Computers, Environment and Urban Systems 65 (2017) 113–125

Contents lists available at ScienceDirect

Computers, Environment and Urban Systems

journal homepage: www.elsevier.com/locate/ceus

A machine learning-based method for the large-scale evaluation of the qualities of the urban environment

Lun Liu [a], Elisabete A. Silva [a], Chunyang Wu [b], Hui Wang [c,*]

[a] Lab of Interdisciplinary Spatial Analysis, Department of Land Economy, University of Cambridge, United Kingdom
[b] Machine Intelligence Laboratory, Department of Engineering, University of Cambridge, United Kingdom
[c] School of Architecture, Tsinghua University, China

ARTICLE INFO

Article history:
Received 13 August 2016
Received in revised form 30 May 2017
Accepted 14 June 2017

Keywords:
Machine learning
Physical quality
Street view image
Urban design
Architecture

ABSTRACT

Given the present size of modern cities, it is beyond the perceptual capacity of most people to develop a good knowledge about the qualities of the urban space at every street corner. Correspondingly, for planners, it is also difficult to accurately answer questions such as 'where the quality of the physical environment is the most dilapidated in the city that regeneration should be given first consideration' and 'in fast urbanising cities, how is the city appearance changing'. To address this issue, in the present study, we present a computer vision method that contains three machine learning models for the large-scale and automatic evaluation on the qualities of the urban environment by leveraging state-of-the-art machine learning techniques and wide-coverage street view images. From various physical qualities that have been identified by previous research to be important for the urban visual experience, we choose two key qualities, the construction and maintenance quality of building facade and the continuity of street wall, to be measured in this research. To test the validity of the proposed method, we compare the machine scores with public rating scores collected on-site from 752 passers-by at 56 locations in the city. We show that the machine learning models can produce a medium-to-good estimation of people's real experience, and the modelling results can be applied in many ways by researchers, planners and local residents.

1. Introduction

As a city grows, it becomes hardly possible for its dwellers, as well as planners, to gather a complete knowledge about how it looks at every street corner and in every narrow alley (Lynch, 1960). Theoretically, the human perception of the urban environment is inherently incomplete, discontinuous and distorted, as depicted by research on cognitive mapping (Downs & Stea, 1973) and the city's image (Lynch, 1960). It is especially the case, given the overwhelming size of modern cities. Questions such as 'which are the worst-looking places in the city where regeneration should be given first consideration' and 'in fast urbanising cities, how is the city appearance changing' are hard to answer.

For the past many years, several studies have attempted to measure a city's appearance in a consistent manner on a larger scale (Harvey, 2014). The dominant method is by sending human auditors to the field to observe and record the city's appearance (Brownson, Hoehner, Day, Forsyth, & Sallis, 2009). However, this method is quite limited in terms of sample size because its manual nature makes it inherently expensive and derives few economy of scale (Harvey, 2014). Recently, the availability of online street view images, which have an unprecedentedly wide coverage on the built environment, provides a new methodological opportunity for this topic (Dubey, Naik, Parikh, Raskar, & Hidalgo, 2016; Hara, Le, & Froehlich, 2013; Hwang & Sampson, 2014; Kelly, Wilson, Baker, Miller, & Schootman, 2013; Sun, Fan, Bakillah, & Zipf, 2015; Zhou, Liu, Oliva, & Torralba, 2014). When combined with computer vision techniques, there is a possibility for the large-scale automatic evaluation of various high-level judgements of the urban environment (Doersch, Singh, Gupta, Sivic, & Efros, 2012; Lee, Maisonneuve, Crandall, Efros, & Sivic, 2015; Naik, Philipoom, Raskar, & Hidalgo, 2014; Ordonez & Berg, 2014; Quercia, O'Hare, & Cramer, 2014; Salesses, Schechtner, & Hidalgo, 2013).

Our goal in this paper is to explore this possibility in terms of the physical quality of the urban environment. We refer to architectural and urban design theories (explained in Section 2) and choose two physical qualities, the construction and maintenance quality of building facade and the continuity of street wall, to be measured in this study. Beijing, a fast-growing city with quite diverse urban environment, is chosen as the case study area.

However, the use of street view images and computer vision is challenged by several issues in producing an appropriate estimation of people's real experience. First, we used the method of expert rating to label images and train the models. Although we attempted to make

* Corresponding author at: School of Architecture, Tsinghua University, Haidian District, Beijing 100084, China.
E-mail address: wh-sa@mail.tsinghua.edu.cn (H. Wang).

http://dx.doi.org/10.1016/j.compenvurbsys.2017.06.003

the rating standard as objective as possible, there may be a gap between the experts' opinions and the public's preference (e.g. in terms of what makes a good quality and what is considered a bad condition). Moreover, the expert rating is based on static and two-dimensional images instead of the on-site, dynamic, three-dimensional experience (e.g. in real settings, how people judge the immediate urban environment may be affected by what he/she experienced few seconds ago, so the experience is dynamic).

Regarding the validity of using street images in place of field survey, there have been a few studies that compare the results of observational field audits and street view image-based audits and show that there is generally an agreement between them (Hara et al., 2013; Kelly et al., 2013). However, what these studies deal with are usually quite objective and straightforward variables such as the building height and the existence of obstacles on the sidewalk, while the physical qualities that we are looking at are integrated judgements. To test the validity of our proposed method, we conducted a field survey on 752 passers-by at 56 locations in Beijing and compared the public's rating scores with the machine rating scores.

The research questions that we aim to explore include

- How is the performance of machine learning models in judging the physical qualities of the urban environment based on street view images? Is it possible to apply this method as a replacement to conventional labour-intensive manual audits?
- How is the correlation between image-based machine rating and the public's in-situ rating?

The results show that our machine learning models can reach a mean squared error (MSE) of 0.61 on the task of rating the construction and maintenance quality of building facade (rating scale: 1–4) and an accuracy of 75% on the task of judging the continuity of street wall. When compared with the public's in-situ rating scores, the street view image-based machine rating scores show a Spearman's correlation coefficient of 0.66 ($p < 0.0001$) with the public's rating scores on the former task and 0.71 ($p < 0.0001$) on the latter task.

The rest of the paper is organised as follows: Section 2 provides the conceptual framework; Section 3 reviews the long-lasting efforts in measuring the qualities of urban environment and the recent progresses in applying machine learning on extracting high-level information from city images; Section 4 explains the definitions and impacts of the physical qualities modelled in this study; Section 5 introduces the data and methodology; Section 6 presents the performance of the machine learning models and the validation results and the urban physical quality maps of Beijing produced from the model results; Section 7 concludes and discusses the potential directions of research.

2. The conceptual framework

The research question tackled in this paper stems from the larger conceptual framework that links the objective physical environment with individual's subjective experience (Fig. 1). The framework is based on the notion that specific physical features of buildings are mediated by a number of more abstract qualities and then the perceptual processes to shape the experience in the urban space. Wohlwill (1976, p. 108) argued that affect has often been found unrelated to individual physical features unless features are combined in a more meaningful composite measure, which makes the physical qualities such as order and enclosure. However, unlike the specific physical features such as building height and width, these qualities are not easily measured directly with a physical measure (Nasar, 1983). Physical measures of different parts of the scene would have to be combined to arrive at visual prominence (Nasar, 1983). The conceptual framework points to several issues in relation to a meaningful understanding of the physical environment: what are the key qualities that affect people's perceptions,

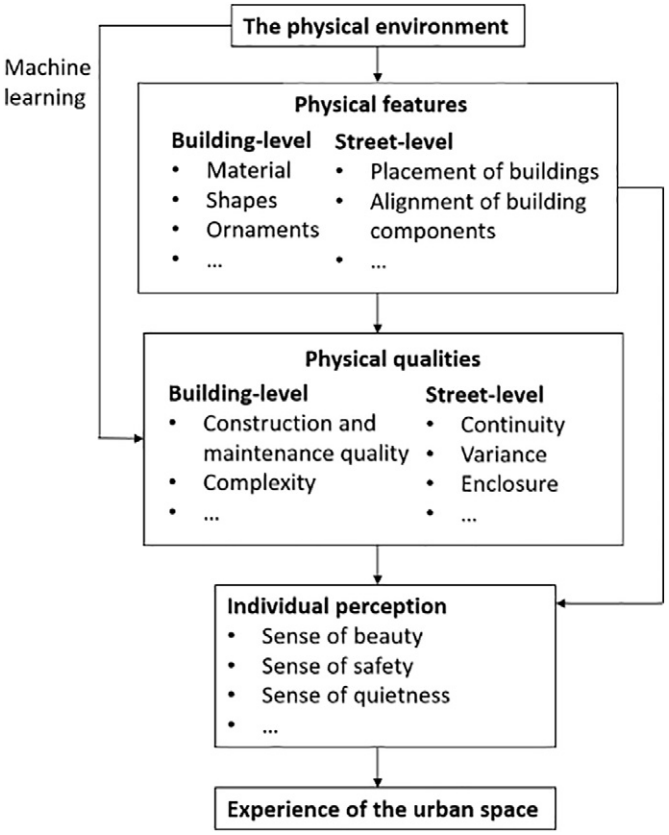

Fig. 1. The conceptual framework.
Partly adapted from Ewing and Handy (2009).

how can these qualities be measured from raw materials of the physical environment and how do they impact perceptions.

