Machine Vision and Applications (2024) 35:7
https://doi.org/10.1007/s00138-023-01484-1

RESEARCH

Visual complexity of urban streetscapes: human vs computer vision

Pietro Florio[1] · Thomas Leduc[2] · Yannick Sutter[2] · Roland Brémond[3]

Received: 22 June 2023 / Revised: 7 September 2023 / Accepted: 28 October 2023 / Published online: 2 December 2023

Abstract
Understanding visual complexity of urban environments may improve urban design strategies and limit visual pollution due to advertising, road signage, telecommunication systems and machinery. This paper aims at quantifying visual complexity specifically in urban streetscapes, by submitting a collection of geo-referenced photographs to a group of more than 450 internet users. The average complexity ranking issued from this survey was compared with a set of computer vision predictions, attempting to find the optimal match. Overall, a computer vision indicator matching comprehensively the survey outcome did not clearly emerge from the analysis, but a set of perceptual hypotheses demonstrated that some categories of stimuli are more relevant. The results show how images with contrasting colour regions and sharp edges are more prone to drive the feeling of high complexity.

Keywords Visual complexity · Streetscapes · Computer vision · Perception

1 Introduction

1.1 Context

Vision in urban environments is certainly a complex task, given the diversity of visual stimuli offered by cities. The dynamicity of urban scenes, in which observers are ecologically immersed [1, 2], solicits their visual system in a sequence of revelations and serial images [3], ultimately forming "the image of the city" [4].

In the context of urban areas and streetscapes, visual complexity may be intended as the visual richness or diversity of the built environment. A diversity of factors concurs in determining the complexity of an urban view, based on visual perception [5]. Some aspects are linked with the difficulty of processing the sensory information in relation with the physiological limits of vision, like the angular size of target objects and the luminance contrast between target objects and their background: we call these bottom-up (or low level) processing aspects [6]. Some other factors refer to semantic structures built by cognition and experience, namely the meaning a target object represents in its context: these are top-down (or high level) aspects, as opposed to the previous ones [7]. Too many equally meaningful target objects in a scene confuse human senses and the attention is diverted to more prominent stimuli, based on saliency, dynamics and motion of target objects: this exercise of selective attention [8], given also the activity and behaviour of the observer, constitutes another layer of complexity.

Specifically in urban environments, the height and size of buildings, as well as the variety of textures and patterns are likely bottom-up aspects. The presence of landmarks, the density and arrangement of built volumes are linked with top-down processing. Perception is constrained to a maximum rate of usable information [9]; reported complexity depends on expectations, anticipations of open and closed spaces, linked with the human prospect of refuge [10] or with the quest for variety [11]. Complexity is identified as

✉ Pietro Florio
pietroflorio@gmail.com

Thomas Leduc
thomas.leduc@crenau.archi.fr

Yannick Sutter
yannick.sutter@crenau.archi.fr

Roland Brémond
roland.bremond@univ-eiffel.fr

[1] Renewable Energies Cluster ENAC, École Polytechnique Fédérale de Lausanne (EPFL), Station 18, EPFL, Route Cantonale, CH-1015 Lausanne, Switzerland

[2] Nantes Université, ENSA Nantes, École Centrale Nantes, CNRS, AAU-CRENAU, UMR 1563, F-44000 Nantes, France

[3] PICS-L lab, CoSys Department, Université Gustave Eiffel, 16 Boulevard Newton, F-77420 Champs-sur-Marne, France

one of the fifty urban design qualities related to walkability [12].

Atmospheric variables also influence the perception of the outdoor urban space, modifying the view depth and the lighting conditions of the scene [13]: from more diffuse with cloudy weather to harsher in contrast under direct sun [14]. These aspects, like the seasonal-dependent presence of vegetation [15, 16], have psychological effects on the observer too [17, 18].

Everyday experiences of the mentioned perceptual processes can be the detection of solar modules, or similar camouflaged visual pollution, in urban landscapes (bottom-up) [19], the visual search for road indications and signs (top-down) [20], the attention captured by advertisement (selective attention) [21].

In the scientific literature, a multitude of metrics for the quantification of visual complexity in urban contexts extends from city planning, fractal geometry and information theory [22] to optical and lighting physics, through neurology, psychology and physiology. Recently, visual perception trials with human participants often accompany algorithms of computer vision [23], which mimic some perceptual processes. The complexity ranking of image samples made by panels of participants is typically matched with computational metrics based on image contrast [24], edge detection, entropy and image Compression [25]. In some cases, the fractal dimension is specifically investigated [26]. Some studies employ artificial neural networks to train and predict the complexity of generic images as perceived by humans [25, 27], or adopt neuron activations in hidden layers of image segmentation algorithms as proxy for image complexity [28]. Other studies try to implement attention mechanisms in computer vision, imitating the high-level extraction of key information [29, 30].

Within research focusing on streetscapes, worth to note are techniques estimating visual complexity based on taxonomic labelling of visible features in images [31], or on the "noise" introduced by specific features like signage [32], that increase the feeling of complexity [33]. Highly dense and cluttered urban environments can lead to cognitive load, stress, fatigue, inducing reduced visual clarity, decreased legibility and impaired wayfinding.

The semantic labelling of streetscapes associated with machine learning showed also a promising potential [34]. Interestingly, there is neurological evidence of perceived complexity in streetscapes [35], leaving a trace in the electroencephalogram. In some cases, visual complexity enhances the experiential quality of the built environment by providing visual interest, stimulation and diversity.

Table 1 Experimental sample size of participants for the cited studies and relevance of the investigated image collection to the urban environment

Research	Number of participants involved	Images collection in urban environment
Cavalcante et al. [24]	40	Yes
Machado et al. [25]	240	No
Gunawardena et al. [31]	20	Yes
Kacha et al. [35]	6	Yes
Gunawardena et al. [33]	50	Yes
Nagle and Lavie [27]	53	No
Saraee et al. [28]	1687	No
Ma et al. [26]	0	Partially (gardens only)
Guan et al. [34]	68	Yes
Kawshalya et al. [36]	78	Yes

2 Research objective

Visual complexity is a multifaceted concept that has both positive and negative effects on human perception and behaviour in urban areas. Its definition and measurement can vary. In this study, the objective is to derive an empirical definition of visual complexity from a web survey, by submitting a set of images to the public.

Overall, the existing literature lacks studies corroborated by a large sample of participants (e.g., more than 200 subjects). From Table 1, it also emerges that visual complexity has been seldom investigated thoroughly in urban environments. In contrast, the present study leverages on crowd-sourced visual complexity ranking of images portraying exclusively urban streetscapes; the outcome is compared with various metrics issued from computer vision algorithms. Due to the pursued research objective, the concept of complexity is not explicitly defined beforehand, but rather evoked through a series of abstract artworks, shown to the visitors of the survey website (Fig. S. 1). By working on a sample of more than 450 internet users, this study aims at bringing statistical significance to the match between human and computer-based visual complexity ranking, from a collection of 25 urban streetscape images gathered in different parts of the world. The outcome is expected to facilitate the complexity modelling in urban areas, by providing contextual information to the addition of new objects to the scene, in simulations for visual pollution generated for instance by solar modules, antennas, advertisement signs or any other nuisance.

