3D Saliency for Finding Landmark Buildings

Nikolay Kobyshev* and Hayko Riemenschneider* and András Bódis-Szomorú* and Luc Van Gool*†
* Computer Vision Laboratory, ETH Zurich, Switzerland
†K.U. Leuven, Belgium
{nk,hayko,bodis,vangool}@vision.ee.ethz.ch

Abstract

In urban environments the most interesting and effective factors for localization and navigation are landmark buildings. This paper proposes a novel method to detect such buildings that stand out, i.e. would be given the status of 'landmark'. The method works in a fully unsupervised way, i.e. it can be applied to different cities without requiring annotation. First, salient points are detected, based on the analysis of their features as well as those found in their spatial neighborhood. Second, learning refines the points by finding connected landmark components and training a classifier to distinguish these from common building components. Third, landmark components are aggregated into complete landmark buildings. Experiments on city-scale point clouds show the viability and efficiency of our approach on various tasks.

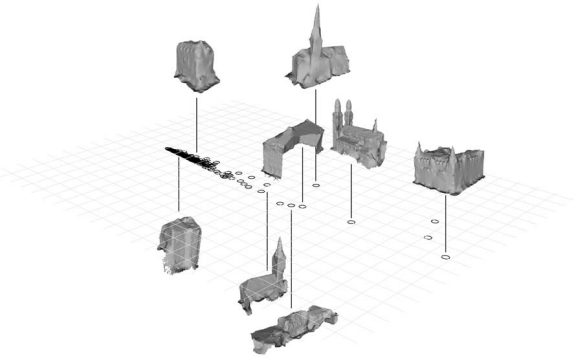

Figure 1. Chart for landmark buildings – what makes a building a landmark? Using our our landmarkness measure, we can find a distinction between buildings. The majority of (ordinary) buildings are grouped together (high density circles on left), and the landmarks stand apart.

1. Introduction

Across-the-board visualization of urban data is not necessarily the best way to aid people navigating, be it as 2D maps or 3D models [34]. Indeed, in order to maximally assist people, alternative visualizations may be preferable. A good example are tourist maps, which may not be metrically correct, but include the visually salient parts, like landmark buildings that stand out. One could consider them a mix of 2D and 3D visualizations.

Buildings are the single most interesting and representative objects in a city. They are often linked to pivotal moments in history or are just stunning to look at. What makes a building a landmark is not so well-defined though in general. [22] define landmark buildings as 'uniquely memorable in the context of the surrounding environment'. On top of visual aspects such as unique structural design, level of decoration, or monumentality (visual), a landmark can also derive from historical or societal connotations (cognitive), or simply from an exquisite location (structural) [40].

Only a few papers have appeared that ease the production of such maps. Grabler et al. [22] produce tourist maps automatically, based on a wide gamut of criteria. Landmark buildings are suggested by tourist websites, and are then graded on the basis of multiple characteristics, including color, location, but also shape. The shape features are rather straightforward though, including not filling a rectangular box well or exhibiting irregular triangular meshes. We contribute a method that looks for 3D shapes that stand out, based on the local context. Indeed, however special a building may be, if there are many similar buildings around, it may lack all those features and become the salient one.

A second strand of related work consists of methods to detect interest points and features in 3D models to be able to identify important 3D parts. The work of Shtrom et al. [38] in this area defines a local saliency to identify areas in 3D point clouds which may be useful. This approach determines unique parts however not at the scale of buildings and results in local responses.

We propose a method to analyze the 3D models of buildings and to rank them in terms of how special their structure is. First, our method takes such contextual influences into account evaluating across an entire city. Second, our approach aims at automatically discovering such salient build-

978-1-5090-5407-7/16 $31.00 © 2016 IEEE
DOI 10.1109/3DV.2016.35

ings, instead of depending on websites listing them. Hence, our method also works for smaller towns for which the websites needed by [22] cannot be found.

Admittedly, this will not capture all buildings regarded as landmarks, but nonetheless captures most of those landmarks that an average tourist would like to see due to their special structure.

In the context of this paper, 'special' is defined in terms of high similarity yet global rarity. We give a high score for similar points that only occur within a local neighborhood (low scatter) and a low score to those that occur at a large distance (high scatter). For this we propose a novel saliency measure for localized component extraction, a discriminative learning designed to localize additional components in a city, and finally a scheme for building-wise aggregation. In Fig. 1 we give a preview of our chart which separates buildings extracted from a city point cloud into landmarks and ordinary buildings.

To the best of our knowledge we are the first to look at 3D structural patterns of 3D buildings and at the scale of an entire city. Our results show a clear benefit of our proposed method over more directly related work in the field of point cloud analysis. This paves the way for navigating around the interesting landmarks [34, 41] as well as fully automatic visual tourist map generation [22].

2. Related Work

Since this work touches various fields, we highlight the closest topics, i.e. finding discriminative unique elements in a given dataset. The literature here is related to methods for detecting outliers and mining discriminative patches. In the following we highlight the cornerstones of each field.

2.1. Outlier Detection

Outlier detection methods try to find statistically those elements which stand apart. Datta and Wand [11] propose a *familiarity* feature as the average distance of a test image to the k-closest training images. The higher this distance, the less familiar (more novel) is an image.

Zhong et al. [43] train a model from frequent observations and simply label resultant outliers as special. In this vein, the Local Outlier Factor method proposed by Breunig et al. [8] uses a neighborhood and computes the degree of an outlier.

Saliency detection is a mature field itself and elaborate surveys such as [5, 16, 24] cover more than 250 different methods. The definition of saliency however is not so clear and varies across papers. In general, the saliency of a data point within a large set of other data points is defined as to what makes that point unique and discriminant w.r.t. its surrounding context.

