Ultra-wide Baseline Facade Matching for Geo-localization

Mayank Bansal[1,2], Kostas Daniilidis[1], and Harpreet Sawhney[2]

[1] GRASP Lab., University of Pennsylvania, Philadelphia PA, USA
{mayankb,kostas}@cis.upenn.edu
[2] Vision Technologies Lab., SRI International, Princeton NJ, USA
harpreet.sawhney@sri.com

Abstract. Matching street-level images to a database of airborne images is hard because of extreme viewpoint and illumination differences. Color/gradient distributions or local descriptors fail to match forcing us to rely on the structure of self-similarity of patterns on facades. We propose to capture this structure with a novel "scale-selective self-similarity" (S^4) descriptor which is computed at each point on the facade at its inherent scale. To achieve this, we introduce a new method for scale selection which enables the extraction and segmentation of facades as well. Matching is done with a Bayesian classification of the street-view query S^4 descriptors given all labeled descriptors in the bird's-eye-view database. We show experimental results on retrieval accuracy on a challenging set of publicly available imagery and compare with standard SIFT-based techniques.

1 Introduction

In this paper, we propose a novel method for matching facade imagery from very different viewpoints – like from a low flying aircraft and from a street-level camera. The scenario we address entails a database of pre-processed bird's-eye-view (BEV) images and street-view (SV) queries. Such images are characterized by unmitigated differences in local appearance which render any comparison of bags of visual words infeasible. A visual comparison of this imagery ever after rectification testifies to the hardness of the problem. Moreover, a vast majority of facades contain repetitive patterns which make correspondence estimation highly ambiguous. We rather have to rely on comparing the structures of the facade patterns and still account for any transformations between such structures.

The key idea in this paper is to avoid direct matching of features to solve this extreme case of wide-baseline matching. Thus, we formulate the problem as "embeddings" within each respective dataset (SV and BEV) so that large variations are incorporated within the structure of embeddings. This idea has not been explored before especially in the context of air-ground matching. We make the following contributions to the state of the art: (a) we introduce an approach for matching image regions with significant appearance, scale, and viewpoint variations based on a novel *Scale-Selective Self-Similarity (S^4)* feature

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part I, LNCS 7583, pp. 175–186, 2012.

that combines intrinsic scale selection with self-similarity descriptors, and (b) we demonstrate a novel system for matching street-level queries to a database of birds-eye views. We show experimental results on the retrieval accuracy from our technique and compare our performance with standard SIFT-descriptors.

We approach the facade detection and matching problem from a combined statistical and structural viewpoint. While other approaches model the lattice structure explicitly [1], we capture the statistical self-similarity (or dis-similarity) of a local patch to its neighbors. By avoiding using a specific feature like SIFT, MSER, or line segments, we can capture this structure at any point – in implementation we do it on a randomly jittered grid. In addition, the self-similarity descriptor also captures the dis-similarity between neighboring elements ignored in lattice approaches but still observed e.g. in [2]. The challenge with self-similarity is to capture the intrinsic local scale governed by the periodicity/generator group of a lattice. We estimate the scale by discovering the closest most salient repetition of a patch which can be centered anywhere. With the exception of [3], other approaches rely on the robustness of interest point or line segment detectors. Having obtained the intrinsic scale enables us to compute the scale-invariant S^4 descriptor and also allows us to **detect** facades as clusters of such points in space that have similar scale and descriptors. Similar descriptors are obtained from the query street-level image as well. At this point, instead of lattice or graph matching [3,2], we apply a labeling approach that labels each query descriptor with the most probable facade label (cluster) in a naive-Bayes sense. This way, we match local lattice structures rather than global ones and the most likely closest database facade is obtained.

