3D Model Matching with Viewpoint-Invariant Patches (VIP)

Changchang Wu,[1]Brian Clipp[1], Xiaowei Li[1], Jan-Michael Frahm[1] and Marc Pollefeys[1,2]

[1]Department of Computer Science
The University of North Carolina at Chapel Hill, NC, USA
{ccwu,bclipp,xwli,jmf}@cs.unc.edu

[2]Department of Computer Science
ETH Zurich, Switzerland
marc.pollefeys@inf.ethz.ch

Abstract

The robust alignment of images and scenes seen from widely different viewpoints is an important challenge for camera and scene reconstruction. This paper introduces a novel class of viewpoint independent local features for robust registration and novel algorithms to use the rich information of the new features for 3D scene alignment and large scale scene reconstruction. The key point of our approach consists of leveraging local shape information for the extraction of an invariant feature descriptor. The advantages of the novel viewpoint invariant patch (VIP) are: that the novel features are invariant to 3D camera motion and that a single VIP correspondence uniquely defines the 3D similarity transformation between two scenes. In the paper we demonstrate how to use the properties of the VIPs in an efficient matching scheme for 3D scene alignment. The algorithm is based on a hierarchical matching method which tests the components of the similarity transformation sequentially to allow efficient matching and 3D scene alignment. We evaluate the novel features on real data with known ground truth information and show that the features can be used to reconstruct large scale urban scenes .

1. Introduction

In recent years, there have been significant research efforts in fast, large-scale 3D scene reconstruction from video. Recent systems show real time performance [14]. Large scale reconstruction from only video is a differential technique which can accumulate error over many frames. To avoid accumulated errors, the reconstruction system must recognize previously reconstructed scene parts and determine the similarity transformation between the current and previous reconstructions. This similarity transformation is equivalent to the accumulated drift. Our novel feature can be used to establish these recognition based links. Traditionally image-based matching is used to provide the loop closing constraints to bundle adjustment. The irregularity of 3D structure makes matching 3D models using only texture information difficult or impossible for large changes in viewing direction. In urban modeling for example, a video's path often crosses at intersections where the viewing direction differs by about 90°.

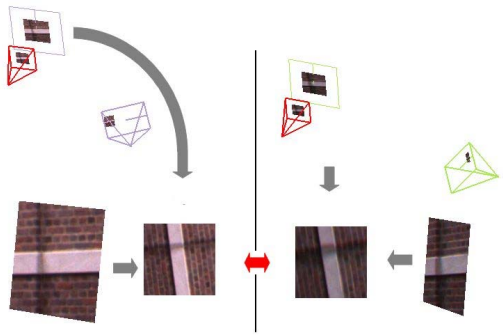

Figure 1. Two corresponding VIPs. The green and grey view frustums are original camera poses. Red view frustums are viewpoint normalized cameras. Lower left and right show patches in the original images while center patches are the ortho-textures for the feature rotationally aligned to the dominant gradient direction.

We propose the novel viewpoint invariant patch (VIP) which provides the necessary properties to determine the similarity transformation between two 3D scenes even under significant viewpoint changes. VIPs are extracted from images using their known local geometry, and the detection is performed in a rectified image space to achieve robustness to projective distortion while maintaining full knowledge of texture. This is an essential advantage over invariant mappings. For example, our method is able to distinguish between squares and rectangles which are indistinguishable using affine invariant methods. As less of the local texture variation is sacrificed to achieve invariance, more is left for discrimination.

In our method image textures are rectified with respect to the local geometry of the scene. The rectified texture can

978-1-4244-2243-2/08/$25.00 ©2008 IEEE

be seen as an ortho-texture[1] of the 3D model which is viewpoint independent. This first rectification step is essential to our new concept because rectification using the local geometry delivers robustness to changes of viewpoint. We then determine the salient feature points of the ortho-textures and extract the feature description. In this paper we use the well known SIFT-features and their associated descriptor [10] as interest points . The 3D models are then transformed to a set of VIPs, made up of the feature's 3D position, patch scale, surface normal, local gradient orientation in the patch plane, and a SIFT descriptor. The rich information in VIP features makes them particularly suited to 3D similarity transformation estimation. One VIP correspondence is sufficient to compute a full similarity transformation between two models by comparing the 3D positions of the features, their normals, orientations in the ortho-texture and patch scales. The scale and rotation components of the VIP correspondence are consistent with the relative scale and rotation between the two 3D-models. Moreover, each putative correspondence can be tested separately facilitating efficient, robust feature matching. These advantages lead to a Hierarchical Efficient Hypothesis Testing (HEHT) scheme which delivers a transformation, by which 3D textured models can be stitched automatically.