Among the top 2D approaches is Hou et al. [23] who analyze 2D images in a residual spectral domain, which is difficult to adopt for 3D point clouds.

Other methods use the rarity of a feature [17] which uses the unpredictablility of local attributes like color and orientation to compute the entropy of local patches. Pop-out features [19, 20] try to maximize the response of a given salient patch w.r.t. the background – and in [12] also via neural nets.

Recently, deep learning is employed to more effectively learn the mapping of visual features to a saliency score [27]. However this requires massive amounts of training data for per-eye fixation data, whereas we work in a completely unsupervised manner.

Other recent methods also include depth information into the saliency extraction (RGB+D). Desingh et al. [13, 25] show that depth is a useful key for saliency detection by computing a local saliency descriptor based on the distribution of normals in a segmented depth image.

3D keypoint detectors evaluate saliency over the entire spatial context and then select salient locations for later description. Salti et al. [37] provide an overview to evaluate the compatibility of these detectors and descriptions. For example, [29] detect salient features on point clouds by comparing the differences of a point's normal within its neighborhood on different scale levels (with the goal of point cloud alignment). The main limitation is the purely local detection, without taking account of the wider context.

In the domain of methods that work on 3D data, most similar to ours are the works of Akman and Jonker [1] on small scenes and Shtrom et al. [38] on a medium-scale city block. The latter defines a low-level and high-level distinctiveness that operates on two fixed scales and are finally linearly aggregated. By common definition a point is considered to be salient if has a different appearance from the points close by. We, however, suggest defining a point salient if it looks similar to its local neighborhood, and the appearance of this local neighborhood is not found in other parts of the point cloud. This approach leads to identifying larger salient components and is significantly more robust to outliers, compared to [38].

2.2. Discriminative Mining

Another approach for finding unique parts is viewing this task as an unsupervised discriminative mining or a clustering (i.e. reduction of the data points) problem.

Generally, standard clustering algorithms like K-means, affinity propagation [18], etc. can be used, however they suffer from the frequency curse. That is, very frequent data points dominate over less frequent ones, hence important rare discriminative points may be missed.

Discriminative clustering approaches like [10, 9] start by sampling random data points and training discriminative prototypes. These prototypes are used in ensemble classifications to determine the feature proximity between data

points and hence cluster them more effectively. In a similar fashion, [21] partition entire images, resulting in suitable discriminative classifiers. A suitable classifier has the properties of class separation, class balance and appropriate classifier complexity. Related is also [28], who use curriculum learning [3] for unsupervised object discovery by starting with the easiest clusters and gradually increasing the complexity of clusters.

As a different approach, discriminative mining starts by a few data samples and iteratively refines these. Typically initial clustering of data [42] is followed by learning a discriminative classifier for each cluster. Based on the discriminatively-learned similarity, new cluster memberships can be computed by reassigning data points to each cluster. This idea inspired our work and we formulate the memberships as actual physical parts of landmark components and buildings. However, it will not work for our problem as landmark points occur very infrequently compared to non-discriminative components, and second it is infeasible to cluster a city-scale dataset.

In [15, 39], one can see examples for such discriminative mining to identify patches which are repeating enough to be useful but not too frequent to still be discriminative. They use a compact linear classifier and careful cross-validation to filter out non-discriminative patches and to avoid overfitting. In [14] the idea was further refined as a discriminative mode seeking. They discover visually coherent clusters that are maximally discriminative given weak labels.

In our problem scenario we start completely unsupervised and determine distinctive clusters by their initial saliency which we refine by learning. In summary, our method has the following contributions and benefits:

- first work on landmark building identification;
- saliency in city-scale 3D point cloud data;
- novel saliency-seeking discriminative neighbors;
- iterative refinement of the feature/spatial neighbors;
- completely unsupervised; without manual labeled data;
- scale and context independent, unsupervised, hence general for cities;
- experiments on noisy image-based 3D reconstruction and LIDAR data.

3. Our Approach for Landmark Identification

The goal of our method is to find an unsupervised city-independent landmarkness score. For this we first define a score for each 3D point – based on its uniqueness. Second, we refine this landmark score by discriminative learning across neighborhoods. Finally, we aggregate the individual 3D points into complete and coherent buildings. Only for this we exploit OpenStreetMap and extract building footprints to collect scores from all 3D points within a building.

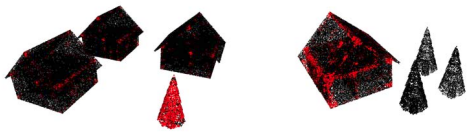

Figure 2. Context-dependency of the notion of landmark, illustrated with our distinctiveness measure for landmarks (red points). Left: almost all unique points belong to the tree since houses are common. Right: the majority of unique features belong to the house as trees are common.

Overall, the notion of landmark is dependent on the context. For architecture, it is the particular city that determines whether an object is a landmark or not. For example, a building having a facade-wide balcony with wrought iron railing cannot be considered a landmark in Paris, since it occurs frequently grace to the Haussmannian renovations. On the other hand, such a balcony would make the building distinctive in Manhattan. A toy example is shown in Fig. 2. In a village with many houses and only a single tree, the tree is special. Whereas a single house in a forest lets the house, not the trees, stand out.

Moreover, most landmark buildings can be identified as landmark due to their unique local features such as special iron railing, ornamentation, windows, ledges, or tower-tops. Exploiting this, we detect unique local features, then aggregate them to find landmark components (such as roofs, walls, towers) and, ultimately, to landmark buildings.

Hence, we define our terminology as follows. A landmark building consists of landmark components. A landmark component comprises distinctive landmark points (salient points, more generally), each characterized by a descriptor of its local neighborhood.