2 Related Work

In the discussion of related work, we emphasize two main aspects: **detection** of facades/lattices and **matching**. Chung et al. [2] extract MSER regions in multiple scales which are then clustered w.r.t similarity. Local histograms of gradient similarity, area ratio, and configuration entropy are used to build adjacency matrices which are matched by using a spectral approach comparing only the graph structure. The commonality with our approach is that we never use any direct comparison of appearance across images. On the other hand, their query and model graph structures have to match globally while our approach uses the statistics of the edges of these graphs represented by the self-similarity descriptor and hence exploits the redundancy in features better. Moreover, the self-similarity descriptor is more general and implicit than the concatenation of several neighborhood descriptions (HoG, area ratio, entropy). Park et al. [1] model the lattice discovery as a multi-target tracking problem using Mean-Shift Belief Propagation. Candidates for lattice vertices are interest points that are obtained through clustering. Hays et al. [3] randomly select regions and search for their repetition in two directions in their immediate neighborhood. Lattice discovery is formulated as a graph matching problem with higher-order constraints that model the lattice structure of the region repetitions. The advantage of [1,3]

is that they can deal with deformed lattices in the detection step while almost all other approaches including ours remove projective and sometimes affine distortions using vanishing points and ratio constraints. Schindler et al. [4] detect lattices by mapping quadruples of SIFT features to the projective basis and checking the consistency of the rest of the points with respect to this basis. They combine multiple 2D-to-3D pattern correspondences and recover the camera orientation and location as an intersection of the family of solutions obtained using each correspondence.

Recently, Bansal et al. [5] established the feasibility of matching highly disparate street view images to aerial image databases to precisely geo-localize SV images without the need for GPS or camera metadata. Doubek at al. [6] match the similarity of repetitive patterns by comparing the grayscale tiles, the peaks in color histogram, and the sizes of the two lattices. In [7], corners are extracted and grouped according to consistency with the geometric transformations corresponding to the generators of the lattice. Kosecka et al. [8] extract rectangle projections by grouping line segments according to vanishing point consistency. Using [9] they match a query street-view image to a database of geo-tagged street-view images using wide-baseline matching. In [10] and [11], a query street-view image is again matched to a database of street-view images and then used to compute the camera pose. They assume the query image camera internal parameters to be known and use a pyramid to match at multiple scales using geometric consistency. In [12], a viewpoint normalization of planar patches is followed by SIFT computation of the rectified patch. We close our discussion with [13] where omnidirectional views are matched to building outline maps by detecting the tallest vertical corners of the buildings which are matched through 2D to 1D projection.

3 Scale-Selective Self-similarity Features

The viewpoint and appearance difference between oblique Bird's-Eye-View (BEV) and street-view (SV) imagery is too large to be captured by direct matching of descriptors like SIFT and MSER. Therefore, we propose to create a descriptor that captures the structure of repetition of patterns or more generally the relative similarity between local patches within facades. Instead of modeling the structure with a graph or lattice and relying on the robustness of the detection of their nodes, we define a new feature which we call the *Scale-Selective Self-Similarity* or S^4 feature. This feature improves upon the well-known self-similarity descriptor from Shechtman et. al [14] by adding a SIFT-like scale-normalization to allow characterization of the self-similar structure in a scale-invariant manner.

Using the same notation as [14], for a given pixel q, the local self-similarity descriptor d_q is computed as follows. A local image patch of width w_{ss} (e.g., 5 pixels) centered at q is correlated with a larger surrounding image region of radius r_{ss} (e.g., 40 pixels), resulting in a local internal 'correlation surface'. The correlation surface is then transformed into a binned log-polar representation

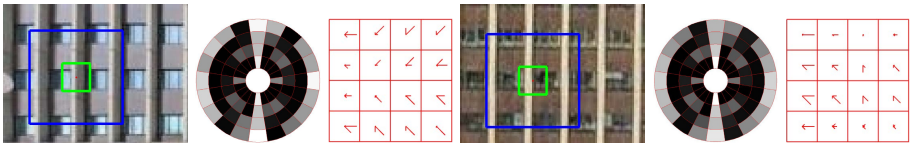


Fig. 1. Example self-similarity and SIFT descriptors for corresponding facades from SV and BEV images respectively

which accounts for increasing positional uncertainty with distance from the pixel q, accounting, thus, for local spatial affine deformations.

