Regularity-Driven Building Facade Matching between Aerial and Street Views

Mark Wolff Robert T. Collins Yanxi Liu
School of Electrical Engineering and Computer Science
The Pennsylvania State University
University Park, PA. 16802, USA
wolff@psu.edu, {rcollins, yanxi}@cse.psu.edu

Abstract

We present an approach for detecting and matching building facades between aerial view and street-view images. We exploit the regularity of urban scene facades as captured by their lattice structures and deduced from median-tiles' shape context, color, texture and spatial similarities. Our experimental results demonstrate effective matching of oblique and partially-occluded facades between aerial and ground views. Quantitative comparisons for automated urban scene facade matching from three cities show superior performance of our method over baseline SIFT, Root-SIFT and the more sophisticated Scale-Selective Self-Similarity and Binary Coherent Edge descriptors. We also illustrate regularity-based applications of occlusion removal from street views and higher-resolution texture-replacement in aerial views.

1. Introduction

With the increasing availability of Google maps and other online mapping tools, geolocating consumer images has become a popular yet challenging task. As a step in this direction, we are interested in matching aerial view facades, such as those automatically detected by the method in [19], with a set of street-view facades to identify the same buildings. This is a challenging problem due to large differences in viewpoint and lighting (Figure 1), temporal disparities between aerial and street-level image collection, and perspective deformations at the street-view level due to the camera's close proximity to each building. Occlusions also complicate the problem – lower levels of a building may be blocked by other buildings in aerial views, and street-view images can be occluded by trees, street-lights, cars, and pedestrians, as well as by other buildings.

Urban facade feature-level matching is inherently ambiguous due to pattern regularities. Even though many existing works in computer vision and computer graphics have exploited such regularities computationally (see Section 2),

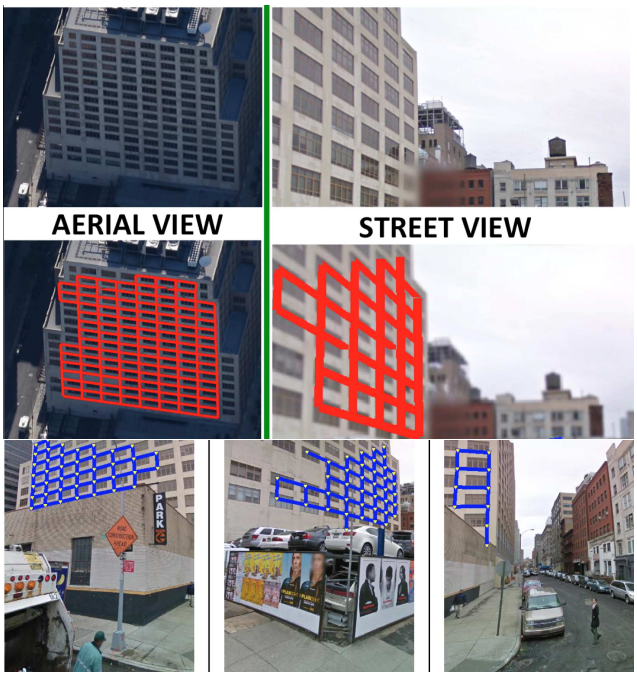

Figure 1. Aerial-view (top-left) and street-view (top-right) images from the **same facade** of an NYC building (image data provided by Google). Our facade matching pipeline finds corresponding facades in spite of drastic variations in viewpoint and lighting. Our method is regularity-driven, using features induced from the automatically detected lattices. Bottom row shows additional matched street-views of this facade. Note that a facade can be matched correctly even when the detected street-view lattice does not overlap with the aerial view lattice.

matching between aerial and street views of the same facade poses technical challenges beyond generic image patch matching and even beyond ground-level-only wide-baseline facade matching. Furthermore, very little work (*e.g.* [4, 30]) has explored a *regularity-driven* approach for urban scene segmentation and matching at the facade level.

We propose to use a lattice and its associated median tiles (motifs) as the basis for matching widely differing aerial

and street-level facade views. Using a lattice tile/motif as a novel, regularity-based descriptor for facades immediately distinguishes this work from all local descriptor-based methods, since regularity is not a local property [21, 22]. We formulate the facade matching problem as a joint regularity optimization problem, seeking well-defined features that reoccur across both facades to serve as match indicators. Match costs based on edge shape contexts, L*a*b color features, and Gabor filter responses are used to find the best one-to-one matching of sampled patches between two roughly aligned motifs, yielding an effective cost function for matching widely disparate facade views (Figure 1).

2. Related Work

It is well known that generic local features such as HOG [8] or SIFT [23] are difficult to match across extreme changes in illumination, viewing angle and image resolution. More robust patch matching features have been proposed, based on feature descriptor normalization to reduce descriptor variance, e.g. Root-SIFT [2] and edge contrast normalization [36], or by methodically trying combinations of feature transforms and binning layouts while learning parameters to maximize matching performance [33]. However, even with the use of robust generic patch descriptors, matching architectural facades is inherently difficult due to an ambiguity in finding the correct correspondence among self-similar patches [28]. These correspondence ambiguities lead in turn to difficulties in estimating planar homographies, fundamental matrices, camera locations, and other quantities computed in a typical structure from motion pipeline [13, 17].

Approaches to wide-baseline facade matching in the literature can be broken roughly into three strategies. The first strategy is to correct for the differences in viewing angle, allowing view-dependent matching using traditional local features to proceed. This is commonly achieved by applying an orthorectification preprocessing step that transforms an arbitrary perspective view of a planar facade into a frontal view where repetition of pattern elements occurs along the horizontal and vertical image axes [4, 17, 34]. This can be done by discovering vertical and horizontal vanishing points and solving for the camera rotation that unwarps the view [34]. The vanishing line of a planar surface can also be estimated from change of scale of repeated pattern elements in the image [7], allowing affine rectification, while rotation and reflection among the elements introduces further constraints that allow solving for a true frontal view (up to similarity transform) [26].