The fields of architecture and urban design have made many efforts in identifying the key qualities that contribute to people's experience. For instance, Moughtin (2003, p. 59) wrote that 'order, unity, balance, symmetry, scale, proportion, rhythm, contrast and harmony are among the important tools used to define good architecture'. In urban design, rules of enclosure, coherence, variety and so on are widely acknowledged and discussed in many design handbooks as well as governments' design codes (see for instance, Ewing et al., 2013, p. 8; American Planning Association, 2006, p. 165; Parolek, Parolek, & Crawford, 2008, p. 41 for the narratives on enclosure).

Our work focuses on the second question mentioned above and aims to explore the potential of machine learning algorithms in measuring the physical qualities from street view images. Our approach is different from relevant works by Quercia et al. (2014) and Ordonez and Berg (2014), which directly measured people's perceptions from street images. While appreciating their works, we argue that our approach is of particular importance for at least two reasons. First, the physical qualities are more operational than perceptual variables for urban planning and design practice, which themselves point to specific measures to improve. Second, our approach could facilitate further research on the relationship between physical qualities and human perceptions by providing consistently measured inputs.

3. Related works

3.1. Measuring the qualities of urban environment

Over the last four decades, there have been constant efforts in measuring the physical qualities of the urban environment that would potentially be perceptually meaningful. According to Stamps (2000, p.

preface), there had been '275 relevant empirical studies, covering over 12,000 stimuli and more than 41,000 respondents' by 2000. In a more recent review on this topic, Ewing and Handy (2009) found 51 perceptual-related qualities that were analysed.

Wohlwill (1976, p. 61) referred to two ways of measuring physical attributes of the environment: the physical approach and the judgemental approach. Because the perceptual-related qualities are usually more qualitative ones, they are more commonly measured using the judgemental approach, which resorts human judges to assess these qualities (Nasar, 1983; Wohlwill, 1976, p. 61). In an attempt to quantify and operationalise these human-judged qualities, Ewing and Handy (2009) employed an innovative method and produced a series of models linking concrete physical features with more abstract perceptual qualities. Our work is similar to that of Ewing and Handy in that we also aim to model the expert judgements on the qualities of urban environment based on very basic attributes. However, the proposed big-data-based machine learning method can be more automatic and labour saving.

3.2. Applying machine learning algorithms to urban image

Previously, most computer vision algorithms related to places focused on technical tasks such as scene classification or parsing scene images into constituent objects and background elements (Madhavan et al., 2006; Ordonez & Berg, 2014). Building upon that, a few interesting research studies into the perceptual and cultural aspects of urban images have emerged in recent years.

In the seminal work of '*What makes Paris look like Paris*', Doersch et al. (2012) dealt with the identification of local architectural identity by proposing a discriminative clustering approach that automatically discovers geographically representative elements from Google Street View images. With regard to that, there are also studies on the automatic classification of architectural styles by capturing the morphological characteristics, which can be further applied to the identification of architectural style mix and style transformation over time (Goel, Juneja, & Jawahar, 2012; Lee et al., 2015; Shalunts, Haxhimusa, & Sablatnig, 2011, 2012; Xu, Tao, Zhang, Wu, & Tsoi, 2014).

The most relevant works to the present study are those that aim at understanding people's perceptions of urban scenes, which are usually analysed by crowd-sourcing rating on urban images. Quercia et al. (2014) identified several aesthetic informative elements that positively (e.g. the amount of greenery) or negatively (e.g. broad streets, fortress-like buildings) affect people's perception of beauty, quietness and happiness. Ordonez and Berg (2014) modelled the perception for wealth, uniqueness and safety judged from street view images and validated the results against local income and crime statistics. The perception of safety was also modelled by Naik et al. (2014) and Porzi, Rota Bulò, Lepri, and Ricci (2015) and was proved to be consistent with the actual socio-economic indicators (Naik, Kominers, Raskar, Glaeser, & Hidalgo, 2015).

4. Physical qualities selected for this study

More specifically, we select one building-level quality (construction and maintenance quality of building facade) and one street-level quality (continuity of street wall) to be modelled in this analysis. We do not mean to argue that the two selected qualities are the most important or the most suitable. They are just used as the starting points for this line of research, which can be extended to include other qualities in the future. However, these two qualities are slightly more advantageous in that their perceptual implications are generally more straightforward and easier to interpret. For instance, too much contrast may produce disorder and lack of clarity (Moughtin, 2003), while to the authors' knowledge, there is hardly any argument that high construction and maintenance quality or high level of enclosure could have a negative perceptual impact.

4.1. Building level: construction and maintenance quality of building facade

The building-level quality measured in the present analysis is the construction and maintenance quality of building facade. The term 'construction and maintenance quality' is more commonly used in the context of engineering (Atkinson, 2003, p. 4; Brandt & Rasmussen, 2002). In our analysis, we shift the focus of this term away from the engineering domain and emphasise the specific elements that would affect the final appearance of the building facade. The construction- and maintenance-related elements that contribute to the appearance of building facade include

- **Building material**: whether the materials used are of high quality and fine textured;
- **Industrial precision and craftsmanship**: whether the facade is carefully constructed with high level of industrial precision and craftsmanship;
- **Maintenance**: whether the facade is free from cracks, bulges, broken components, deterioration, corrosion, dirt and stain, hanging wires, messy add-ons, etc.

Although this quality seems to be technical oriented, its impacts are not limited to the technical realm. In the book 'Sense of Beauty', Santayana (1955, p. 51) wrote highly of the aesthetic importance of material, saying that 'the beauty of material is thus the ground work of all higher beauty'. Leading modern architects such as Walter Gropius, Le Corbusier and Mies van der Rohe were inspired by what they saw as the great beauty of technical perfection (Voordt & Wegen, 2005). The famous saying of 'God is in the details' is also a reminder of the importance of technical perfection on the overall architectural quality. Dilapidation in the environment has been found to be related to negative affect frequently, and there is no compelling reason to expect different results (Nasar, 1983).

Furthermore, the construction and maintenance quality of building facade also has obvious social effects. According to the famous theory of 'broken window' on urban appearance and social effects, neighbourhood appearances drive the reality of neighbourhood safety: one broken window leads to another broken window and, in turn, to future crimes (Quercia et al., 2014). In less extreme situations, deterioration in the physical environment may not necessarily lead to crime but may very possibly affect the image and identity of a place (Said, Zubir, & Rahmat, 2014) and the economic development potential available to it. Therefore, modelling results can not only help understand the physical conditions of the urban space but also help identify areas vulnerable to social disorder and economic deprivation.

4.2. Street level: continuity of street wall

The street wall refers to the interface formed by building facade along a street. A continuous street wall is formed when buildings stand directly on the edges of their parcels (Lehnerer, 2009, p. 28). To be more specific, a continuous street wall requires the following:

- No 'dead spaces' between buildings, which include vacant lots, parking lots, drive ways or setbacks of a large building (Ewing & Handy, 2009)
- No solid and blank wall blocking the sight and activities from the street to the buildings, specifically in the context of China where most residences and work compounds are gated and surrounded by walls. However, if the wall itself is carefully designed and visually attractive, it may also be perceived as a continuous flow of the street interface.