Our method consists of four major steps, as shown in Fig. 3. First, we use a novel measure for distinctiveness (i.e. how likely is a point to be a landmark point; we also call it 'Kobyshev score'). Second, we update the spatial neighborhoods to find landmark components. Third, we refine the measure discriminatively to highlight landmark points by updating the feature neighborhoods. Finally, we propose various ways of aggregation of the Kobyshev point scores into a consistent building score.

3.1. Distinctiveness Measure for Landmarks

In this section, we introduce a simple yet powerful measure of distinctiveness. Later in Section 4 we will demonstrate its benefits over existing saliency measures in the task of landmark identification.

Our intuition is that landmark points are rare unique points, which are locally similar, yet do not occur every-

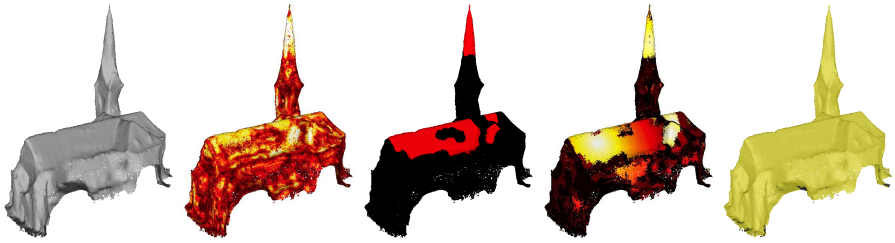

Figure 3. Our pipeline (shown on one object extracted from the city-scale point cloud). Left to right: Input, landmark score $\mathcal{L}$, components $\mathcal{C}$, refined landmark score $\mathcal{L}^*$, building aggregation. The color changes from black (low) to white (high score).

where across the context. For example, descriptors of points on a landmark tower are locally similar, yet are different from descriptors of other points in the 3D city model.

Let $\mathcal{P} = \{p_i\}_{i=1}^{N}$ denote a set of 3D points $p_i = (\mathbf{p}_i, \mathbf{f}_i)$, each characterized by a position $\mathbf{p}_i$ in 3D space $\mathbb{E} = \mathbb{R}^3$ and a feature vector (a.k.a. descriptor) $\mathbf{f}_i$ in feature space $\mathbb{F} = \mathbb{R}^m$ describing the local geometry around location $\mathbf{p}_i$. We use FPFH [36] for a descriptor, but the method is generic enough to accommodate other 3D point feature descriptors. For every point p_i we define:

- *feature neighborhood* $\mathcal{N}^{\mathbb{F}}(\mathbf{f}_i)$, or $\mathcal{N}_i^{\mathbb{F}}$ in shorthand: the set of indices $j_1, j_2, \ldots, j_K$ of K points with the closest feature descriptors $\mathbf{f}_{j_1}, \mathbf{f}_{j_2}, \ldots, \mathbf{f}_{j_K}$ (by Euclidean distance) to the descriptor $\mathbf{f}_i$ of the i-th point;

- *spatially close points*: out of points from feature neighborhood $\mathcal{N}^{\mathbb{F}}(\mathbf{f}_i)$, the set $p_{j_1}, p_{j_2}, \ldots, p_{j_K}$ of points that are spatially close to the query point p_i. The spatial proximity is defined by the proximity measure $w(p_i, p_j)$, as discussed later (see Eq. 1);

- *average feature distance*: for the points within the feature neighborhood $\mathcal{N}^{\mathbb{F}}(\mathbf{f}_i)$, mean of the Euclidean distances from the descriptor $\mathbf{f}_i$ to each of the descriptors in its feature neighborhood $\{\mathbf{f}_{j_2}, \ldots, \mathbf{f}_{j_K}\}$;

- *average spatial distance*: for the points within the feature neighborhood $\mathcal{N}^{\mathbb{F}}(\mathbf{f}_i)$, mean of the Euclidean distances from $\mathbf{p}_i$ to each of the points in the set $\{\mathbf{p}_{j_1}, \mathbf{p}_{j_2}, \ldots, \mathbf{p}_{j_K}\}$ and those of the query point.

To introduce the Kobyshev score, we consider an example of a building shown in Fig. 4. We have computed the above-mentioned neighborhoods and distances on a larger dataset, and have cropped out one building to demonstrate the distribution of point properties.

In the figure, for every point of the considered building we compute the average feature and spatial distances and plot them against each other. The distribution is star-shaped and has three characteristic areas:

Landmark points (red in the figure): these points have sufficiently similar feature descriptors in their feature neighborhood. Additionally, the points from their feature neighborhood are also spatially close (can be inferred from the low average spatial distance). We consider these points to be landmark points as they form a set of similarly looking points that don't occur in the other parts of the point cloud (otherwise, the feature neighborhood will contain points from other parts of the cloud resulting in the increase of the average feature distance).

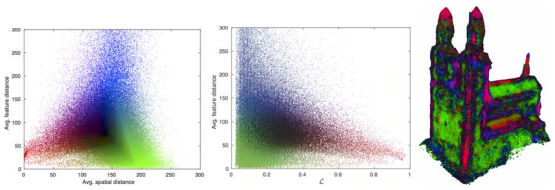

Figure 4. Distinctiveness within feature and spatial neighborhood. Left: a distribution of the average spatial distance vs. the average feature distance of the K nearest neighbors. The color maps red to landmark points (similar and local), green to ubiquitous points like walls (similar but everywhere) and blue to noise (different and wide spread). Middle: our measure $\mathcal{L}$ reweighs the distribution where the landmark points are all clearly separated on the right side. Right: a landmark, where the red towers and roof structures stand out against the green walls and blue outlier edges.