More generally, the authors of [35] note that repeated patterns form *low-rank textures* and present an algorithm called TILT that performs automatic orthorectification of intensity patterns in user-defined regions. Orthorectification greatly simplifies subsequent translation and reflection symmetry analysis [34], allows the use of more discriminative local features such as upright SIFT [3], and reduces the degrees of freedom needed to align two facade views [17].

An alternative to orthorectification is to warp one view into approximate alignment with another oblique view, prior to matching. In [31], ground based multi-view stereo is used to produce texture-mapped depth maps that are then re-rendered based on known camera pose information to synthesize the approximate appearance of the building as seen in the target aerial view. The work of [1] aligns a dominant plane between two oblique aerial views by introducing into the patch matching process an explicit search over affine transformations that simulate the range of patch distortions expected due to viewpoint changes. A recent paper by [18] uses range data and camera parameters from Google street views to warp the dominant building surface plane to appear approximately like a 45% aerial view in order to collect a cross-view patch dataset for deep learning.

A second broad strategy for wide-baseline facade matching is to form feature descriptors specialized for describing self-similar symmetric patterns. A Scale-Selective Self-Similarity (S^4) descriptor is developed in [4] to capture local self-similarity of a patch to its surrounding region, computed at an intrinsic scale proportional to the spatial wavelength of repetition of the pattern. The similarity descriptor for an image patch is formed as a binned log-polar representation of its local autocorrelation surface, computed at the intrinsic scale. Computed over a grid of patches, these descriptors are clustered to detect and segment facades, and to form a set of visual words for naive Bayes matching of facades. The work of [12] densely scores local horizontal and vertical reflection symmetries and local 2n-fold rotational symmetries at all locations and scales in an image. Being based on local symmetry rather than photometry, the resulting descriptors can match facades across large changes of image appearance (e.g. day vs night, drawing vs photo, and modern vs historical view).

The third strategy for facade matching is to explicitly treat the facade as a near-regular texture and to isolate and match unique tiles representing the underlying translated pattern element. One-dimensional frieze patterns and two-dimensional wallpaper patterns are generated when a fundamental pattern element is shifted by integer multiples of one (frieze) or two (wallpaper) generator vectors to form a lattice. However, any translational offset of the lattice defines an equally good partition of the facade pattern into repeating elements, thus there is an inherent ambiguity in determining a unique tile for matching.

Recent work by Ceylan *et al*. [6] requires a user to outline the fundamental repeating element of a pattern, while our application requires an automated solution. In [9], unique tiles are defined by finding the lattice offset such that the Fourier transform of the repeated pattern has phase co-

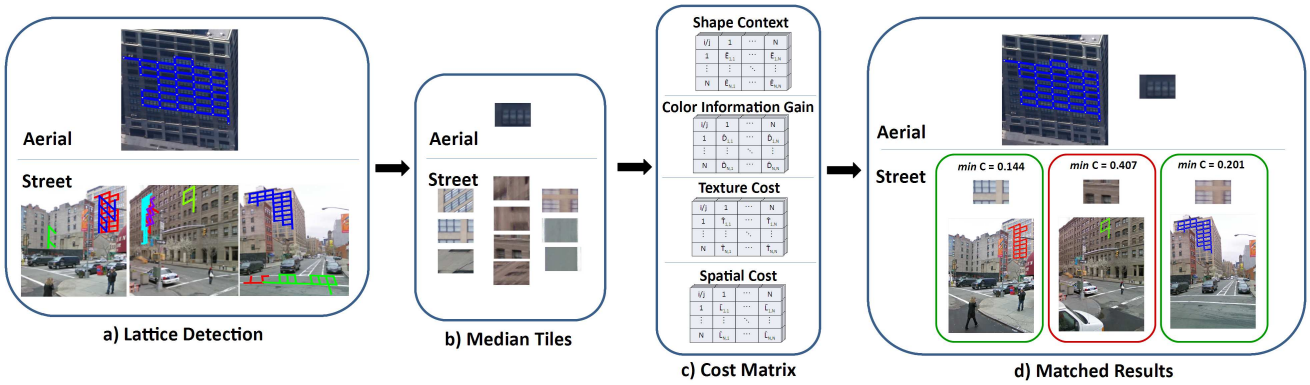

Figure 2. Flowchart showing the overall process of the proposed method on the NYC dataset. a) Lattices are extracted from a street-view (bottom) database and the aerial facade (top) in question. Detected lattices are pruned based on their estimated world-coordinate normal vector to keep only vertical facades. b) Each lattice is represented by the median tile of its translational symmetry. c) A cost matrix is computed from all potential point correspondences for each street-to-aerial pair. Each motif pair will have a cost matrix for each of the four feature costs (shape context, color information gain, texture, spatial smoothness). d) A match cost for each street to aerial facade pair is computed as the sum of its optimal point correspondence set costs. Positive/negative matches are determined by a threshold, learned by maximizing precision/recall on a separate training set.

efficients of zero at its fundamental frequencies in the horizontal and vertical directions. Extracted mean tiles are then matched based on similarity of their grayscale patterns and of the largest two peaks in their RGB color histograms. The work of [20] defines a *motif* of a repeated pattern as a tile that locally exhibits the same rotation and reflection symmetries that characterize the entire periodic pattern. This idea is used in [30] to match facades based on normalized cross-correlation of their respective motifs.