Fig. 1 shows a pair of (ortho-rectified) SV and BEV images of a facade that have been manually normalized to the same image scale, and compares how well their self-similarity descriptors match relative to their SIFT descriptors. The self-similarity descriptor at the center of the green ROI (local patch) is computed by correlating within the surrounding support region (blue ROI). The computed descriptors are noticeably quite similar even with the large appearance difference between the images themselves. In comparison, the SIFT descriptors computed using the same support region are dissimilar.

Scale-Selection. While it is clear that the inherent self-similar structure in building facades can serve as a good matching criterion, it is not clear how that structure can be matched if the building is seen at different scales. The basic self-similarity descriptor discussed above assumes a distance binning which is not scale invariant. To account for feature scale differences, Shechtman et al. [14] suggest computing the self-similarity descriptors on a Gaussian image pyramid representation and then searching for the template object across all scales. For the purposes of retrieval, however, such an approach would not work. In particular, for building facades, capturing the self-similar structure at all scales will reduce the discriminability evident at the fundamental scale of the facade. Instead, we would like a SIFT like normalization so that the descriptors between differently scaled buildings can still be matched. The repetitive structure of building facades provides one such normalization scale. However, building facades typically also exhibit *local* periodicity. While recovering this scale will serve the purpose of a valid normalizing scale, it may compromise on the overall discriminability of the computed descriptor by (a) being too local, and (b) by being too dependent on the inherent image scale (the smallest scale structure will be lost first in a noisy query image).

In this paper, we focus on recovering the *motif scale*. We define the motif scale at a pixel in the facade as the smallest wavelength at which any patch in this pixel's local neighborhood repeats. Defined this way, a local window scale would be ignored if it is not consistent with a few other window pixels in its neighborhood – thus making this scale robust against local pattern noise. This motif scale can be measured independently in both horizontal and vertical directions; in our implementation, we have only used the horizontal scale (denoted as λ_x), but the approach is symmetric with respect to using either of the two. Given the motif-scale λ_x value at any pixel, the S^4 descriptor is defined as the

self-similarity descriptor computed by setting the patch size w_{ss} to the estimated motif scale λ_x and the correlation radius to $r_{ss} = 2\lambda_x$.

Our approach for motif scale-selection is based on the peaks in the autocorrelation surface in a local neighborhood surrounding a pixel. Consider a pixel (x, y) inside an image $\mathcal{I}$ exhibiting periodic structure and let λ_x be its scale along the x-direction. Now consider a small $w \times h$ patch of pixels around this pixel and correlate it with patches extracted at various offsets (r, θ) in a polar representation. To capture the correlations most relevant to the self-similarity descriptor, we measure the correlation profile using the following SSD measure. Let $\mathcal{J}(s, t) = \mathcal{I}(x + s, y + t)$, then:

$$q(r,\theta) = \sum_{\substack{t_y = \\ -\frac{h}{2}}}^{\frac{h}{2}} \sum_{\substack{t_x = \\ -\frac{w}{2}}}^{\frac{w}{2}} \left(\mathcal{J}(t_x, t_y) - \mathcal{J}(t_x + r\cos(\theta), t_y + r\sin(\theta))\right)^2 \tag{1}$$

Then, the correlation profile $p_{(x,y)}(r)$ is computed by integrating the scores $q(r, \theta)$ in a 20^o lobe ($\theta_0 = 10^o$) around the horizontal direction:

$$p_{(x,y)}(r) = \exp\left(-\frac{1}{2\theta_0 + 1} \sum_{\theta=-\theta_0}^{\theta_0} q(r,\theta)\right) \tag{2}$$