To test the generalisability of the approach, the selected images include heterogeneous and diverse streetscapes from disparate locations worldwide, with a prevalence in Europe. However, the approach can be replicated to any set of geo-referenced images, including those available in street view repositories, to obtain complexity maps as produced in other studies [36].

3 Overview

This research focuses on the visual aspects of complexity perceived in urban streetscapes. Human and computer vision were specifically compared with a perceptual approach.

The overall workflow relies on (i) the selection of a set of 38 images of streetscapes distributed around the globe. (ii) Streetscapes were manually tagged to identify several significant features and classify the images semantically. (iii) The set of streetscapes was ranked by more than 450 participants via an online interactive form created ad-hoc: despite several limitations and drawbacks discussed in the forthcoming sections, this turned out to be the most viable solution. After extracting an average ranking from the human-based classification, (iv) a series of hypotheses leveraging on certain perceptual aspects that may impact on the quantification of complexity is laid down (Table 2).

These hypotheses led to a series of computer vision algorithms (last column of Table 2), used to rank the images in the collection. This set of computer vision-based indicators was selected after a thorough literature review, but it may not be fully exhaustive yet: however new algorithms can be easily added to the collection. Finally, (v) the fitting of complexity ranking based on computer vision was assessed using human ranking as ground truth in a fivefold cross-validation. This fitting allowed estimating the best correlations, and thus the most relevant hypotheses among those tested. The process included comparing bottom-up (H1 to H6) and top-down (H7) components in the human appreciation of streetscapes complexity. Results are explained and discussed in the following sections.

4 Methodology

4.1 Image database

First, a set of 38 images portraying as many streetscapes distributed in different parts of the world was constituted by consulting free online photo repositories, which include geographic positioning. Several criteria were adopted for the selection of the images: (i) the Creative Commons license was imperative for processing and reusing the image, (ii) each photograph had to be taken from the medial axis of the street, pointing along its direction, at pedestrian level. The scene should portray an urban streetscape, without prominent objects in the foreground; (iii) the target scene was illuminated by clear or partly cloudy sky, bringing both direct and diffuse daylight.

A manual image tagging operation was performed by the authors, which led to a first complexity ranking and helped to preserve a meaningful variety while reducing the dataset. Before assessing their visual complexity, the images were reduced to a set of 25 elements, to avoid making the online ranking survey too demanding for respondents (Fig. 1). Repetitive images were filtered out when containing common elements that constituted redundant scenes and were kept, ensuring a significant diversity in materials, objects, scenes and shading.

4.2 Online survey

The refined image set containing 25 elements was published online to collect a crowd-sourced ranking of their visual complexity. To confirm or reject any correspondence between human and computer-based complexity ranking, the statistical ground truth had to be as representative as possible of the average human response. Instead of an expensive field survey, an online form based on photographs could reduce spatial constraints and open to a more diversified public. To ensure the necessary quality of respondents' submissions, a set of convenient practices aimed at: limiting the survey time to 10 min to avoid distractions; securing the website to minimize the risk of remote attack or corruption; designing a self-explanatory, attractive and ergonomic interface, with easy interaction (drag and drop) and navigation (back and review buttons).

The survey was presented to the user in five steps. First, a static web page introduced the survey in a few words and hosted two optional links: to a 3′30 presentation video and to a photo gallery containing the 25 images of the survey. In the second step, a web page asked the user to group each of the 25 images by three categories of complexity (low, medium or high), leveraging on JavaScript drag and drop functionalities. The third step requested sorting the images by increasing complexity within the three different categories, each category being presented in a dedicated table. The fourth step profiled the respondent in an anonymous form, by retrieving the name of the base city, the gender, the age and the field of expertise (five choices were proposed). The fifth and last step was a message to summarize the information collected anonymously and thank the respondent.

The survey has been based on a 3-tier architecture, collecting responses directly on the user's device through a cookie and feeding a database, at the end of each survey session, by means of a Google-app for the data access layer. The survey data was arranged in a simple spreadsheet with 12 columns:

Table 2 Hypotheses considered for assessing streetscapes complexity through computer vision

Hypothesis	To estimate complexity:	Associated indicators
H0	All components of the image are considered as a whole	Descriptive statistics on the image (or its tiles) after different colour space conversion
H1	The colour component may be neglected	Descriptive statistics on the image (or its tiles) after greyscale conversion
H2	The luminance component may be neglected	Descriptive statistics of the image (or its tiles) after conversion into the Lab colour space and setting the luminance component to zero
H3	Low frequencies are the most important factor	Descriptive statistics on the image (or its tiles) after various blurring or various alterations by wavelet or Fourier transforms
H4	Objects edges are the most important factor	Descriptive statistics on the image (or its tiles) after edge detection via the Scharr or Sobel filters
H5	The capture of visual's attention is the most important factor	Descriptive statistics on saliency maps (spectral or fine-grained), maps from interest point detection (via ORB or SIFT) or UAE maps
H6	Fractal dimension is the most important factor	Analysis of the fractal dimension of the B&W image
H7	The classes identified in a panoptic segmentation are the most important factor	Indicator derived from the panoptic segmentation of the image

three of them represent the different complexity categories and feature a list of image labels sorted by increasing complexity.

In summary, the output of the online survey gathered the complexity category attributed to each photograph by the user as a first step, then the complexity ranking of the photographs from the least to the most complex. Overall results from the whole set of respondents were aggregated by averaging the position in the complexity ranking.

4.3 Computer vision methods (computational indexes)

To test hypotheses H0-H7, it was necessary to compute, from each image, indicators that may correlate with the level of complexity. This set of indicators–a kind of complexity predictor–was compared with the ranking obtained from the online survey subsequently. To automate the assessment of complexity, various options, inspired by the state of the art, have been explored. The process was globally achieved in four steps: a first one (i) was dedicated to the conversion of the image into several colour spaces available in the OpenCV library [37]. A second step (ii) determined the actual transformation of the image (low-pass filtering, edge detection, high-pass filtering, etc.), and in a third step (iii), the image was subdivided into smaller tiles. The last step (iv) consisted in an extraction of descriptive statistics from the transformed image and its tiled subdivisions. The computation followed the sequence of operations summarised in Fig. 2 and illustrated with more detail in Fig. S. 3.

4.3.1 First image transformation

Colour space (hypothesis 0, 1, 2 and baseline for further hypotheses) A first iteration of the computer vision algorithms was calculated on standard images, cropped and resized all to the resolution of 400 × 600 pixels (2/3 proportion), encoded in different colour spaces: this phase was planned to find the colour space that best explains the complexity ranking from the online survey. Then, the scope encompassed the importance of the colour vs. the luminous intensity of each pixel. To achieve this de-coupling, the CIELAB colour space was exploited, also referred to as L*a*b*.

