Accurate Geo-registration by Ground-to-Aerial Image Matching

Qi Shan[†], Changchang Wu[⋆], Brian Curless[†],
Yasutaka Furukawa[⋄], Carlos Hernandez[⋆], and Steven M. Seitz[†⋆]

[†]*University of Washington* [⋄]*Washington University in St. Louis* [⋆]*Google*

{*shanqi,curless,seitz*}*@cs.washington.edu*
{*ccwu,chernand*}*@google.com* *furukawa@wustl.edu*

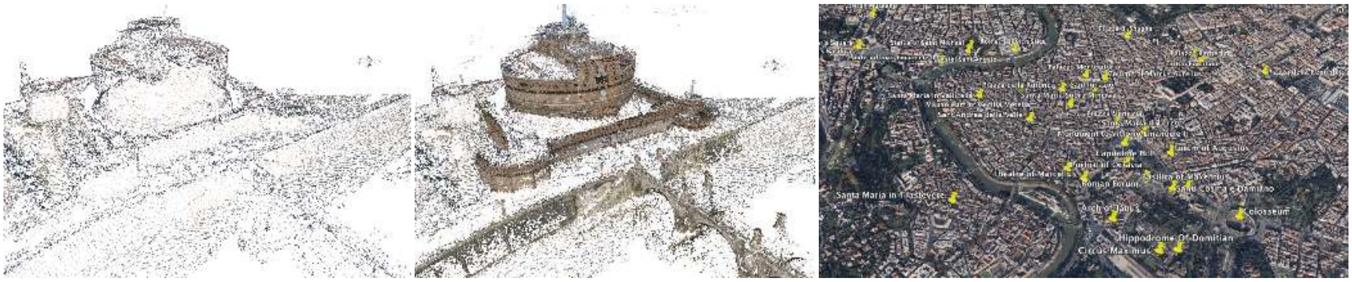


Figure 1. Accurate geo-registration of ground based Multi-view Stereo (MVS) models. Left: an MVS model from geo-located aerial images. Middle: the ground model is successfully geo-registered using the proposed method. Landmark: Castel Sant'Angelo. Right: an overhead view of the Roman landmarks that have been geo-registered, as part of our large scale evaluation.

***Abstract*—We address the problem of geo-registering ground-based multi-view stereo models by ground-to-aerial image matching. The main contribution is a fully automated geo-registration pipeline with a novel viewpoint-dependent matching method that handles ground to aerial viewpoint variation. We conduct large-scale experiments which consist of many popular outdoor landmarks in Rome. The proposed approach demonstrates a high success rate for the task, and dramatically outperforms state-of-the-art techniques, yielding geo-registration at pixel-level accuracy.**

***Keywords*-Registration, Pose Estimation, 3D Modeling**

I. Introduction

In the past several years, 3D reconstruction from Internet photo collections has shown impressive improvements in both scalability and accuracy [3][7][16][21][22][23]. As these Internet photos are mostly taken from the ground, the reconstructed multi-view stereo (MVS) models are highly detailed, but are often disconnected due to the lack of photo coverage in less popular areas. At the same time, these ground-level models are usually not accurately geo-located in a global coordinate system, making it difficult for them to support applications such as digital mapping and autonomous navigation. In contrast, commercial products like Google Earth, Apple's 3D Maps, and Bing maps use geo-located aerial imagery for more uniform 3D reconstruction. The aerial 3D models are complete but much less detailed than the ground-level models. A natural question to ask is: can we achieve the best of both worlds by using the aerial and ground imagery together in 3D reconstructions?

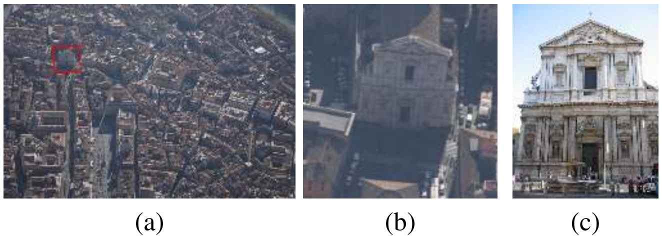


Figure 2. A typical scenario for the ground-to-aerial image registration problem. (a) An aerial image shows part of the city of Rome. The red rectangle highlights the Sant'Andrea della Valle. Even for human vision, it is difficult to find the target geometry from the aerial view. (b) A close-up view of (a). (c) A ground image of Sant'Andrea della Valle.

However, it is difficult to directly match ground and aerial images together, due to the large differences in their camera viewpoints and imaging conditions. Figure 2 illustrates the challenges. First, in the case of aerial images, the scene is observed from much greater distances and at very different angles than in the case of ground images. Typically, landmarks roughly corresponds to 400 × 400 pixels in high resolution aerial images. Second, depending on the direction of the sunlight, certain facades appear very dark in the aerial images, making standard feature detection and matching difficult. In addition, most previous wide-baseline feature-matching methods rely on dominant planar

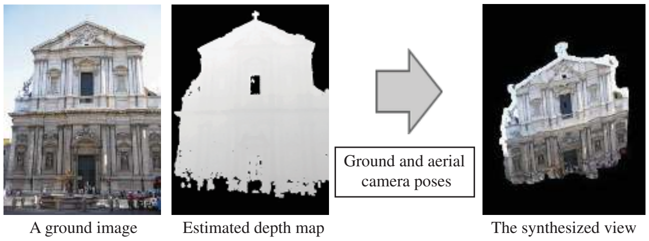


Figure 3. Warping the ground-level image into target view using depth maps and corresponding camera poses.

structures [19][29], but the actual 3D geometry can be more complicated, an assumption that fails for many landmarks (e.g., Figure 1).

