A New Approach to Assessing Perceived Walkability:

Combining Street View Imagery with Multimodal Contrastive Learning Model

Xinyi Liu
Department of Civil, Environmental and Geomatic Engineering; The Bartlett Centre for Advanced Spatial Analysis
University College London
London, The United Kingdom
x.liu.22@ucl.ac.uk

James Haworth
SpaceTimeLab; Department of Civil, Environmental and Geomatic Engineering
University College London
London, The United Kingdom
j.haworth@ucl.ac.uk

Meihui Wang
SpaceTimeLab; Department of Civil, Environmental and Geomatic Engineering
University College London
London, The United Kingdom
meihui.wang.20@ucl.ac.uk

ABSTRACT

Walkability is becoming increasingly important in urban planning, public health, and environmental protection. Traditional assessment tools like streetscape images and semantic segmentation focus on objective factors, while questionnaires as the main tool for perceived walkability are limited by cost and scale. This study introduces a new method using the Multimodal Contrastive Learning Model, CLIP, to assess perceived walkability by analysing both tangible and subjective factors such as safety and attractiveness. The method compares perceived with physical walkability by scoring street view images with a customized scale. Initial results indicate CLIP can identify pedestrian-friendly streetscapes that might score low on physical metrics. While its accuracy needs more evaluation, CLIP offers a cost-effective alternative without needing extensive labelled datasets. This method can be combined with objective pedestrian assessment methods to serve as reference information for various industries such as real estate, transportation planning, and tourism.

CCS CONCEPTS

• Deep learning • Vision-Language Model • Street View Imagery

KEYWORDS

Perceived Walkability, Zero-shot Learning, Street View Imagery, Vision-Language Model

ACM Reference format:

Xinyi Liu, James Haworth, and Meihui Wang. 2023. A new approach to assessing perceived walkability: Combining Street View Imagery with Multimodal Contrastive Learning Model. In *2nd ACM SIGSPATIAL International Workshop on Spatial Big Data and AI for Industrial Applications (GeoIndustry '23), November 13, 2023, Hamburg, Germany. ACM, New York, NY, USA, 4 pages.* https://doi.org/10.1145/3615888.3627811

GeoIndustry '23, November 13, 2023, Hamburg, Germany

ACM ISBN 979-8-4007-0350-8/23/11.
https://doi.org/10.1145/3615888.3627811

1 INTRODUCTION

As global urbanization accelerates, walkability has moved beyond a purely transportation function to become a function of community connectivity, public health, and environmental protection. More cities are beginning to promote walking in practical ways, pushing for pedestrian-friendly neighbourhood environments.

Research on walkability has focused on objective environmental factors of walking. Early studies identified mesoscale factors such as residential density and land use as crucial determinants of walking behaviour[7]. Technological advances such as streetscape imagery and semantic segmentation have enabled measuring street-scale peripatetic features [8, 11]. However, walking behaviour is also influenced by subjective walking intentions. For example, some narrow neighbourhoods may have actual high walkability due to cultural attractiveness but perform poorly on objective metrics (e.g., sky openness, degree of greenery, percentage of sidewalks, etc.). Therefore, perceived walkability should be considered in addition to the objective context when considering overall walkability.

The current method of assessing perceived walkability is primarily questionnaires, and relatively authoritative questionnaire scales have been developed in this area such as NEWS and its derivatives [4, 13, 14], LWI[10], and PANES[3]. These scales have been widely recognized and used. Streetscape imagery has also played a role in this area. The Place Pulse dataset, released by MIT Media Lab, is a pairwise comparison dataset collected through web-based research. Version 1.0[15] contains three subjective dimensions, and version 2.0[5] contains six. However, Place Pulse is positioned as a dataset about urban perceptions and does not fully reflect willingness to walk. A recent study [9] used a similar approach to constructing the Place Pulse dataset to publish a street view image dataset about concerns about walking preferences in Jeonju City, South Korea, developing a deep learning model to assess perceived walkability. However, the generalization ability of this model has yet to be validated. To summarize, the limitations of previous studies are apparent. Regarding physical walkability, the accuracy of the semantic segmentation model also needs to be further improved, and the

walkability assessment cannot rely solely on semantic segmentation techniques. Regarding perceived walkability, the questionnaire survey method is challenging to apply widely due to geographical limitations, high time, and cost. Developing and training pairwise comparison datasets and derived deep learning models through web research are costly and cannot investigate the detailed factors affecting perception. Evaluating walkability by combining semantic segmentation models and object detection models is one solution idea, but the multimodal comparative learning model, CLIP, seems able to provide a more flexible and efficient solution. Furthermore, its zero-shot learning capability dramatically reduces the training cost of the model, offering the possibility of rapid deployment in a variety of urban scenarios.