To test hypothesis 1, images were altered into grey levels without the colour component, using the L* channel alone (an alternative would have been to study the "Intensity" component of the HSI colour space). Conversely, hypothesis 2 was tested by setting to zero the luminance component L*, to return "a*b*" images. From the initial set of 25 colour images, several distinct groups of 25 images were derived using the colour conversion codes provided by the

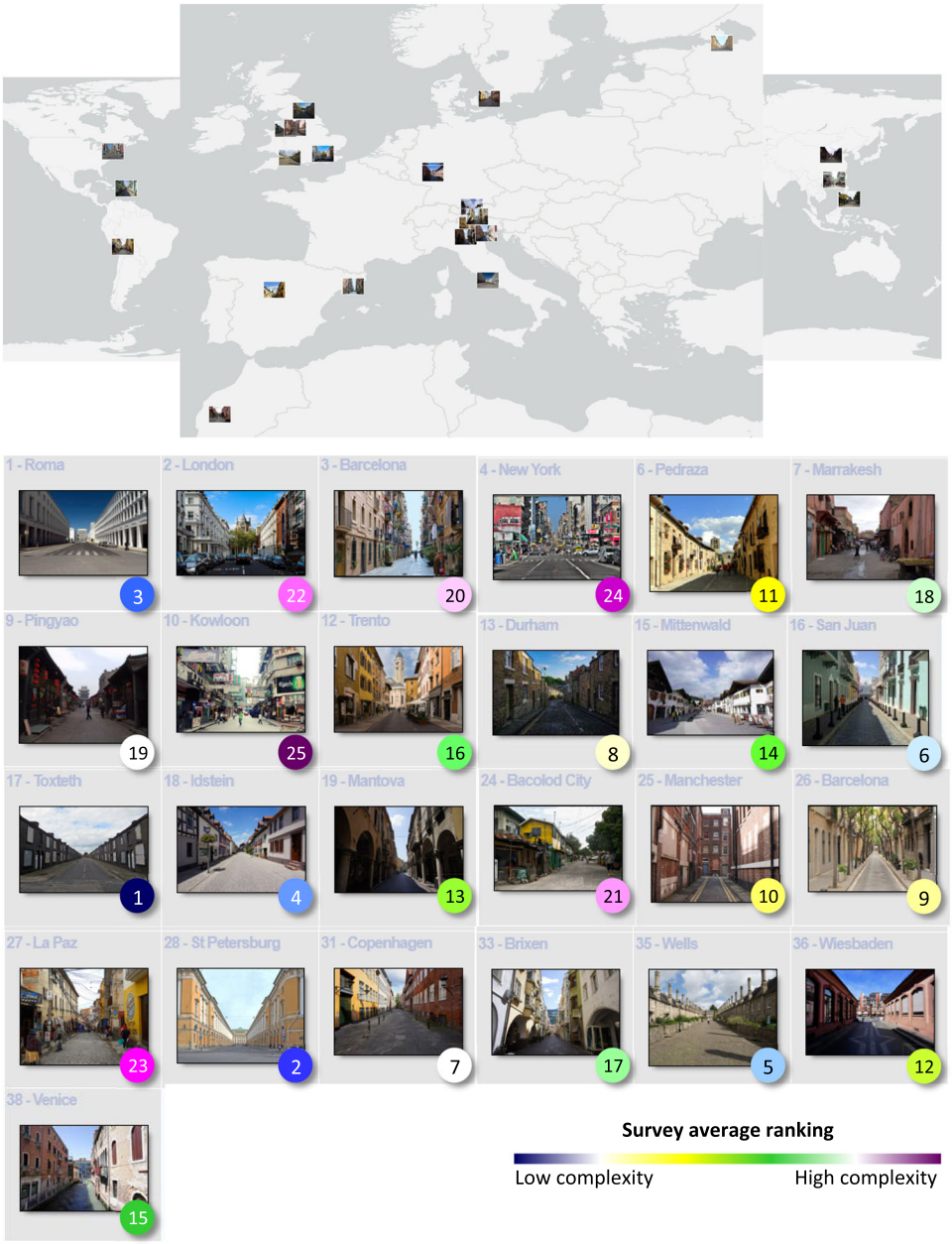

Fig. 1 Map and gallery view of the 25 selected photographs used in the online ranking survey. The average complexity ranking position issued from survey respondents in ascending order is indicated in the round marker at the bottom-right corner of each photo, and encoded on a colour scale (blue–yellow–green–purple). All selected photographs are licensed Creative Commons: geographic locations, authors and links to the photographs are included in the data annex

OpenCV library [37]. The following colour spaces were considered: HLS, HSV, L*a*b*, L*u*v*, XYZ, YCC and YUV, in addition to the original RGB and to the above-mentioned luminance-only (L*) and colour-only (a*b*).

The image processing algorithms described below were applied to the resulting images, taking inspiration by Machado et al. [25], to study the relevance of each hypothesis (H0-H7) and explain the perceived complexity. Following Gunawardena [31], black and white images derived from binarised luminance maps allowed for estimating their fractal dimension under hypothesis 6. Once these groups were constituted, the different treatments described below were carried out, to study the influence of each major image feature on the average perceived complexity.

Spatial frequency (hypothesis 3) This step aimed at testing hypothesis 3, through the degradation of each image, by removing the details and keeping only the low-frequency features, also called trends. The underlying assumption states that a high density of trends is characteristic of a high complexity of the image. To extract these trends, three approaches have been tested. First, the reduction of the image noise through a Gaussian blur, mimicking the physical filter of the image per means of a translucent screen. This low pass smoothing or blurring filter could be applied