Psychologically, a continuous street wall offers 'a sense of enclosure' (Ewing et al., 2013) and 'majesty and controlled uniformity' (Lyon,

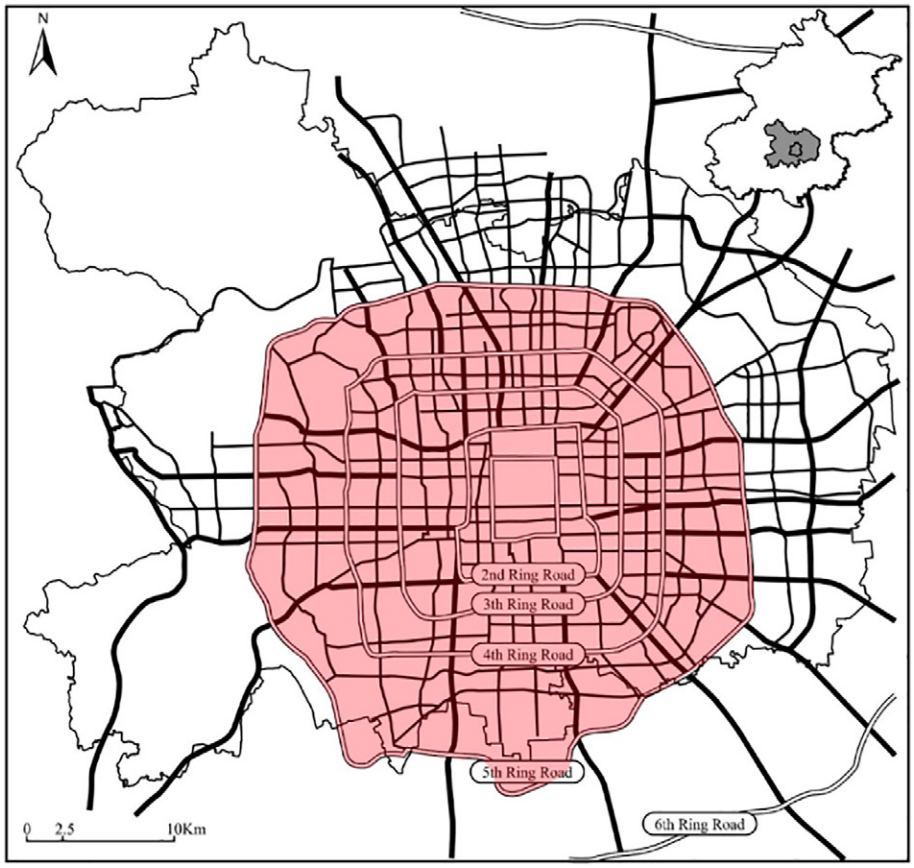

Fig. 2. Base map of Beijing (red coloured area indicates the study area).
Source: adapted from Zhao (2011).

1978). It positively affects the experience of urban space by 'giving a psychological security' (Lang, 1994, p. 324), 'instilling a sense of position, of identity with the surroundings' and 'embodying the idea of hereness' (Cullen, 1961, p. 29). Behaviourally, it draws pedestrians and activities and 'sustains a vital urban district' (Marcus & Francis, 1997, p. 19), and hence, it is considered to be one of the key rules for place making (Bain, Gray, & Rodgers, 2012, p. 7).

As early as the 15th century, relevant rules had appeared in street design codes in Nuremberg, Germany, which required buildings to be lined up to create an 'undeviating building line' (Kostof, 1999). Presently, it is addressed in numerous planning codes and guidelines, e.g. the American Planning Association (APA) Planning and Urban Design Standards requires infill projects to 'maintain ground floor facade to define a consistent street edge' (American Planning Association, 2006).

5. Data and methodology

5.1. Case study area

We chose Beijing as the case study area, which has undergone dramatic transformation from the imperial capital to the administrative centre, and even at present, to a hotspot of global investment. The cityscape is a complex mosaic of traditional and super-modern malls and giant structures. In addition, its rapid expansion in the recent decade has resulted in considerable amount of poorly constructed buildings at the urban fringe, where the cityscape is much different from that of the city centre. The highly diversified physical environment makes Beijing a vivid example for our analysis. We focused on the area within the 5th ring road, which covers most of the built-up areas (Fig. 2). The study area is approximately 670 km^2 and resides around 10.54 million people

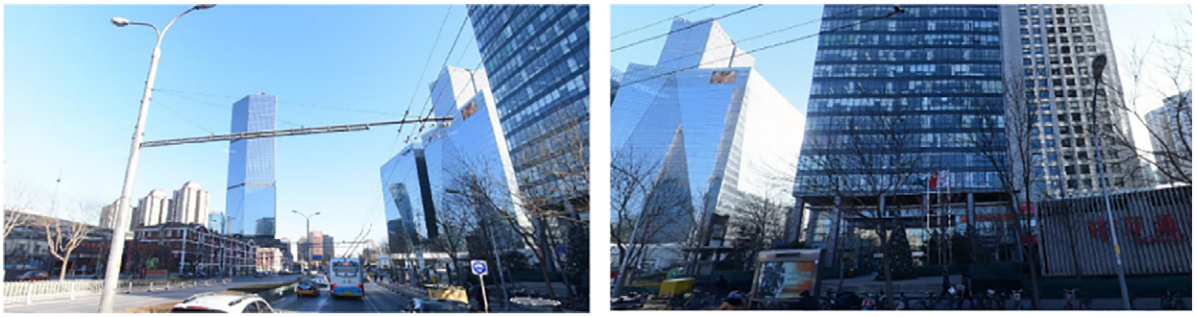

Fig. 3. Camera facing the street (left) and facing the buildings (right).

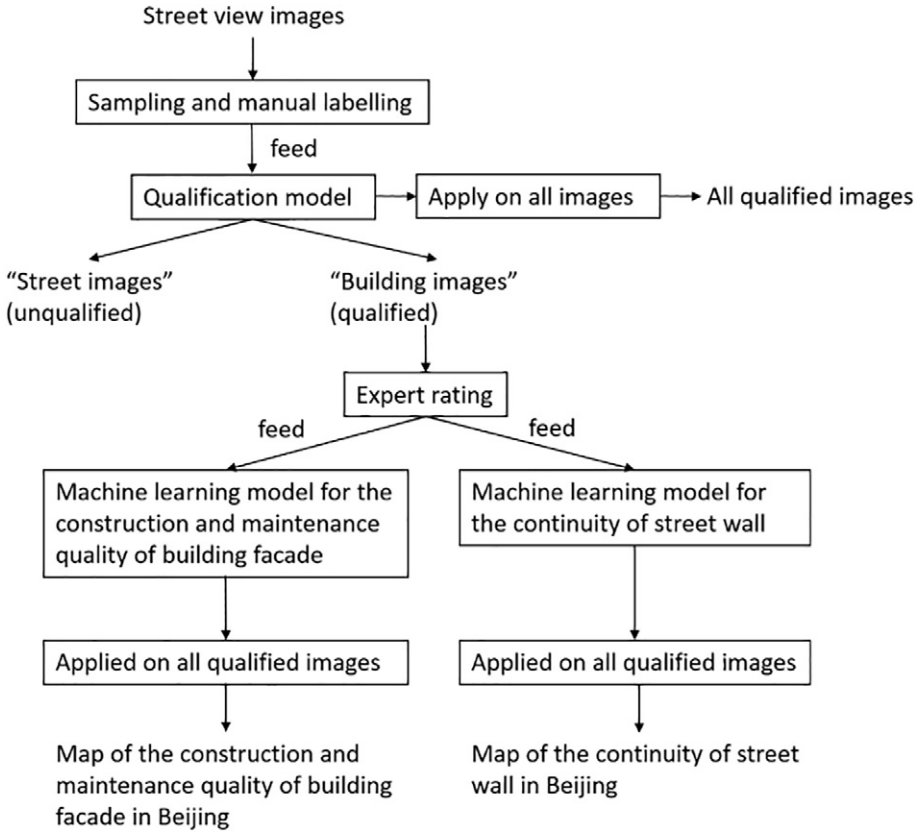

Fig. 4. Work flow diagram.

(Beijing Municipal Bureau of Statistics & National Bureau of Statistics Survey Office in Beijing, 2015).

5.2. Data and framework

We used street view images obtained from Baidu Map, the Chinese equivalent of Google Map. The images were requested at an interval of 200 m along all the streets in the city in February 2016, resulting in 360,796 images (800 * 500 pixels). Different from most existing studies that focused on the entire streetscape and used images taken with the camera facing the street, we emphasised more on the building facade and set the camera facing the buildings so that the buildings cover a larger proportion of the image (see Fig. 3). However, approximately 30% of the images were still streetscape images, which were taken around street corners or entrances. Therefore, a machine learning model was developed to discern streetscape images from building images to screen out unqualified images.

We followed a two-step approach to develop the machine learning models and three models are developed in total (see Fig. 4). In the first step, we randomly sampled 3500 images from the database and manually labelled them as 'building images' (2575) and 'street images' (925) as shown in Fig. 3. These images were then used to train a 'qualification' model to decide whether the content of an image is appropriate to be included in the analysis. In the next step, the qualified 'building images' were labelled through expert rating on the two qualities. The two scores were then fed to develop the models of construction and maintenance quality and continuity. We then applied the two models on all the qualified images from the entire study area.

Table 1
Rating standard for the construction and maintenance quality of building facade.