The remainder of the paper is organized as follows: Related work is discussed in Section 2. Section 3 introduces the viewpoint-invariant patch and discusses its properties. An efficient VIP detector for urban scenes is discussed in Section 4. Section 5 describes our novel hierarchical matching scheme. The novel algorithms are evaluated and compared to existing state of the art features in Section 6.

2. Related Work

Many texture based feature detectors and descriptors have been developed for robust wide-baseline matching. One of the most popular is Lowe's SIFT keypoints [10]. The SIFT detector defines a feature's scale in scale space and a feature orientation from the edge map in the image plane. Using the orientation, the SIFT detector generates normalized image patches to achieve 2D similarity transformation invariance. Many feature detectors, including affine covariant features, use the SIFT descriptor to represent patches. We also use the SIFT-descriptor to encode the VIP. However, our approach can also be applied with other feature descriptors. Affine covariant feature go beyond only achieving invariance to affine transformations. Mikolajczyk et al. give a comparison of several such features in [13]. Our proposed feature detection and description method goes beyond affine invariance to robustness to projective transformations. Critically, our features are not invariant to projective transformations but they are stable under projective transformations. Whereas affine invariant approaches can not distinguish between a square and a rectangle, our feature representation is able to distinguish between the two. Our representation has fewer intrinsic ambiguities which improves matching performance.

Recent advances in Structure from Motion (SfM) and active sensors have generated increased interest in the alignment of 3D models. In [2] Fitzgibbon and Zisserman proposed a hierarchical SfM method to align local 3D scene models from connective triplets. The technique exploits 3D correspondences from common 2D tracks in consecutive triplets to compute the similarity transformation that aligns the features. Their technique works well for the small viewpoint changes between triplets typically observed in video.

Snavely et al. proposed a framework for the registration of photo collections downloaded from the internet in [16]. Their framework also uses SIFT features to automatically extract wide baseline salient feature correspondences from photo collections. Robust matching and bundle adjustment are used to determine camera positions. The method relies on a reasonably dense set of viewpoints. Finally, the cameras and images are used to provide an image alignment of the scene. These methods are based only on texture. Goesele et al. [3] introduced a method to use the camera registration and 3D feature points from [16] to compute the scene geometry. This geometry is bootstrapped from small planar patches and then grown into a larger model. Our novel features could use the small local patches to improve the feature matching for the global registration of local camera clusters.

Other approaches are based entirely on geometry and ignore texture information. Iterative closest point (ICP) based methods can be used to compute the alignment by iteratively minimizing the sum of distances between closest points. However, ICP requires an initial approximate scene alignment and local reconstructions which are more accurate than are typically available.

Another purely geometric approach is to align 3D models with the extracted geometric entities called spin images [5]. Stamos and Leordeanu used mainly planar regions and 3D lines on them to do 3D scene alignment [17]. The approach uses a pair of matched infinite lines on the two local 3D geometries to extract the in-plane rotation of the lines on the planar patches. The translation between the models was computed by estimating it as the vector that connects the mid-points of the matching lines. In general, two pairs of matched 3D lines give a unique solution and so can be used efficiently in a RANSAC scheme.

There are also methods based on both texture and geometry. Liu et al. in [9] extended their work to align 3D points from SfM to range data. They first register several images independently to range data by matching vanishing

[1] Ortho-texture: Representation of the texture that is projected on the surface with orthogonal projection.

points. Then the registered images are used as common points between the range data and a model from SfM. In the final step a robust alignment is computed by minimizing the distance between the range data and the geometry obtained from SfM. After the alignment, photorealistic texture is mapped to 3D surface models. An extension of the approach is discussed in [8]. King et al. [6] align laser range scans with texture images by first matching SIFT keypoints extracted directly from texture images and backprojecting those keypoints onto the range measurements. A single backrojected keypoint correspondence defines the transformation between two models. A region growing variant of ICP is used to refine the model alignment while detecting outlier correspondences.