In this paper, we address the problem of registering ground-level models to aerial imagery. To this end, we introduce a new viewpoint-*dependent* matching technique to establish pixel accurate feature correspondences between aerial and ground imagery. Our approach helps mitigate the problems caused by the large discrepancies in viewing angles and image resolutions that have frustrated prior efforts. As a result, we can now achieve pixel-level accuracy in geo-registration (within a few centimeters in many cases). This is a significant improvement over the ~ 5.5-meter accuracy attainable using GPS or text labels in prior approaches [17][32].

Our contributions are: 1) a novel viewpoint-dependent matching method that handles large viewpoint changes; 2) a fully automated geo-registration pipeline for matching ground-level photos to aerial imagery; and 3) a large-scale geo-registration evaluation which consists of the most popular outdoor landmarks in Rome, demonstrating an approximately 70% success rate with the proposed system. Aligning ground-based models enables adding dramatically more details to aerial reconstructions.

II. Related Work

Ground-to-aerial image matching for geo-registration is difficult, and standard feature matching techniques often fail (Figure 4a), necessitating manual intervention [6][24]. Very few approaches have demonstrated fully automatic matching and reconstruction of aerial and ground imagery. Shan et al. [21] obtain best results to date using SIFT feature matching [18]. However, their experiments were limited to two landmarks, the Colosseum in Rome and San Marco Square in Venice, which appear relatively large even in the aerial views. In particular, they employ imagery from low-altitude helicopters and use overlook views from the tall towers (i.e., semi-aerial views) to bridge the gap between aerial and ground viewpoints. Unfortunately, such semi-aerial views are not available for most landmarks. The proposed method in this paper works for a much broader range of landmarks and does not rely on semi-aerial views.

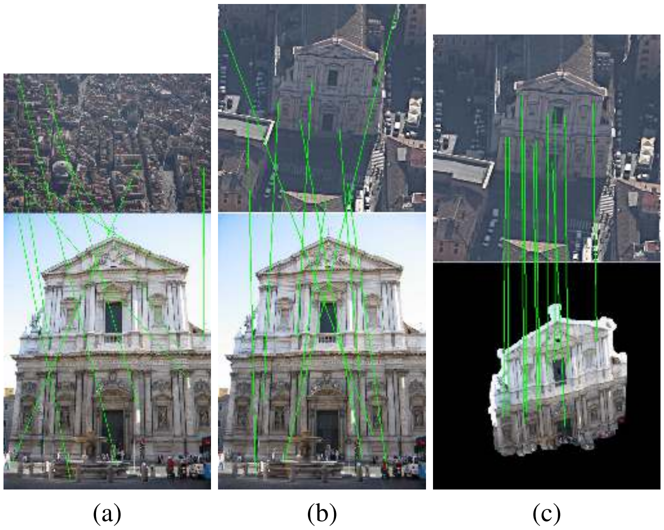


Figure 4. Two-view matching of ground and aerial images. (a) Matching the whole aerial photo with the ground image produces mostly useless feature pairs. (b) Matching an automatically cropped and sharpened aerial photo with the ground image also fails. (c) More reliable feature correspondence is obtained by matching the cropped aerial photo with the synthesized target view from the ground image.

Coarse geo-registration of ground-level models is possible using photo meta-data. GPS and text labels are commonly used to estimate rough geographic location. For more accurate registration, early attempts focused on matching aerial image edges (or map edges) to 2D projections of ground models (projection along the "up" vector) [9][14][32]. However, reliable matching by these methods requires multiple facades (or multiple map edges) in the ground-level models, which is not always the case. Furthermore, their altitude estimation is often less accurate. Recently, researchers have looked into using other sources of geo-location proxies, for example, matching ground images to geo-located Google Street View photos [8][27], or looking for GPS tagged images with similar appearance [11][15], matching to geo-located ground-level 3D points [17]. Nevertheless, it is difficult for these methods to achieve high precision; for example, the average error is about 20 meters in [32], and 5.5 meters in [17]. In this paper, we establish feature matches between aerial and ground imagery for geo-registration with pixel-level (centimeter) accuracy.

Invariant features (e.g., SIFT) are typically used to tackle viewpoint changes. Beyond the invariance to scale changes, there exists a rich body of work on affine or perspective invariance for improved robustness to large viewpoint changes [25][19][29][30][4]. These techniques usually assume dominant planar geometry for simulating different views with a homography or affine transformation, and as a result, their performance suffers when the scene geometry is complex or has many foreground occluders. Furthermore, the viewing angle variation between the aerial and ground imagery is so dramatic (45°) that it usually falls outside the operating range of most image matching methods. Our experiments with state-of-the-art methods ([19] and [29]) show that they are insufficient for our aerial to ground registration task.

An alternative to image matching is direct 3D model alignment using 3D feature points [5][13][31][20]. These approaches assume meshes with similar resolutions and with a substantial amount of overlap. Unfortunately, aerial models are much coarser than ground models (meter versus centimeter resolution), and it is difficult to extract accurate mesh features for matching. Furthermore, the aerial and ground models usually do not overlap enough. Geometry that is visible in aerial views, e.g., rooftops, rarely appears in ground images, and vice versa. Therefore, it is difficult to achieve pixel-level accuracy via 3D feature based techniques.