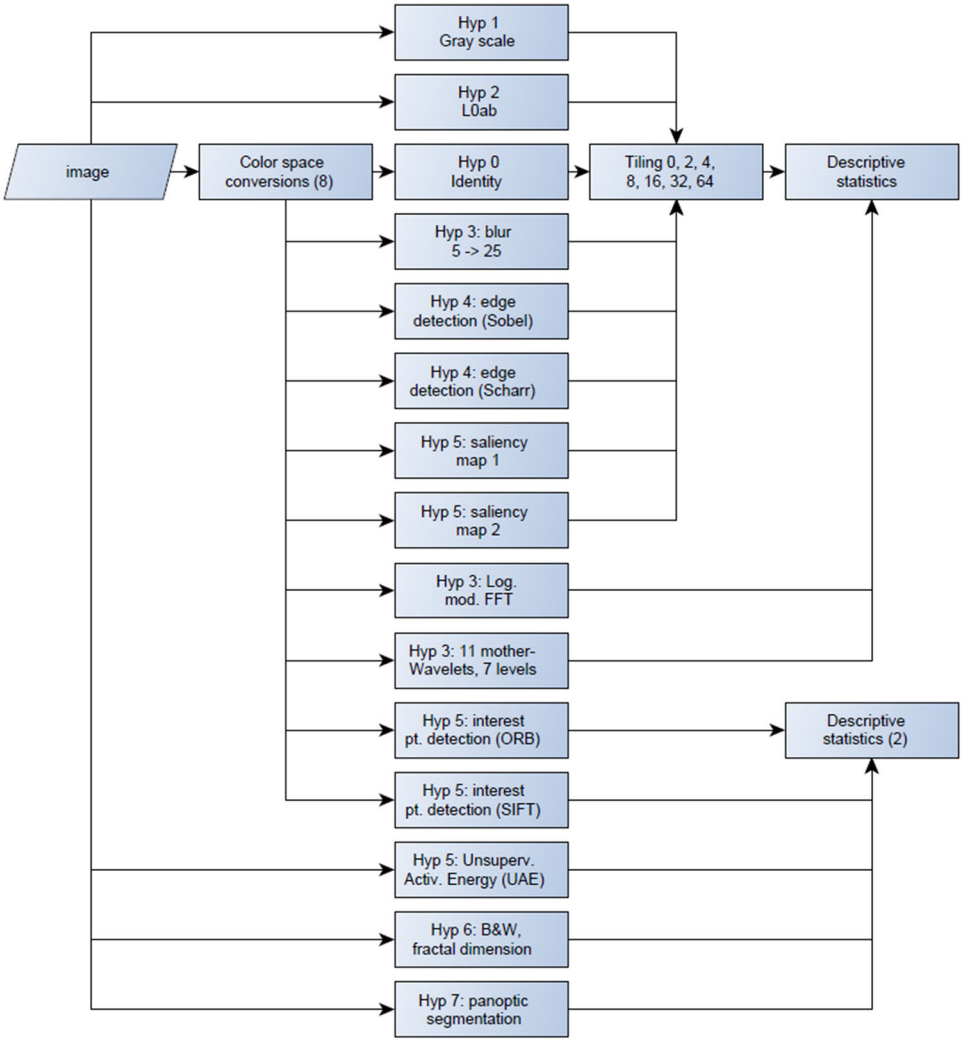

Fig. 2 This workflow synthesizes the computer vision processing of the images. A colour image given as input (centre) was first converted into an appropriate colorimetric space before undergoing a first transformation (blurring, edge detection, etc.). The various indicators were computed from this transformed image, using appropriate descriptive statistics

iteratively to gradually increase the blurring effect. In practice, five different fuzzy levels were produced (i.e., after 5, 10, 15, 20 and 25 iterations of a kernel of size 9 × 9 and standard deviation equal to 20 in both directions). An alternative, second approach, consisted in a wavelet-based compression. This involved using a 2D multi-level decomposition-reconstruction technique which removed, in an incremental way (and up to eight iterations), the items corresponding to the "detail coefficients". With this technique, different discrete built-in wavelets could be tested. These two low-pass filtering strategies (Gaussian blur and wavelet-based compression) have been applied to all the colour space images.

The third approach employed Fourier transform to represent a signal—in this case an image—in the frequency domain by means of a development on a base of exponentials. Characterizing an image by its frequency spectrum allows highlighting the importance of the fundamental harmonic, as well as the more or less rapid decrease in the amplitude of the harmonics of higher ranks. By removing the high spatial frequencies from the image in the Fourier domain, it was possible to build a low frequency image, based on the power of the signal, as the modulus of the Fourier transform (neglecting the phase).

Edge detection (hypothesis 4) The Sobel and Scharr derivatives operators allowed testing hypothesis 4, by highlighting the edges of the images, namely the regions where the brightness of the pixels changes abruptly. These brightness gradients may characterize the boundaries of an object or part of an object, like the boundaries of shading or illumination areas, textures, etc. As edges represent discontinuity in the image, they participate in its structuring and thus contribute to its perceived complexity. Compared to the lossy compression techniques mentioned previously, edge detection relies on high frequency information.

As before, these two edge detection strategies (Sobel and Scharr filters) have been applied to all the colour space images.

Visual attention (hypothesis 5) The purpose of saliency maps is to highlight regions of an image that are important to human vision, as they capture visual attention. A classical image processing algorithm [20] allows for testing hypothesis 5: each pixel of the image is rated according to its "meaningfulness" (in a sort of biomimetic-based interpretation). This work is based on two saliency algorithms derived from Itti's seminal paper: the Fine-Grained Saliency assesses centre-surround differences [38], while the Spectral Residual approach analyses the log spectrum of the image [39]. These methods have been applied to all the colour space images.

Feature detection algorithms can be equally relevant to test hypothesis 5. They help to derive the prominent visual content from the local analysis of an image, as independently as possible from scale, framing, viewing angle and exposure. The detection of points of interest with multiple invariances facilitates image stitching and therefore content-based image retrieval. This technique was used here, under the assumption that the distribution of such points can inform on the dispersion or clustering of the complexity. Two key points detection methods have been considered: the Scale Invariant Fourier Transform (SIFT) [40], and the Oriented FAST and Rotated BRIEF (ORB) algorithm [40, 41]. These two algorithms have been applied to all the colour space images.

The recent spread of deep learning algorithms may also help testing hypothesis 5. Saraee et al. exploit information about the complexity of images, carried by the intermediate convolutional layers of deep neural networks [28]. In practice, the algorithm accounts for the energy map corresponding to the fourth max-pooling layer of image prediction, by the VGG-16 architecture, trained for a scene recognition task. Descriptive statistics of this energy map have been computed as in the other computational approaches described above.

Fractal dimension (hypothesis 6) A figure is said to be fractal when it has a similar structure at all scales. The fractal dimension of an object is computed by measuring the fragmentation of its contours, which characterizes a form of entanglement. To test hypothesis 6, we followed Gunawardena et al. (Gunawardena et al., 2015), and assumed that the fractal dimension of black and white luminance images was an indicator of its perceived complexity. In practice, the implementation of this metric is made available by the software *ImageJ* [42], the binarization of the image being previously implemented by an Otsu's threshold method.