2 CLIP-BASED ASSESSMENT METHODOLOGY

2.1 The Potential for Perceived Walkability Assessment

Traditional walkability assessment methods, like semantic segmentation and questionnaires, are hindered by issues of cost, efficiency, and adaptability in complex environments. In the domain of deep learning, while models like ViLBERT and LXMERT[2] attempt to integrate vision and language understanding, it's the CLIP model[12] that stands out due to its superior zero-shot learning capability. The power of the CLIP model lies in its contrastive learning approach, which capitalizes on existing knowledge to decipher complex urban dynamics without extensive labelled data. Having been pre-trained on numerous image-text pairs, CLIP can deeply understand semantic relationships. This makes it possible to evaluate urban scenes using natural language cues like "wide sidewalks" or "good walking facilities" to more accurately reflect real-world settings. The versatility of the CLIP model and its minimal data preparation needs set it apart from traditional methods. Nevertheless, challenges persist. These include handling systematic tasks such as counting and distance calculations, differentiating object types, and providing precise semantic similarity values for text-image pairings. Moreover, the model's sensitivity to phrasing means that iterative "just-in-time" optimization is essential for optimal performance. Despite its challenges, the CLIP model's adaptability and versatility ensure its relevance in the face of rapidly evolving urban environments.

2.2 Architecture and Training Approach of CLIP

The CLIP model consists of a visual encoder, often using Vision Transformer or ResNet, and a text encoder based on the Transformer architecture. During pre-training, these encoders align image and text pairs in a shared space by maximizing their similarity. This training allows CLIP to relate visual and textual data seamlessly. In inference, the model matches input labels to images, assessing the best fit. CLIP's design facilitates zero-shot learning, enabling it to handle new images without labelled data, reducing the need for vast datasets and increasing its versatility across tasks.

2.3 Perceived Walkability Assessment

Ewing's theory[6] suggests that willingness to walk is based on objective environmental factors. They argue that physical features can directly influence individual reactions or indirectly affect them through urban design qualities, ultimately determining the overall walkability. This provides a theoretical basis for the development of a perceptual scale based on detailed features of the physical environment in this study.

Figure 1 depicts the computational process for assessing perceived walkability scores. Constructing an assessment scale based on objective environmental factors is crucial in the process. The