III. Algorithm Overview

Given ground-level MVS reconstructions, our goal is to accurately and automatically align these MVS models to the aerial images, which have been geo-referenced already. The following is an overview of the proposed algorithm.

We first obtain an approximate geo-referenced ground-based MVS model by performing GPS-based geo-registration using the EXIF tags of ground images. The ground-level images are collected from Flickr [1], of which roughly 10% have GPS tags [7]. As many of the GPS tags are inaccurate (due to poor reception, user tagging, etc.), the RANSAC process typically can locate the 3D models only within a 20 meter range [32].

Based on the estimated geo-location of the ground models, we retrieve oblique aerial views from Google Maps [2].[1] The oblique images are captured from 4 different directions, east, south, west, and north. Our method finds feature matches between the ground and aerial images to geolocate the ground models to pixel-level accuracy.

Section IV proposes a new viewpoint-dependent matching method, which effectively deals with the large viewpoint differences between the ground-level and the aerial images. Section V presents our aerial view-selection algorithm, which leads to efficient and robust matching. The final 3D transformation is recovered by applying RANSAC to the feature matches. In our experiments, 41 out of 59 outdoor landmarks in Rome were successfully registered, a 70% success rate.

IV. Viewpoint-Dependent Feature Matching

We consider the problem of finding accurate feature matches between two sets of images with large viewpoint changes. In ground-to-aerial matching, we have accurate geo-reference information for aerial images, while the location of the ground MVS reconstructions can be recovered only approximately from GPS tags.

SIFT is sufficient for small viewpoint changes, as the local transformations are close to similarity transforms. For larger viewpoint changes, affine invariance can be achieved on planar structures [19]. When accurate, dense 3D reconstructions of both models are available, improved invariance can be achieved with viewpoint-invariant patches extracted from synthesized local orthogonal views [29]. Unfortunately, the ground-to-aerial registration problem has (i) drastic viewpoint changes, (ii) very complicated geometry, and (iii) sparse and noisy reconstructions from aerial imagery. Therefore, none of the above techniques are applicable.

Instead of seeking invariant feature detections, we propose to match view-dependent features by exploiting approximate alignment information and underlying 3D geometry. Consider matching a ground MVS model (source) and an aerial image (target). We assume that dense depth maps exist for the source images, and that approximate alignment information is available, from which we can synthesize the source images rendered from the target viewpoint. Standard small-baseline features can then be applied to match the synthesized views with the target image. Note this is fundamentally different from VIP matching [29], which needs to synthesize invariant views for both source and target images. In fact, the proposed matching method only requires the dense 3D geometry of the source imagery.

For the target view synthesis, we compute MVS reconstructions of the source imagery and create dense depth maps with a bilateral filter-based interpolation process as described in [12]. The depth maps are first computed by back-projecting visible MVS points to each view, and then interpolated with a bilateral filter to fill in possible holes (window radius is 10 pixels and the regularization parameter is 0.16). We then further smooth the depth maps with a Gaussian filter of size 11 to reduce warping artifacts.

Given the recovered depth maps, we are able to synthesize the target view for each source (ground level) image by depth-based warping . After that, we use SIFT to match with the aerial views. See Fig. 4(c) for an example. Experiments show that the viewpoint-dependent feature matching works well for the large viewpoint and scale changes in the aerial and ground matching, where direct matching will typically fail.

A key advantage of our viewpoint-dependent feature matching over [29] is the ability to handle large scale changes and exploit the approximate alignment information. By warping the ground level images into the aerial views, and matching at the resolution of the aerial views, our algorithm naturally ignores the small 3D structures that are invisible in the aerial views. In fact, we found that the failures of [29] in our problem are often due to feature matching at the wrong scales.

[1] Note that we don't need aerial 3D geometry for geo-registering the ground models.

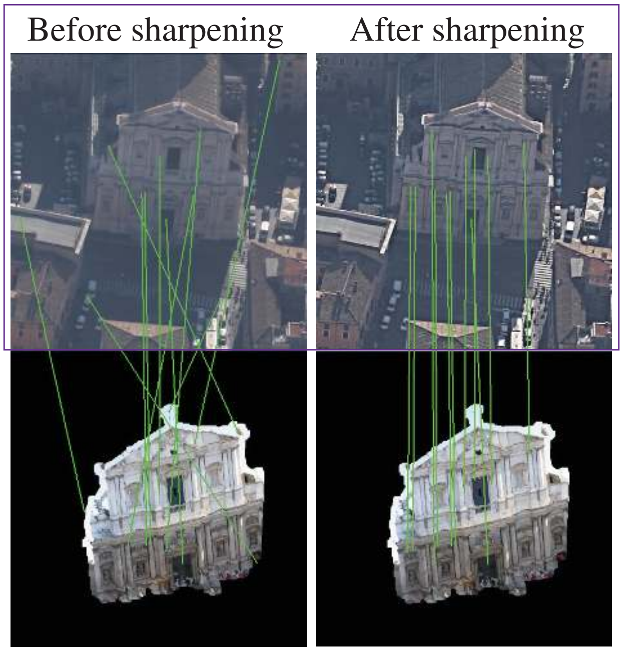


Figure 5. Image sharpening brings up the contrast level of aerial views in shadow, improving the feature matching. Note that the aerial crop we show in Figure 4 (b,c) is after sharpening.