Our proposed approach in this paper is also based on extracting the motif of a lattice to use as a descriptor for facade matching. However, unlike [9] and [30], our matching is based on filtering out candidates using a progressively more discriminative pipeline of features, starting with coarse lattice-structure (geometric) filtering, followed by filtering based on illumination/shadow insensitive color distributions, and finishing with filtering based on features that capture the spatial layout of motif pattern edges.

3. Regularity-based Matching Approach

We propose a regularity-based matching pipeline to identify corresponding facades across aerial and street-level views (Figure 2). High resolution aerial views are first processed by the method in [19] to extract a set of near-regular building facades. Lattices are extracted for each aerial facade and for a set of candidate street-view images that are potential matches, using the translational symmetry detection algorithm developed in [25]. To reduce computational complexity when searching for corresponding street-view images for a detected aerial facade, approximate camera pose information available with both aerial and street-view images is used. Specifically, by backprojecting viewing rays into a UTM ground coordinate system to estimate the approximate ground location for the aerial facade, we select one hundred street-view camera locations that are in close proximity to the estimated aerial facade location. Each street view location yields eight camera shot directions, giving a total of 800 candidate images which are further pruned by the orthogonality between the estimated normal vectors of the corresponding lattice and that of the ground plane.

These lattices facilitate ortho-rectification of the aerial and street-view facades and provide a basis for extracting motifs summarizing their appearance. Each lattice partitions an image region into tiles, which are brought into alignment and then fused by computing the pixel-wise median [20]. This *median tile* or *motif* summarizes the scene facade in terms of regularity and appearance. However, different views of the same facade will still result in orthorectified tiles with slightly different appearances due to projective distortion and differences in scene illumination. One way to look at our method is to consider each computed median tile to be a sample from the entire facade distribution generated under different geometry and lighting conditions. The median tile, as a representative of that distribution, allows us to compare local distributions generated from aerial and street-view samples to identify whether they belong to the same whole-facade distribution.

The main technical contribution of our work is to define a matching cost function to compare a street-view motif to an aerial-view motif based on similarity of color, texture and edge-based context features. The remainder of this section describes in detail this cost function, the features that comprise it, and the sample-based matching procedure that produces a final motif-pair matching score.

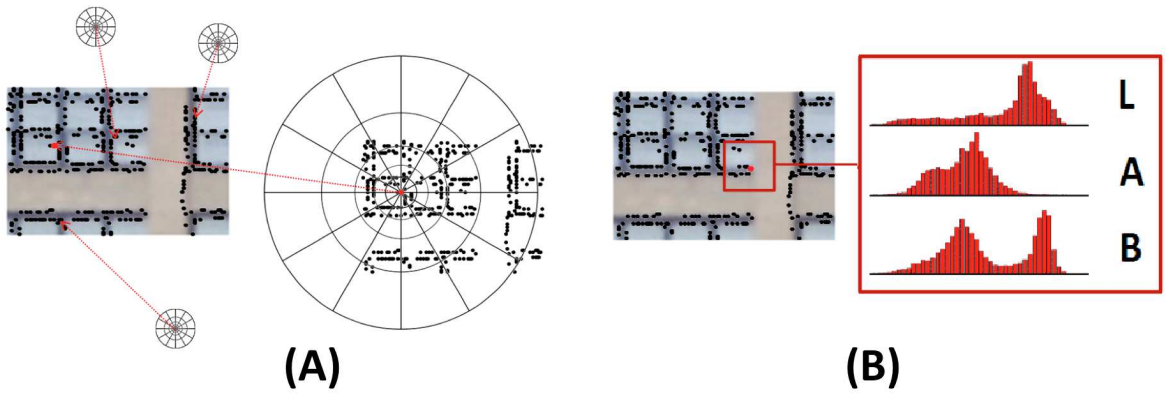

Figure 3. Each motif's edge image is sampled, and patches around each sample are described by shape context, color and texture features. The first two features are illustrated here. (a) **Each** sampled point is described by its own log-polar histogram and shape context scores. For each attempted match, a cost matrix is formed from the SSD scores between all pairs of possible point correspondences between the two motifs. (b) The color cost term is computed from the information gain score between the two equalized LAB color-spaces. A 32x32 patch at each of the sampled points is used to obtain each distribution. The motif shown is from the NYC dataset.

3.1. Motif Cost Function

We characterize a motif by randomly sampling at most 400 points from its high-gradient (edge) pixels. Given two motifs extracted from two facades, we compute their similarity in the form of a pairwise, point-to-point cost function formed as a weighted combination of four terms: (1) local shape context [5], (2) color, (3) texture, and (4) location proximity:

$$C_{i,j} = W_E \hat{E}_{i,j} + W_D \hat{D}_{i,j} + W_T \hat{T}_{i,j} + W_L \hat{L}_{i,j} \quad (1)$$

where $\hat{E}_{i,j}$ is the edge-based shape context cost function for matching sampled pixels i and j, and $\hat{D}$, $\hat{T}$, and $\hat{L}$ are the corresponding color similarity, texture similarity, and location proximity cost functions, respectively.

Since any offset in the translational lattice yields a valid motif tile, we first roughly align each street-view facade motif with the aerial motif before comparison by circularly shifting it to the offset that yields the maximum normalized cross correlation (NCC) score.