Ubiquitous points (green in the figure): for each point of this category, in the K-neighborhood of the feature descriptors, points are very close to the query point (resulting in a low average feature distance score). However, the average spatial distance is high because the points are spread all around the point cloud. This is a common case for repeating patterns, such as walls or roofs that are wide-spread all around the data.

Noise (blue in the figure): these are points who within their k-nn feature neighborhood have many points whose feature descriptors are dissimilar (which leads to larger average feature distance). Although the fact that the point has a distinctive neighborhood can make it special (as it is assumed, for example, in [38]), in case of considerably noisy datasets this is a characteristic of a single noise point.

Having identified the properties of the landmark points on the 2D chart, we aim to find a scalar measure that describes how likely is a point to be on a landmark component. To do so, we use the concept of *spatially close points* defined above. We introduce the proximity measure between any two points p_i and p_j using Gaussian weights vanishing

with distance:

$$w(p_i, p_j) = \exp\left\{-\frac{||\mathbf{p}_i - \mathbf{p}_j||^2}{\sigma^2}\right\} \quad (1)$$

where the parameter σ encodes our concept of locality. It reflects the expected size of a landmark component. This weighting gives scores close to 1 for points that are spatially close to the query point, dropping to 0 if the point is significantly far away. Such a transformation is more robust to the points that are far away from the query point: if the point is not within the expected landmark component size defined by σ, it gets a score close to 0.

We now can define the Kobyshev score of point p_i by averaging the point's proximity measures to its feature neighborhood. More formally,

$$\mathcal{L}(p_i, \mathcal{N}_i^{\mathbb{F}}) = \sum_{j \in \mathcal{N}_i^{\mathbb{F}}} \frac{w(p_i, p_j)}{|\mathcal{N}_i^{\mathbb{F}}|} \quad \in [0, 1], \quad (2)$$

where the cardinality of $\mathcal{N}_i^{\mathbb{F}}$ is $|\mathcal{N}_i^{\mathbb{F}}| = K$ (nearest neighbors), $\forall i$. Since Eq. (2) averages the spatial distance weights of points with feature vectors similar to that of point p_i, the more such points lie close to p_i the higher the score.

The value of $\mathcal{L}(p_i, \mathcal{N}_i^{\mathbb{F}})$ depends only on points from the feature neighborhood that are spatially close to the query point (other points will contribute with a result close to 0). That makes the scores of noisy or ubiquitous points equally low, while keeping the score for landmark points high. This can be seen in Fig. 4. We color-code the distribution on the left-most plot, and then change the x-axis to $\mathcal{L}(p_i, \mathcal{N}_i^{\mathbb{F}})$, as demonstrated in the middle plot. One can infer that the points can be separated just by looking at the $\mathcal{L}(p_i, \mathcal{N}_i^{\mathbb{F}})$. From now, $\mathcal{L}(p_i, \mathcal{N}_i^{\mathbb{F}})$ is a landmarkness score per point.

Our measure in Eq. (2) has the following interesting properties. First, it gives a high score for similar points (in $\mathbb{F}$) that only occur within a local neighborhood (low scatter) and a low score to those that occur at a large distance (high scatter). Second, unlike with ball search in $\mathbb{F}$, it allows us to choose a large enough neighborhood K to achieve a uniform statistical significance for the averaging.

The closest work [38] measures low-level distinctness by averaging a ratio of the distance in feature space $\mathbb{F}$ and the distance in 3D space $\mathbb{E}$ over all pairs of points in a small ball-neighborhood retrieved in $\mathbb{F}$. Our distinctiveness measure differs in many ways, since a) we count the similar points leading to more robustness w.r.t. minor differences in similar descriptors and b) the introduction of the notion of scale σ defines a local spatial neighborhood for similar points. Further, we additionally introduce c) a specific notion of spatial context, d) our final saliency is result of a discriminative learning procedure, e) works on city-scale datasets. As it will be shown in Sec. 4, our approach improves significantly over the measures of [38] in the task of landmark building identification.

3.2. Unsupervised Discriminative Refinement

In this section we show how to refine the initial distinctiveness measure by unsupervised discriminative learning, which does not require manual annotations.

First, we identify landmark components, i.e. parts of a landmark building which have a high distinctiveness score. Second, these landmark components are then discriminatively learned to refine their distinctiveness. Both stages are completely unsupervised. No ground truth training data is needed – making our method generic for other types of data.

3.2.1 Landmark Component Identification

In this section, we aim to identify coherent, local components of groups of points covering a distinctive architectural component, such as a special tower or roof, i.e. a distinctive part of a landmark building. These groups of points are used as training examples to learn how the landmark components look like.

First, we evaluate the distinctiveness of every point in the dataset via Eq. (2). Next, we propose an optimization to identify points that belong to landmark components. We formulate it as a binary segmentation which assigns binary labels $x_i \in \{0, 1\}$ to points $p_i \in \mathcal{P}_i$, where $x_i = 1$ indicates a landmark point. We denote by $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ the K-NN graph over spatial locations $\mathbf{p}_i$ of the point set, where $\mathcal{V}$ is the set of vertices and $\mathcal{E}$ the set of edges in $\mathcal{G}$. Our segmentation is driven by the energy

$$E(\mathbf{x}, \mathcal{P}) = \sum_{i=1}^{N} \Theta(p_i, x_i) + \beta \sum_{(x_i, x_j)\mathcal{E}} \Psi(p_i, p_j, x_i, x_j), \quad (3)$$

where $\mathbf{x} = (x_1, \ldots, x_N)$ is a complete labeling over the point cloud $\mathcal{P}$, β is a balance between the unary $\Theta(p_i, x_i)$ and the pairwise term $\Psi(p_i, p_j, x_i, x_j)$, the latter being defined for any pair of points $(p_i, p_j) \in \mathcal{E}$.