Ratings	Rating standard
Four points	Built with high quality, fine-textured materials; Built with high industrial precision or fine craftsmanship, e.g. building components and material pieces are well aligned, small gaps between material pieces unless they seem to be designed wide, etc.; Well maintained without obvious cracks, breakage, corrosion, dirt and stain or messy add-ons such as rusty iron rails on windows, hanging/loose wires
Three points	Built with lower quality, not very fine-textured materials; Do not show high level of industrial precision or craftsmanship, e.g. material pieces may not be well aligned and may have wide gaps in between; May have a few obvious cracks, breakage, corrosion, dirt and stain or messy add-ons but generally present a neat and clean look
Two points	Built with low quality, not very fine-textured materials; Built with low-level industrial precision or craftsmanship; Show a lot of cracks, breakage, corrosion, dirt and stain or messy add-ons
One point	Built with low-quality materials, in many cases, bare cement and colour plate[a]; Built with low-level industrial precision or craftsmanship, sometimes seem unfinished; Seriously deteriorated with a lot of cracks, breakage, corrosion, dirt and stain or messy add-ons

[a] We do not mean that these two materials are in themselves of low quality, but they are often used in low-quality buildings in Beijing.

Table 2
Rating standard for the construction and maintenance quality of building facade.

Ratings	Rating standard
Continuous	Building facades progress through the image without any interruption, blockage or significant setback, at least at the eye height.
Discontinuous	There is a wide gap between two adjacent buildings. There is a significant setback of a wide building. There is a solid wall blocking the building from the street; however, if the wall is carefully designed and visually attractive, it can be considered continuous.

5.3. Expert rating

As mentioned in Section 3.1, expert ratings have been frequently employed in research that involves the measurement of qualities of the urban environment. The judgemental approach was considered a simple way of measuring the qualities here (Nasar, 1983). Despite the element of subjectivity in the rating scale and the categorisation methods that are relied upon in this approach, the reliability of the resulting values has generally been found to be acceptable and, in some cases, quite high (Wohlwill, 1976, p. 63).

Ideally, experienced experts in the field should be invited to make judgements. However, given the size of the task in our research (each expert needs to rate several hundreds of images), it was difficult for us to invite experienced architects, urban designers or scholars to do this job. Therefore, we chose to recruit eight graduate students who have received architectural training for >5 years to accomplish this task. Although the validity of expert rating is supported by the virtue of their specialised expertise (Ewing & Handy, 2009) and there is usually little reason to expect that their assessments would differ systematically from other such professionals (Nasar, 1983), we also took extra measures to reduce potential bias as much as possible. First, we held a training and discussion session with the recruited students to make an agreement on the rating standard for each quality (Tables 1 & 2, Fig. 5), which linked the judgement of the qualities with more concrete features. Second, we held a practice session in which all students rated a same sample group of images until in most cases they made same judgements.

5.4. Machine learning

In the field of computer vision, there are many approaches for image representation. For our work, we evaluate three features: the conventional SIFT histogram (Lowe, 1999) and two state-of-the-art deep convolutional networks, namely AlexNet (Krizhevsky, Sutskever, & Hinton, 2012) and GoogLeNet (Szegedy et al., 2015). AlexNet and GoogLeNet outperformed all other features in the 2012 and 2014 ImageNet Large Scale Visual Recognition Competition, respectively. Compared with conventional image techniques, which are dominated by low-level features such as edges and corners, the deep convolutional networks can capture both local- and high-level image characteristics. We used the output of the last hidden layer of the two pre-trained neural networks and trained a SVR (Support Vector Regression) classifier for each of the scene attributes.

The labelled data set was randomly sampled into three subsets: the training set, the development set and the test set. For each task, the development set and the test set were equally and randomly sampled in each labelled class, and the rest of the images were used as the training set. For example, for the visual quality task, 40 images were randomly sampled in each of the four scoring groups for the development set and 60 images each for the test set. The hyper parameters of SVM, namely the regularisation constant and the regression epsilon width, were optimised through grid searching on the development set. In terms of the evaluation of model performance, we used F1 score for the classification models (the qualification model and the continuity model) and MSE for the construction and maintenance quality model, which were calculated using the following equations:

$$Recall = \frac{TP}{P}$$

$$Precision = \frac{TP}{TP + FP}$$

$$F1 = \frac{2TP}{2TP + FN + FP} = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

$$MSE = \frac{1}{n}\sum (y_i - t_i)^2$$

where P (positive), TP (true positive), FP (false positive) and FN (false negative) denote the number of the images that are qualified/continuous, both labelled and predicted to be qualified/continuous, labelled unqualified/discontinuous but predicted to be qualified/continuous and labelled true but predicted to be false, respectively, and y_i and t_i denote the machine rating and expert rating for each image, respectively.

The models with the best performance for the three tasks were to be applied to the entire image database of the research area. We then calculated the average scores for each street segment and produced the urban physical quality maps of Beijing.

5.5. Validation survey

As mentioned in the Introduction section, to test the validity of the proposed method, we conducted a field survey to collect the public's

Fig. 5. Rating examples.

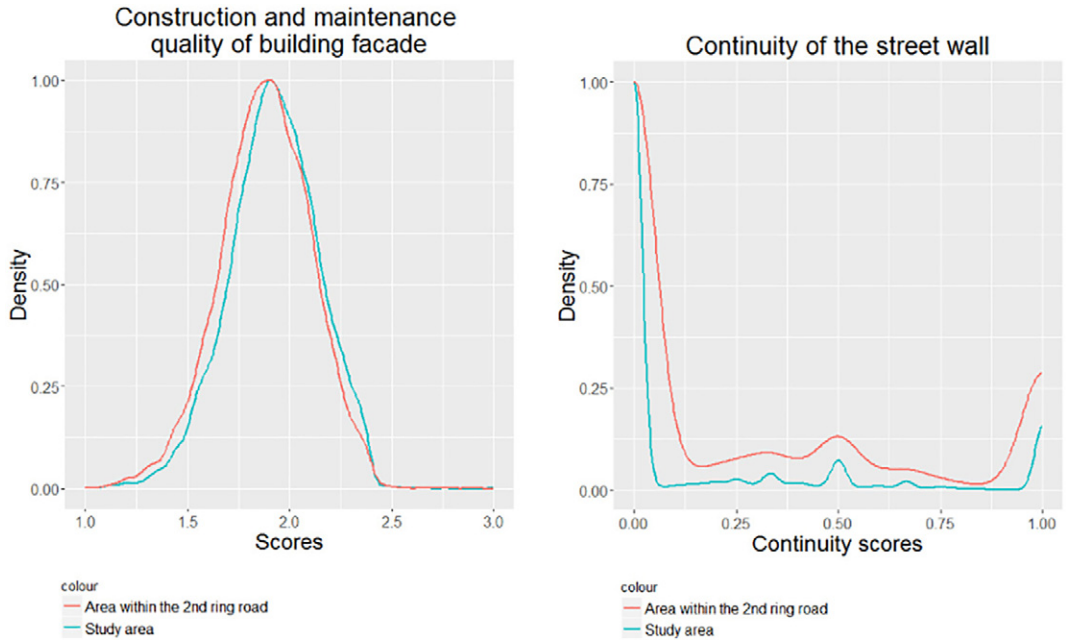

Fig. 6. Density distribution of machine scores within the second ring road and in the entire study area.

in-situ opinion on the two physical qualities and compared the results with the machine scores. The survey was conducted on 56 street segments in March 2016. The street segments were sampled from the area within the second ring road, which is the most diverse area of the city in terms of physical environment. We sampled the street segments from this area instead of the entire city for the following reasons. First, the area chosen is a relatively small area; therefore, the efficiency of the survey can be largely enhanced. Second, it is a good representative of the physical environment of the city because all types of physical environment (traditional vs modern, brand new vs dilapidated, high density vs low density, etc.) can be found within this area. The distribution of the machine scores on the two qualities within the second ring road does not deviate much from that in the whole city (Fig. 6), and the distribution of the continuity scores is even more balanced. To avoid the potential bias on people's judgement caused by different architectural styles (traditional vs modern), half of the street segments were sampled from traditional areas, and the other half were from modern areas.

Eight surveyors were recruited to help with the survey. Each of them was assigned six to eight street segments. To reduce the bias caused by demographic differences, the surveyors were required to keep a balance in the demographic profile of their interviewees in accordance to the demographic distribution of the whole city (Beijing data from the sixth National Population Census, Table 3). However, for each street segment, because of the relatively small sample size, only the balance in the distribution of gender and age was required. From each street segment, 10–15 interviewees were surveyed, and the total sample size was 752. For validation, the Spearman's correlation coefficient was calculated for the machine scores and the average public rating score for each street segment.

Table 3
Descriptive statistics of the interviewees.

