序列进行分析，该研究发现：

① 上海市的总体情感倾向偏正，均值0.76，标准差0.36；

② 在2022年新型冠状病毒感染疫情期间，情感值出现了显著的负向波动；

③ 情感分布呈现极化特征，即正、负评价同时放大，反映出短视频平台在社会情绪表达中的即时性与放大效应。

与同期微博数据对比显示，两者情感趋势高度一致，说明短视频数据具有较高的代表性和社会感知敏感度。

研究结论

1）城市空间结构特征。

通过短视频的地理信息可视化分析发现，上海市城市意象呈现“核心集聚-外围扩散”的空间结构：

① 9个核心意象区主要分布在外滩、陆家嘴、人民广场、徐家汇等地标区域；

② 3条主要的城市发展轴线沿黄浦江两岸、世纪大道和中环线延展；

③ 10个重点意象区覆盖迪士尼、虹桥综合交通枢纽、张江高科技园区等区域，体现了城市功能、空间与象征意义的融合。

这一空间结构模式与林奇的城市意象理论高度契合，验证了“地标—节点—区块”对城市认知的深远影响。

2）景观意象的多元化结构。

通过视觉聚类，该研究识别出4类景观主题(见表1)。这一分类不仅展现了上海城市景观的多样性，也揭示了短视频中不同景

表1 4类景观主题

Tab.1 Four types of landscape themes

景观类型	特征要素	意象特征	典型区域
现代都市	高楼、交通、商业设施	国际化、现代感强	陆家嘴、浦东新区CBD
传统建筑	古建筑、水体、街巷	历史文化底蕴	豫园、外滩、老城厢
自然景观	植被、水系、天空	生态宜居	世纪公园、辰山植物园
消费空间	商业街区、公共设施	活力、时尚	南京路、新天地

观类型的主导地位。尤其是“现代都市”和“消费空间”占比最高，表明公众更倾向于通过影像表达现代化和消费文化特征。

3）社会意象与活动空间。

社会意象分析揭示了上海城市生活的活力与多中心性。餐饮与娱乐活动高度集中在市中心，呈现城市生活方式的浓缩空间；而户外旅游与公共服务活动则展现了城市功能网络的扩展性。这一分析从“人-活动-空间”的角度，丰富了传统城市意象研究中的物理空间认知框架。

4）情感意象的时序变化。

情感分析显示，公众对上海市的总体评价虽是积极的，但在重大社会事件(如新型冠状病毒感染疫情)期间，负面情感会显著增加。这种情感值的剧烈波动，揭示了城市集体情绪的脆弱性，同时也揭示了短视频平台在社会情绪表达中的放大效应。以疫情封控期间为例，情感值的显著下降，正是社会韧性监测中一个值得关注的潜在预警信号。

研究总结

1）理论贡献与方法创新。

理论拓展：该研究在林奇的空间可读性理论基础上，提出“空间-景观-社会-情感”4个维度城市意象框架。与传统研究主要关注城市物理结构不同，该框架强调了公众情绪和社会活动对城市形象的影响，拓宽了城市意象的研究视野。

数据创新：短视频的多模态应用。该研究创新性地将视频图像、文本、音频与地理信息结合，构建了多模态感知分析方法。这为未来基于用户生成内容(UGC)的城市研究提供了新的思路和范例。

方法创新：深度学习与情感分析的结合。该研究首次将U-Net语义分割、K-means聚类和朴素贝叶斯情感分析相结合，形成高效、精确的自动化城市感知分析流程。这一方法为城市意象研究提供了全新的技术支持。

应用启示：从认知到规划的反馈机制。通过识别情感热点和空间聚集区，该研究为城市公共空间设计与品牌传播提供了量化依据。积极情感区域可以作为景观优化和旅游推广的重点，而负向情感区域则为城市治理和环境改进提供了指引。

2）研究局限。

数据来源单一。本研究仅依赖抖音平台，受算法推荐和用户画像的影响，可能存在样本偏向。未来可以整合微博、小红书、哔哩哔哩等多平台数据，以提高数据的代表性和多样性。

空间覆盖不均。核心城区的影像较为密集，而郊区和工业区的覆盖较少，这可能导致城市意象的表达不够均衡。未来可以扩大数据采集范围，提升空间覆盖的全面性。

跨城市对比不足。该研究仅以上海市为例，缺乏与其他城市的数据对比。未来可进行跨城市比较，验证模型的普适性和可迁移性。

情感分析的细化不足。现有情感分析模型主要聚焦于正、负情感的分类，未能深入捕捉更细微的情感类型(如怀旧、惊喜、焦虑等)。未来可结合大语言模型，以提升情感识别的精确度与分析粒度。

3）研究展望。

未来研究可以沿以下方向展开：融合空间句法与社会网络分析，探索城市认知路径和情感传播机制；构建实时城市意象监测系统，支持城市品牌管理和应急响应；探讨短视频与虚拟现实和增强现实技术的结合，分析沉浸式城市体验。

综上，该研究推动了城市意象研究在数据来源、方法框架和理论体系上的重要进展；同时结合深度学习、情感分析和地理空间计算，展示了短视频如何从娱乐媒介转变为城市认知和公众情感分析的有力工具。该研究不仅验证了林奇经典理论在数字时代的适用性，也为城市规划、品牌建设和社会情绪管理提供了新的量化分析方法。随着多模态感知技术和大语言模型发展的推进，未来的城市意象研究将变得更加精准、灵活和以人为本。

资料来源：Chen Minxin, Wei Zhen, Cao Kai. Multi-perspective perception of city image in Shanghai via massive short videos[J]. Applied earth observation and geoinformation, 2025, 144: 104844.