Semantic segmentation (hypothesis 7) A deep learning method called panoptic segmentation was used to test hypothesis 7. This adaptation of the "Masked-attention Mask Transformer" [43] can identify, in each image, various types of instances such as cars, bicycles, people, streets, portions of sky, pavement and buildings. An indicator for a top-down model to explain complexity, in accordance with the ranking from the online survey, relied on the classes resulting from the panoptic segmentation. This indicator constituted a predictor of complexity, constructed as follows. For each image, classes were selected where the surface ratio r_j of pixels P_j in the class j was greater than or equal to 2% (of the total number of pixels in the image S), as per Eq. 1. This process was intended to remove irrelevant objects from the complexity computation. Then, the algorithm calculated the complexity ratio c, as the number of classes x (discarding those below the $r_j \leq 2\%$ threshold), divided by the median of the surface ratios r_j for such selected classes (above the $r_j > 2\%$ threshold). For example, if a photograph was segmented into 4 classes, of which only 3 cover an area that exceeds 2% of the total photograph area, the ratio c would be equal to 3 divided by the median of the 3 surface ratios r_j, namely the surface ratio r_j which is not the maximal nor the minimal. This complexity ratio c was the one used to rank the photographs in the framework of semantic segmentation. Various threshold values were tested (1%, 2%, 3%, etc.), along with various ratios (as well as standard deviations, percentiles, maximum values, etc.), before proposing the one presented in Eq. 2 (2%), which led to the best correlation in the Pearson sense, relatively to hypothesis 7 ($\rho = 0.74$). The outcome of the ranking operated in this step can be consulted in Table S. 1.

$$r_j = \frac{P_j}{S} \text{ and } S = \sum_{j=1}^{x} P_j \tag{1}$$

$$J = \left\{ j \in [1, x]; r_j \geq 2\% \right\} \quad c = \frac{\text{card}(J)}{\text{Median}\left(\left\{ r_j;\ j \in J \right\}\right)} \tag{2}$$

4.3.2 Spatial distribution analysis using image tiling

The tiling of an image is a way to study the complexity through spatial decomposition. This set of localised analyses makes it possible to study the overall variability of a given indicator, throughout the global image, as well as its distribution in individual tiles. The scope here was to determine the spatial distribution of complexity over the image, i.e., if some regions of the image were, or not, more complex than others, and in what proportion. This recursive tiling technique was applied iteratively to produce many small tiles (we stopped this recursive tiling at the 6th iteration with 64 $\times$ 64 thumbnails of 9 $\times$ 12 pixels size). Tiling was applied to all the colour space images and algorithms.

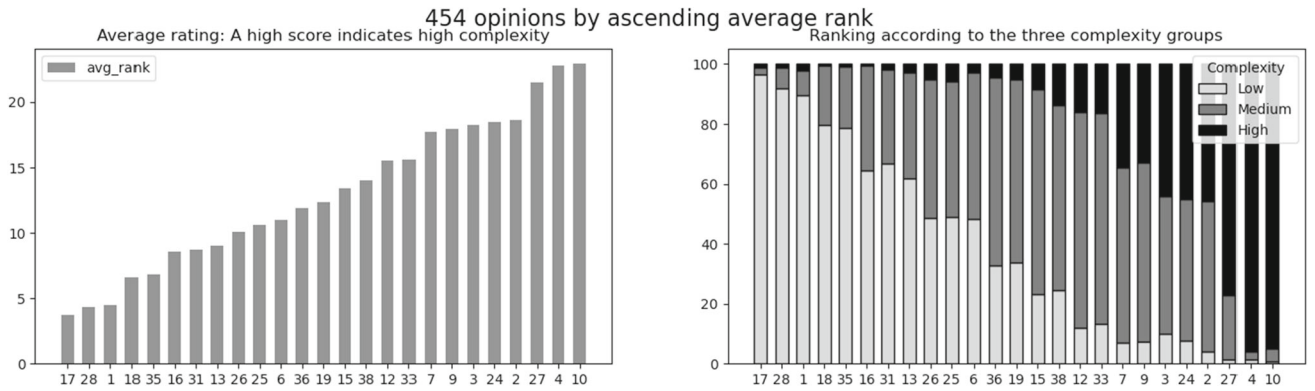

Fig. 3 (Left) Average position from all respondents' ranking; (right) breakdown of complexity categorization by share of respondents

4.3.3 Descriptive statistics on images

Descriptive statistics have been computed from various images (either the original, testing hypothesis 0, or those resulting from the transformations presented above). For the greyscale images, statistics included the mean, median, range of values, standard deviation, skewness, Kurtosis, entropy and signal-to-noise ratio of the pixel values. Each of these eight parameters was considered as a candidate predictor of the perceived complexity from the survey, and tested for correlation. For colour images (CIELAB or RGB colour spaces in particular), a variability indicator around the average colour was also computed, as a possible index of the image complexity. This variability (denoted v) consisted in the mean Euclidean distance from the colour coordinates of each pixel to the average colour coordinates of the image. In addition, after each tiling, the entropy and standard deviation (resp. variability indicator around the average colour) for the greyscale images (resp. colour space images) was carried out. In the case of key point detection (ORB and SIFT algorithms), instead of using these descriptive statistics which do not apply to binary images, two alternative indexes were employed. The first one equals the number of key points, the second one is the surface coverage ratio of the associated disks of a given radius, relatively to the image size, after a spatial union operation. Similarly, descriptive statistics were not relevant to the fractal algorithm and to the UAE algorithm, as they directly return quantitative values as proxy of the image complexity.

5 Results

5.1 Average complexity ranking by human respondent

The crowd-sourced survey and the results are available online. Out of the 458 submitted responses, 4 were filtered out as the respondent filled the survey in less than 5 min, thus leaving a total of 454 accurate submissions. Figure 3 shows that there is consistency between the average position from all respondents' ranking (left) and the breakdown of complexity categorization by share of respondents (right). The variability around the average ranking is shown in Fig. S. 2. On the right, it seems photographs are clustered in couples or triplets with a rather homogeneous categorization. For the subsequent analyses, the average complexity ranking is assumed as the ground-truth to which computational indexes are compared against.

5.2 Statistical match between human and computer-based complexity ranking

The complexity rankings issued from the computer vision indicators mentioned in Sect. 3.3 have been tested for correlation with the average complexity ranking from human respondents (Fig. 4). A linear regression model was set-up to this purpose (simple linear regression after a systematic centring and scaling), using each indicator as an independent variable, respectively, and the average complexity ranking from human respondents as the predicted (dependent) variable. To evaluate the performance of each of the 7932 linear regressions, the corresponding coefficients of determination (R^2) and Pearson's correlations have been calculated. This correlation in absolute value is greater than or equal to 0.63 for 461 indices (nearly 6%), which is not surprizing as the presented approach was testing a large number of candidate indexes and statistics: the maximum of 0.79 is reached for an H0 indicator (derived from a 16×16 tiling of an HLS image). Indexes that do not reach the 0.63 threshold in absolute value on correlation have been discarded from further analysis.