V. Aerial View Selection and Matching

The initial alignment for the ground models is obtained by using a GPS tag based RANSAC, which typically has an accuracy of around 20 meters. This precision allows us to automatically select the proper aerial views for matching.

First, given an approximately aligned ground model, we identify the aerial images that contain the model in their viewports. Specifically, for each of the four aerial oblique viewing directions, we select the aerial image, whose center is the closest to the center of the ground model on the image.

Second, each ground model usually corresponds to a small fraction of an aerial image, due to the large scale changes. For the purpose of efficiency, our system automatically crops the aerial images to narrow down the search space (Figure 4b). Specifically, based on the 3D bounding box of the ground model, we extract the sub-images that cover the desired region of interest by projecting the bounding box into the aerial views. In our implementation, we use 5 pre-defined crop sizes: 401×401, 601×601, 801×801, 1001×1001, and 1201×1201. To account for the error in the initial registration, we choose one crop size that is approximately twice the size of the region of interest.

Third, we apply contrast adjustment to deal with landmarks in shadow. The aerial images are all taken on sunny days in Rome, Italy. Therefore, north-facing facades are always in shadow, resulting in low contrast aerial views. The contrast mismatch has a significant effect in SIFT matching. To address it, we apply an unsharp mask to enhance these north facing views. That is, $I_s = (1+a)I - aI * g$, where I_s is the sharpened image, I is the original cropped aerial image, g is a 7×7 Gaussian filter with standard deviation $\sigma = 1$, $*$ is the convolution operator, and a is the sharpening ratio (0.25 in our implementation). The aerial view enhancement produces better two-view matching (Figure 5), leading to more accurate geo-registration.

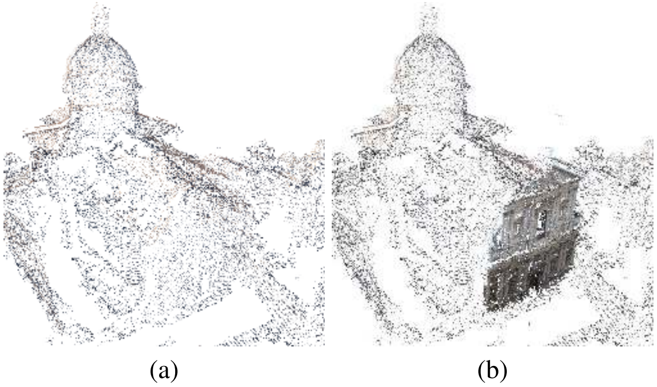


Figure 6. Applying the estimated similarity transform to the ground model. (a) The aerial MVS model. (b) The transformed ground model on top of the aerial model. Note that accurate geo-registration has been achieved.

Feature matches between the ground and aerial images are then converted into a list of 3D point pairs, denoted $\{(P_i^a, P_i^g)\}$. P_i^a is one (feature) point in the aerial view, back-projected into 3D according to the underlying aerial geometry. P_i^g is the matched ground point, which is back-projected based on the interpolated depth map. The ground-to-air alignment is sought by finding the optimal 3D similarity transformation between the two sets of 3D points. The error to be minimized can be written as $\sum_i \left\| P_i^a - (s\mathbf{R}P_i^g + \mathbf{T}) \right\|^2$, where s is the scale factor, $\mathbf{R}$ is the rotation matrix, and $\mathbf{T}$ is the translation vector. The closed form solution for the similarity transform is given by [26]. For robustness to outliers, we use a RANSAC process to find the transformation with the largest number of inliers. We set the distance threshold to 5 centimeters. To account for possible low inlier ratios, we empirically let the RANSAC process take $100,000$ iterations, where each iteration randomly picks 3 pairs of points. This would guarantee a success probability of 0.999996 even if the inlier ratio is 5%. Finally, the estimated similarity transform is applied to the ground model for geo-registration (Figure 6).

VI. Evaluation

The proposed system is evaluated on popular landmarks in the city of Rome. We downloaded ground images from Flickr, and ran a standard 3D reconstruction pipeline (VisualSFM [28] followed by PMVS [10]). After removing indoor scenes and small models of less than 20 images, we keep 59 datasets for the geo-registration experiment. The number of images in each dataset varies from 28 (Santa Croce in Gerusalemme) to 5000 (Colosseum). 12 of the 59 datasets

