Navigation using Special Buildings as Signposts

Julien Weissenberg  
ETH Zurich  

Michael Gygli  
ETH Zurich  

Hayko Riemenschneider  
ETH Zurich  

Luc Van Gool  
ETH Zurich, K.U. Leuven

ABSTRACT

Navigation has been greatly improved by positioning systems, but visualization still relies on maps. Yet because they only represent an abstract street network, maps are sometimes difficult to read. Conversely, Tourist Maps, which are enriched with landmark drawings, have been shown to be much more intuitive to understand. However, outside of a city centre, major landmarks are too sparse to be helpful. In this work, we present a method to automatically augment maps with drawings of the most locally prominent landmarks, at every scale. Further, we generate a characterization which helps emphasize the special attributes of these buildings. Descriptive features are extracted from facades, analyzed and re-ranked to match human perception. To do so, we collected a total number of over 5900 human annotations to characterize 117 facades across 3 different cities. Finally, the characterizations are also used to produce natural language descriptions of the facades.

Categories and Subject Descriptors

I.4.9 [Computer Graphics]: Image Processing And Computer Vision Applications

Keywords

map, city, facade, building, automatic description, visual summarization, visual perception, search, navigation

1. INTRODUCTION

During the last decade, navigation aids have undergone a genuine revolution. Rather than having to trace their trajectories on maps, people get real-time information about where they are and what to do next to reach their goal. Nonetheless, much of this novel technology is still grounded in the use of traditional maps. Unfortunately, many people find maps difficult to use [3]. As perception research has shown [24], people tend to remember trajectories on the basis of decisions relative to landmarks, rather than guidelines like ‘first take the second street on the left...’. What makes navigation still difficult? First, once on the road, street view navigation (which would help localize those landmarks) is too cumbersome to continuously use. Second, maps rely on street names and house numbers, which are not always visible. Last, navigational aids would be most needed where the language or even alphabet are different, precisely when maps are the most difficult to use.

As a matter of fact, traditional tourist maps offer a good compromise for this. They combine an - often metrically incorrect - map with drawings of major landmarks. The creation of such maps was not a coincidence, as it offers people the appearance-based reference they would not find on traditional maps. Such maps can be generated automatically by mining information about landmarks from the web [9].

Figure 1: Our automatically mined local landmark buildings maps (top) vs. Street View (middle) vs. Street network (bottom) maps. The last two are respectively hard to browse or very abstract. We mine atypical buildings and highlight them at any given scale, even if there are no major landmarks.
Nevertheless, traditional tourist maps are only useful in city centres, where landmarks are dense. The visual aid quickly starts to lose its effect once the user enters the maze of smaller streets, where there are no tourist landmarks such as an Eiffel tower or a Big Ben.

In this work, we propose to use an adaptive tourist map that automatically shows the most helpful buildings to navigate in any area. The destination and all important buildings on the user’s path can be highlighted, at every scale.

Since they are visible from street level, we work directly at the facade level. We analyze each facade and produce a semantic description of its main architectural characteristics. This description is based on the differences to others in the neighbourhood. In particular, our work learns the saliency of different facade attributes, to emphasize what is important in the eyes of humans.

Last but not least, our method also is capable to produce language-based descriptions of the facades. Therefore, even when it is dangerous to take one’s eyes off from the street when walking or driving, a spoken voice can provide descriptions of the landmarks of interest. Our contributions are:

- A method for discovering atypical buildings in an area.
- A method to produce a visual characterization and describe a facade in natural language relative to its surroundings.
- We study the relevance of different features as perceived by humans, provide a statistical analysis and adapt our tool to match human perception.
- We introduce a new facade dataset for the city of Zurich, and a dataset of 5904 human annotations for unusualness of 117 facades across Graz, Paris and Zurich.

2. RELATED WORK

The topic most related to our work is facade saliency for navigation. In [19, 16], the ideas of selecting local landmarks is defined and features covering many aspects such as building visibility and function are used. The main differences to our work is we focus on the visual and perceptual aspect. We use extra visual features and a novel outlier detection pipeline to match with the human perception of saliency. In [27] simple linear weights between different saliency features are established experimentally. [18] focuses on evaluating facade saliency depending on the user context. Finally, [21] presents challenges in automatic facade mining.

Concerning maps, the two works most related to ours are [9, 8]. In the first, maps are augmented with visual abstractions of landmarks. The position and pictures of landmarks are mined from the internet. In our work, we also augment a map with unique buildings displayed for reference, which are mined following a unusualness analysis for facades. The seminal work “What makes Paris look like Paris?”[8] solves our inverse problem, i.e. identifying the patterns that are typical for a city. In contrast, we find the oddities of facades within a style.