The unary cost $\Theta(x_i)$ encodes the likelihood of point p_i to be a landmark point, irrespective of the labels x_j in its neighborhood in 3D space $\mathbb{E}$. The initial unary is composed as

$$\Theta(p_i, x_i) = \begin{cases} \Gamma(p_i), & x_i = 1 \\ 1 - \Gamma(p_i), & x_i = 0 \end{cases}, \quad (4)$$

where

$$\Gamma(p_i) = 1/(1 + \exp\{-\gamma(\mathcal{L}(p_i) + t)\}), \quad (5)$$

where t is a soft threshold for our measure $\mathcal{L}$ in Eq. (2).

The pairwise cost is the weighted Potts-penalty

$$\Psi(p_i, p_j, x_i, x_j) = \begin{cases} 0, & x_i = x_j \\ e^{-||\mathbf{p}_i - \mathbf{p}_j||^2/(2\sigma_s^2)}, & x_i \neq x_j \end{cases}, \quad (6)$$

which enforces spatial smoothness of the labeling solution $\mathbf{x}$. The penalty vanishes with distance between the two points, and the parameter σ_s controls its rate.

We can solve for the global optimum efficiently via graph cuts [6]. As a result, we obtain spatially coherent groups of points marked as landmark points.

Next, we consider the subgraph $\mathcal{G}_l \subset \mathcal{G}$ that only contains nodes p_i labeled as landmark and edges between these, and perform a connected component search to identify landmark components, denoted by $\mathcal{C}_k$ ($k = 1, 2 \dots C$). These components are used for refining our feature neighborhood, hence, our distinctiveness measure for landmarks.

3.2.2 Refinement of Feature Neighborhood

Since the original distance measure is distinctive yet not necessarily discriminative, we learn a component-specific distance measure by leveraging the segmented components as training data for a discriminative classifier. Our goal is to reinforce and extend the detected landmark components by updating the neighborhoods $\mathcal{N}^{\mathbb{F}}(\mathbf{f}_i)$ in feature space $\mathbb{F}$ and further boosting the distinctiveness parts. For this purpose, we make use of discriminative learning to optimize the weighting in function of distance between descriptor pairs. We employ a random forest classifier [2, 7] to learn how each landmark component $\mathcal{C}_k$ looks like.

For each component $\mathcal{C}_k$ ($k = 1, 2 \dots C$), we consider the feature descriptors of all points assigned to $\mathcal{C}_k$ as positive samples, and the descriptors of all other points in the dataset as negative examples to train our classifiers.

Then, we run a binary classification which results in a $C \times N$ matrix $\mathbf{P}$, where C is the number of landmark components, and N the number of points. Matrix $\mathbf{P}$ contains the prediction for any of the points p_i belonging to any component $\mathcal{C}_k$. Note that one could train a multi-classification classifier, however, this would increase the memory footprint significantly.

After these predictions are obtained, we update the initial sets of points having similar descriptors $\mathcal{N}^{\mathbb{F}}(\mathbf{f}_i)$ originally obtained by K-NN search in feature space $\mathbb{F}$. For every point, we consider the classifier that has given it the highest prediction score, and take the indices of its best K predictions as the indices of the nearest neighbors. This way we form the new set of most similar points $\mathcal{N}_i^{*\mathbb{F}}$ for each point p_i. These new sets yield a new distinctiveness measure based on Eq. (2). Namely, our updated measure is

$$\mathcal{L}^*(p_i, \mathcal{N}_i^{*\mathbb{F}}) = \sum_{j \in \mathcal{N}_i^{*\mathbb{F}}} \frac{w(p_i, p_j)}{|\mathcal{N}_i^{*\mathbb{F}}|} \quad \in [0, 1], \qquad (7)$$

Finally, all points in the dataset are evaluated against this updated distinctiveness measure.

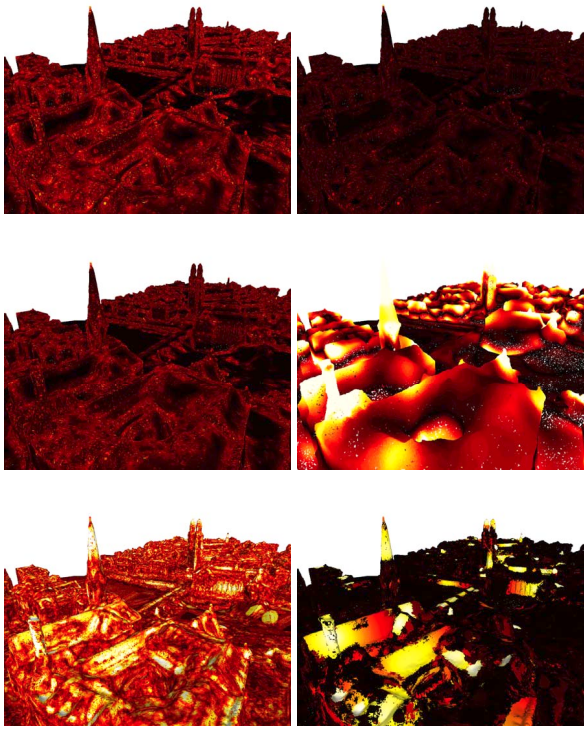

Figure 5. Point-wise scores (left to right, top to bottom): $\mathcal{D}_{\text{low}}$, $\mathcal{D}_{\text{high}}$, $\mathcal{D}_{\text{aggr}}$ from [38], curvature, $\mathcal{L}$, refined $\mathcal{L}^*$. Each method delivers a different result: [38] detects local edges, curvature redundantly finds shape changes, whereas our method identifies only unique landmark components.