In [24], Zhao and Nistér proposed a technique to align 3D point clouds from SfM and 3D sensors. They start the method by registering two images, fixing a rough transformation, and use ICP for alignment. ICP is effective because of the precision of 3D laser range. Vanden Wyngaerd et al. proposed a method to stitch partially reconstructed 3D models. In [23], they extract and match bitangent curve pairs from images using their invariant characteristics. Aligning these curves gives an initialization for more precise methods such as ICP. In an extension of this work [21], they use the symmetric characteristics of surface patches to achieve greater matching accuracy. In [22], texture and shape information guide each other while looking for better regions to match. Additionally, Rothanger et. al. [15] proposed a matching technique which finds matches between affine invariant regions and then verifies the matches based on their normal directions.

Concurrent with this research Koeser and Koch [7] developed a very similar approach to ours. The main difference between our approaches is that they extract MSER in the original images, backproject these regions onto a depthmap and then extract normalized images using cameras with optical axis parallel to the surface normal. They too use SIFT descriptors as their final invariant patch descriptor. We find keypoints directly in textures from orthographic virtual cameras with viewing direction parallel to the surface normals.

3. Viewpoint-Invariant Patch (VIP)

In this section we describe our novel features in detail. Viewpoint-Invariant Patches (VIPs) are features that can be extracted from textured 3D models which combine images with corresponding depth maps. VIPs are invariant to 3D similarity transformations. They can be used to robustly and efficiently align 3D models of the same scene from video taken from significantly different viewpoints. In this paper we'll mostly consider 3D models obtained from video by SfM, but our method is equally applicable to textured 3D models obtained using LIDAR or other sensors. Our robustness to 3D similarities exactly corresponds to the ambiguity of 3D models obtained from images, while the ambiguities of other sensors can often be described by a 3D Euclidean transformation or with even fewer degrees of freedom.

Our undistortion is based on local scene planes or on local planar approximations of the scene. Conceptually, for every point on the surface we estimate the local tangent plane's normal and generate a texture patch by orthogonal projection onto the plane. Within the local ortho-texture patch we determine if the point corresponds to a local extremal response of the Difference-of-Gaussians (DoG) filter in scale space. If it is we determine its orientation in the tangent plane by the dominant gradient direction and extract a SIFT descriptor on the tangent plane. Using the tangent plane avoids the poor repeatability of interest point detection under projective transformations seen in popular feature detectors [13].

The next sections will give more details about the different steps of the VIP feature detection method. The first step in the feature detection is to achieve a viewpoint normalized ortho-texture for each patch.

3.1. Viewpoint Normalization

Viewpoint-normalized image patches need to be generated to describe VIPs. Viewpoint-normalization is similar to the normalization of image patches according to scale and orientation performed in SIFT and normalization according to ellipsoid in affine covariant feature detectors. The viewpoint normalization can be divided into the following steps:

1. **Warp the image texture** onto the local tangential plane. Non-planar regions are warped to a local planar approximation to the surface which causes little distortion over small surface patches.

2. **Project the texture** into an orthographic camera with viewing direction parallel to the local tangential plane's normal.

3. **Extract the VIP descriptor** from the orthographic patch projection. Invariance to scale is achieved by normalizing the patch according to local ortho-texture scale. Like [10] a DoG filter and local extrema suppression is used. VIP orientation is found based on the dominant gradient direction in the ortho-texture patch.

Figure 1 demonstrates the effect of viewpoint normalization. The 2nd and 3rd column in the figure are the normalized image patches. The normalized image patches of a matched pair are very similar despite significantly different original images due to the largely different viewing directions.

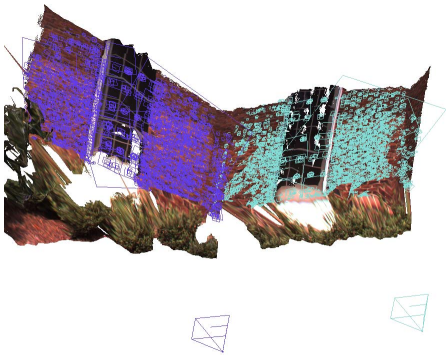

Figure 2. VIPs detected on the 3D model, the cameras corresponding to the textures are shown at the bottom

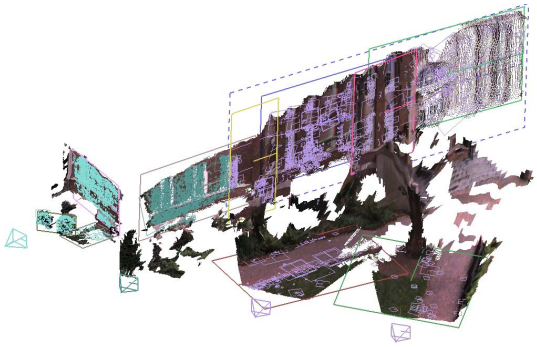

Figure 3. VIPs detected on dominant planes.