where the subscript (x, y) makes explicit the fact that the profile was obtained by correlating the patch around pixel (x, y). The angular integration provides robustness against image distortions and ortho-rectification errors. The value of r is varied such that $r \in \{1, \ldots, S_{max}\}$, where S_{max} is a pre-defined maximum scale value we expect the structure in the input image to exhibit. The correlation profile thus obtained captures the periodicity of the structure by producing the highest correlation for $r \in \{\lambda_x, 2\lambda_x, \ldots\}$. However, depending on the starting location (x, y), the correlation profile can exhibit peaks at r values which are non-integral multiples of λ_x. This will be the case if the patch contains a sub-motif of the facade which is locally periodic at a higher frequency. The illustration in Fig. 2 depicts this happening for the green and blue profiles obtained from the (black) 1-D signal. The wavelength of both these curves is smaller than the motif scale λ_x by our definition above. To alleviate this issue, we compute multiple correlation profiles by varying the starting offset in an interval $\mathcal{O} = \{(x, y), (x + 1, y), (x + 2, y), \ldots, (x + m, y)\}$. The maximum offset $(x + m, y)$ is set so that the patch around it covers the structure at the maximum scale S_{max} from the starting position i.e. $m + w/2 \geq S_{max}$. The correlation profiles are combined into a single profile $p_{avg}(r)$ by integrating across the offsets, i.e. $p_{avg}(r) = \sum_{o \in \mathcal{O}} p_o(r)$. This removes the higher-frequency peaks in the individual profiles, leaving only the peaks corresponding to the actual wavelength λ_x as depicted in Fig. 2. Furthermore, the scale estimation becomes independent of the choice of the patch dimensions w and h.

To be robust against shallow peak responses, we measure a peakness measure around each peak in the profile $p_{avg}(r)$ and prune peaks which are shallower than a threshold t_{peak}. This threshold is set empirically by running the scale-estimator on textureless and non-repetetive structures. From the locations of

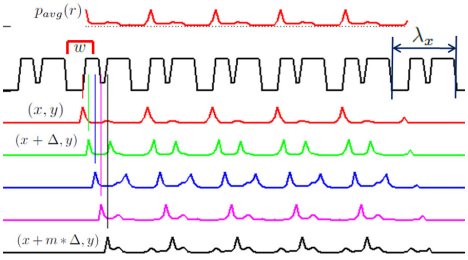


Fig. 2. Scale selection. To determine the scale λ_x of the (black) 1D signal in the second row, if we autocorrelate a patch of width w, we get one of the profiles shown in rows 3-7 depending on the starting offset. However, for a poor offset choice (green and blue curves), one can get comparable peaks in the correlation profile for scale values $< \lambda_x$ making it difficult to extract the correct scale. Integrating across these profiles, however, resolves this issue and results in a well defined profile $p_{avg}(r)$ shown in the first row. The high peaks now correspond to the correct wavelength λ_x.

the remaining peaks, the scale value λ_x can be readily obtained by a discrete Fourier transform. In the absence of any peaks the underlying structure is labeled aperiodic (assigned scale *zero*) – this removes most of the non-facade pixels and serves as an effective building detection mechanism.

4 Facade Extraction and Segmentation

We now describe our general approach for extracting building facade regions which is applicable to both BEV and SV images. The key idea is to exploit the self-similar structure of building facades: ortho-rectify the image, compute motif scales at sampled locations in the given image, compute S^4 descriptors at the computed scales and then cluster the descriptors to group similar structures together.

Motif Scale Computation. In the rectified image, we sample a grid of pixel locations every $\sigma_f = 5$ pixels apart and add uniformly random spatial jitter of amplitude $\sigma_f/2$ at each sample location. This jitter allows us to capture a good sampling of the feature distributions expected from this facade structure at the matching stage. At each sample location, we compute the motif scales using the approach discussed in section 3. An example result at this stage is shown in the left half of Fig. 3 . Note that the scale selection has removed the non-building areas almost completely by labeling them with a *zero* scale value (shown as red dots in the figures). Also note the wide range of motif scales seen across buildings stressing the importance of proper scale selection. At this point, we need a way to segment out individual facades into disjoint groups so that a matching approach can predict labels at the building level.