4. Experiments

We demonstrate the effectiveness of our method in finding landmark components and entire buildings which are groups of points consistently identified as interesting, distinctive, and discriminative. We show results for comparison with baselines and over novel building-wise results which enabled example applications like tourist navigation, landmark comparison and level-of-detail rendering.

As the method is designed for large-scale point clouds, we run it on city-scale point cloud datasets, however, it is generic to work on any 3D point cloud. To the best of our knowledge, we are the first to provide such city-wide results on 3D point clouds.

The types of datasets for these experiments are (1) an image-based multi-view stereo (MVS) reconstruction on aerial images which effectively results in a 3D point cloud (Zurich-MVS, 1.2 km^2, 81M points), (2) two image-based aerial 3D reconstructions which effectively are a 2.5 DSM (Digital Surface Model) image converted into a 3D point cloud (Toronto-DSM and Vaihingen-DSM, the latter has uneven point density), and (3) an airborne LIDAR scan, which also is a 3D point cloud (Amsterdam-LIDAR). Please see supplemental material for detailed images. For the pur-

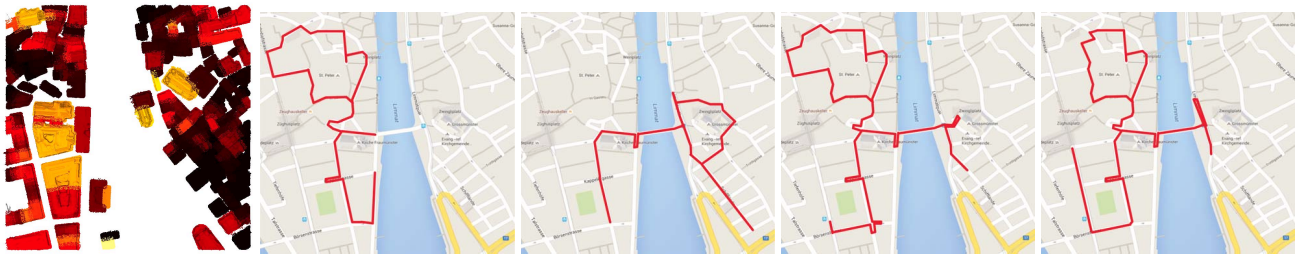

Figure 6. Landmark-visiting tour on the Zurich-MVS dataset (left to right): relative building-wise scores by our method; route planned for scores aggregated from scores of [38], curvature scores, our method's scores; a professional city tour plan.

Figure 7. Interactive 3D tourist maps (left to right): buildings thresholded by 90-, 75-, 50-percentile of landmarkness value on a 3D map - giving a view of the most important landmarks.

pose of context processing, the 3D models are split into 50x50m tiles.

We discuss the major parameters of our method (and provide additional experiments on the parameters, scale and context in supplemental material), and show comparisons to closest related work and for landmark building identification, as well as the novel 2D building chart.

4.1. Implementation details and runtime

Each point's local geometry is described as a histogram of normals by Fast Point Feature Histogram (FPFH) [36]. We have experimented with various support radii for the descriptor and the radius of 3 meters performed best.

We have evaluated the runtime of the method on an 8-core machine with a 3.50 GHz CPU. Calculating the FPFH descriptors takes around 17 minutes for an area of 400m^2 (64 tiles of 50m^2) with 8.3M points. The current bottleneck is the search for nearest neighbors on FPFH. For 50 nearest neighbors it takes about 45 minutes. Training a classifier on a single core for one component takes around 1 minute, but the process can be parallelized on multiple machines. The rest of the pipeline takes 7 seconds to compute. That results in a processing time of less than one second per cubic meter.

Landmark components are segmented to identify salient building parts and to provide training data for the refinement learning. The initial segmentation is based on using Eq. (3), which uses the initial landmarkness score $\mathcal{L}$. In a study, we found that t=0.3 and γ=0.1 provide the best overall performance (see supplemental materials for more details). Once the landmark components are segmented, we can refine $\mathcal{L}$ to obtain $\mathcal{L}^*$ by training a discriminative classifier. The values of $\mathcal{L}^*$ are shown in Fig. 5 (bottom right) and 8 (right pictures in image pairs). It can be seen that our discriminative $\mathcal{L}^*$ has much lower values on non-landmark points, while strengthening the high values on the points that belong to landmark components.

4.2. Point-wise landmark identification

Since our point-wise scoring resembles the notion of saliency, we compare to the closest measures [38] which provide a way to score individual points. We further compare to standard curvature estimation, which is not saliency but a local notion of local change of shape. However, please note that our landmark score is aimed at quite a different purpose than point-wise saliency, i. e. finding unique and interesting landmark component and buildings.

We show the results of our method and the baselines in Fig. 5. The $\mathcal{D}_{\text{low}}$ score from [38] indicates the points that are unique in a very local spatial neighborhood, resulting in every edge being highlighted. $\mathcal{D}_{\text{high}}$ from [38] considers how different the point is from the spatially far points. That results in the highest scores given to outliers and lower scores spread around the entire point cloud.

In contrast, our $\mathcal{L}$ scoring gives the highest values to the parts of the point clouds that are globally distinctive components, such as the tower top or unusual roofs. The results of $\mathcal{L}$ on a whole city tile are shown in Fig. 8 (top left).

Our measure best identifies landmark components and buildings. The related methods either focus on very small elements like corners and cannot identify landmark components, or (e.g the curvature) can identify landmark components yet has no notion of distinctiveness and redundantly produces the same components across the entire city. See suppl. material in Fig. 8, 9, and 10 for further comparisons.