Variables	Frequency	%Share	%Share (from the 6th National Census)
Gender			
Male	377	50.13	51.6
Female	375	49.87	48.4
Age			
<18	62	8.24	8.6 (0–14)
18–40	272	36.17	82.7 (15–64)
41–60	254	33.78	
60+	164	21.81	8.7 (65+)
Residence			
Beijing resident	249	66.89	
Visitor	503	33.11	
Education			
Elementary school and under	39	5.19	10.0
Junior school	177	23.54	31.4
High school and equivalent	267	35.50	21.2
Bachelor's degree and equivalent	252	33.51	31.5 (included above)
Master's degree and above	17	2.26	

6. Results

6.1. Results of expert rating

In terms of the construction and maintenance quality, the expert rating returns 485 four-point images (18.8%), 1079 three-point images (41.9%), 809 two-point images (31.4%), and 202 one-point images (7.8%). In terms of continuity, the expert rating identifies 1069 'continuous' images (41.5%) and 1506 'discontinuous' images (58.5%) (Table 4).

Table 4
Distribution of expert rating.

Rating criteria	Proportion%
Qualification	
Qualified	73.6
Unqualified	26.4
Total	100
Construction and maintenance quality	
4 points	18.8
3 points	41.9
2 points	31.4
1 point	7.8
Total	100
Continuity	
Continuous	41.5
Discontinuous	58.5
Total	100

Table 5
Performance of the qualification model.

	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
SIFTHist + SVR	79.2	45.1	71.3	55.2
AlexNet + SVR	89.3	**48.2**	85.9	61.8
GoogLeNet + SVR	**90.0**	48.1	**86.3**	**61.8**

Bold numbers indicate the highest performances.

Table 7
Performance of the visual continuity model.

	Accuracy%	Precision%	Recall%	F1%
SIFTHist + SVR	72.0	45.0	72.0	55.4
AlexNet + SVR	**75.0**	**48.0**	**72.0**	**57.6**
GoogLeNet + SVR	**75.0**	**48.0**	**72.0**	**57.6**

Bold numbers indicate the highest performances.

6.2. Machine learning performance

Table 5 shows the performance of the SIFT, AlexNet and GoogLeNet features on the test set of the qualification task. The deep convolutional networks, AlexNet and GoogLeNet, performed better than the traditional SIFT features. GoogLeNet achieved a slightly higher F1 score than AlexNet, which indicated a more balanced performance between recall and precision. Tables 6 and 7 show the performance on the other two tasks. Similar to the task of qualification, deep features outperformed the SIFT features. GoogLeNet showed the best capability of generalisation with the lowest MSE on the development set on the task of construction and maintenance quality. GoogLeNet and AlexNet showed almost the same levels of capability on the task of continuity. On the basis of these results, we chose the GoogLeNet model for large-scale application.

To better estimate the capability of the models, we took a closer look at the machine rating results on the test sets and compared with the expert rating scores. Fig. 7 shows that the machine scores generally fall into a narrower range than the expert rating scores (average score of 'one-point' images = 2.0, of 'two-point' images = 2.3, of 'three-point' images = 2.9, and of 'four-point' images = 3.4). There were a number of overlaps between the machine scores within the groups of low-quality building facades (the 'one-point' and 'two-point' images) and high-quality facades (the 'three-point' and 'four-points' images). However, there was little overlap between these two big groups (the lower quartile of the machine scores for 'three-point' images was higher than the higher quartile of the machine scores for the 'two-point' images). The results indicate that the model performs well in discriminating high quality from low quality but that it tends to produce more errors in identifying the nuanced differences within the two big groups.

Regarding the task of continuity, we manually analysed 60 images that were wrongly classified (false positive or false negative). Two major types of errors for false positive and three major types of errors for false negative were identified (Fig. 8). For false positive, the major types of error were failing to identify an unattractive wall that disrupts the continuity (12%) and failing to identify the gap between buildings because of perspective (76%). For false negative, the major types of error are failing to identify dilapidated buildings as a building (20%), failing to identify a continuous street wall because of blockage by trees and cars (30%) and failing to identify a continuous street wall when the picture was taken from a distance (35%), usually from the opposite side of a wide street. These errors are mainly because of the lack of labelled data to train the model to be aware of relevant situations. Although the total number of labelled images was >2000, when it comes to a very specific type of situation, the relevant sample size could be <50. Therefore, the model performance may be further enhanced by collecting more labelled datax.

Table 6
Performance of the visual quality model.

MSE	Training set	Development set	Test set
SIFTHist + SVR	0.36	0.84	0.84
AlexNet + SVR	**0.22**	0.64	**0.62**
GoogLeNet + SVR	0.28	**0.61**	0.64

Bold numbers indicate the highest performances.

6.3. Validation

The distribution of machine scores and survey scores on the two measured qualities for the sample street segments is shown in Fig. 9. To validate the models, a correlation analysis was performed, which showed that the machine scores and survey scores were moderately-to-highly correlated for both physical qualities (Spearman's $r = 0.66$, $p < 0.0001$ for the construction and maintenance quality of building facade, Spearman's $r = 0.71$, $p < 0.0001$ for the continuity of street wall).

Because of the lack of similar works, we could not directly compare our results against prior research and judge the 'goodness' or 'badness' of our results. However, the work by Ordonez and Berg (2014) took a similar approach in comparing ground truth statistics with the qualities of the urban environment judged by machine learning models from street view images, which could provide some hints for the understanding on the magnitude of the correlation. In their work, Ordonez and Berg (2014) compared the perceptual scores of wealthiness with household income statistics and compared the scores of safety with homicide statistics. The Pearson correlation coefficients were 0.51 for the former task (can increase to 0.61 when only counties with large sample sizes are included) and −0.36 for the latter task (can increase to −0.47 when only counties with large sample sizes are included). We acknowledge that the tasks in Ordonez and Berg's work may have involved a higher level of uncertainty and complexity; therefore, a weaker correlation was reported. Nevertheless, it lends some evidence that our models can provide a medium-to-good approximation to the public's visual experience in the real urban environment.

Looking closer at the results, we can identify a few interesting issues. In terms of the construction and maintenance quality, the correlation is stronger when both the machine scores and the public rating scores are high but weaker when the machine scores are lower. This indicates that there is generally more consensus in terms of what makes a 'good' quality but more divergence in terms of how 'bad' is 'bad'. The experts make judgements depending on an overview of the conditions of the entire city, whereas the public may have heterogeneous standards, for instance, using the conditions of their residences or work places as the benchmark. While acknowledging the value of public opinion in understanding the quality of the physical environment, the result also indicates the advantage of consistency of expert rating. In terms of continuity, the public generally rate higher than the model. A possible explanation could be that in real settings, the public's judgement of continuity is affected by the presence of trees, which help visually diminish the feeling of discontinuity. This finding indicates that we may need to train another model that considers the influence of trees in the future.

6.4. Urban physical quality maps of Beijing

By calculating the average score for each street segment, we developed scoring maps for the two physical qualities (Figs. 10 and 11). The maps can be used in various ways:

- Research: Researchers can further link large-sized and consistently measured qualities with human perceptions and behaviour in the urban environment and phenomena such as crime and economic deprivation.
- Planning: Planners and city managers can achieve a more detailed and comprehensive understanding of the urban built environment

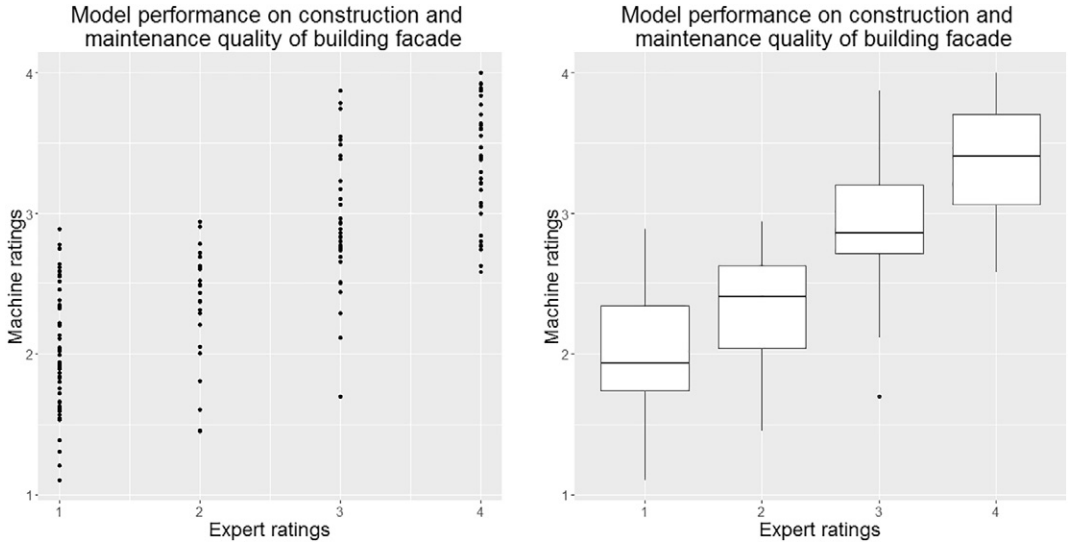

Fig. 7. Comparison between machine scores and expert rating scores on the task of construction and maintenance quality.

by reading the maps. It could provide support for policy making in areas such as urban renewal, neighbourhood revitalisation and city branding by helping identify the areas that are most in need of change or most successful.