Given the high number of selected indicators (461) compared to the low number of photographs (25), the risk of overfitting was limited by sub-setting the dataset into a test set

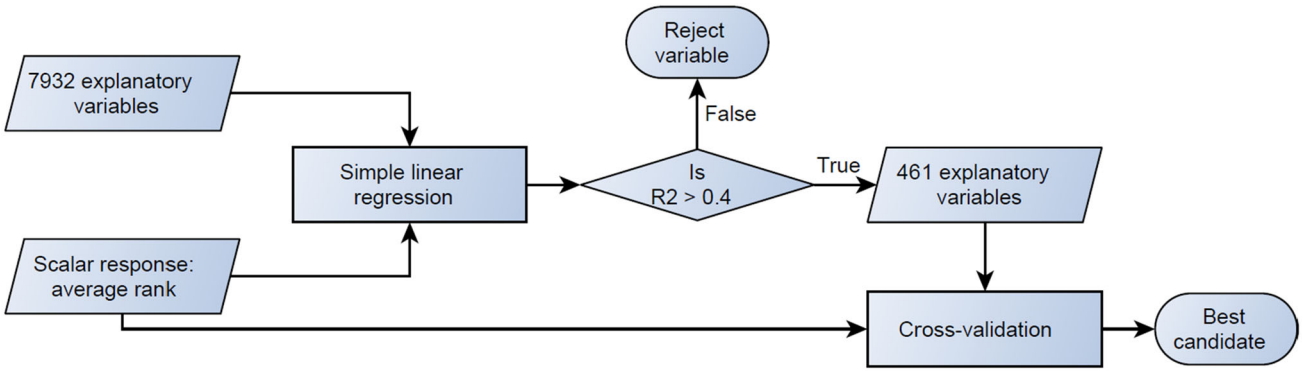

Fig. 4 Procedure to select the computer vision indicators for cross-validation, indicating the best candidates

of 7 photographs, shuffled and extracted pseudo-randomly 5 times for cross-validation.

With the "shuffle split" cross-validation technique, the performance of the indicator (the mean R^2 score) decreases in comparison with its overall correlation with the whole dataset, as it is tested across multiple (namely 5) subdivisions of the dataset. As an example, the cross-validated mean R^2 score of a simple regression model based on the UAE indicator is 0.488 (with a standard deviation of 0.142), compared to an overall R^2 score of 0.578 calculated on the whole dataset (without cross-validation). Figure 5 illustrates the most well-performing (top) indicators, based on their mean Pearson's correlations and coefficients of determination (R^2 scores). From the 461 indices, only 384 were retained, with an average R^2 greater than 0.2 after cross-validation. From these 384 indexes, 19 involve L*a*b* images, 3 involve grey images, 7 involve colour (a*b*) images, 88 involve blurring, 353 involve tiling, 120 involve saliency maps, etc. The best performance after cross-validation is $R^2 = 0.620$ (SD = 0.197) for an indicator that measures the skewness of the variability of colour values around their mean in each HLS component of images split into 16 × 16 thumbnails, which corresponds to hypothesis H0. The top R^2 scores and their standard deviation across test sets are detailed in Table S. 2 (Supplementary materials).

6 Discussion

This massive computational exploration of a large set of indicators from computer vision did not allow identifying with certainty a precise indicator of complexity. However, it allowed observing that some hypotheses (H2, H6 and H7) present very low scores, while others are clearly more conclusive (H0, H4, H5, H3 and even H1). Hypothesis H0 correlated best when associated to the variability of the pixel colours. This may be an unexpected result as the associated index does not use any sophisticated algorithm that extract features (low spatial frequencies, edges, key points, number of object classes, etc.) to represent the complexity. This aspect led to the conclusion that the tested pre-processing altering the image to better underline one of its specificities, subtracts at least one component from this complexity and confirms, in practice, the rather plural character of the latter.

Only a few indicators showed significant correlation with the survey (Fig. 5): in particular, thirty-four indices exceeded the cross-validated R^2 score of 0.5, meaning moderate correlation: the associated hypotheses captured some significant aspects of the subjectively perceived complexity–if not all. Still, none of the tested indicators correlated perfectly with the ranking issued from the average responses of the online survey.

Some interesting results emerged from the analysis of the indicators above a cross-validated R^2 score of 0.5: as explained above, and surprisingly, the family of indicators that correlated best with the survey is a simple variability around the average colour of the image, after tiling, derived from hypothesis H0. The edge detection, under hypothesis H4, follows closely. Subsequently, low frequency features are of particular significance (hypothesis H3), as well as visual attention-grabbing features (hypothesis H5). When separating the colour components (a, b) from the luminance component (L) in CIELAB images, the luminance component alone explains the results of the survey with higher correlations (hypothesis H1), through the given set of computer vision indicators, compared with colour alone (hypothesis H2). The fractal dimension (hypothesis H6) has no correlation at all with the reported complexity. Overall, the presence of many separated image regions, with clear boundaries, independently from their semantics, seems to be linked with the complexity ranking issued from the online survey. In such case, it is difficult for the observer to form a synthetic impression of the image as a whole.

The H7 hypothesis was not much supported by the results, suggesting that an effective mix of bottom-up and top-down

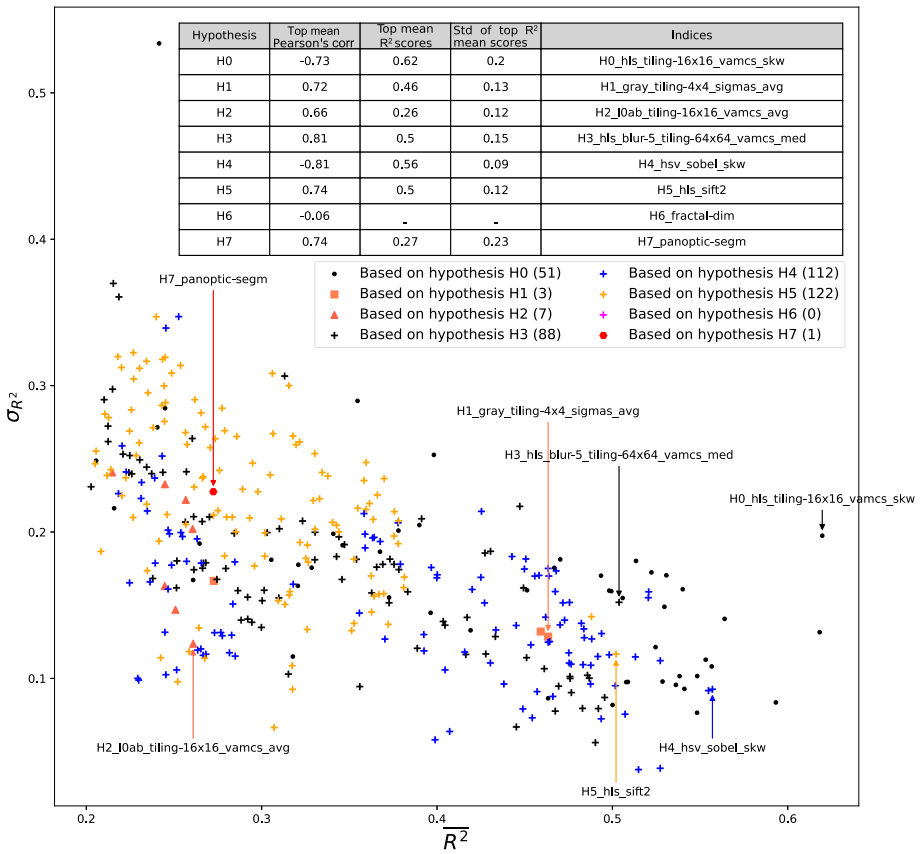

Fig. 5 Stability of cross-validated correlation scores between computer vision indicators and human-based ranking from the online survey. The coefficient of determination R^2 is shown on the horizontal axis against its standard deviation across test sets on the vertical axis. Top mean (cross-validated) scores for R^2 and their relative standard deviation under each hypothesis are shown in the table, along with their Pearson's correlation with the average survey ranking

indicators that correlate soundly with human appreciation of complexity is yet to be found. This may be due to the small number of items in the image collection (25). Enriching the image database would probably make the results more robust, but it would require a different and more effective design of the crowd-sourced survey, to ensure respondents are not bored by manually ranking too many scenes.