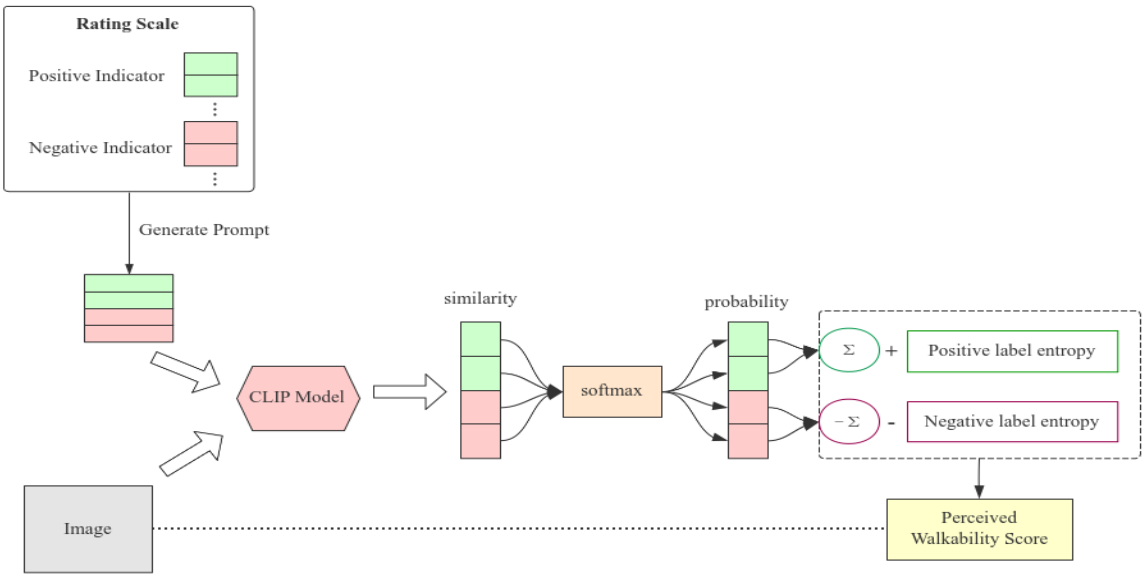

Figure 1 Flow Chart for Calculation Perceived Walkability Score

assessment scale customized for this study was based on positive and negative perceived walkability impact factors collated from previous research. These factors were uniformly rewritten as descriptive prompts and input to the CLIP model along with the street view images. CLIP calculates a similarity score for each image based on each prompt and transforms the score into a more interpretable probability distribution using the SoftMax function. In other words, the larger probability value obtained means that the image better matches the corresponding metric. To avoid the single dependency of the calculation results, the positive and negative indicator entropy is additionally introduced as an adjustment strategy. Its specific calculation is as follows:

$$H(X) = -\sum p(x_i)\log_2 p(x_i) \quad (1)$$

$$S = \sum_{i=1}^{N} d_i p(x_i) + H(X_{positive}) - H(X_{negative}) \quad (2)$$

In equation (1), $H(X)$ is the entropy of indicator, $p(x_i)$ is the probability that the image corresponds to the *i*-th indicator. In equation (2), d_i is the direction of the *i*-th indicator (+1 for positive direction and -1 for negative direction) and $p(x_i)$ is the image corresponding to the i-th indicator. $H(X_{positive})$ is the entropy of the positive indicator and $H(X_{negative})$ is the entropy of the negative indicator.

The purpose of this adjustment strategy is to give an additional reward to those images that satisfy multiple positive metrics simultaneously and to impose a corresponding penalty on those images that satisfy multiple negative metrics. It ensures that images with higher scores are not just higher because of a better match to one indicator but because of a better match to multiple positive indicators.

3 CASE STUDY

Mapillary is an open source street view image platform which provides image metadata[1] that allows users to filter images by timestamps. Experimental results of the method are shown for the Centrum district of Amsterdam. Street View images are collected at 30-meter intervals on the road network. A total of 5,669 images were collected. To ensure consistency in the urban landscape, the image was limited to April through October of each year, and the images were taken within the last five years. Figure 2 shows some examples of the perceived walkability score. The Top 3 Labels with Probability are the three labels in customized rating scale (Appendix 1) that the model thinks best fits/ describes the image.

In order to show more clearly the sensitivity of the present method to perceived factors, this study additionally calculated the physical walkability of the study area using a traditional semantic segmentation method – the DeepLab V3 model has been used here. Four main factors (visual crowdedness, greenery, sky openness and sidewalk ratio) with calculation formulas are shown in Appendix 2. The weights of the four indicators were assigned using hierarchical analysis (Appendix 3). The calculated physical

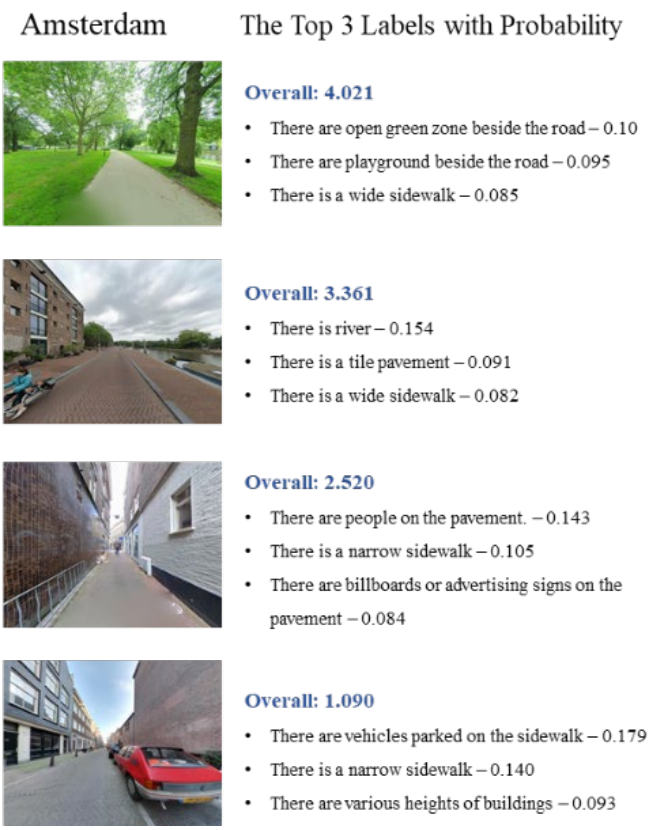

Figure 2 Examples of Perceived Walkability Calculation

and perceived walkability score results were visualized as a heat map (Figure 3).