4.3. Building-wise landmark identification

Here the goal is to identify how likely a buildings is to be a landmark building, which is interesting due to its unusual geometry and rare appearance elsewhere in the city-scale 3D point cloud.

Each city is unique in its landmarkness patterns and also what type of buildings are considered distinctive. In Fig. 8 we show an overview of the results for the 4 different datasets. For instance, in Zurich-MVS mainly different shapes of churches, in Toronto-DMS the remarkable town

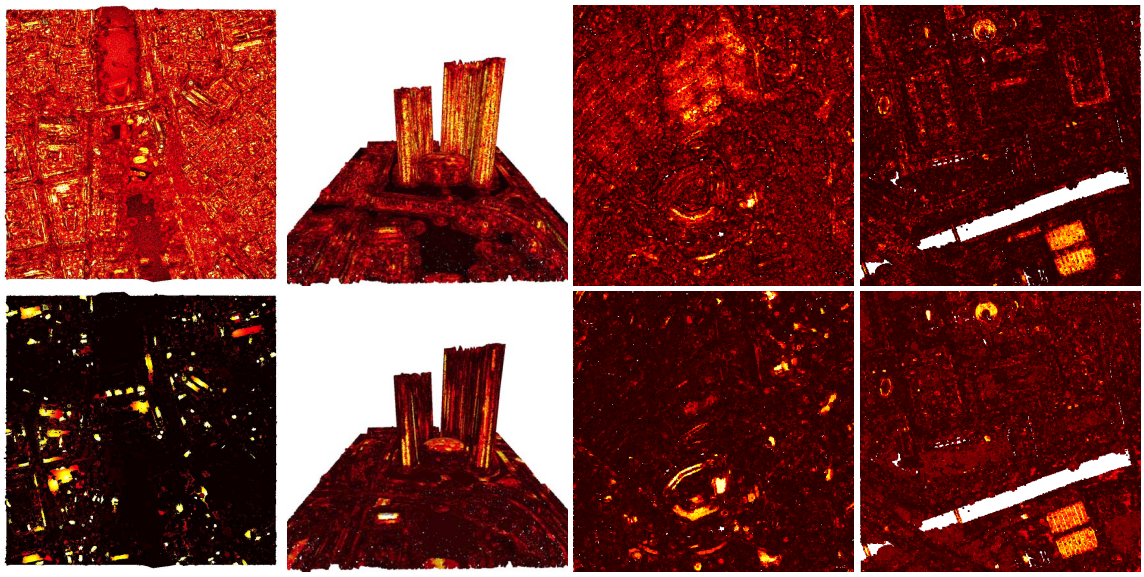

Figure 8. This shows the initial $\mathcal{L}$ (top) and refined $\mathcal{L}^*$ (bottom) for the four datasets (Z-MVS, T-DMS, V-DMS, A-LIDAR). The context is calculated over 400m giving a good overview of strong landmark buildings.

hall and buildings with special roof structures, in Vaihingen-DMS the vineyard structures (that are cleaned up after iterative updates) and the house on the hill, and in Amsterdam-LIDAR the train station and the Oude church are identified.

Given a point to building assignment, one can aggregate the individual point or component scores into building scores. For the aggregation we took the 95-th percentile of landmarkness scores in every building as a robust score.

As an example for navigation we show an automatic tourist path and a 3D rendering of the most interesting buildings for the Zurich-MVS dataset. In Fig. 6 (left) the aggregated scores for buildings in Zurich-MVS using OpenStreetMaps castrate outlines. Further, we generate tourist tours along the most interesting buildings, as shown in Fig. 6 (middle). We formulate the tour optimization problem as gaining as much landmarkness score during the walk that is limited by its distance (1.8km). If the building is already visited, its landmarkness is set to zero. We tackle it as a branch-and-bound search problem and can generate the route in less than 30 seconds. Fig. 6 also contains the tourist tour based on various baselines, e.g. D_{aggr} [38], curvature and our tourist tour. Fig. 6 (right) shows an example of a professional tourist city tour. The two baselines identify mostly local structures and hence collect more redundant local areas, which leads to shorter tourist tours in the same already. Our method due to its global context search is able to identify special landmarks more effectively. Our tourist path is the most similar to the professional city tour.

Finally, in Fig. 7 we also render a 3D map with buildings with the highest landmarkness score at various thresholds.

We further show how the landmark scoring for identifying buildings can be used to create a novel 2D chart for buildings, tourist navigation around the most interesting landmarks, and a benefit for point cloud registration. In Fig. 1 we show our novel 2D chart for buildings which separates the landmark buildings from common buildings. To aggregate the buildings into the chart, we compute pairwise building-to-building distance by comparing the histograms of $\mathcal{L}^*$ values per building, binned into 10 intervals. Then, we use multidimensional scaling to represent the buildings in two dimensions. It shows a good distinction between standard and landmark buildings.

5. Conclusions

In this work we proposed a new method for finding of landmark buildings in large, city-scale 3D point sets. Our method identifies components of landmarks (e.g. towers) that locally stand out and help to identify landmark buildings. To that end we introduced a novel saliency distance that outperforms measures with similar goals from related work significantly. This is confirmed by our results in terms of qualitative point clouds, building-wise aggregation, 2D building comparison charts as well as applied to tourist city tours and interactive 3D city maps based on context of landmarkness. This is the first work to automatically generate a selection of landmark buildings in a city-scale. In future work we plan to incorporate more visual and semantic cues from street-side images [35, 32], reason about other types and decomposition of landmarks [26, 4, 31], and compare further manual tourist guides and guidance systems [30, 33].

Acknowledgments. This work was supported by the European Research Council project VarCity (273940). We thank Michal Havlena, Wilfried Hartmann and Konrad Schindler for the Zurich aerial dataset.