3.2. VIP Generation

With the virtual camera, the size and orientation of a VIP can be obtained by transforming the the scale and orientation of its corresponding image feature to world coordinates. A VIP is then fully defined as ($\mathtt{x}$, σ, $\mathtt{n}$, $\mathtt{d}$, $\mathtt{s}$) where

- $\mathtt{x}$ is its 3D position,
- σ is the patch size,
- $\mathtt{n}$ is the surface normal at this location,
- $\mathtt{d}$ is texture's dominant orientation as a vector in 3D, and
- $\mathtt{s}$ is the SIFT descriptor that describes the viewpoint-normalized patch. Note, a sift feature is a sift descriptor plus it's position, scale and orientation.

The above steps extract the VIP features from images and known local 3D geometry[2]. Using VIPs extracted from two models we can then find areas where the models represent the same surface.

3.3. VIP Matching

Putative VIP matches can be obtained with a standard nearest neighbor matching of the descriptors or other more scalable methods. After obtaining all the putative matches between two 3D scenes, robust estimation methods can be used to select an optimized scene transformation using the 3D hypotheses from each VIP correspondences. Since VIPs are viewpoint invariant, given a correct camera matrix and 3D structure, we can expect the similarity between correct matches to be more accurate than a transformation derived from viewpoint dependent matching techniques.

The richness of the VIP feature allows computation of the 3D similarity transformation between two scenes from a single match. The ratio of the scales of two VIPs expresses the relative scale between the 3D scenes. Relative rotation is obtained using the normal and orientation of the VIP pair. The translation between the scenes is obtained by examining the rotation and scale compensated feature locations. The scale and rotation needed to bring corresponding VIP features into alignment is constant for a complete 3D model. We will use this property later to set up an Hierarchical Efficient Hypothesis Testing (HEHT) scheme to determine the 3D similarity between models.

[2]Local geometry denotes the geometry that is recovered from the images by using SfM and multi-view stereo methods for example. Please note that the local geometry is usually given in the coordinate system of the first camera of the sequence with an arbitrary scale w.r.t. the real world motion.

4. Efficient VIP Detection

In general planar patch detection needs to be executed for every pixel of the image to make the ortho-textures. Each pixel (x, y) together with the camera center C defines a ray, which is intersected with the local 3D scene geometry. The point of intersection is the corresponding 3D point of the feature. From this point and its spatial neighbors we then compute the tangential plane Π_t at the point, which for planar regions coincides with the local plane. For structures that only slightly deviate from a plane we retrieve a planar approximation for local geometry of the patch. Then the extracted plane can be used to compute the VIP feature description with respect to this plane. This method is generally valid for any scene.

VIP detection for a set of points that have the same normal can be efficiently done in a single pass. Considering these VIPs, the image coordinate transformations between them are simply 2D similarity transformations. This means that the VIP detection for points with the same normal can be done in one pass on a larger planar patch, on which all the points are projected, and the original VIP can be recovered by applying a known similarity transformation.

Figure 3 illustrates a result of detecting VIPs on dominant planes. The planes here compensate for the noise in the reconstructed model, and improve VIP localization. Figure 4 shows an example of a viewpoint normalized facade.

Figure 4. Original image (left) and its normalized patch (right)

5. Hierarchical Estimation of 3D Similarity Transformation

A hierarchical method is proposed in this section to estimate the 3D similarity transformation between two 3D models from their putative VIP matches. Each single VIP correspondence gives a unique 3D similarity transformation, and so hypothesized matches can be tested efficiently. Furthermore, the rotation and scaling components of the similarity transformation are the same in all inlier VIP matches, and they can be tested separately and efficiently with a voting consensus.