In recent times, much effort has been devoted to augmenting maps with 3D models. These are however very large and still prone to reconstruction errors. As a result, semantic labelling has been used as a way to address these two issues. For example, image-based modelling [28] and procedural modelling [15] allow to enhance the reconstruction result and reduce the amount of data. Yet, the resulting maps are only not accurate and visually pleasing, and not a fundamentally better tool for navigation. In this work, we aim at filtering out the excessive amount of information pertaining to buildings to help users navigate.

Oddities, i.e. unusualness can be defined from different angles. In general, an event is usual if it can be explained by prior knowledge, and unusual if this contradicts. From a statistical perspective, we call an unusual event an outlier. As there is no unique and formal definition of an outlier, we use the one of [1]: “An observation which appears to be inconsistent with the remainder of that set of data.” Methods for unusualness detection rely on diverse techniques including statistical models, clustering, entropy-based methods etc. We refer the reader to the survey by [5]. We use in particular the Local Outlier Factor (LOF) [4], which gives a measure of the relative degree of isolation of a sample.

Despite all this, what is unusual may not necessarily match the human perception of what is important in an image. In particular, the works of [25, 2] focus on determining the human-centric importance of attributes. Compared to [25], we are able to find the most unusual attribute for a single instance.

Automatically mining atypical buildings and establishing what makes a facade special come with challenges that have not been dealt with before. First, we must find what makes a facade special compared to its neighbours, bearing in mind that two facades from the same street often look very similar. Then, we have to rely on differences which are easily spottable by a non-specialist: Pointing out that a facade is neo-classical will most likely and regrettably not help.

3. APPROACH OVERVIEW

Fig. 2 gives an overview of our approach to discover atypical facades and what attributes make them special. We start with ortho-rectified facades and semantic labellings of each architectural asset. Automated solutions, which have steadily and rapidly improved in the last few years, exist [23, 14] to obtain such labelling. We denote the following terms: Asset is the main architectural elements of a facade, e.g. windows, doors, balconies. Attribute types are the properties of the assets, e.g. size, color, material, shape. Attribute values are the numerical or fuzzy values, e.g. ‘tall’ or ‘blue’.

First, features are computed for each facade and for each asset within the facades. An outlier score is estimated for each value with respect to the neighbourhood. When a feature value is an outlier with respect to the distribution of feature values for other facades in the neighbourhood, it is
potentially perceived as an unusual feature. In this work, we tackle two questions: mining unusual facades and characterizing a facade by describing its unusual features.

4. WHAT MAKES A FACADE UNIQUE?

To spot what makes a facade unique, we need to identify its characteristic features, i.e. those which help discriminate between different facades. These discriminative features correspond to the most unusual ones (see Section 2 for the definition of unusualness). For our purpose, the challenges are numerous. First, abnormality is a contextual notion. For instance, a balcony on a facade may be perceived as unusual or very usual depending on the style, the street and the neighbouring facades. Then, the boundary between usualness and unusualness is fuzzy. This makes it difficult to obtain ground truth data, which will also be fuzzy in nature. Finally, in our case, what is measured as statistically unusual must be matched with the human perception of unusualness. In particular, some features may be perceived by humans as more important than others, or easier to perceptually distinguish. Therefore, the list of statistically most unusual elements needs to be re-ranked such that the resulting list matches with what humans perceive as unusual.

In this section, we first introduced notations and a method to mine most unusual facades and what is special within a facade.

4.1 Notations

We are given a set of rectified facade images $\mathcal{I}$ and their associated labellings $\mathcal{L}$. Each labelling consists of a set of bounding boxes $\mathcal{B}$. Each bounding box $b \in \mathcal{B}$ is associated with a set of features $\mathcal{A}$. Let $a_m \in \mathcal{A}$ denote a feature. As a result, each facade is associated with a matrix $\mathcal{F}$ comprising $|\mathcal{A}|$ columns and $|\mathcal{B}|$ lines, where $|\mathcal{A}|$ and $|\mathcal{B}|$ designate the number of elements in $\mathcal{A}$ and $\mathcal{B}$ respectively.

4.2 Features

Each feature needs to comply with the following condition:

1. If an attribute value is an outlier with respect to its distribution, it must be perceived by humans as an outlier.

2. The feature must be visually abstractable and translatable in words, such that it is immediately understandable by a human with no training.