The limitations of working with a limited set of images is due to the design of the experiment, which involves a people-based ranking as definition of complexity. The list of computer vision algorithms may also be non-exhaustive, especially on top-down indicators, even if many combinations covering most perceptual aspects were tested here. Adding more indicators to the list, together with more assessed images, would make the results more robust and enable an extensive automation of the process, relying for instance on street view services with large geographic coverage.

7 Conclusion

This paper presents the comparison between crowd-sourced and computer vision-based complexity ranking of urban streetscapes. The average complexity ranking issued by human respondents did not match perfectly with the one derived from computer vision indicators in the wide range selected here. However, correlations show that fragmented colour regions enclosed by sharp edges are evaluated more complex, on average. Among low-level, or bottom-up features, contrasts seem to have a relatively higher importance compared to colours. These results may inform on the perception of urban environments by pedestrians and citizens, driving design strategies to make streetscapes more appreciated. The analysis could potentially be extended to street view services with large geographic coverage.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00138-023-01484-1.

Acknowledgements The authors would like to acknowledge the 458 respondents to the online survey, who dedicated some time to make this research work progress.

Author contributions P.F. conceptualised the work, edited the manuscript, and prepared the dataset. T.L conceptualised the work, wrote parts of the manuscript, and analysed the dataset using the computer vision algorithms. Y.S. supervised, reviewed the manuscript, and provided further analyses. R.B. conceptualised the work, supervised, and reviewed the manuscript.

Funding Open access funding provided by EPFL Lausanne.

Data availability The data that support the findings of this study are openly available in FigShare at the following link: https://doi.org/10.6084/m9.figshare.c.6662132.v1.

Declarations

Competing interests The authors declare no competing interests.

Ethical approval Image rankings and responses from the online survey were collected in complete anonymity, and by no means, the responders can be identified or linked to their submissions. The survey participants were informed about the scope of the research, the anonymity of the data collection and its storage in the opening page of the web survey, by the authors as the data owners. In this setting, this research did not require any ethics approval by the institution which collected the data (Nantes Université, ENSA Nantes, École Centrale Nantes, CNRS, AAU-CRENAU, UMR 1563, F-44000 Nantes, France).

References

1. Gibson, J.J.: The Ecological Approach to Visual Perception Classic. Psychology press, London (1979)
2. Gibson, J.J.: The perception of the visual world. Cambridge (1950)
3. Cullen, G.: The Concise Townscape. (1961)
4. Lynch, K.: The image of the environment-The city form. In: The Image of the City. App A&B: Cambridge (1960)
5. Donderi, D.C.: Visual complexity: a review. Psychol. Bull. **132**, 73–97 (2006). https://doi.org/10.1037/0033-2909.132.1.73
6. Gibson, J.J.: The Senses Considered as Perceptual Systems. Houghton Mifflin, Oxford, England (1966)
7. Gregory, R.L.: The Intelligent Eye. Mcgraw-Hill, London (1970)
8. Treisman, A.M.: Selective attention in man. British Med. Bull. **20**, 12–16 (1964). https://doi.org/10.1093/oxfordjournals.bmb.a070274
9. Rapoport, A., Hawkes, R.: The perception of urban complexity. J. Am. Inst. Plann. **36**, 106–111 (1970). https://doi.org/10.1080/01944367008977291
10. Appleton, J.: The Experience of Landscape. Wiley, London (1975)
11. Elsheshtawy, Y.: Urban complexity: toward the measurement of the physical complexity of street-scapes. J. Arch. Plann. Res. **14**, 301–316 (1997)
12. Ewing, R., Handy, S., Brownson, R.C., Clemente, O., Winston, E.: Identifying and measuring urban design qualities related to walkability. J. Phys. Act. Health **3**, S223–S240 (2006). https://doi.org/10.1123/jpah.3.s1.s223
13. Seinfeld, J.H., Pandis, S.N.: Atmospheric Chemistry and Physics: From Air Pollution to Climate Change. Wiley, New Jersey (2016)
14. Higuchi, T.: The Visual and Spatial Structure of Landscapes. MIT Press, Cambridge-Mass. & London (1983)
15. Bartie, P., Reitsma, F., Kingham, S., Mills, S.: Incorporating vegetation into visual exposure modelling in urban environments. Int. J. Geogr. Inf. Sci. **25**, 851–868 (2011). https://doi.org/10.1080/13658816.2010.512273
16. Llobera, M.: Modeling visibility through vegetation. Int. J. Geogr. Inf. Sci. **21**, 799–810 (2007). https://doi.org/10.1080/13658810601169865
17. Półrolniczak, M., Kolendowicz, L.: The influence of weather and level of observer expertise on suburban landscape perception. Build. Environ. **202**, 108016 (2021). https://doi.org/10.1016/j.buildenv.2021.108016
18. Stigsdotter, U.K., Corazon, S.S., Sidenius, U., Kristiansen, J., Grahn, P.: It is not all bad for the grey city–a crossover study on physiological and psychological restoration in a forest and an urban environment. Health Place **46**, 145–154 (2017). https://doi.org/10.1016/j.healthplace.2017.05.007
19. Florio, P., Peronato, G., Perera, A.T.D., Di Blasi, A., Poon, K.H., Kämpf, J.H.: Designing and assessing solar energy neighborhoods from visual impact. Sustain. Cities Soc. **71**, 102959 (2021). https://doi.org/10.1016/j.scs.2021.102959
20. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Machine Intell. **20**, 1254–1259 (1998). https://doi.org/10.1109/34.730558
21. Wilson, R.T., Casper, J.: The role of location and visual saliency in capturing attention to outdoor advertising: how location attributes increase the likelihood for a driver to notice a billboard ad. J. Adv. Res. **56**, 259 (2016). https://doi.org/10.2501/JAR-2016-020
22. Boeing, G.: Measuring the complexity of urban form and design. Urban Des Int. **23**, 281–292 (2018). https://doi.org/10.1057/s41289-018-0072-1
23. Zhang, B.: Computer vision vs. human vision. In: 9th IEEE International Conference on Cognitive Informatics (ICCI'10). pp 3–3. IEEE, Beijing, China (2010)
24. Cavalcante, A., Mansouri, A., Kacha, L., Barros, A.K., Takeuchi, Y., Matsumoto, N., Ohnishi, N.: Measuring streetscape complexity based on the statistics of local contrast and spatial frequency. PLoS ONE **9**, e87097 (2014). https://doi.org/10.1371/journal.pone.0087097
25. Machado, P., Romero, J., Nadal, M., Santos, A., Correia, J., Carballal, A.: Computerized measures of visual complexity. Acta Physiol. (Oxf.) **160**, 43–57 (2015). https://doi.org/10.1016/j.actpsy.2015.06.005
26. Ma, L., He, S., Lu, M.: A measurement of visual complexity for heterogeneity in the built environment based on fractal dimension and its application in two gardens. Fractal Fract. **5**, 278 (2021). https://doi.org/10.3390/fractalfract5040278
27. Nagle, F., Lavie, N.: Predicting human complexity perception of real-world scenes. R. Soc. Open Sci. **7**, 191487 (2020). https://doi.org/10.1098/rsos.191487
28. Saraee, E., Jalal, M., Betke, M.: Visual complexity analysis using deep intermediate-layer features. Comput. Vis. Image Underst. **195**, 102949 (2020). https://doi.org/10.1016/j.cviu.2020.102949
29. Guo, M.-H., Xu, T.-X., Liu, J.-J., Liu, Z.-N., Jiang, P.-T., Mu, T.-J., Zhang, S.-H., Martin, R.R., Cheng, M.-M., Hu, S.-M.: Attention mechanisms in computer vision: a survey. Comp. Visual Media. **8**, 331–368 (2022). https://doi.org/10.1007/s41095-022-0271-y
30. Li, Y., Zhang, C., Wang, C., Cheng, Z.: Human perception evaluation system for urban streetscapes based on computer vision algorithms with attention mechanisms. Trans. GIS **26**, 2440–2454 (2022). https://doi.org/10.1111/tgis.12882