5.1. 3D Similarity Transformation from a Single VIP Correspondence

Given a VIP correspondence of $(\mathtt{x}_1,\ \sigma_1,\ \mathtt{n}_1,\ \mathtt{d}_1,\ \mathtt{s}_1)$ and $(\mathtt{x}_2,\ \sigma_2,\ \mathtt{n}_2,\ \mathtt{d}_2,\ \mathtt{s}_2)$, the scaling between them is given by

$$\sigma_s = \frac{\sigma_1}{\sigma_2} \tag{1}$$

The rotation between them satisfies

$$(\mathtt{n}_1, \mathtt{d}_1, \mathtt{d}_1 \times \mathtt{n}_1)R_s = (\mathtt{n}_2, \mathtt{d}_2, \mathtt{d}_2 \times \mathtt{n}_2). \tag{2}$$

The translation between them is

$$T_s = \mathtt{x}_1 - \sigma_s R_s \mathtt{x}_2 \tag{3}$$

A 3D similarity transformation can be formed from the three components as $(\sigma_s R_s, T_s)$.

5.2. Hierarchical Efficient Hypothesis-Test (HEHT) Method

The scale, rotation and translation of a VIP is covariant with the global 3D similarity transformation, and the local feature scale change and rotation are the same as the global scaling and rotation. Solving these components separately and hierarchically increases accuracy and dramatically reduces the search space for the correct similarity transformation.

The 3D similarity estimation in this paper is done hierarchically in three steps starting from a set of putative VIP correspondences. First, each VIP correspondence is scored by the number of other VIP correspondences that support it's scaling. All VIP correspondences which are inliers to the VIP correspondence with most support are used to calculate a mean scaling and outliers are removed from the putative set. Second, the same process is repeated with scoring based on support for each correspondence's rotation and the putative set is again pruned of outliers. Third, the same process is repeated scoring according to translation to determine the final set of inlier VIP correspondences. A nonlinear optimization is run to find the scaling, rotation, and translation using all of the remaining inliers.

5.3. Using RANSAC with VIP Features

It is worth note that in our experiments all possible hypotheses are exhaustively tested, which is very efficient because each VIP correspondence generates one hypothesis and the whole sample space is linear with the number of putative VIP matches. The method described above can be easily extended to a RANSAC scheme by checking only a small set of hypotheses. It is know that the RANSAC requires $N = \frac{\log(1-p)}{\log(1-(1-e)^s)}$ random samples to get one inlier sample free of outliers, where e is ratio of outliers, p is the expected probability, and s is number of matches to establish a hypothesis [4]. In our case, $s = 1$, so that $N = \log_e(1-p)$. For example, when outlier ratio is 90%, 44 random samples are enough to get at least one inlier match with probability 99%. This leads to an even more efficient estimation of 3D similarity transformations. However, in the cases where there are many outliers, an exhaustive test of all transformation hypotheses is the most reliable and still very efficient.

6. Experimental Results and Evaluation

This section compares viewpoint invariant patches to other corner detectors in terms of the number of correct correspondences found and the feature re-detection rate. In addition we apply the VIP-based 3D alignment to several reconstructed models to demonstrate reliable surface alignment and perform SfM of a large scene completing a loop around a large building using VIPs.

6.1. Evaluation

To measure the performance of the VIP feature we performed an evaluation similar to the method of [13]. Our test data [1] is a sequence of images of a brick wall taken with increasing angles between the optical axis and the wall's normal. Each of the images of the wall has a known homography to the first image, which was taken with image plane fronto-parallel to the wall. Using this homography we extract a region of overlap between the first image and each other image. We extract features in this area of overlap and use two measures of performance, the number of inlier correspondences and the re-detection rate, to evaluate a number of feature detectors. The number of inliers is

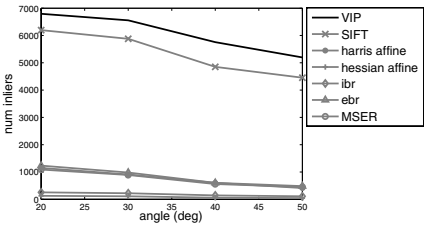

Figure 5. Number of inliers under a projective transformation

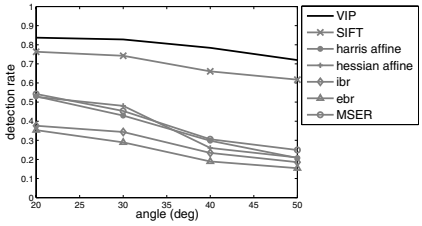

Figure 6. Re-detection rate of features under a projective transformation

the number of feature correspondences which fit the known homography. Re-detection rate is the ratio of inlier correspondences found in these overlapping regions to the number features found in the fronto-parallel view. The number of inliers is shown in Figure 5 and the re-detection rate is shown in Figure 6.