As shown in Figure 3(a), the "ARTIS" zoo in southeast Amsterdam is an important physical walkability hotspot due to its vast open green spaces. Other hotspots are concentrated on major urban arteries and intersections, such as Amsterdam Central Station, while other areas have relatively similar walkability scores. In contrast, the Figure 3(b) shows a more complex distribution. In addition to the ARTIS Zoo, the city centre business district and the famous "red light district" are also perceived walkability hotspots. This suggests that the methodology captures micro urban features that influence perception, such as commercial activities and landmarks, which are difficult to detect in a purely physical assessment.

4 DISCUSSION AND CONCLUSION

The above results show that the perceived walkability approach proposed in this study has significant advantages in the following two aspects:

1. It can provide a more comprehensive assessment of walkability due to its ability to identify more details of the streetscapes, especially for streets where cannot be measured uniformly using objective metrics. This perceptual perspective-based approach to walkability assessment can

Figure 3 Heatmap of Physical (a) and Perceived (b) Walkability Result

complement the traditional objective walkability assessment approach based on semantic segmentation techniques by providing more detailed insights into the improvement of the walking environment from both subjective and objective dimensions.

2. It takes full advantage of the CLIP model's strengths in zero-shot learning and contrastive learning, greatly simplifying the process of model construction and dataset training. Meanwhile, the deterministic nature of the CLIP model ensures its robustness and reproducibility, meaning it consistently produces the same results for the same input, highlighting its reliability. This approach also avoids the time costs associated with traditional questionnaires, making the evaluation process more efficient.

Although this method demonstrates its potential, some aspects still need to be improved. First, the accuracy of the computational results is limited by the performance of the CLIP model, and a reasonable method needs to be developed to assess the performance of the model. In addition, there is still room for optimization of the evaluation scale of perceived walking ability. In addition to determining the direction of each indicator, its weight allocation can be refined to make the assessment results more targeted. Finally, considering the diversity of perception, specialized assessment scales can be designed for different populations and cultural backgrounds in the future.

Benefiting from the training strategy of contrastive learning, the CLIP model is equipped with zero-shot learning capability, thus demonstrating excellent generalization ability, allowing it to adapt to different urban scenarios. Considering its core features of efficiency and low cost, this lightweight assessment method can be considered for future integration into websites or applications. This not only provides a real-time walkability assessment tool, but also opens up new research directions in urban planning, transportation engineering, and other related industry fields.

It is worth emphasizing that the core idea of this method can be widely applied to a variety of perception studies based on the generation of objective factors as long as reasonable and scientific assessment criteria are developed.

REFERENCES

[1] Mapillary. 2023. API Documentation. https://www.mapillary.com/developer/api-documentation. Accessed: 2023-08-25.

[2] Cafagna, M. et al. 2021. What Vision-Language Models `See' when they See Scenes. arXiv.

[3] Calise, T.V. et al. 2019. Food access and its relationship to perceived walkability, safety, and social cohesion. Health Promotion Practice. 20, 6 (2019), 858–867.

[4] Cerin, E. 2007. Measuring perceived neighbourhood walkability in Hong Kong. Cities. 24, 3 (2007), 209–217. DOI:https://doi.org/10.1016/j.cities.2006.12.002.

[5] Dubey, A. et al. 2016. Deep learning the city: Quantifying urban perception at a global scale. In Computer Vision–ECCV 2016: 14th European Conference. Part I 14, (2016), 196–212.

[6] Ewing, R. and Handy, S. 2009. Measuring the unmeasurable: Urban design qualities related to walkability. Journal of Urban design. 14, 1 (2009), 65–84.

[7] Frank, L.D. et al. 2005. Linking objectively measured physical activity with objectively measured urban form: findings from SMARTRAQ. American journal of preventive medicine. 28, 2 (2005), 117–125.

[8] Hsieh, I.-H. et al. 2020. Outdoor walking guide for the visually-impaired people based on semantic segmentation and depth map. In 2020 international conference on pervasive artificial intelligence (ICPAI) (2020), 144–147.