31. Gunawardena, G.M.W.L., Kubota, Y., Fukahori, K.: Visual complexity analysis using taxonomic diagrams of figures and backgrounds in Japanese residential streetscapes. Urban Stud. Res. **2015**, 173862 (2015). https://doi.org/10.1155/2015/173862
32. Tomczuk, P., Wytrykowska, A.: Digital billboards dynamic luminance measurements. MATEC Web Conf. **231**, 04013 (2018). https://doi.org/10.1051/matecconf/201823104013
33. Department of Town and Country Planning, Faculty of Architecture, University of Moratuwa, Sri Lanka, Gunawardena, G.M.W.L.: Evaluation of streetscape complexity created by streetscape signage using different objective analysis techniques. In: Presented at the International Conference on Arts and Humanities February 26 (2019)
34. Guan, F., Fang, Z., Wang, L., Zhang, X., Zhong, H., Huang, H.: Modelling people's perceived scene complexity of real-world environments using street-view panoramas and open geodata. ISPRS J. Photogramm. Remote. Sens. **186**, 315–331 (2022). https://doi.org/10.1016/j.isprsjprs.2022.02.012
35. Kacha, L., Matsumoto, N., Mansouri, A.: Electrophysiological evaluation of perceived complexity in streetscapes. J. Asian Arch. Build. Eng.. **14**, 585–592 (2015). https://doi.org/10.3130/jaabe.14.585
36. Kawshalya, L.W.G., Weerasinghe, U.G.D., Chandrasekara, D.P.: The impact of visual complexity on perceived safety and comfort of the users: a study on urban streetscape of Sri Lanka. PLoS ONE **17**, e0272074 (2022). https://doi.org/10.1371/journal.pone.0272074
37. Bradski, G.: The OpenCV Library. Dr. Dobb's Journal of Software Tools. (2000)
38. Montabone, S., Soto, A.: Human detection using a mobile platform and novel features derived from a visual saliency mechanism. Image Vis. Comput. **28**, 391–402 (2010). https://doi.org/10.1016/j.imavis.2009.06.006
39. Hou, X., Zhang, L.: Saliency detection: a spectral residual approach. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8. IEEE, Minneapolis, MN, USA (2007)
40. Lowe, D.G.: Object recognition from local scale-invariant features. In: Proceedings of the Seventh IEEE International Conference on Computer Vision. pp. 1150–1157, Kerkyra, Greece (1999)
41. Rublee, E., Rabaud, V., Konolige, K., Bradski, G.: ORB: an efficient alternative to SIFT or SURF. In: 2011 International Conference on Computer Vision. pp. 2564–2571. IEEE Barcelona: Spain (2011)
42. Schneider, C.A., Rasband, W.S., Eliceiri, K.W.: NIH image to image J: 25 years of image analysis. Nat. Methods **9**, 671–675 (2012). https://doi.org/10.1038/nmeth.2089
43. Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R.: Masked-attention Mask Transformer for Universal Image Segmentation. (2021) doi https://doi.org/10.48550/ARXIV.2112.01527

Pietro Florio holds a PhD from the Solar Energy and Building Physics Lab at EPFL. He has been working for several years on climate resilience and renewable energies planning in cities. His areas of expertise include building energy modelling and monitoring. Energy Poverty, for which he has been part of the COST Engager Action and had an active role in research and development both in Italy and France; Parametric Architecture, for which he collaborated in a workgroup within the COST RESTORE community; Solar Energy and Urban Planning, for which he has been expert and discussion leader in several IEA Tasks.

Thomas Leduc is a CNRS research engineer at Nantes Université, School of Architecture of Nantes. An applied mathematician by training, he holds a Ph.D. in computer science from the Université Pierre et Marie Curie (now renamed Sorbonne University). His research is devoted to understanding urban forms and spatial dynamics, with a particular focus on forms captured or experienced by the pedestrian through visibilities in particular. This understanding embraces various registers of form analysis including landscape, bioclimatic, layout, or interrelationships between components of the urban fabric.

Yannick Sutter is a lecturer in lighting and bioclimatic design at the School of Architecture of Montpellier and a researcher at Nantes Université, School of Architecture of Nantes. He is a building physics engineer with a PhD on daylighting and visual comfort. His research focuses on the study of characterization and perception of lighting in architectural and urban environments.

Roland Brémond has an Engineer degree from the Ecole des Ponts ParisTech and holds a PhD in image processing from the Ecole des Mines ParisTech. He is a full-time researcher at the Université Gustave Eiffel (France). His field of research is visual perception applied to practical outdoor situations, such as visibility on the road, urban lighting, glare and visual attention. His expertise includes human and sensor vision, image processing and computer graphics, and he is co-author of more than 50 papers in international peer review journals.