Figure 5 shows that the VIP generates a significantly larger number of inliers over a wide range of angles than the other detectors. The other detectors we compare to are SIFT [10], Harris-Affine [12], Hessian-Affine [12],Intensity Based Region (IBR) [19], Edge Based Region (EBR) [19], and Maximally Stable Extremal Region (MSER) [18]. Our novel VIP feature also has a significantly higher re-detection rate than the other detectors as seen in Figure 6. This high re-detection rate is a result of the detection of features on the ortho-textures. Even under large viewpoint changes which often result in a projective transformation between images the VIP performs well.

6.2. Experiments

For our experimental evaluation of the novel detector we used several image sequences of urban scenes. For each image sequence we used SfM to compute its depths map and camera positions. We used two image sequences of each scene with different viewpoints and camera paths. Camera positions were defined relative to the pose of the first camera in each sequence. The first scene, shown in Fig. 7, consists of two facades of a building reconstructed from two different sets of cameras with significantly different viewing directions (about $45°$). The cameras moved along a path around the building. One can observe reconstruction errors due to trees in front of the building. An offset was added to the second scene model for visualization of the matching VIPs. The red lines connect all of the inlier correspondences. Rotation and scaling have been corrected using transformations calculated using VIPs in this visualization. The HEHT determined 214 inliers out of 2085 putative matches. The number of putative matches is high because putative matches are generated between all features in each of the models.

The second evaluation scene shown in Figure 8 consists of two local scene models, with camera paths that intersect at an angle of 45 degrees. The overlapping region is a small part of the combined models, and it is seen from very different viewpoints in the two videos. Experiments show that our 3D model alignment method can reliably detect the small common surface and align the two models. Videos in the supplemental materials illustrate the details of our algorithm.

We match models reconstructed from camera paths which cross at a $90°$ angle in Figure 9. Note the large difference in viewing directions between the cameras on the left and right in the image. This shows that the VIP can match features reconstructed from widely different viewpoints.

Table 1 shows quantitative results of the HEHT. Note that scale and rotation verification remove a significant portion of the outliers. For evaluation we first measure the distances of the matched points after the first 3 stages and after the nonlinear refinement. To measure the quality of surface alignment, we check the point distances between the overlapping parts of the models. The models are reconstructed with scale matching the real building and so the error is given in meters. The statistics in table 1 demonstrate the performance of our matching.

Scenes	#1 (Fig 7)	#2 (Fig 8)	#3 (Fig 9)
Viewing direction change	$45°$	$45°$	$90°$
Putative #	2085	494	236
Inlier #	1224/654/214	141/42/38	133/108/101
Error after first 3 stages	0.0288	0.0230	0.114
Error after nonlinear ref.	0.0128	0.018	0.0499
Surface alignment error	0.434	0.135	0.629

Table 1. HEHT running details in 3 experiments. The second row gives the approximate viewing direction change between two image sequences. The third row the number of putative matches, and the fourth row shows the number of inliers in each stage of the 3-stage HEHT. The errors in the following 3 rows are median errors in meter. The large difference between the surface alignment error and feature matching show the large noise in stereo reconstruction.

Additionally we compared the VIP-based alignment with SIFT feature and MSER. For SIFT and MSER, the 2D feature locations are projected to the 3D model surface to

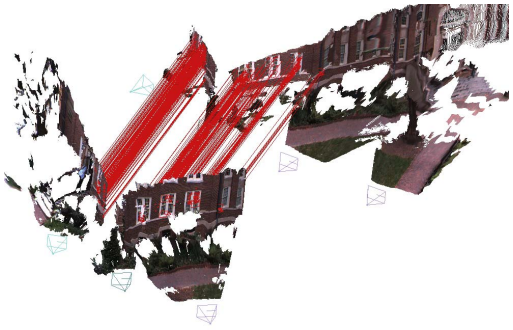

Figure 7. Scene 1: Matched 3D models with 45° viewing direction chage.