Facade Segmentation. At each sample location, we compute the S^4 descriptor ($n_\theta = 20$ angular bins and $n_r = 4$ distance bins) by setting the patch size w_{ss}

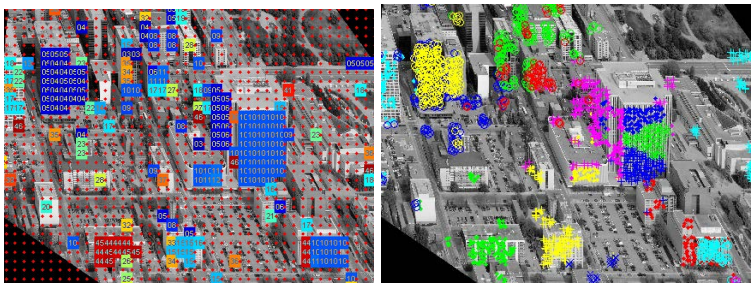


Fig. 3. Facade Extraction and Segmentation. Rectified BEV images showing, left: the selected horizontal scales with red dots at the locations assigned zero scale value and, right: cluster assignments after K-means.

to the estimated motif scale λ_x and the correlation radius to $r_{ss} = 2\lambda_x$. Now, we perform K-means clustering in this S^4 feature space using L_1 norm as our distance measure. To avoid descriptor grouping across different buildings, we penalize clustering of descriptors which were sampled from far off locations. The desired number of clusters N is set as follows. We manually mark the boundaries of a small number of buildings (5 in our case) in the BEV image and initialize $N = N_0$. Now, we iteratively run K-means with decreasing value for N as long as the following invariant is maintained: clusters on the marked buildings are contained within the marked boundaries. At the end of this process, we obtain a clustering that has the fewest number of clusters within each building and does not merge two different buildings into a single cluster (note that this is not guaranteed for unmarked buildings in general, but due to the descriptor-based grouping, we have not seen any merging of separate buildings into a single cluster in our experiments). For our test BEV set, we typically obtain 1-3 clusters per facade after this procedure. The right half of Fig. 3 shows an example of the clusters obtained after K-means clustering.

Notation. In the following, we will denote the S^4 descriptor vectors obtained from the entire set of BEV imagery by words $\mathcal{V} = \{v_1, v_2, \ldots, v_m\}$, the cluster labels as $\mathcal{C} = \{c_1, c_2, \ldots, c_N\}$ and the labeling function mapping each word to its cluster assignment by the function $\mathcal{L} : \mathcal{V} \rightarrow \mathcal{C}$.

5 Facade Matching

Given a query street-view image, we would like to retrieve facades from our BEV database that match the dominant facade(s) in the query. Sec. 6.3 and Fig. 7 illustrate the key steps in our SV-to-BEV matching pipeline. After ortho-rectification, motif scale selection and S^4 descriptor computation, we obtain a set of descriptor vectors $\mathcal{W} = \{w_1, w_2, \ldots, w_n\}$ from the query. For each of these words, we would like to estimate the probability $p(C = c_k|w_i)$ of being assigned to one of the clusters c_k in $\mathcal{C}$. The problem of finding the closest cluster label for each word w_i can be formulated in a Bayesian settings as follows. By Bayes' theorem,

Algorithm 1. BEV processing

1. Ortho-rectify BEV image using vanishing points.
2. Compute motif-scale λ_x at a jittered grid of pixel-locations on the BEV.
3. Compute S^4 descriptors v_i at locations with non-zero scales.
4. Cluster S^4 descriptors v_i using K-means to obtain label-set $\mathcal{C}$ and labeling function $\mathcal{L}$.

Algorithm 2. SV processing

1. Ortho-rectify SV image using vanishing points.
2. Compute motif-scale λ_x at a jittered grid of pixel-locations on the SV.
3. Compute S^4 descriptor-set $\mathcal{W} = \{w_j\}$ at locations with non-zero scales.
4. Compute labels $\mathcal{L}(w_j)$ using Eqn.3.
5. Best matching BEV facade: Facade containing cluster $\mathcal{L}(\mathcal{W})$ (Eqn.6).
6. Top matching facade set: For threshold t, return facades containing clusters k s.t. $f(k) > t$ (Eqn.5).