3.1.1 Shape Context

Spatial edge layout is a useful measure for discriminating between different window shapes/sizes, as well as weakly discriminating between buildings with different surface textures, *e.g.* uniform texture *vs.* brick texture. Each sampled edge point is characterized by a local shape context [5], using a normalized log-polar histogram, as shown in Figure 3a. The normalized cost of matching two sampled points, i and j, is given by

$$\hat{E}_{i,j} = \sum^{K} \frac{(h_i(k) - h_j(k))^2}{h_i(k) + h_j(k)} \quad (2)$$

where k is a bin belonging to a log-polar histogram, h.

3.1.2 Color

We characterize the color appearance of a building by the color distribution of the motif of the repeated facade pattern. Color distribution of the motif is measured in CIELab color space to account for potential differences in lighting or the presence of shadows. Work done in [10, 11, 29] shows that the CIELab color space is effective at detecting/segmenting despite shadows, since the presence of a shadow will linearly shift each of the three CIELab color space dimensions by a proportional amount depending on the strength of the shadow. We describe the overall texture of a motif by its L, a*, and b* distributions, $f_L(x)$, $f_{a*}(x)$, and $f_{b*}(x)$ respectively.

When comparing two motifs, we first shift the L space distribution of the street-view motif so that its mean value matches the mean of the aerial-view motif. We then shift the a* and b* distributions by the L space shift, ΔL, multiplied by a corresponding proportionality constant, γ, effectively obtaining a shadow-invariant color space. The shifting process is described by the equation

$$\overset{*}{f}_d(x) = f(x - \gamma_d \Delta L) \quad (3)$$

where d is the color dimension, either a* or b*. In our experiments we set γ_a=.135 and γ_b=.435, learned from a training set of street-to-aerial facade matches separate from the ones used for evaluating the PR curves.

To compare color distributions, our approach uses *information gain*, also known as the Kullback-Leibler divergence, D_{KL} [16]. Information gain effectively measures the overall difference between two distributions by measuring the loss of information that occurs when one probability distribution is used to approximate another. In our case, we use information gain to measure how well the aerial-view

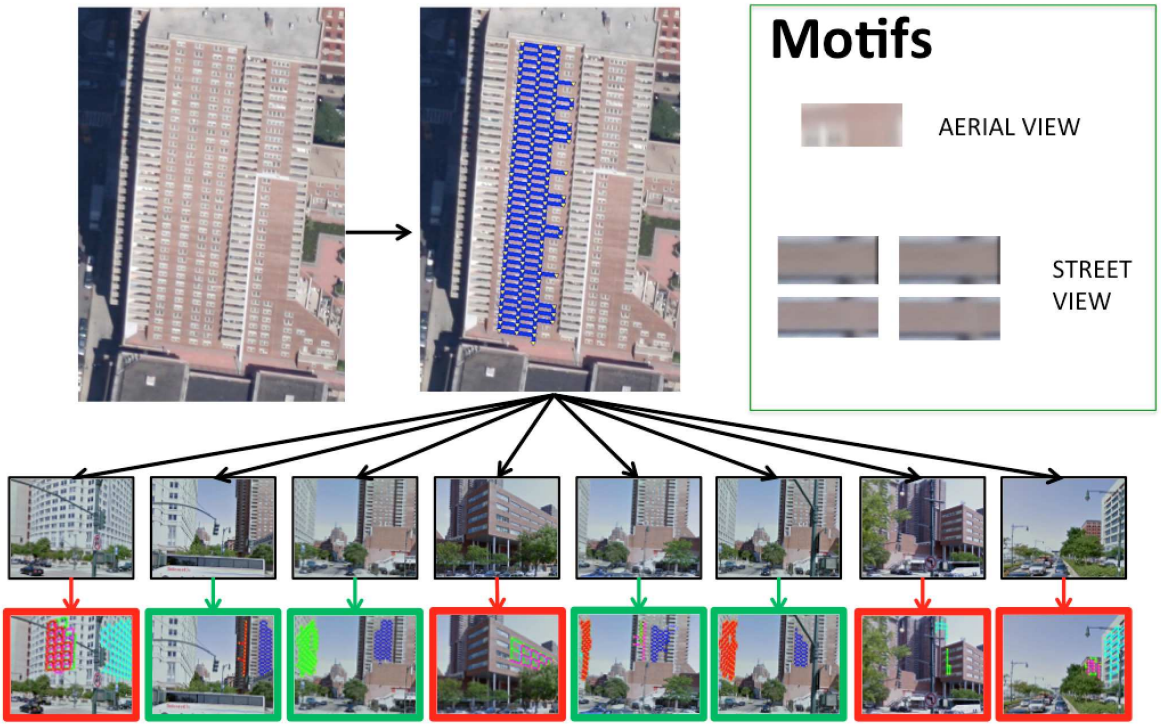

Figure 4. Sample images selected from positive/negative matching results as determined by our matching pipeline (green/red borders, respectively) on the NYC dataset. Matched facades within the positive images are colored blue. Sample motifs from the aerial and matched street view facades are also shown.

patch describes the street-view patch, as given by

$$D_{KL}(f_d^A, \overset{*}{f}_d^S) = \sum_x f_d^A(x) \ln \frac{f_d^A(x)}{\overset{*}{f}_d^S(x)} \quad (4)$$

where $\overset{*}{f}_d^S$ is the shadow-corrected street-view motif distribution, and f_d^A is the aerial-view motif distribution. Two identical distributions result in a score of 0. We normalize D_{KL} by

$$\hat{D}_{KL} = 1 - exp(-D_{KL}) \quad (5)$$

To obtain the cost associated with the color similarity, we apply the Kullback-Leibler divergence to a 32x32 image patch at pixels i and j for each of the color spaces, as shown by Figure 3b. The cost $\hat{D}_{i,j}$ is the average divergence over the three CIELab color dimensions.