[9] Kang, Y. et al. 2023. Assessment of Perceived and Physical Walkability Using Street View Images and Deep Learning Technology. ISPRS International Journal of Geo-Information. 12, 5 (May 2023), 186. DOI:https://doi.org/10.3390/ijgi12050186.

[10] Leyden, K.M. 2003. Social capital and the built environment: The importance of walkable neighborhoods. American Journal of Public Health. 93, 9 (2003), 1546–1551.

[11] Nagata, S. et al. 2020. Objective scoring of streetscape walkability related to leisure walking: Statistical modeling approach with semantic segmentation of Google Street View images. Health & Place. 66, (2020), 102428.

[12] Radford, A. et al. 2021. Learning transferable visual models from natural language supervision. In International conference on machine learning (Jul. 2021), 8748–8763.

[13] Rosenberg, D. et al. 2009. Neighborhood Environment Walkability Scale for Youth (NEWS-Y): Reliability and relationship with physical activity. Preventive Medicine. 49, 2–3 (2009), 213–218.

[14] Saelens, B.E. et al. 2003. Neighborhood-based differences in physical activity: an environment scale evaluation. American journal of public health. 93, 9 (2003), 1552–1558.

[15] Salesses, P. et al. 2013. The collaborative image of the city: mapping the inequality of urban perception. PloS one. 8, 7 (2013), 68400.

A APPENDICES

A.1 Perceived Walkability Rating Scale

No.	Index	Indicator	Prompt	Direction
1	Sidewalk condition	Obstacles	There are vehicles parked on the sidewalk	-1
2			There are scooters parked on the sidewalk	-1
3			There are bicycles, and motorcycles parked on the sidewalk	-1
4		Sidewalk width	There is a wide sidewalk	1
5			There is a narrow sidewalk	-1
6		Sidewalk construction	There is a fenced sidewalk	1
7			There is a heightened sidewalk	1
8		Sidewalk materials	There is a tile pavement	1
9			There are cracks, depressions, and flooded sidewalks.	-1
10	Traffic safety	Presence of traffic control devices	There is road name/direction signs	1
11			There is a pedestrian symbol marked on the pavement	1
12			There are green belts and fences between sidewalks and vehicle lane	1
13		Presence of pedestrian facilities	There is crosswalk	1
14			There is a pavement traffic light	1
15		Unkeep /maintenance issue	There are scaffolding or construction sites on the sidewalk	-1
16		Traffic flow	There are many vehicles on the road	-1
17			There are people on the pavement.	1
18		Train	There are railways along the road	-1
19		Lighting	There are streetlights on the pavements	1
20	Security	Crime possibility	There are police officers along the road	1
21			There are security cameras	1
22	Comfort	Building's height	There are skyscrapers	-1
23			There are various heights of buildings	1
24		Wall Graffiti	There is graffiti on the walls	-1
25		Cleaning	There are billboards or advertising signs on the pavement	-1
26			There are garbages on the road	-1
27		Tree and shadow area	There are trees	1
28			There are shadow areas	1
29		Seating facilities	There are benches along the sidewalks	1
30		Animals	There are unleashed dogs on the sidewalk	-1
31			There are birds	1
32	Attractiveness	Commercial zone	There are shops along the road	1
33			There are cafes along the road	1
34		Institutional zone	There are public buildings such as hospitals, schools, libraries, and office complexes	1
35		Parks and open sapces	There are open green zone beside the road	1
36			There are playground beside the road	1
37		Accessibility to public transportation	There are bus or underground stations	1
38		Landmark	There are landmarks	1
39		Landscape and nature	There is river	1
40			There are flowers	1

A.2 Physical Walkability Indicators and Formula

Index	Data	Formula*
Visual crowdedness	SVI	Σ('car'+'bicycle'+'truck'+'person'+'train'+'bus'+'motorcycle')/ 'total_pixels'
Sidewalk ratio	SVI	'sidewalk' / 'total_pixels'
Greenery	SVI	Σ('vegetation'+'terrain')/ 'total_pixels'
Sky openness	SVI	'sky' / 'total_pixels'

* 'xx' means the number of pixels in this label

A.3 Weighting Results of AHP analysis

Indicators	Eigenvector	Weights (%)	Eigenvalue	CI
sidewalk ratio	1.819	45.467	4.01	0.003
greenery	0.565	14.114		
crowdedness	1.052	26.305		
sky openness	0.565	14.114		