Table 1. Parameter settings

w	h	S_{max}	σ_f	w_{ss}	r_{ss}
13 px	13 px	48 px	5 px	λ_x	$2\lambda_x$
n_θ	n_r	N_0	$\sigma_\mathcal{K}$		
20	4	100	2.5		

Table 2. Facade detection performance

Scene	TP Rate	# Buildings	# FPs
BEV-1	86%	29	8
BEV-2	91%	33	3
BEV-3	86%	21	5

(a) Satellite coverage and sample BEV (b) Sample queries

Fig. 4. Pittsburgh dataset

$$p(C = c_k|w_i) = \frac{p(w_i|C = c_k)p(C = c_k)}{\sum_{j=1}^{N} p(w_i|C = c_j)p(C = c_j)} \tag{3}$$

For each word w_i, we estimate the likelihoods $p(w_i|C = c_k)$ by kernel density estimation using a Gaussian kernel $\mathcal{K}(w_i, v_j)$ with wavelength parameter $\sigma_\mathcal{K}$. The likelihood is then computed as:

$$p(w_i|C = c_k) = \frac{1}{|c_k|} \sum_{\mathcal{L}(v_j)=c_k} \mathcal{K}(w_i, v_j) \tag{4}$$

where $|c_k|$ denotes the cardinality of cluster k. The prior probability $p(C = c_k)$ is simply set from the sample proportions: $p(C = c_k) = \frac{|c_k|}{m}$. For each word w_i, we estimate the MAP estimate of the label by choosing the label k with the maximum a-posteriori probability: $\mathcal{L}(w_i) = \arg\max_k p(C = c_k|w_i)$. Given

the above word assignments, we can now compute the most probable label for the entire query facade by accumulating the word assignments from each word:

$$f(k) = \sum_i \delta(\mathcal{L}(w_i) = c_k) \tag{5}$$

$$\mathcal{L}(\mathcal{W}) = \arg\max_k \{f(k) \,|\, k = 1, \ldots, N\} \tag{6}$$

where $\delta(.)$ is the indicator function. The label $\mathcal{L}(\mathcal{W})$ identifies a cluster $c^* \in \mathcal{C}$ which, by construction of the clustering algorithm, identifies a single BEV facade.

6 Experiments and Results

Algorithm Parameters. In Table-1, we list all the parameter settings we used in our implementation. The scale estimation process was found robust against different choices of patch-size parameters w and h. S_{max} was set to a number greater than the maximum horizontal building scale for our BEV dataset (manually eyeballed). The S^4 values for n_θ and n_r were set the same as in [14].

BEV and SV Imagery Datasets. Our dataset comprises of BEV imagery (2000×1500 pixels) downloaded using Microsoft's Bing service for an area approximately 2 Km×1.2 Km in size (Fig. 4(a)) in downtown Pittsburgh, PA, USA. This dataset is challenging due to a large number (approx. 40) of buildings and very similiar facade patterns. This dataset also covers a much larger area than used in related works in air-ground-based localization e.g. $440m \times 440m$ in [13]. Street-view images downloaded using Panoramio, Flickr, Google Street-View(screenshots), and Microsoft Bing's Streetside(screenshots) were used as queries. For ground-truth purposes, only the SV imagery with geo-tags or visually identifiable facade correspondence (with the BEV) was retained.

Imagery Rectification. We rectify BEV to an orthographic view aligned with the dominant city-block direction. Similarly, the SV imagery is rectified to an orthographic view of the dominant facade in the scene using the Geometric Parsing based vanishing point estimation approach and code [15,16].

6.1 Scale Selection Results

To characterize our scale selection algorithm, we selected a test set of 10 building facades extracted from the Pittsburgh BEV dataset. We manually measured the ground-truth horizontal scale(s) for each facade and compared them to those estimated by our approach. Since we densely estimate these scale values over the facade, we computed a histogram of the estimated scale values and the normalized histogram values are shown as the blue circles (with radii proportional to the histogram values) in the bubble plot of Fig. 5. The red pluses denote the ground-truth scale values – multiple in cases where the facade exhibits more than one motif scale. The comparison shows the accuracy of our scale estimation and the presence of very few outliers.