- Daily life: Local residents and visitors may also benefit from these maps by knowing more about high-quality places to stay in the city and planning more enjoyable routes for daily travel.

The purpose of this section is not to discuss every detail of the two maps and their policy implications but to present a few examples of the implications that can be drawn from this work. It needs to be noted that although the analysis on the model performance indicates a relationship between the model scores and people's in-situ experience of the built environment, the scores should not be considered absolutely accurate but as estimations with errors. For instance, the red coloured street segments may not always be of higher visual quality than the orange ones, but these are in most cases of higher quality than the blue ones.

We here consider the whole city, the major avenues and the blocks as the three levels of analysis, from which different types of patterns can be identified. For instance, in terms of the visual quality, there is an apparent pattern at the city scale that the northern part of the city (Zone A) generally scores higher than the southern part (Zone B), especially at the urban fringes between the 4th and the 5th ring roads. The street view image showed that while most areas between the north 4th ring road and the north 5th ring road maintain a modern urban look, many areas in the south resemble more of a dilapidated village than a city. It indicates that greater emphasis should be given to the southern city in the agenda of urban renewal, which could involve a range of measures from removing stains on the facade to overall facade improvement and the demolition and reconstruction of very degraded structures. Regarding the major avenues, for example, it draws attention that the north-south central axis (Zone C), which is considered the heritage of the ancient city and is given great importance, does not

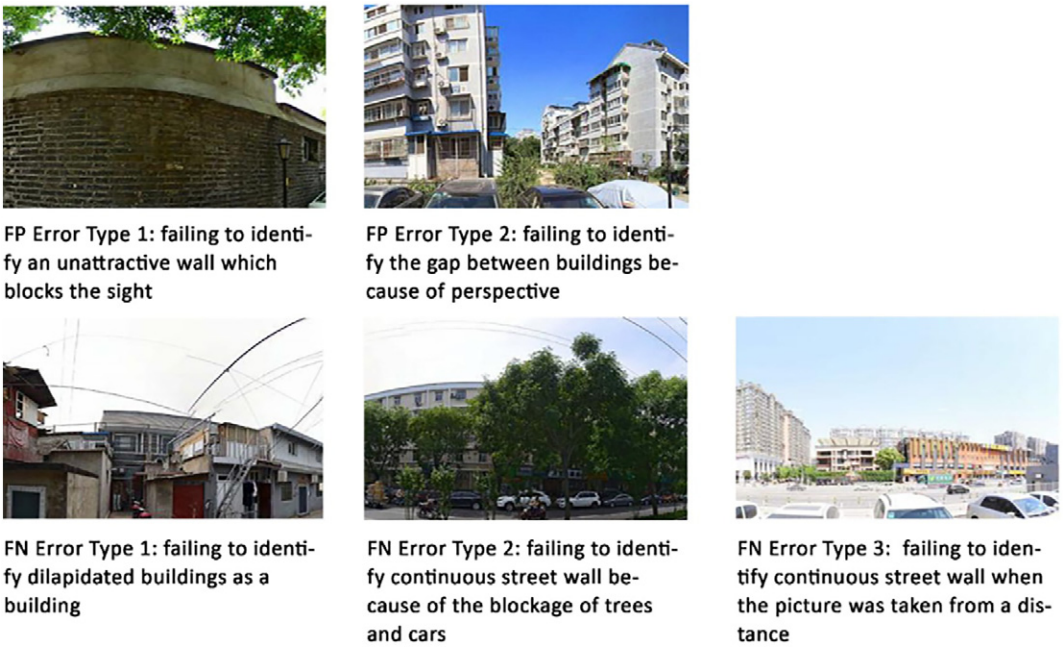

Fig. 8. Examples of errors in the continuity task.

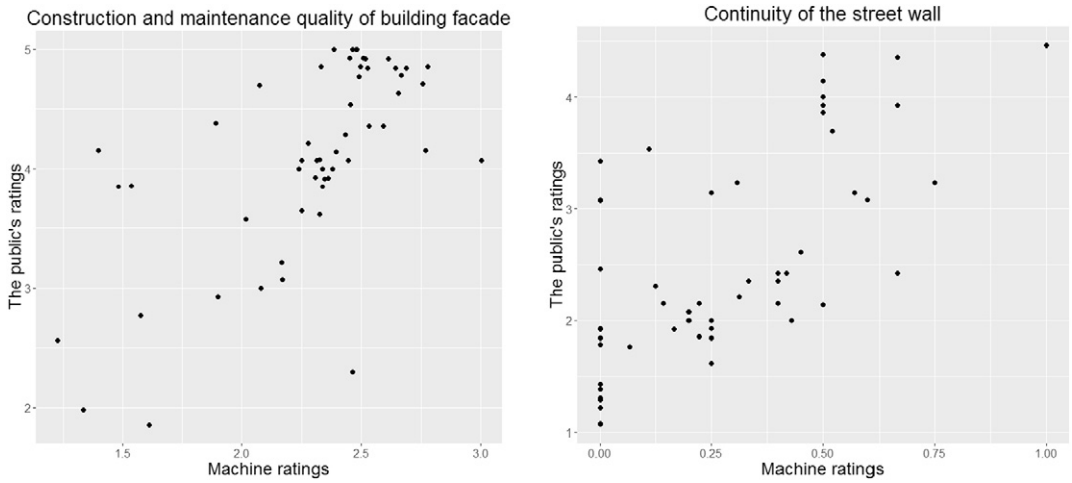

Fig. 9. Comparison between machine scores and the public in-situ ratings (colour print not needed).

seem to present an outstanding architectural visual quality. Instead, the west-east axis (Zone D) appears to be more visually appealing. It therefore indicates the need of more measures to be taken in the making of this central axis. At the block level, small concentrations of high scores and low scores can be identified. For instance, most key development areas score above the average and form a hotspot of warm colours on the map, which proves the success of place making in these areas, such as the CBD (Zone E) and the second CBD (Zone F). However, some of them appear to be isolated from the surrounding areas because of the sudden drop of score in the surroundings. For example, Zone G is a business park with decently designed office buildings, but on the other side of the adjacent railway lie shabby village houses (Zone H). Such an imbalance in the development of the built environment also needs to be alleviated in planning practice.

In terms of the visual continuity rating, at the city scale, it is apparent that a large proportion of the street walls in Beijing are not continuous. The historical areas within the 2nd ring road (Zone I), where the streets take the form of traditional Hutongs, score much higher than elsewhere in the city, which reveals one of the major differences between the visual environment in the historical areas and modern developments. It reminds that if the city is going to keep its urban identity, not only the architectural styles but also this kind of structural feature needs to be

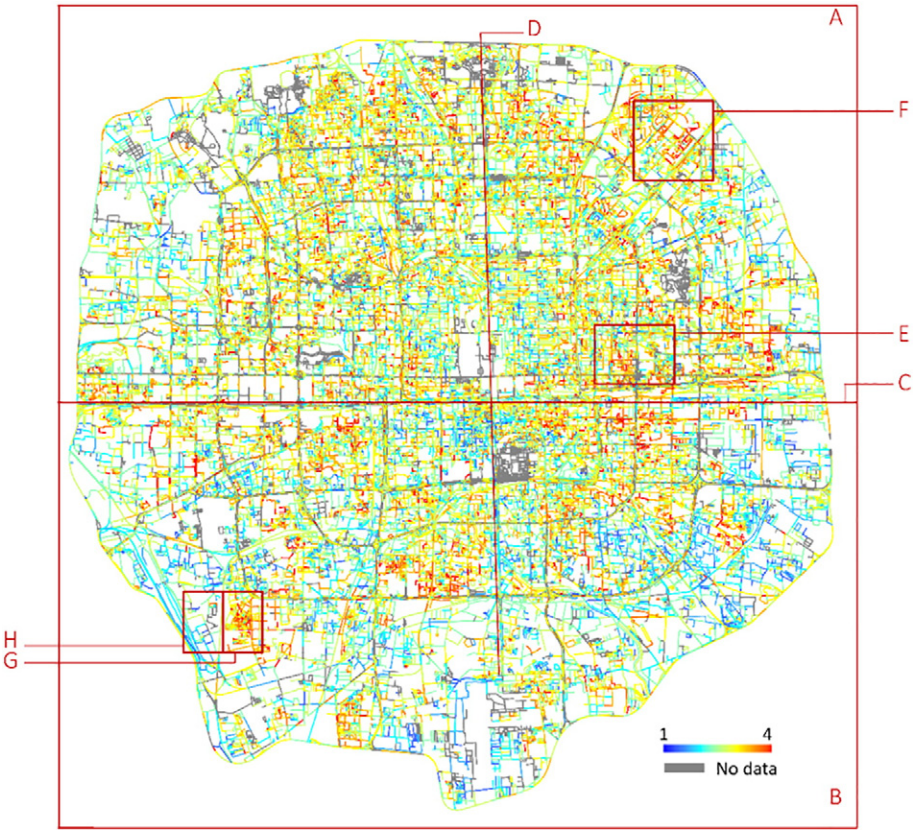

Fig. 10. Map of the construction and maintenance quality of building facade in Beijing.