3.1.3 Texture

Gabor filter bank responses have been shown to be effective descriptors for many datasets [24, 27]. While urban facade datasets are not as sparse as previously tested datasets, texture features can be useful discriminators for building facades.

When comparing two motifs, we apply four 1-wavelength Gabor filters to each motif at 0°, 45°, 90°, and 135°. The texture cost $\hat{T}_{i,j}$ for each pair of sampled points associated with matching two motifs is the sum of each filter response's SSD (sum squared difference).

3.1.4 Location Proximity

Due to the rigid structure of building facades, relative locations of corresponding motif pixels are expected to vary smoothly, e.g. according to affine deformations due to viewpoint. Therefore, we include an additional location change cost in order to bias the overall solution by this smoothness constraint. The cost is given by $\hat{L}_{i,j}$, which is the relative distance between the two matched points as a ratio to the maximum possible distance (diagonal of motif).

3.2. Matching by Cost Minimization

Optimal correspondences between the sets of sampled points from two motifs are solved by minimizing the cost function of Equation 1 over all 1-1 point correspondences, solved as a bipartite matching problem using the Hungarian algorithm [15]. Weights for the four component cost matrices are $W_E = W_D = W_T = 0.3167$ and $W_L = 0.05$. This gives edge-based, color-based and texture-based appearance features equal weighting, while spatial similarity has a lower weight that adds a slight smoothing bias. An

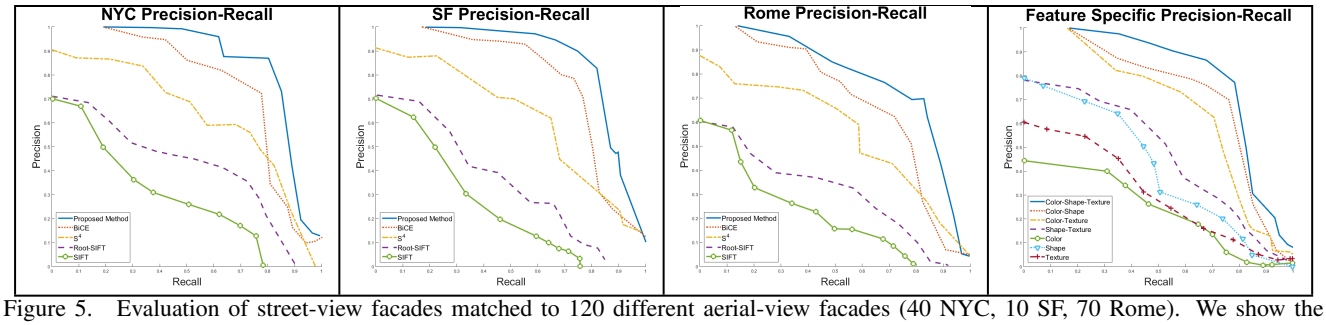

Figure 5. Evaluation of street-view facades matched to 120 different aerial-view facades (40 NYC, 10 SF, 70 Rome). We show the Precision-Recall curves for our proposed method against (1) a baseline approach using SIFT descriptor matching in orthorectified views, (2) Root-SIFT [2], a renormalization of SIFT that outperforms SIFT for retrieving building facades across large view changes, (3) Binary Coherent Edge Descriptors [36], a generic patch matching descriptor applied to our extracted motifs, and (4) S^4 [4], a sophisticated symmetry-based feature designed for facade matching between aerial and ground-level views. The far right panel shows results of our method using different combinations of the 3 major feature spaces (shape context, color, texture) used in our motif matching cost function.

overall matching score for the pair of motifs is given by the sum of the costs returned by the Hungarian algorithm with the optimal point-match set. All potential pairs of motif matches are ranked based on their matching scores, from which positive/negative matches are then determined.

4. Experimental Results

Figure 4 shows qualitative matching results for an aerial facade. Given an aerial facade and its automatically detected lattice, samples of some of the candidate street-view images are shown.

A quantitative evaluation of our method is carried out on a set of 120 aerial facades. Each facade is visible in 10-15 street images, giving us over 1000 total potential facade matches. We have hand labeled all street-view facades corresponding to each aerial facade in the dataset. These labeled facades are treated as the ground truth during our evaluation. A **true positive** match from a street-view facade to an aerial facade occurs when their motifs achieve the highest ranking matching score and they are from the same scene facade. Such a motif-based match can occur even in cases where the two detected facade lattices do not have any spatial overlap. This type of match is still 1-to-1 (albeit not pixel to pixel) since only the best-scoring lattice/motif pair is chosen, one from an aerial view image and one from a street-view image, and thus the Precision-Recall curve is well-defined.

Figure 5 shows a quantitative evaluation based on 120 aerial facade examples. Four different sets of precision-recall curves are shown. The first three show our method compared with other matching methods on 40 NYC, 10 SF, and 70 Rome facades, respectively. To make this comparison fair, street view and aerial view facades were first orthorectified using their detected lattices before computing SIFT descriptors, since it is known that SIFT features are not able to match well across large, oblique viewpoint changes. Even with that help, SIFT and Root-SIFT matching are not as effective at matching facades as our proposed method, or the other sophisticated methods. Finally, we compare the average results of different combinations of our cost function feature spaces across all three cities. Although color alone is not an effective tool for discriminating between different facades, it still adds improvement when used in conjunction with other features. In Figure 6, a 3D cost space is shown for the shape, color, and texture feature costs computed when matching to a particular NYC aerial facade. Blue/red stars are used to indicate whether a street facade is a ground truth match/non-match to the reference aerial facade. The decision made by our matching process is depicted by a green or red dashed line for a match or non-match, respectively.