Figure 8. Scene 2: 3D models matched with very small overlap.

Figure 9. Scene 3: Matched 3D models from camera paths crossing at 90°.

get 3D points. The putative match generation for them is the same as the VIP matching since they all use SIFT descriptors. Then a Least-Square method [20] and RANSAC is used to evaluate the 3D similarity transformation between the point matches. Table 6.2 shows the comparison between SIFT, MSER and VIP. The results show that VIP can handle the large viewpoint changes for which SIFT and MSER do not work.

The advantages of the VIP for wide baseline matching are perhaps best demonstrated by a large scale reconstruction. We collected video of a building with footprint approximately 37 by 16 meters where the camera's path completes the loop by crossing at an angle of approximately ninety degrees. Matching correspondences across this wide angle using previous methods is difficult or impossible. However, using VIP patches we were able to complete the loop, generating an accurate 3D point model of the building.

Scene 1	SIFT	MSER	VIP
#Feature(M1/M2)	8717/12244	2254/3410	5947/5553
#Putative Matches	1600	420	2085
#Inlier Matches	176	22	214
Successful	Y	Y	Y
Scene 2	SIFT	MSER	VIP
#Feature (M1/M2)	12951/16664	4071/5024	9015/4828
#Putative	641	67	278
#Inlier	11	0	203
Successful	N	N	Y
Scene 3	SIFT	MSER	VIP
#Feature (M1/M2)	2363/733	1128/342	2713/1804
#Putative	131	0	90
#Inlier	12	0	61
Successful	N	N	Y

Table 2. Comparison with SIFT and MSER in the 3 scenes. SIFT and MSER work in the first scene but fail in the other two. It can be seen that VIP also gives the highest rate of inlier number to feature number in the first one. It is worth noting that VIP work in scene where there is a 90 degree viewing direction change.

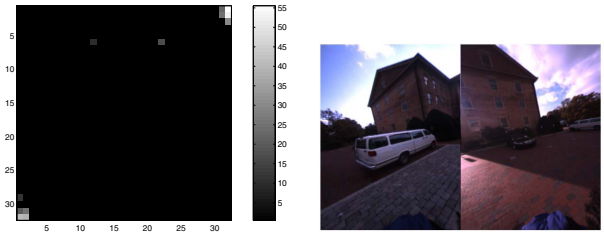

Figure 10. (left) VIP matches found in camera path circling building. VIPs are only extracted between frames where KLT features could not be tracked in between. Note the matching features at either end of the sequence where the loop completes. (right) First and last frame in video circling building.

Our reconstruction is done in three steps. First we estimated the camera path using SfM with KLT [11] feature measurements, bundle adjusting the result. Figure 11 shows the scene points and camera path before applying the VIP correspondences. We then extracted VIP correspondences between key frames in the initial reconstruction. The number of matches found between key frames is shown in Figure 10. Matching these correspondences allowed us to measure the accumulated error over the reconstruction. Using VIP correspondences we compensated for this error by linearly distributing it through the camera poses in the sequence and the 3D feature estimates. The VIP features were added to the set of 3D features and measurements from the first bundle adjustment and we ran the bundle adjustment again. The final result of Structure from Motion using VIPs with the completed loop is shown in Figure 11. In Figure 11 the region of double representation is circled and the angle between the first and last camera poses is shown.

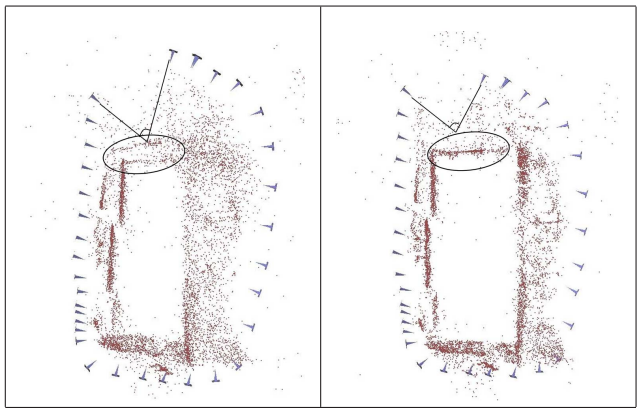

Figure 11. (left)Camera path and sparse points before loop completion with VIPs. (right) Loop around building completed with VIP correspondences. Note the angle between the first and last cameras.