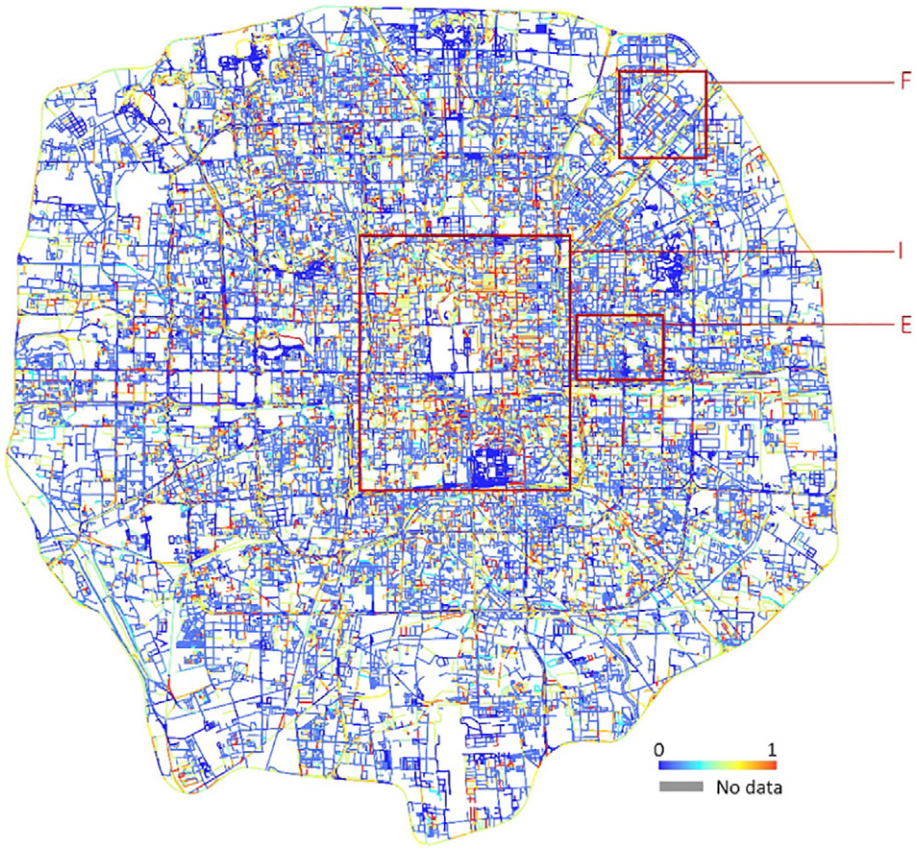

Fig. 11. Map of the continuity of street wall in Beijing.

preserved and inherited. In addition, the street walls along the ring roads are generally more continuous because they are considered the gateways of the city, and the streetscape is given more emphasis. The key development areas turn out to be much less outstanding in Fig. 11 than in Fig. 10, which indicates that the high-quality individual buildings fail to provide a feeling of continuity as a whole. In summary, the visual continuity rating demonstrates the need to incentivise infill development and more aggressively regulate shallow setbacks through urban planning and design guidelines and policies so that the feeling of enclosure and appeal by the street wall can be re-established.

7. Conclusion and discussion

Our aim in this paper was to develop and test a machine learning method, which contains three machine learning models, to automatically evaluate the urban visual environment in a large scale. We chose two key features as the starting points of this research line: the visual quality of architectural facade and the visual continuity of street wall. The method can be further extended to evaluate other built environment features that shape the visual experience, such as the architectural style, building scale and relationship between adjacent buildings.

By applying the state-of-the-art deep convolutional networks, we could achieve a satisfying machine learning performance on the expert-rated data sets. The MSE for the visual quality task was 0.61 on a rating scale of one to four, and the accuracy for the visual continuity task was 75%. In the next step, we conducted a field survey on the public's opinions of the built environment and found a moderate-to-high correlation between the machine rating and the public's rating (Spearman's $r = 0.66$ for visual quality, 0.71 for visual continuity), which shows that the present method produces a good approximation to the real experience in the urban environment.

The main contributions of our paper are as follows:

- Machine learning models for the measurement of two physical qualities of the urban environment, which is one of the key issues involved with a meaningful understanding of the physical environment. Previously, in most cases, these qualities were measured using expensive and labour-intensive conventional methods such as field audit or image-based audit.
- Validation of the proposed models against the public's opinions collected from a field survey.
- Two full-coverage and consistently measured maps on the two physical qualities in Beijing.

Our work is also limited in several aspects. First, the size of expert-labelled data set is not quite large, so we may not have achieved a maximum performance of the algorithm. To tackle this problem, we have set up a website (www.urbanvisionstudy.com) that showcases the project and advocates for crowdsourcing the labelling task to obtain a larger data set. Second, although the convolutional neural network can capture more 'global' features (i.e. responsive to a larger region of pixel space), it may not still grasp all the visual cues that contribute to the judgements as mentioned in Section 3, which remains an open problem in the field of deep learning. Third, similar to the opinion of Quercia et al. (2014) that the computer algorithm is 'a tool, not a directive', we would like to say that the present method provides evidence but not decision. When it comes to the complex issue of urban planning and design, a one-size-fits-all solution does not exist, and a high score in the algorithm does not always suggest the best condition. For instance, although the continuity of street wall contributes to the sense of enclosure and appeal, interruptions at certain points are also necessary to provide

variety and a rest for the eyes. In addition, revolutionary designs and historical structures, which should be valued, may be lowly rated by the algorithm because they do not take a 'normal' look (Quercia et al., 2014). Therefore, it should not be oversimplified that a high-scored streetscape is good enough and that a low-scored one needs change. To translate this evidence into appropriate decisions, more work is needed to understand the aesthetic cognition of the built environment through cognitive experiments, physiological psychology and son on.

However, it should also be noted that the variability in the built environment also involves a socio-economic dimension. The co-evolution between the physical appearance and social composition of cities has gained considerable interest of scholars for centuries and underpinned most architectural and urban planning movements (Naik et al., 2015). Moreover, the recent progresses in computer vision-based large-scale measurement of the built environment lend further evidence to the co-evolution relationship. For instance, Ordonez and Berg (2014) found a Pearson's correlation coefficient of 0.51 between household income and built environment-based judgement of wealthy and a coefficient of −0.36 between homicide statistics and built environment-based judgement of safety. Naik et al. (2015) found that the education level of a neighbourhood strongly predicts changes in the physical environment. Considering this, one should be very careful in drawing policy suggestions from the computer ratings. Instead of saying that 'all places should be wealthy and highly scored', which is unreasonable and raises substantial ethical questions, one should interpret the ratings within the local contexts and evaluate the performance of the built environment against the socio-economic background. Future development of machine learning models may take socio-economic variability into consideration and train separate models for different levels of affluence, which corresponds to questions such as 'how would a successful non-wealthy neighbourhood look like'.

This paper serves as the first step in profiling the cityscape with computational methods. We propose that this line of research can be extended in several ways. First, as mentioned before, more urban design features can be fed into the machine learning algorithm to produce a more comprehensive profile of the urban visual environment, such as the building material, architectural style, building scale and design of details. Moreover, the relationship between adjacent buildings is also an important factor that shapes the streetscape, including the consistency and diversity in the use of materials, colour, style, scale and details, and structural features, such as the alignment of cornice and belt course lines.

Second, the long-term vision is that with the regular update of street view images and the growing volume of geo-tagged images online, we can consistently monitor the transformation of the cityscape at a large scale. The urban planning issues that have been analysed case by case depending on a limited database in the past will be easily reviewed on the city scale, e.g. 'which areas of the city are upgrading and which are decaying' and 'how do new built projects complement existing buildings' geometry, scale and/or quality of detail' (Parolek et al., 2008).

Third, cross-city and cross-regional comparison can also be an interesting direction. The cross-regional comparison is somewhat linked to the research area of *computational geo-cultural modelling* proposed by Doersch et al., which serves to provide stylistic narratives to explore the diverse visual geographies of our world (Doersch et al., 2012). Following our proposed research line, the regional differences in urban design cultures can be evaluated by comparing the aforementioned features, which may provide a deeper insight to the design cultures. Regarding the cross-city comparison, a direct next step can be applying the algorithms developed in the present paper to all the Chinese cities and produce city rankings in terms of the visual environment. In this case, the primate or the most economically developed cities may not win over lower tier cities. We expect such comparisons to provide a more experience-oriented and quality-of-life-oriented perspective towards urban development, in addition to the measurement of hard numbers such as GDP and road network density.