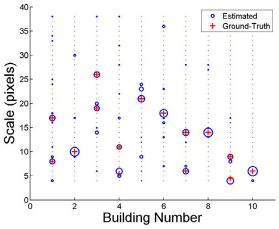


Fig. 5. Evaluation of scale estimation accuracy for 10 BEV building facades

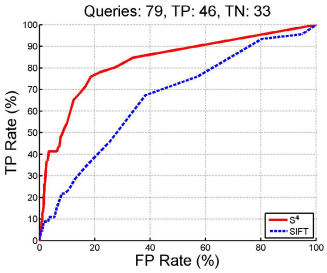
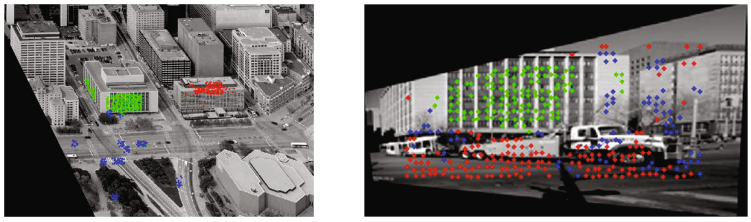


Fig. 6. ROC curve for BEV-to-SV matching on Pittsburgh dataset

(a) Query SV image, and ortho-rectified SV with extracted motif-scales



(b) Matching result with BEV with correspondingly matching clusters shown in same colors.

Fig. 7. Example Street-view (SV) processing

6.2 Facade Detection Evaluation

Table-2 shows results from our facade detection algorithm. For each BEV scene, we looked at the computed horizontal scales – points with non-zero scale values are treated as potential facades. We quantify the performance as follows: for each building facade, if at least 50% of its visible area was assigned a non-zero scale, then we count it as a true detection. If in any 4×4 sub-grid of sampled locations not on a building facade, at least 25% are assigned a non-zero scale, then we count it as a false-positive.

6.3 SV to BEV Matching

Fig. 7 illustrates our typical query SV processing pipeline. The algorithmic steps are outlined in Algorithm-2.

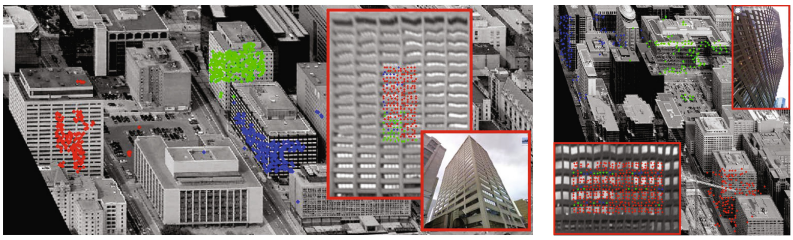


Fig. 8. Qualitative Matching Results. The main tiles show rectified BEV images. The insets show the original and rectified query street-view facades. On the rectified inset, the colored points are a subset of the words $w_1, w_2, \ldots, w_n$ with the top three most frequent recovered labels $\mathcal{L}(w_i)$ shown as red, green and blue points respectively; similarly colored points in the BEV image are words v_j which belong to these three clusters.

Fig. 6 shows the retrieval performance of our approach (along with a comparison with SIFT – details in Sec. 6.4) with a query set of 79 images including 33 true negatives i.e. buildings which were either not part of the BEV database or were significantly occluded. The query set contains challenging images with significant uncorrected image distortions, urban clutter and varied zoom range. A third of these images are high-resolution pictures from Flickr and Panoramio and the remaining are low-resolution screenshots from Google Street-View and Bing Streetside. A few samples from the query set are shown in Fig. 4(b). For generating the ROC curves, instead of using the most probable label from Eqn.6 directly, we treat the vector of frequency of each label $f(k) = \sum_i \delta(\mathcal{L}(w_i) = c_k)$ as a probability distribution. Then, to get a point on the ROC curve, we pick a value between 0.0 and 1.0 and select all the labels with probabilities higher than this value. This becomes our retrieval set which is compared with the ground-truth facade set to compute the TP and FP rates in the usual manner.