To determine the set of features which are suitable for the task, we set up the following user studies. Ten participants were asked to tell what is special about a facade with respect to two other facades in a maximum of 40 characters, for the Zurich dataset described in Section 5. In Table 1, the free text line presents an overview of the analysis of the 190 descriptions. Note that structural differences are more important than others, or easier to perceptually distinguish. Therefore, the list of statistically most unusual elements needs to be re-ranked such that the resulting list matches with what humans perceive as unusual.

In this section, we first introduced notations and a method to mine most unusual facades and what is special within a facade.
In essence, we want to predict the unusualness probability \( u_{\text{predict}}(q_M) \) for mining and \( u_{\text{predict}}(q_C) \) for characterization, where \( q_M \) and \( q_C \) are the feature vectors. We refer to \( p_M \) and \( p_C \) as the perceived importance, which are used to re-rank the scores from the LOF outliers, as described in the following section. The regression, which is learnt per dataset, helps moving from statistical outlier detection to perceived unusualness.

**Regression training** The regression \( r \) is trained using the Decision Tree regression from [13] to predict the importance \( p \):

\[
    u_{\text{predict}}(q, o) = r(q, o) \tag{2}
\]

We now detail the vector design and the importance matching.

**Vector design** The tuple \((q, o)\) is the vector containing the LOF score \( o \) per facade or asset concatenated with raw feature values \( q \).

**Importance matching** We define the perceived importance \( p \), which we want to regress:

\[
    p = \frac{n_q}{N} \tag{3}
\]

where \( n_q \) is the number of times a facade image or a feature value/element type combination was selected, \( N \) is the total number of times a facade image was shown or the total number of cited feature value/element type combinations.

A grid search indicated that performance is robust with respect to the number of trees and highest with a moderate number of trees. We keep the number of trees fixed to 200 trees for both regressions and all datasets. For training, we employ a leave-one-out training. Since we are interested in retrieving the most salient facades, the regression is also trained such that it gives more emphasis to the training examples with a high importance. To do so, we applied a linear weight equal to \( p \) to minimize the error of the facade images or features and assets which were perceived as most important.

Fig. 3 shows the effect of the mapping within the facade, which improves our result. Note the very large difference between the two heatmaps: statistical unusualness needs to be re-mapped to match with human perception of unusualness.

The large difference is to be expected, as the statistical degree of different attributes cannot be directly compared without re-weighing according to human perception. In the example, the window at the top is statistically the most unusual element, but is not perceived so by humans.

### 4.4 Clustering

Clustering is needed for natural language characterization, in order to group elements when referring to them. For example, if all windows are large, they should be referred to as a single entity (“the large windows”) rather than picking one of them (“the large window at the top-left”). Therefore, the facade elements are clustered according to their colours, widths and heights using Mean-shift [7]. The advantage of Mean-shift is that the number of clusters or variances does not need to be known beforehand, and only the bandwidth needs to be set. In our experiments, we set the bandwidth
Figure 3: Heatmap representation of measured unusualness. Left: facade image, Middle: using LOF results, Right: after mapping according to human perception.

Figure 4: Map highlighting the most unusual building around an itinerary in Graz.

to 17, after visual inspection of a few examples.

5. EXPERIMENTAL EVALUATION

For the experiments, we use three different datasets, comprising 50, 47 and 20 facade images respectively. Each dataset corresponds to a single street or area, meaning the styles are expected to be similar within each dataset. Each ortho-normal image is associated with a ground truth labelling. Labels are Door, Wall, Sky, Window, Shop, Balcony, Roof and sometimes Shutters.

The Graz50 [20] and ECP2011 [23] datasets have originally been designed to assess the quality of facade parsing and consist of labelled rectified pictures. In addition, we introduce the Zurich20 dataset. In all experiments, the data was split into training and testing in a leave-one-out fashion. Overall, 423 participants took part in the AMT experiments. 62% and 38% of the participants were female and male subjects respectively. Their ages ranged from 18 to 65 (average 31.9).

5.1 Special facade mining

The unusual facades are mined using (2) and the method presented in Section 4. From there, we generate a map showing the most unusual buildings. We gathered ground truth data for Graz, Paris and Zurich. Table 2 gives a summary of the collected data and the performance of the regression. Note that the regression does not need to be very precise for low scored facades, as in the end the top results are displayed. Therefore, we also report some of the TopK scores in Fig. 7. The TopK score quantifies how well the computer-ranked top K images agree with the human ranking [10]. The score ranges from 0 to 1, where a perfect agreement of the two rankings leads to a score equal to 1.