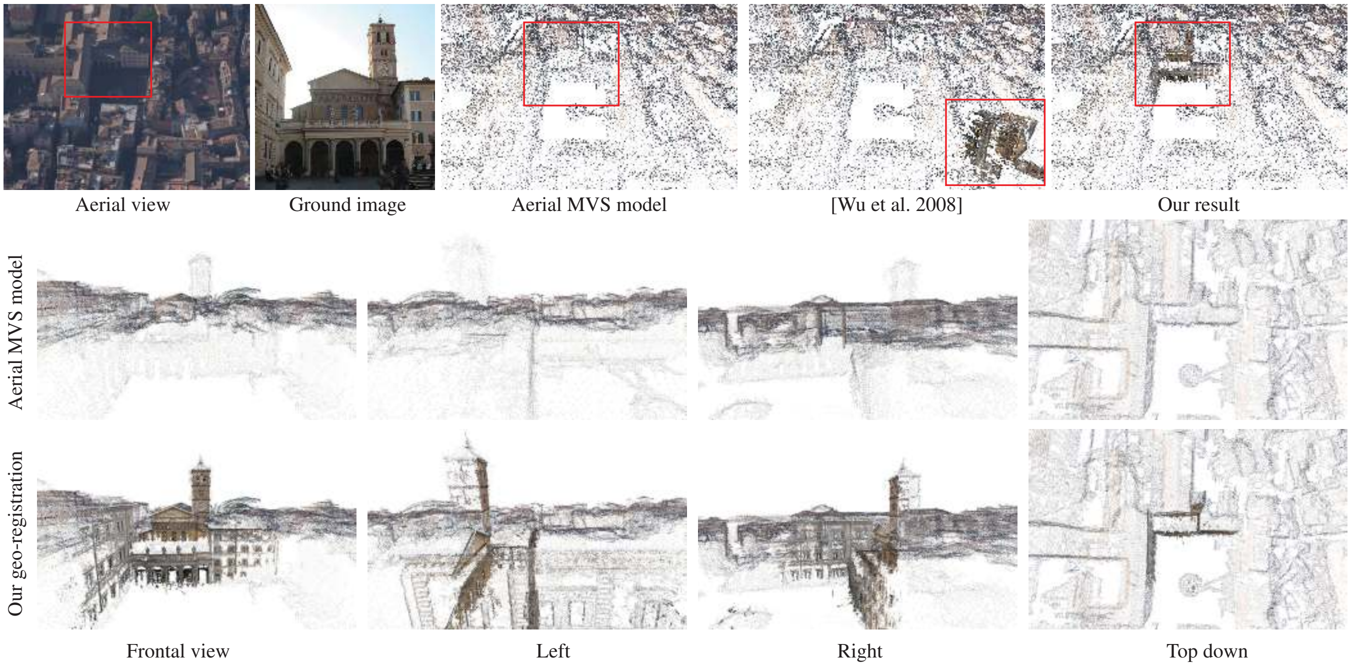


Figure 9. Comparing against the VIP matching in [29]. Landmark: Santa Maria in Trastevere.

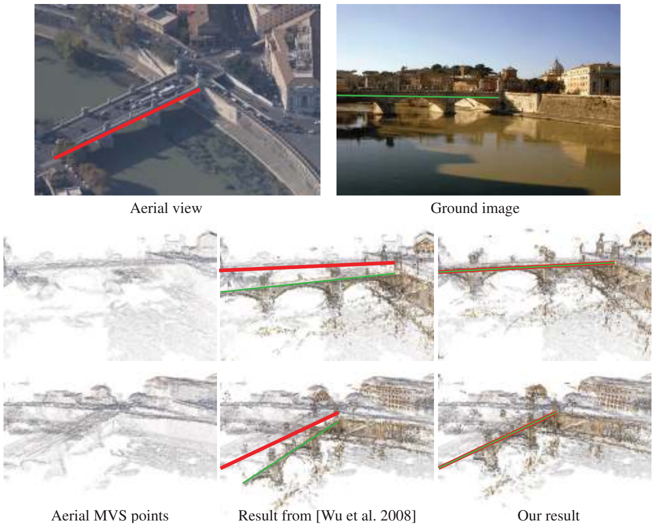


Figure 7. Comparing against the VIP matching in [29]. For better visualization, we manually place red and green lines to highlight the bridge beams in aerial and ground models, respectively. Note that the geo-registration from the proposed method is more accurate as the bridge beams from the aerial and ground model overlap. Landmark: Ponte Vittorio Emanuele II.

have more than 1000 images. As the geo-registration target, we collected $31,891$ aerial images that cover the entire city of Rome. Most of the computation time is spent in the pre-processing steps computing SfM [28] and MVS [10]. In particular, it takes about a day for each of the biggest datasets with a distributed reconstruction system. The geo-registration process takes less than 20 minutes on a single machine with 8 threads.

The proposed method successfully registers 41 out of the 59 landmarks in Rome, which gives a 69.5% success rate. Qualitative results can be found in Figure 7, 8, 9, and 10, which demonstrate the accuracy of the proposed registration method. Our method shows clear improvements over the initial GPS-based geo-registration.

Our view-dependent feature matching approach is critical to handling large viewpoint and scale changes. For comparison, we also run the geo-registration pipeline using standard SIFT [18], ASIFT [19] and VIP matching [29]. As expected, SIFT and ASIFT are not capable of handling the drastic viewpoint changes (See Fig. 8). One may suspect that a possible reason of the failure is the ratio-test. We experimented with this hypothesis by disabling the ratio-test. It indeed increases the number of putative feature matches between the ground and aerial views. However, the feature matches become much noisier. In the end, disabling the ratio-test does not increase the number of successfully registered landmarks. VIP matching works reasonably well for a few datasets where the scene geometry is relatively simple, but fails in most cases. One of the successful examples of VIP matching is shown in Figure 7, where the ground MVS model of the bridge has been registered to the aerial model. However, the registration error from VIP matching is significantly larger. Two failure examples are shown in Figure 8 and 9. VIP detection and matching relies heavily on correctly parsing local scene geometries, e.g., the plane detection process. The performance varies depending on various thresholds which need to be tuned