References

[1] O. Akman and P. Jonker. Computing saliency map from spatial information in point cloud data. In *ACIVS*, 2010.

[2] Y. Amit, G. August, and D. Geman. Shape quantization and recognition with randomized trees. *Neural Comp.*, 1996.

[3] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *ICML*, 2009.

[4] A. Bódis-Szomorú, H. Riemenschneider, and L. Van Gool. Efficient edge-aware surface mesh reconstruction for urban scenes. *CVIU*, 2015.

[5] A. Borji, M. Cheng, H. Jiang, and J. Lis. Salient Object Detection: A Survey. *TIP*, 2014.

[6] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *PAMI*, 23(11):1222–1239, 2001.

[7] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[8] M. Breunig, H. Kriegel, R. Ng, and J. Sander. LOF : Identifying Density-Based Local Outliers. *ACM Sigmod*, 2000.

[9] D. Dai, M. Prasad, C. Leistner, and L. Van Gool. Ensemble partitioning for unsupervised image categorization. In *ECCV*, 2012.

[10] D. Dai and L. Van Gool. Ensemble projection for semi-supervised image classification. In *ICCV*, 2013.

[11] R. Datta and J. Z. Wang. Studying Aesthetics in Photographic Images. In *ECCV*, 2006.

[12] M. de Brecht and J. Saiki. A neural network implementation of a saliency map model. In *Neural Networks*, 2006.

[13] K. Desingh, M. Krishna, D. Rajan, and C. Jawahar. Depth really matters: Improving visual salient region detection with depth. In *BMVC*, 2013.

[14] C. Doersch, A. Gupta, and A. A. Efros. Mid-level visual element discovery as discriminative mode seeking. In *NIPS*, 2013.

[15] C. Doersch, S. Singh, A. Gupta, J. Sivic, and A. A. Efros. What makes paris look like paris? *SIGGRAPH*, 31(4):101:1–101:9, 2012.

[16] K. Duncan and S. Sarkar. Saliency in Images and Video: A Brief Survey. *IET Computer Vision*, 2012.

[17] S. Escalera, P. Radeva, and O. Pujol. Complex salient regions for computer vision problems. In *CVPR*, 2007.

[18] B. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[19] D. Gao and N. Vasconcelos. Bottom-up saliency is a discriminant process. In *ICCV*, 2007.

[20] S. Goferman, L. Zelnik-Manor, and A. Tal. Context-aware saliency detection. In *CVPR*, 2010.

[21] R. Gomes, A. Krause, and P. Perona. Discriminative clustering by regularized information maximization. In *NIPS*, 2010.

[22] F. Grabler, M. Agrawala, R. W. Sumner, and M. Pauly. Automatic generation of tourist maps. *SIGGRAPH*, 27(3):100:1–100:11, 2008.

[23] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *CVPR*, 2007.

[24] T. Judd, F. Durand, and A. Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012.

[25] G. Kim, D. Huber, and M. Hebert. Segmentation of salient regions in outdoor scenes using imagery and 3d data. In *WACV*, 2008.

[26] N. Kobyshev, A. Bódis-Szomorú, H. Riemenschneider, and L. Van Gool. Architectural Structural Element Decomposition. *CVIU*, 2016.

[27] S. S. Kruthiventi, K. Ayush, and R. V. Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *CoRR*, abs/1510.02927, 2015.

[28] Y. Lee and K. Grauman. Learning the easy things first: Self-paced visual category discovery. In *CVPR*, 2011.

[29] X. Li and I. Guskov. Multi-scale features for approximate alignment of point-based surfaces. In *Symposium on Geometry Processing*, 2005.

[30] A. Locher, M. Perdoch, H. Riemenschneider, and L. Van Gool. Mobile Phone and Cloud – a Dream Team for 3D Reconstruction. In *WACV*, 2016.

[31] A. Mansfield, N. Kobyshev, H. Riemenschneider, W. Chang, and L. Van Gool. Frankenhorse: Automatic Completion of Articulating Objects from Image-based Reconstruction. In *BMVC*, 2014.

[32] A. Martinović, J. Knopp, H. Riemenschneider, and L. Van Gool. 3D All The Way: Semantic Segmentation of Urban Scenes from Start to End in 3D. In *CVPR*, 2015.

[33] M. Mauro, H. Riemenschneider, A. Signoroni, R. Leonardi, and L. Van Gool. A unified framework for content-aware view selection and planning through view importance. In *BMVC*, 2014.

[34] K.-F. Richter and S. Winter. Landmarks, 2014.

[35] H. Riemenschneider, A. Bodis-Szomoru, J. Weissenberg, and L. Van Gool. Learning Where To Classify In Multi-View Semantic Segmentation. In *ECCV*, 2014.

[36] R. Rusu, N. Blodow, and M. Beetz. Fast Point Feature Histograms (FPFH) for 3D Registration. In *ICRA*, 2009.

[37] S. Salti, A. Petrelli, F. Tombari, and L. Stefano. Complex salient regions for computer vision problems. In *3DPVT*, 2012.

[38] E. Shtrom, G. Leifman, and A. Tal. Saliency detection in large point sets. In *ICCV*, 2013.

[39] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *ECCV*, 2012.

[40] M. Sorrows and S. Hirtle. The nature of landmarks for real and electronic spaces. *COSIT*, 1999.

[41] J. Weissenberg, M. Gygli, H. Riemenschneider, and L. Van Gool. Navigation using special buildings as signposts. In *MapInteract*, 2014.

[42] J. Ye, Z. Zhao, and M. Wu. Discriminative k-means for clustering. In *NIPS*, 2007.

[43] H. Zhong, J. Shi, and M. Visontai. Detecting unusual activity in video. In *CVPR*, 2004.