5. Applications

In this section we show two potential applications for regularity-based matching by using the 2D lattice information for image enhancement in both aerial and street-level views. The first application removes foreground objects that occlude an architectural facade of interest (inpainting) and the second replaces low-resolution facade texture with a higher-resolution version (superresolution). Both inpainting [14] and superresolution [25] of a repeated facade pattern have been addressed previously, but those works synthesize a virtual new texture assuming a perfectly repeating pattern, whereas our approach copies actually observed pattern data from a different unoccluded or higher-resolution view. Inpainting work such as [32], also copies information from other views, but the region to be inpainted is chosen by a user. Our approach automatically detects the region of occlusion by analyzing the facade pattern.

5.1. Removal of Street-Level Occlusion

From our set of matched street-view lattices, a *central lattice* is built that collects and associates patches from each facade across all images in which that facade is vis-

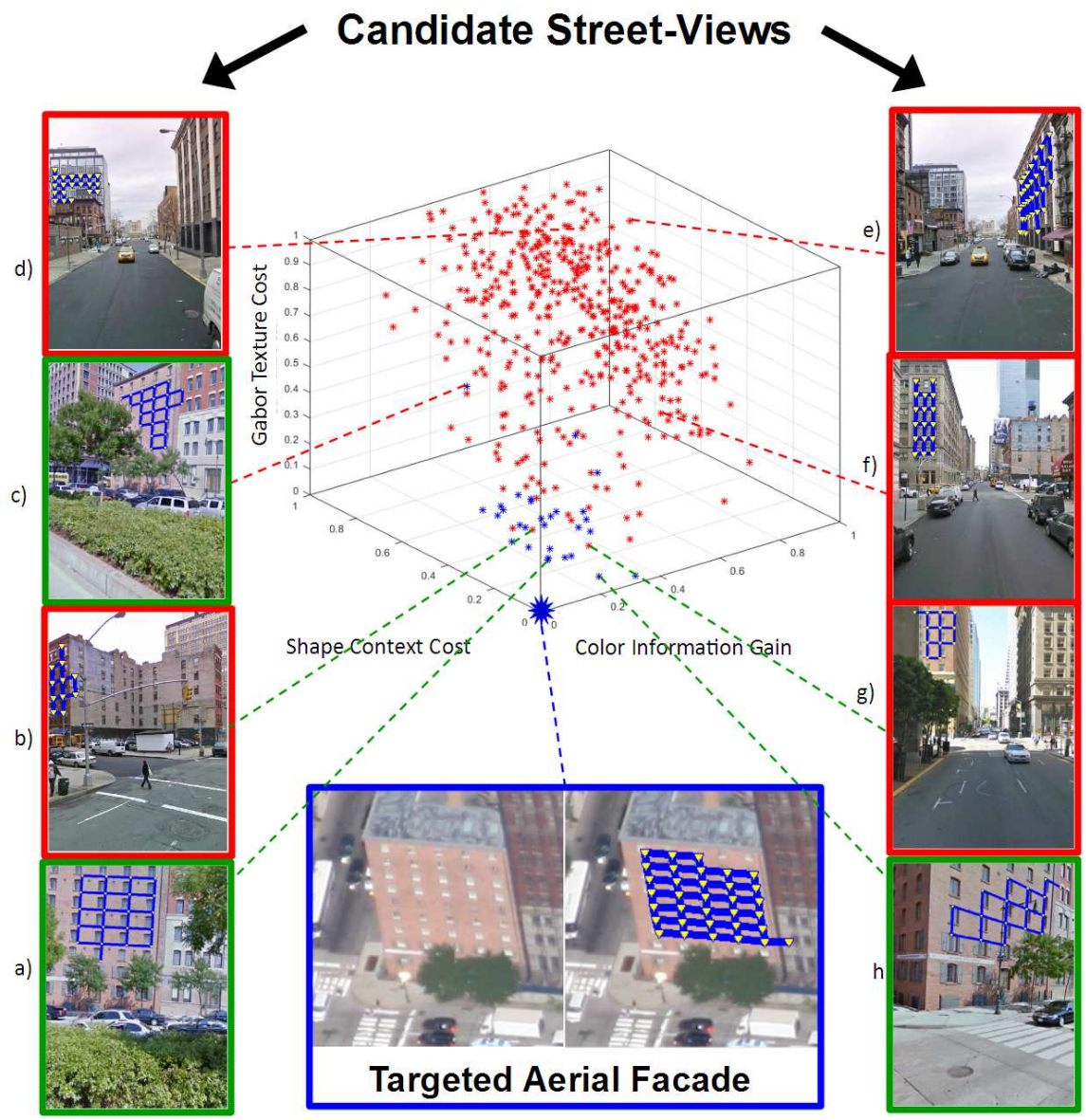

Figure 6. NYC 3D Feature-Cost Space for the three major features used in our proposed method. Blue star = ground truth positive, red star = ground truth negative. Green dashed line = selected as a match by our matching process, red dashed line = selected as non-match. a, h show facades that are matched with low cost to the aerial facade. b, g are examples of facades with similar appearances according to our feature space but are considered false positives, while c is an image that our method does not match well due to significant affine deformations and changes in the window reflection colors. Quantitative results shown in Figure 5

ible. Cross-view matching is performed by correlating each lattice patch set over the patch sets of other images while maintaining the alignments of the two patch sets. The correlation offset location with the highest score is selected as the best matched lattice alignment. That is,

$$loc_{\mathbf{Q}} = \arg\max_{i,j} \ \mathrm{NCC}_{i,j}(\mathbf{P}, \mathbf{Q}) \qquad (6)$$

where $\mathbf{Q}$ is the set of lattice patches currently being considered, $\mathbf{P}$ is the current central lattice patch set, *i,j* is the offset of $\mathbf{Q}$ with respect to the origin of $\mathbf{P}$, and NCC computes the mean normalized cross correlation score between two lattice patch sets at an offset of *i,j* (correlation scores between one or more null patches are not included in the mean score). We leverage the initial aerial view facade by restricting the offset location from causing the central lattice to exceed the aerial-view lattice patch set dimensions. At the end of this process, the central lattice patch set contains patch samples from all matched facades, in their appropriate relative positions with respect to each facade. Note that multiple sample patches may be available for the same relative facade location, when that location is visible in multiple matched street views.