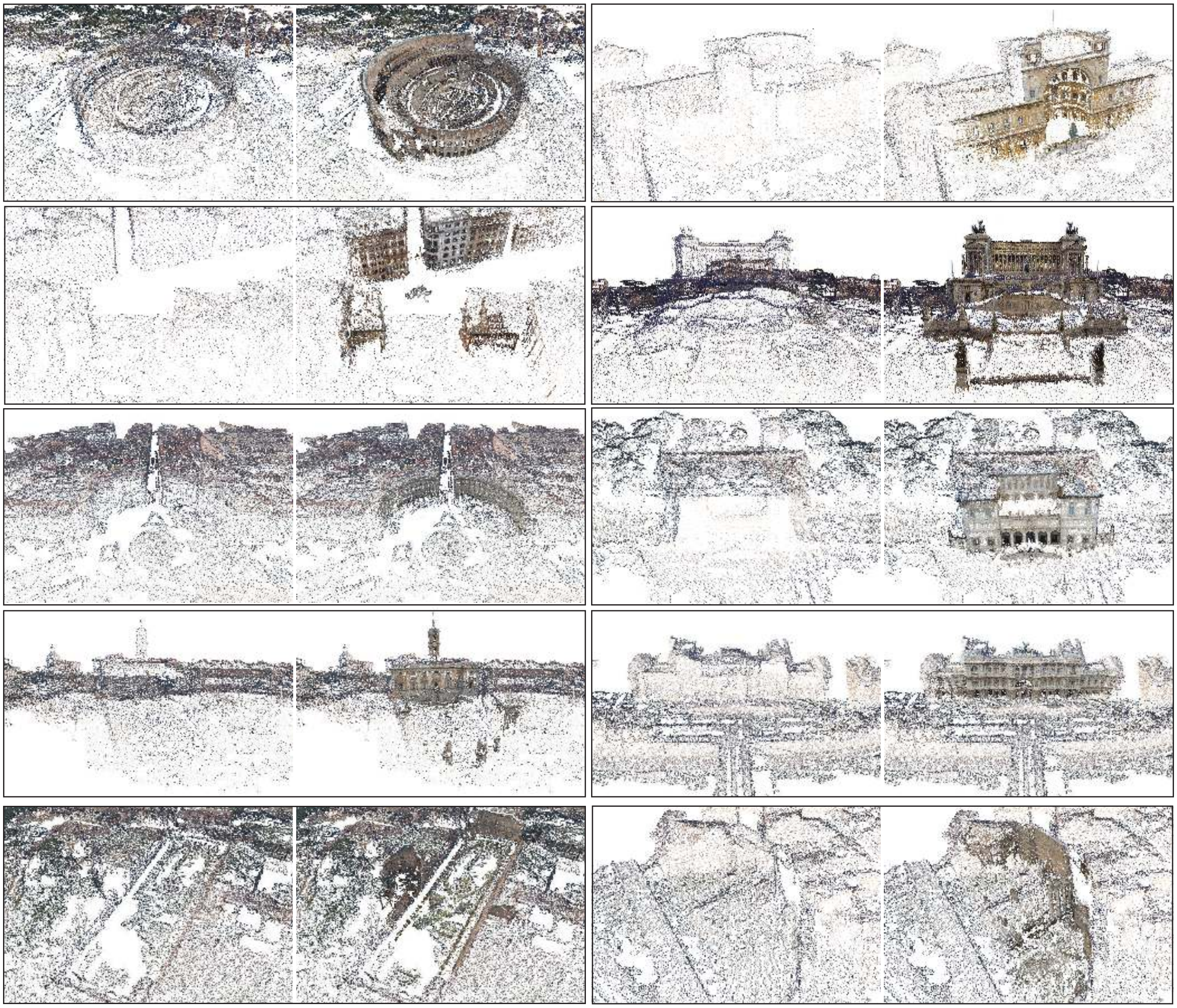


Figure 10. More results on matching the aerial view to the ground image. Left: the point cloud from aerial MVS models. Right: our geo-registered models. Please zoom into the original resolution for best visual quality.

for each landmark. Moreover, the matching tends to get confused by the large number of small 3D features in the ground-level reconstruction.

We tried to conduct quantitative evaluations of the geo-registration accuracy, but it is not clear how to define a good metric. One option is to define a recall score by setting a distance threshold and measuring the percentage of ground MVS points that have aerial reconstructions within the threshold. However, due to the large difference in resolution and coverage between the aerial and ground models, this metric does not necessarily favor the better geo-registration result. For example, a ground model of the frontal facade of a church may be mis-matched to an incorrect planar region in the aerial model, and produce a comparable or higher recall score. We hope to develop a better metric for the problem in future work.

The proposed approach does fail for some landmarks. Figure 11 shows such an example, where the building is under construction in the aerial model, but not in the ground model. When the aerial and ground models are not consistent, the proposed method is not able to find enough feature matches. Another cause of failure is due to noisy ground reconstructions. Since our viewpoint-dependent feature matching relies on the ground geometry for warping, it is vulnerable to noisy ground reconstructions, which result in severely distorted synthesized views.