As can be seen, our mining method yields a moderate to strong correlation for the Graz50 and Zurich20 datasets, while correlation is lesser on the ECP2011 dataset. We notice that in the AMT experiments, the standard deviation of the building importance score is smaller for ECP2011 (0.082) compared to Graz50 and Zurich20 (0.13 and 0.17). Thus, for ECP2011 (which is an extreme example of facade regularity) ranking building importance can be expected to be a harder task than for other datasets. Fig. 4 shows an example of the resulting map.

5.2 Perceptual study for characterization

Experimental setup Participants on AMT were instructed to describe a building to a friend using the features in Table 1. All details are published online.

Results Fig. 5 presents statistics about the Zurich dataset. In our experiments, we considered the two neighbouring facades. We notice that the colour plays an important role in identifying unusual elements to discriminate between facades. In contrast, the Local Outlier Factor (LOF) gives more importance to the size differences. The regression reestablishes the dominance of colour over size. Also, sizes are not used equally by humans: large, tall and wide elements constitute a large majority of the sizes which are stated. This can be explained by the fact that it is easier to identify a large element than a smaller one. Finally, rare architectural elements are relatively more important. For example, although doors are much rarer than windows, they are cited relatively frequently.

Table 3 gives a summary of our performance results. The importance rankings are published online.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Zurich20</th>
<th>Graz50</th>
<th>ECP2011</th>
</tr>
</thead>
<tbody>
<tr>
<td># facades</td>
<td>20</td>
<td>50</td>
<td>47</td>
</tr>
<tr>
<td># subjects</td>
<td>87</td>
<td>53</td>
<td>51</td>
</tr>
<tr>
<td># annotations</td>
<td>1473</td>
<td>2531</td>
<td>2100</td>
</tr>
<tr>
<td>Corr. (LOF)</td>
<td>0.25</td>
<td>0.13</td>
<td>0.07</td>
</tr>
<tr>
<td>Corr. (re-ranked RF)</td>
<td>0.52</td>
<td>0.73</td>
<td>0.57</td>
</tr>
</tbody>
</table>

Table 2: Summary of the facade mining study results. RF: Random Forest, Corr: Pearson correlation. A correlation of 1.0 would mean that we can perfectly predict the distribution of the responses of humans for each facade.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Zurich20</th>
<th>Graz50</th>
<th>ECP2011</th>
</tr>
</thead>
<tbody>
<tr>
<td># facades</td>
<td>20</td>
<td>50</td>
<td>47</td>
</tr>
<tr>
<td># subjects</td>
<td>20</td>
<td>55</td>
<td>53</td>
</tr>
<tr>
<td># annotations</td>
<td>400</td>
<td>1100</td>
<td>1060</td>
</tr>
<tr>
<td>Corr. (re-ranked RF)</td>
<td>0.60</td>
<td>0.40</td>
<td>0.17</td>
</tr>
</tbody>
</table>

Table 3: Summary of the facade characterization study results. Corr: correlation, LOF: Local Outlier Factor, RF: Random Forest
correlation refers to the correlation between the predicted importance scores and the importance scores measured in the experiments. The Pearson correlation is much higher after re-ranking than when using the normalized LOF scores. This explains the large difference in the heatmaps in Fig. 3.

### 5.3 Facade characterization

For each facade, we obtain both an importance map highlighting the main characteristics and a natural language description. These could be used as input to methods such as [6, 11] to emphasize the most discriminative parts.

The natural language description can be turned into speech and is useful in case one does not want to look at a screen or a map, for instance when driving a car. Figure 6 show an exemplar result of a generated sentenceootnote{All results are published online.}.

![Figure 6: Our automatic description of the facade in the middle: “The facade with green wall and wide red windows.” On right, the heatmap again.](image1)

### 6. CONCLUSION

Facades are at the heart of the urban landscape. The next step for digital mapping has arrived: understanding a facade like a human. In this work, we have introduced a method to analyse facades by mining their atypical features. This solves two fundamental problems in digital mapping. First, we can produce more readable maps even if no famous landmark is present in the area. Second, the descriptions are valuable for other media, such as audio or text. Last but not least, this work illustrates a generic method to automatically learn where to place emphasis, which is a crucial issue in visualization. The resulting data tells us which buildings are considered as important to be shown on a map and what is important within these buildings. This constitutes a valuable input to methods such as [6, 11] who need this information for attracting attention or rescaling. For future work, we plan to add more features (for material and shape), as well as use city-scale datasets and 3D models for the buildings to make the resulting maps more readable.

Acknowledgements. This work was supported by the European Research Council (ERC) under the project VarCity (#273940) at www.varcity.eu.
7. REFERENCES