Fig. 8 shows two examples of the top three retrieval matches on representative (screen-captured) Google street-view queries. From the amount of perspective (and distortion) in the SV imagery, it is clear that features like MSER and SIFT would hardly find any correspondences.

6.4 Comparison with SIFT Features

Given the prevalence of SIFT features in wide-baseline matching literature, we present experimental comparison of its performance with our approach. To avoid any bias against SIFT due to perspective distortions (and to preclude comparison with SIFT variants like A-SIFT), we extract SIFT features on ortho-rectified BEV and ortho-rectified SV imagery. Next, we use the building clusters found using our S^4-based algorithm and perform an assignment of the SIFT features to these clusters using a nearest-neighbor association on pixel coordinates thus discarding any features on non-building background clutter. The Bayesian classification from Sec. 5 is used on the SIFT clusters to retrieve matching facades

for the query images and the quantitative results are shown in the ROC in Fig. 6 which illustrates that we achieve significant improvement in performance using S^4 features instead of SIFT features.

7 Conclusion

We have been able to match query street-level facades to airborne imagery under challenging viewpoint and illumination variation by introducing a novel approach of selecting the intrinsic facade motif scale and modeling facade structure through self-similarity.Using the motif scale, we extract and segment lattice-like facades and construct scale-invariant S^4 descriptors. We localize queries by classifying descriptors, thus matching to facades with semi-local lattice consistency.

References

1. Park, M., Brocklehurst, K., Collins, R., Liu, Y.: Deformed lattice detection in real-world images using mean-shift belief propagation. TPAMI 31, 1804–1816 (2009)
2. Chung, Y., Han, T., He, Z.: Building recognition using sketch-based representations and spectral graph matching. In: ICCV (2010)
3. Hays, J., Leordeanu, M., Efros, A.A., Liu, Y.: Discovering Texture Regularity as a Higher-Order Correspondence Problem. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 522–535. Springer, Heidelberg (2006)
4. Schindler, G., Krishnamurthy, P., Lublinerman, R., Liu, Y., Dellaert, F.: Detecting and Matching Repeated Patterns for Automatic Geo-tagging in Urban Environments. In: CVPR (2008)
5. Bansal, M., Sawhney, H.S., Cheng, H., Daniilidis, K.: Geo-localization of street views with aerial image databases. In: ACM-MM (2011)
6. Doubek, P., Matas, J., Perdoch, M., Chum, O.: Image Matching and Retrieval by Repetitive Patterns. In: ICPR (2010)
7. Schaffalitzky, F., Zisserman, A.: Geometric grouping of repeated elements within images. In: Shape, Contour and Grouping in Computer Vision (1999)
8. Kosecka, J., Zhang, W.: Extraction, matching, and pose recovery based on dominant rectangular structures. In: CVIU, vol. 100, pp. 274–293. Elsevier (2005)
9. Zhang, W., Kosecka, J.: Image Based Localization in Urban Environments. In: 3DPVT (2006)
10. Cipolla, R., Robertson, D., Tordoff, B.: Image-based localisation. In: Proceedings of 10th International Conference on Virtual Systems and Multimedia (2004)
11. Robertson, D., Cipolla, R.: An Image-Based System for Urban Navigation. In: BMVC (2004)
12. Wu, C., Clipp, B., Li, X., Frahm, J., Pollefeys, M.: 3d model matching with viewpoint-invariant patches (vip). In: CVPR (2008)
13. Cham, T., Ciptadi, A., Tan, W., Pham, M., Chia, L.: Estimating camera pose from a single urban ground-view omnidirectional image and a 2D building outline map. In: CVPR (2010)
14. Shechtman, E., Irani, M.: Matching local self-similarities across images and videos. In: CVPR (2007)
15. Barinova, O., Lempitsky, V., Tretiak, E., Kohli, P.: Geometric Image Parsing in Man-Made Environments. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part II. LNCS, vol. 6312, pp. 57–70. Springer, Heidelberg (2010)
16. Tardif, J.: Non-iterative approach for fast and accurate vanishing point detection. In: ICCV (2009)