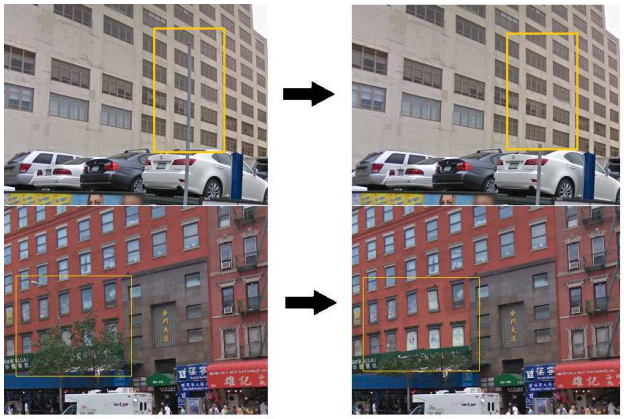

Figure 7. Occlusion removal is performed on this NYC street-view by replacing missing/obstructed lattice patches with available patches from another viewpoint. This is facilitated by construction of a central lattice patch set that brings into alignment corresponding patches from all of the matched street views.

After building a central lattice that contains all visible street-view patches, we are able to automatically remove both major and minor occlusions from a given viewpoint (Figure 7). Minor occlusions are defined as objects that are small and thus minimally affect the occluded lattice patch. Examples include street lamps, sign poles, or electrical wires. These types of occlusions can be automatically detected by comparing the difference to the median patch of this patch to its corresponding median differences from other viewing angles. Patches with minor occlusions are considered those with difference energies several standard deviations above the mean difference energy.

Major occlusions occur when an object obstructs a large portion of the building from some views, affecting the perceived regularity. We can detect these by finding patches in the central lattice patch set that are present in some images, but not present in others even though they fall within that image's field of view.

To correct/replace occluded patches, a mapping of the coordinates from one patch to another, $\mathcal{F}$, is defined by determining the projective transformation between the four corner locations of the occluded patch and the corner locations of a corresponding matched patch. The pixels of the occluded patch are replaced using the mapping $\mathcal{F}$

$$p_o(i,j) = p_m(\mathcal{F}\{i,j\}) \tag{7}$$

where p_o is a pixel in the obstructed patch, p_m is a pixel in a matched patch. We select the image for the patch replacement as the image in closest proximity to the image containing the occlusion in order to minimize perspective distortion.

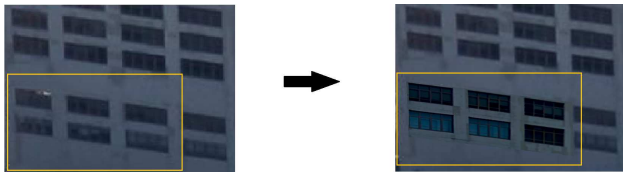

Figure 8. Six aerial lattice patches (from NYC dataset) replaced with corresponding street-view lattice patches after automatically adjusting for lighting differences in CIELab space

5.2. Aerial-Level Image Enhancement through Texture Replacement

As explained in Section 5.1, a central lattice patch set containing all aligned patches detected from street-level views is constructed and can be used to replace pattern tiles that are occluded. Since the central lattice patch set is also aligned with the original aerial image facade lattice, it is also possible to perform texture replacement of patches in the aerial view with patches extracted from the set of street views. Since street views are often of significantly higher resolution than the aerial imagery, this type of texture replacement can be used to generate higher resolution aerial views, as shown in Figure 8.

6. Conclusion

We have addressed the scientific problem of aerial to street-view facade matching. This application poses technical challenges beyond generic image patch matching and even beyond ground-level-only, wide-baseline facade matching. Our results have shown that regularity is an effective tool in extracting discriminative facade features that can be used for matching under challenging viewpoint and lighting changes. By analyzing facade lattice structures, we show that color, shape, and edge-based features combine to form an effective cost function for differentiating between buildings when used within a framework that performs pairwise matching of sample patches summarizing the motif tile of the repeated facade pattern. We also have shown two example applications facilitated by multi-view facade matching and alignment: removal of occlusion from street-level views, and image enhancement of facade texture in aerial views.

7. Acknowledgement

This work is supported in part by NSF grants IIS-1218729, IIS-1144938 (REU), and IIS-1248076 (CREATIV). The Google urban scene data set for research is highly appreciated.