Acknowledgement

This research is funded by the National Natural Science Foundation of China (Grant No. 51478232), Independent Research Project of Tsinghua University (Grant No. 20131089262) and a scholarship from the China Scholarship Council (CSC No. 201306210039). We thank the anonymous reviewers for their many insightful comments and suggestions.

References

American Planning Association (2006). *Planning and urban design standards*. John Wiley & Sons.
Atkinson, G. (2003). *Construction quality and quality standards: The European perspective*. Routledge.
Bain, L., Gray, B., & Rodgers, D. (2012). *Living streets: Strategies for crafting public space*. John Wiley & Sons.
Beijing Municipal Bureau of Statistics and National Bureau of Statistics Survey Office in Beijing (2015). Trends and characteristics of population change in Beijing in 2014. Retrieved from http://www.bjstats.gov.cn/zt/rkjd/sdjd/201603/t20160322_340772.html.
Brandt, E., & Rasmussen, M. (2002). Assessment of building conditions. *Energy and Buildings*, *34*(2), 121–125.
Brownson, R. C., Hoehner, C. M., Day, K., Forsyth, A., & Sallis, J. F. (2009). Measuring the built environment for physical activity: State of the science. *American Journal of Preventive Medicine*, *36*(4), S99–S123.
Cullen, G. (1961). *The concise townscape*. Routledge.
Doersch, C., Singh, S., Gupta, A., Sivic, J., & Efros, A. (2012). What makes Paris look like Paris? *ACM Transactions on Graphics*, *31*(4).
Downs, R. M., & Stea, D. (1973). *Image and environment: Cognitive mapping and spatial behavior*. Transaction Publishers.
Dubey, A., Naik, N., Parikh, D., Raskar, R., & Hidalgo, C. A. (2016). *Deep learning the city: Quantifying urban perception at a global scale. Paper presented at the European Conference on Computer Vision.*
Ewing, R., Clemente, O., Neckerman, K. M., Purciel-Hill, M., Quinn, J. W., & Rundle, A. (2013). *Measuring urban design*. Island Press/Center for Resource Economics.
Ewing, R., & Handy, S. (2009). Measuring the unmeasurable: Urban design qualities related to walkability. *Journal of Urban Design*, *14*(1), 65–84.
Goel, A., Juneja, M., & Jawahar, C. (2012). *Are buildings only instances?: Exploration in architectural style categories. Paper presented at the Proceedings of the Eighth Indian Conference on Computer Vision, Graphics and Image Processing.*
Hara, K., Le, V., & Froehlich, J. (2013). *Combining crowdsourcing and google street view to identify street-level accessibility problems. Paper presented at the Proceedings of the SIGCHI conference on human factors in computing systems.*
Harvey, C. W. (2014). Measuring streetscape design for livability using spatial data and methods. *Doctoral dissertation, The University of Vermont and State Agricultural College.*
Hwang, J., & Sampson, R. J. (2014). Divergent pathways of gentrification racial inequality and the social order of renewal in Chicago neighborhoods. *American Sociological Review*, *79*(4), 726–751.
Kelly, C. M., Wilson, J. S., Baker, E. A., Miller, D. K., & Schootman, M. (2013). Using Google Street View to audit the built environment: Inter-rater reliability results. *Annals of Behavioral Medicine*, *45*(Suppl. 1), S108–S112. http://dx.doi.org/10.1007/s12160-012-9419-9.
Kostof, S. (1999). *The city assembled: The elements of urban form through history*. Thames and Hudson.
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). *Imagenet classification with deep convolutional neural networks. Paper presented at the Advances in neural information processing systems.*
Lang, J. (1994). *Urban design: The American experience*. John Wiley & Sons.
Lee, S., Maisonneuve, N., Crandall, D., Efros, A. A., & Sivic, J. (2015). *Linking past to present: Discovering style in two centuries of architecture. Paper presented at the IEEE International Conference on Computational Photography.*
Lehnerer, A. (2009). *Grand urban rules*. 010 Publishers.
Lowe, D. G. (1999). *Object recognition from local scale-invariant features. Paper presented at the International Conference on Computer Vision.*
Lynch, K. (1960). *The image of the city*. MIT Press.
Lyon, N. B. (1978). *Yonge Street Revitalization Project*. Toronto: N. Barry Lyon Consulting.
Madhavan, B. B., Wang, C., Tanahashi, H., Hirayu, H., Niwa, Y., Yamamoto, K., ... Sasagawa, T. (2006). A computer vision based approach for 3D building modelling of airborne laser scanner DSM data. *Computers, Environment and Urban Systems*, *30*(1), 54–77.
Marcus, C. C., & Francis, C. (1997). *People places: Design guidelines for urban open space*. John Wiley & Sons.
Moughtin, C. (2003). *Urban design: Street and square*. Routledge.
Naik, N., Kominers, S. D., Raskar, R., Glaeser, E. L., & Hidalgo, C. A. (2015,). Do people shape cities, or do cities shape people? The co-evolution of physical, social, and economic change in five major U.S. cities. *National Bureau of Economic Research working paper series, no. 21620*. http://dx.doi.org/10.3386/w21620.
Naik, N., Philipoom, J., Raskar, R., & Hidalgo, C. (2014). *Streetscore—Predicting the perceived safety of one million streetscapes. Paper presented at the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW).*
Nasar, J. L. (1983). Adult viewers' preferences in residential scenes a study of the relationship of environmental attributes to preference. *Environment and Behavior*, *15*(5), 589–614.

Ordonez, V., & Berg, T. L. (2014). Learning high-level judgments of urban perception. *Computer Vision–ECCV 2014* (pp. 494–510). Springer.
Parolek, D. G., Parolek, K., & Crawford, P. C. (2008). *Form based codes: A guide for planners, urban designers, municipalities, and developers*. John Wiley & Sons.
Porzi, L., Rota Bulò, S., Lepri, B., & Ricci, E. (2015). *Predicting and understanding urban perception with convolutional neural networks. Paper presented at the Proceedings of the 23rd Annual ACM Conference on Multimedia Conference.*
Quercia, D., O'Hare, N. K., & Cramer, H. (2014). Aesthetic capital: whatmakes London look beautiful, quiet, and happy? *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing* (pp. 945–955). ACM.
Said, S. Y., Zubir, S. S. S., & Rahmat, M. N. (2014). Measuring physical changes in an urban regeneration scheme. *WIT Transactions on Ecology and the Environment*, *191*, 1165–1174.
Salesses, P., Schechtner, K., & Hidalgo, C. A. (2013). The collaborative image of the city: Mapping the inequality of urban perception. *PloS One*, *8*(7), e68400. http://dx.doi.org/10.1371/journal.pone.0068400.
Santayana, G. (1955). *The sense of beauty: Being the outline of aesthetic theory. Vol. 238*. Courier Corporation.
Shalunts, G., Haxhimusa, Y., & Sablatnig, R. (2011). Architectural style classification of building facade windows. *Advances in visual computing* (pp. 280–289). Springer.
Shalunts, G., Haxhimusa, Y., & Sablatnig, R. (2012). Architectural style classification of domes. *Advances in visual computing* (pp. 420–429). Springer.
Stamps, A. E. (2000). *Psychology and the aesthetics of the built environment*. Springer Science & Business Media.
Sun, Y., Fan, H., Bakillah, M., & Zipf, A. (2015). Road-based travel recommendation using geo-tagged images. *Computers, Environment and Urban Systems*, *53*, 110–122.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). *Going deeper with convolutions. Paper presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.*
Voordt, T., & Wegen, H. B. (2005). *Architecture in use: An introduction to the programming, design and evaluation of buildings*. Routledge.
Wohlwill, J. F. (1976). Environmental aesthetics: The environment as a source of affect. *Human Behavior and Environment: Advances in Theory and Research*, *1*, 37–86.
Xu, Z., Tao, D., Zhang, Y., Wu, J., & Tsoi, A. C. (2014). Architectural style classification using multinomial latent logistic regression. *Computer Vision–ECCV 2014* (pp. 600–615). Springer.
Zhao, P. (2011). Car use, commuting and urban form in a rapidly growing city: Evidence from Beijing. *Transportation Planning and Technology*, *34*(6), 509–527. http://dx.doi.org/10.1080/03081060.2011.600049.
Zhou, B., Liu, L., Oliva, A., & Torralba, A. (2014). Recognizing city identity via attribute analysis of geo-tagged images. *Computer Vision–ECCV 2014* (pp. 519–534). Springer.