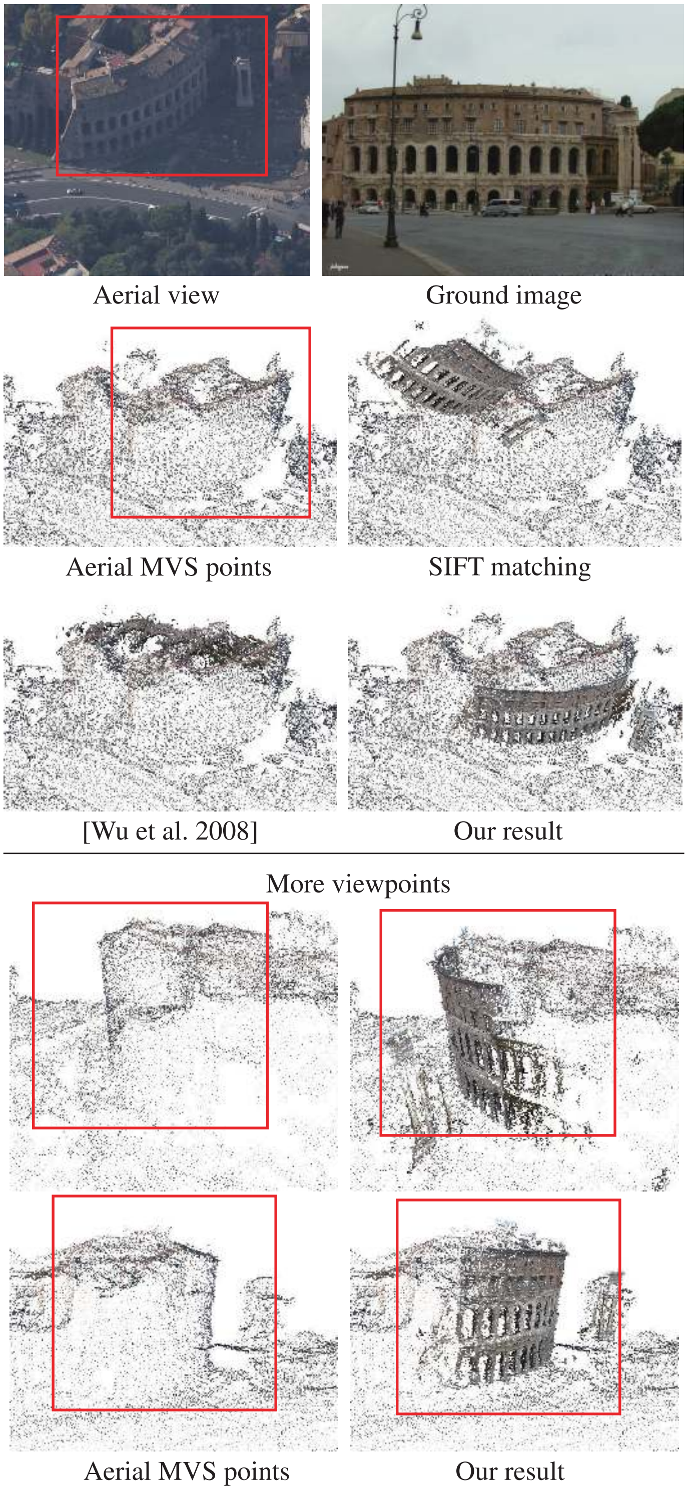
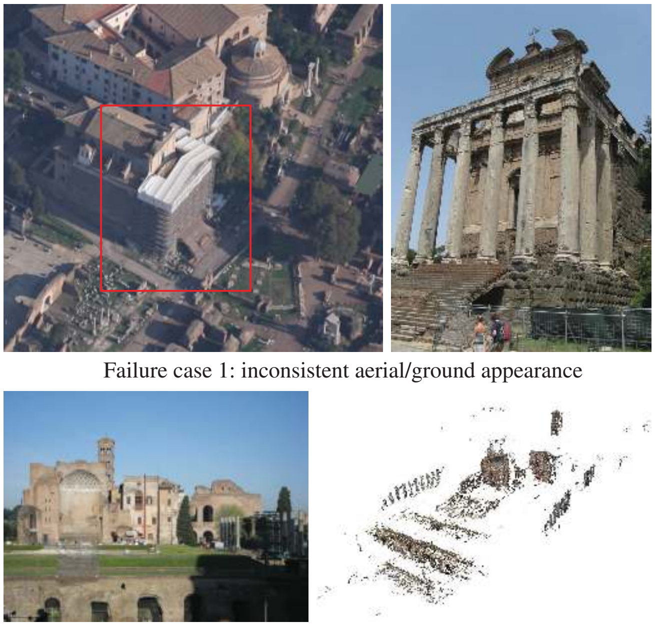


Figure 8. Comparing against the baseline method and the VIP matching in [29]. Landmark: Theatre of Marcellus.


Failure case 1: inconsistent aerial/ground appearance

Failure case 2: noisy ground MVS reconstruction

Figure 11. Failure cases. The proposed method fails when (i) the aerial/ground geometry is not consistent, and (ii) the ground geometry is noisy.

VII. Conclusion

This paper presents a fully automatic system to geo-register ground MVS reconstructions. The system is capable of handling drastic viewpoint variations by adopting a novel view-dependent feature matching approach. We conducted a large scale experiment using 59 popular outdoor landmarks in Rome. Our results are significantly better than existing techniques.

Our approach does have some limitations. It relies on the quality of ground MVS reconstructions, and assumes consistent appearance in both aerial and ground imagery (Figure 11). The registration accuracy degrades at the presence of severe occlusions. Although the ultimate goal is to create high-resolution city-scale 3D models, currently we are only able to achieve this desired resolution at city landmarks where dense ground images are available. One topic of future work is to incorporate another source of ground-level imagery such as Google Street View images. Finally, the proposed approach simply estimates a similarity transformation between the aerial and ground models. Slight mis-alignments are observed in some of the datasets, which could be reduced by global bundle adjustment on all the ground and aerial imagery, incorporating the ground-to-aerial feature matches obtained by the proposed algorithm.