References

[1] H. Altwaijry and S. J. Belongie. Ultra-wide baseline aerial imagery matching in urban environments. In *British Machine Vision Conference (BMVC)*, September 2013. 2

[2] R. Arandjelovic and A. Zisserman. Three things everyone should know to improve object retrieval. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 2911–2918, 2012. 2, 6

[3] G. Baatz, K. Köser, D. M. Chen, R. Grzeszczuk, and M. Pollefeys. Handling urban location recognition as a 2D homothetic problem. In *European Conference on Computer Vision (ECCV)*, pages 266–279, September 2010. 2

[4] M. Bansal, K. Daniilidis, and H. S. Sawhney. Ultra-wide baseline facade matching for geo-localization. In *ECCV Workshop on Visual Analysis and Geo-localization of Large-Scale Imagery*, pages 175–186, October 2012. 1, 2, 6

[5] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(4):509–522, 2002. 4

[6] D. Ceylan, N. J. Mitra, Y. Zheng, and M. Pauly. Coupled structure-from-motion and 3D symmetry detection for urban facades. *ACM Transactions on Graphics*, 33(1):1–15, 2014. 2

[7] O. Chum and J. Matas. Planar affine rectification from change of scale. In *Asian Conference on Computer Vision (ACCV)*, pages 347–360, 2010. 2

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005. 2

[9] P. Doubek, J. Matas, M. Perdoch, and O. Chum. Image matching and retrieval by repetitive patterns. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 3195–3198, August 2010. 2, 3

[10] S. Fukui, N. Yamamoto, Y. Iwahori, and R. J. Woodham. Shadow removal method for real-time extraction of moving objects. In *Knowledge-Based Intelligent Information and Engineering Systems*, pages 1021–1028, 2007. 4

[11] R. Guo, Q. Dai, and D. Hoiem. Single-image shadow detection and removal using paired regions. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 2033–2040, 2011. 4

[12] D. C. Hauagge and N. Snavely. Image matching using local symmetry features. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 206–213, June 2012. 2

[13] J. Heinly, E. Dunn, and J. Frahm. Correcting for duplicate scene structure in sparse 3D reconstruction. In *European Conference on Computer Vision (ECCV)*, pages 780–795, September 2014. 2

[14] T. Korah and C. Rasmussen. Analysis of building textures for reconstructing partially occluded facades. In *European Conference on Computer Vision (ECCV)*, pages 359–372, October 2008. 6

[15] H. W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955. 5

[16] S. Kullback and R. A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, pages 79–86, 1951. 4

[17] M. Kushnir and I. Shimshoni. Epipolar geometry estimation for urban scenes with repetitive structures. *IEEE Pattern Analysis and Machine Intelligence (PAMI)*, 36(12):2381–2395, 2014. 2

[18] T.-Y. Lin, Y. Cui, S. Belongie, and J. Hays. Learning deep representations for ground-to-aerial geolocalization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015. 2

[19] J. Liu and Y. Liu. Local regularity-driven city-scale facade detection from aerial images. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 3778–3785, June 2014. 1, 3

[20] Y. Liu, R. T. Collins, and Y. Tsin. A computational model for periodic pattern perception based on frieze and wallpaper groups. *IEEE Pattern Analysis and Machine Intelligence (PAMI)*, 26(3):354–371, 2003. 3

[21] Y. Liu, H. Hel-Or, C. Kaplan, and L. Van Gool. Computational symmetry in computer vision and computer graphics: A survey. *Foundations and Trends in Computer Graphics and Vision*, 5(1-2):1–199, 2010. 2

[22] Y. Liu, Y. Tsin, and W. Lin. The promise and perils of near-regular texture. *International Journal of Computer Vision*, 62(1-2):145,159, April 2005. 2

[23] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004. 2

[24] B. S. Manjunath and W.-Y. Ma. Texture features for browsing and retrieval of image data. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 18(8):837–842, 1996. 5

[25] M. Park, K. Brocklehurst, R. T. Collins, and Y. Liu. Translation-symmetry-based perceptual grouping with applications to urban scenes. In *Asian Conference on Computer Vision (ACCV)*, pages 329–342, 2010. 3, 6

[26] J. Pritts, O. Chum, and J. Matas. Rectification, and segmentation of coplanar repeated patterns. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 2973–2980, June 2014. 2

[27] F. Riaz, A. Hassan, S. Rehman, and U. Qamar. Texture classification using rotation-and scale-invariant gabor texture features. *Signal Processing Letters, IEEE*, 20(6):607–610, 2013. 5

[28] R. Roberts, S. N. Sinha, R. Szeliski, and D. Steedly. Structure from motion for scenes with large duplicate structures. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 3137–3144, June 2011. 2

[29] E. Salvador, A. Cavallaro, and T. Ebrahimi. Cast shadow segmentation using invariant color features. *Computer Vision and Image Understanding*, 95(2):238–259, 2004. 4

[30] G. Schindler, P. Krishnamurthy, R. Lublinerman, Y. Liu, and F. Dellaert. Detecting and matching repeated patterns for automatic geo-tagging in urban environments. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, June 2008. 1, 3

[31] Q. Shan, C. Wu, B. Curless, Y. Furukawa, C. Hernandez, and S. M. Seitz. Accurate geo-registration by ground-to-aerial image matching. In *International Conference on 3D Vision (3DV)*, pages 525–532, December 2014. 2

[32] O. Whyte, J. Sivic, and A. Zisserman. Get out of my picture! Internet-based inpainting. In *British Machine Vision Conference (BMVC)*, pages 1–11, 2009. 6

[33] S. A. J. Winder and M. Brown. Learning local image descriptors. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2007. 2

[34] C. Wu, J. Frahm, and M. Pollefeys. Detecting large repetitive structures with salient boundaries. In *European Conference on Computer Vision (ECCV)*, pages 142–155, September 2010. 2

[35] Z. Zhang, A. Ganesh, X. Liang, and Y. Ma. TILT: transform invariant low-rank textures. *International Journal of Computer Vision*, 99(1):1–24, 2012. 2

[36] C. L. Zitnick. Binary coherent edge descriptors. In *European Conference on Computer Vision (ECCV)*, pages 170–182, 2010. 2, 6