7. Summary and Conclusions

In this paper we developed a novel method for the alignment of 3D scenes. Our alignment is based on the viewpoint invariant patches(VIP), a novel feature descriptor. The VIP allows scene alignment from only a single correspondence. To match VIPs we introduced a hierarchical efficient hypothesis test which exploits the fact that the different parts of the similarity transformation can be evaluated independently. Through this method we were able to overcome the problems posed by the large amount of uncertainty in the translational alignment of any standard matching method. We evaluated the proposed matching against other feature detectors and used the novel VIPs to align models of a variety of scenes. Our evaluation demonstrates that VIP features are an improvement on current methods for robust and accurate 3D model alignment.

8. Acknowledgements

This research was supported in part by IARPA under the VACE program, NSF Career Award No. 237533 and the David and Lucille Packard Foundation. Thanks to Christopher Zach for his help with bundle adjustment.

References

[1] Wall dataset. http://www.robots.ox.ac.uk/ vgg/research/affine/.

[2] A. W. Fitzgibbon and A. Zisserman. Automatic camera recovery for closed or open image sequences. In *ECCV*, pages 311–326, June 1998.

[3] M. Goesele, N. Snavely, B. Curless, H. Hoppe, and S. M. Seitz. Multi-view stereo for community photo collections. In *ICCV*, 2007.

[4] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[5] A. Johnson, M., and Hebert. Using spin images for efficient object recognition in cluttered 3Dscenes. *TPAMI*, 21(5):433–449, May 1999.

[6] B. J. King, T. Malisiewicz, C. V. Stewart, and R. J. Radke. Registration of multiple range scans as a location recognition problem: Hypothesis generation, refinement and verification. In *3DIM*, 2005.

[7] K. Koeser and R. Koch. Perspectively invariant normal features. In *in Proc. of Workshop on 3D Representation for Recognition*, 2007.

[8] L. Liu and I. Stamos. A systematic approach for 2d-image to 3d-range registration in urban environments. In *In Proc. of workshop for Virtual Representations and Modeling of Large-scale environments*, 2007.

[9] L. Liu, I. Stamos, G. Yu, G. Wolberg, and S. Zokai. Multiview Geometry for Texture Mapping 2D Images Onto 3D Range Data. *CVPR*, 2:2293–2300, 2006.

[10] D. Lowe. Distinctive image features from scale-invariant keypoints. In *IJCV*, volume 20, pages 91–110, 2004.

[11] B. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Int. Joint Conf. on Artificial Intelligence*, pages 674–679, 1981.

[12] K. Mikolajczyk and C. Schmid. Scale and affine invariant interest point detectors. *IJCV*, 60:63–86, 2004.

[13] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *IJCV*, 65:43–72, 2005.

[14] M. Pollefeys, D. Nister, ..., and H. Towles. Detailed real-time urban 3d reconstruction from video. *IJCV*, 2008.

[15] F. Rothganger, S. Lazebnik, C. Schmid, and J. Ponce. 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *IJCV*, 66(3):231–259, 2006.

[16] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3d. In *SIGGRAPH '06: ACM SIGGRAPH 2006 Papers*, pages 835–846, 2006.

[17] I. Stamos and M. Leordeanu. Automated Feature-Based Registration of Urban Scenes of Large Scale. *CVPR*, 2003.

[18] A. Z. T. Kadir and M. Brady. Robust wide baseline stereo from maximally stable extremal regions. In *BMVC*, pages 384–393, 2002.

[19] T. Tuytelaars and L. Van Gool. Matching widely separated views based on affine invariant regions. *IJCV*, 2004.

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

[21] J. Vanden Wyngaerd, , and L. Van Gool. Automatic crude patch registration: Toward automatic 3d model building. *CVIU*, 87:8–26, 2002.

[22] J. Vanden Wyngaerd and L. Van Gool. Combining texture and shape for automatic crude patch registration. In *3DIM*, pages 179–186, 2003.

[23] J. Vanden Wyngaerd, L. Van Gool, R. Koch, and M. Proesmans. Invariant-based registration of surface patches. In *ICCV*, pages 301–306, 1999.

[24] W.-Y. Zhao, D. Nistér, and S. C. Hsu. Alignment of continuous video onto 3d point clouds. *TPAMI*, 27(8), 2005.