Acknowledgments

We thank Google Seattle Lightfield team for helpful discussion and support. This work was done when Qi Shan interned at Google.

References

[1] Flickr. http://www.flickr.com. 3

[2] Google maps. http://maps.google.com/. 3

[3] S. Agarwal, Y. Furukawa, N. Snavely, B. Curless, S. M. Seitz, and R. Szeliski. Building rome in a day. *Communications of the ACM*, 54(14):105–112, October 2011. 1

[4] M. Bansal, K. Daniilidis, and H. Sawhney. Ultra-wide baseline facade matching for geo-localization. *ECCV 2012. Workshops and Demonstrations, Lecture Notes in Computer Science*, 7583:175–186, 2012. 2

[5] P. J. Besl and N. D. McKay. Method for registration of 3-D shapes. *TPAMI*, 14(2):239 – 256, 1992. 3

[6] P. Cho, N. Snavely, and R. Anderson. 3D exploitation of large urban photo archives. In *SPIE Defense, Security, and Sensing*, pages 769714–769714. International Society for Optics and Photonics, 2010. 2

[7] J.-M. Frahm, P. Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys. Building rome on a cloudless day. In *ECCV*, 2010. 1, 3

[8] J.-M. Frahm, J. Heinly, E. Zheng, E. Dunn, P. Fite-Georgel, and M. Pollefeys. Geo-registered 3D models from crowd-sourced image collections. *Geo-spatial Information Science*, 16(1):55–60, 2013. 2

[9] C. Frueh and A. Zakhor. Constructing 3D city models by merging ground-based and airborne views. In *CVPR*, 2003. 2

[10] Y. Furukawa and J. Ponce. Accurate, dense, and robust multi-view stereopsis. *TPAMI*, 32(8):1362–1376, 2010. 4, 5

[11] J. Hays and A. A. Efros. IM2GPS: estimating geographic information from a single image. In *CVPR*, 2008. 2

[12] K. He, J. Sun, and X. Tang. Guided image filtering. In *ECCV*, 2010. 3

[13] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. From structure-from-motion point clouds to fast location recognition. In *CVPR*, 2009. 3

[14] R. S. Kaminsky, N. Snavely, S. M. Seitz, and R. Szeliski. Alignment of 3D point clouds to overhead images. In *Second IEEE Workshop on Internet Vision*, 2009. 2

[15] J. Knopp, J. Sivic, and T. Pajdla. Avoiding confusing features in place recognition. In *ECCV*, 2010. 2

[16] A. Kushal, B. Self, Y. Furukawa, D. Gallup, C. Hernandez, B. Curless, and S. Seitz. Photo tours. In *3DImPVT*, 2012. 1

[17] Y. Li, N. Snavely, D. Huttenlocher, and P. Fua. Worldwide pose estimation using 3D point clouds. In *ECCV*, 2012. 2

[18] D. Lowe. Distinctive image features from scale-invariant keypoints. In *IJCV*, volume 20, pages 91–110, 2003. 2, 5

[19] J.-M. Morel and G. Yu. Asift: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2(2), 2009. 2, 3, 5

[20] R. B. Rusu, N. Blodow, and M. Beetz. Fast point feature histograms (fpfh) for 3D registration. In *ICRA*, pages 3212–3217, 2009. 3

[21] Q. Shan, R. Adams, B. Curless, Y. Furukawa, and S. M. Seitz. The visual Turing test for scene reconstruction. In *3DV*, 2013. 1, 2

[22] Q. Shan, B. Curless, Y. Furukawa, C. Hernandez, and S. M. Seitz. Occluding contours for multi-view stereo. In *CVPR*, 2014. 1

[23] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: Exploring image collections in 3D. In *SIGGRAPH*, 2006. 1

[24] K. Tuite, N. Snavely, D.-y. Hsiao, N. Tabing, and Z. Popovic. Photocity: training experts at large-scale image acquisition through a competitive game. In *CHI*, 2011. 2

[25] T. Tuytelaars and K. Mikolajczyk. Local invariant feature detectors: a survey. *Foundations and Trends® in Computer Graphics and Vision*, 3(3):177–280, 2008. 2

[26] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *TPAMI*, 13(4):376–380, 1991. 4

[27] C.-P. Wang, K. Wilson, and N. Snavely. Accurate georegistration of point clouds using geographic data. In *3DV*, pages 33–40, 2013. 2

[28] C. Wu. Towards linear-time incremental structure from motion. In *3DV*, 2013. http://ccwu.me/vsfm. 4, 5

[29] C. Wu, B. Clipp, X. Li, J.-M. Frahm, and M. Pollefeys. 3D model matching with viewpoint-invariant patches (vip). In *CVPR*, 2008. 2, 3, 5, 7

[30] C. Wu, F. Fraundorfer, J.-M. Frahm, and M. Pollefeys. 3D model search and pose estimation from single images using vip features. In *CVPR Workshops*, pages 1–8, 2008. 2

[31] A. Zaharescu, E. Boyer, K. Varanasi, and R. Horaud. Surface feature detection and description with applications to mesh matching. In *CVPR*, 2009. 3

[32] A. Zamir and M. Shah. Image geo-localization based on multiple nearest neighbor feature matching using generalized graphs. *TPAMI*, 36(8):1546–1558. 2, 3