International Journal of Computer Vision manuscript No.
(will be inserted by the editor)

Physical Scale Keypoints: Matching and Registration for Combined Intensity/Range Images

Eric R. Smith ·
Richard J. Radke ·
Charles V. Stewart

the date of receipt and acceptance should be inserted later

Abstract We present a new framework for detecting, describing, and matching keypoints in combined range-intensity data, resulting in what we call physical scale keypoints. We first produce an image mesh by backprojecting associated 2D intensity images onto the 3D range data. We detect and describe keypoints on the image mesh using an analogue of the SIFT algorithm for images with two key modifications: the process is made insensitive to viewpoint and structural discontinuities using a novel bilinear filter, and a physical scale space is constructed that exploits the reliable range measurements. Keypoints are matched between scans only when their physical scales agree, avoiding many potential false matches. Finally, the matches are rank-ordered using a new quality measure and supplied to a registration algorithm that refines each match into a rigid transformation for the entire scan pair. We report experimental results on keypoint detection and matching and range scan registration and verification in a set of difficult real-world scan pairs, showing that the new physical scale keypoints are demonstrably better than a competing approach based on backprojected SIFT keypoints.

This work was supported in part by the DARPA Computer Science Study Group under the award HR0011-07-1-0016.

110 Eighth Street, Troy, NY USA 12180
smithe4@rpi.edu
rjradke@ecse.rpi.edu
stewart@cs.rpi.edu

1 Introduction

Identifying and matching 3D keypoints are integral steps in many algorithms for range registration and 3D object detection. Regardless of the application, the quality of a keypoint algorithm depends on both the repeatability and the distinctiveness of the keypoints it produces [12]. This means that the description of a keypoint location must be as insensitive as possible to viewpoint, scaling, sampling, and intensity differences, while still being easily distinguished from the descriptions of other keypoints. In this paper, we combine information from calibrated cameras with range scan data to improve the distinctiveness of matching intensity image keypoints, working exclusively in the three dimensional space formed by back-projecting the images into the range data. We call these new keypoints **physical scale keypoints**.

The example shown in Figure 1 illustrates the effectiveness of our approach. The top row shows an automatically detected and matched pair of physical scale keypoints that straddle a depth discontinuity. Using our algorithm, the keypoint formation and matching processes used only information from the foreground surface. In the corresponding two-dimensional images, the intensities from the foreground and background surfaces are in the same neighborhood in one image due to the angle of viewpoint (bottom left); however, this background is occluded in the other image. This makes any keypoint detected in this neighborhood difficult to match, and indeed the best match (bottom right), produced by an existing algorithm that detects SIFT keypoints in the images [12] and backprojects them to 3D prior to matching [19], is incorrect.

Our approach is designed to work with single viewpoint range data acquired from large-scale outdoor scenes. There are typically many complex objects in such scenes, including plants, trees, cars, buildings, street lights and signs, and the data from each object are incomplete and viewpoint-dependent. This means that depth discontinuities are prevalent throughout the data, at large scales from building boundaries, at intermediate scales from edges of cars and tree trunks, and at small scales from window frames. This situation differs from the assumptions of many other related approaches to three-dimensional matching, in which fewer objects are in the data and/or more complete views are available [22,24,7]. Thus, the ability to work with discontinuities plays a central role in the design of our algorithm. On the other hand, an advantage to working with LiDAR range data as opposed to multiview reconstructions, is that there is no scale ambiguity. This means that we can attach reliable physical distances to image measurements. In particular, this implies that we can create keypoints for a surface of exactly the same scale regardless of our scanner's position relative to that surface.

Our approach starts by backprojecting the image(s) onto a mesh formed from the range data. We develop algorithms for keypoint detection and description on this mesh, paralleling several steps of the original SIFT algorithm [12], but exploiting the 3D geometry to make the algorithms insensitive to viewpoint and discontinuities. This invariance is more substantial than the affine invariance of keypoint algorithms working in 2D [14] because the working neighborhood over which a keypoint is detected and described is in itself invariant to viewpoint so long as it does not become obscured. Next, a "physical" version of scale space is formed by choosing a variety of smoothing scales, measured in centimeters, at which to work. Matching occurs by computing keypoint descriptor distances, but only between keypoints detected at the same physical scale. These advantages — the 3D invariance, the ability to form reliable keypoints near discontinuities, and the physical scale keypoint formation and matching — are the primary contribution of our work to the keypoint algorithm literature.

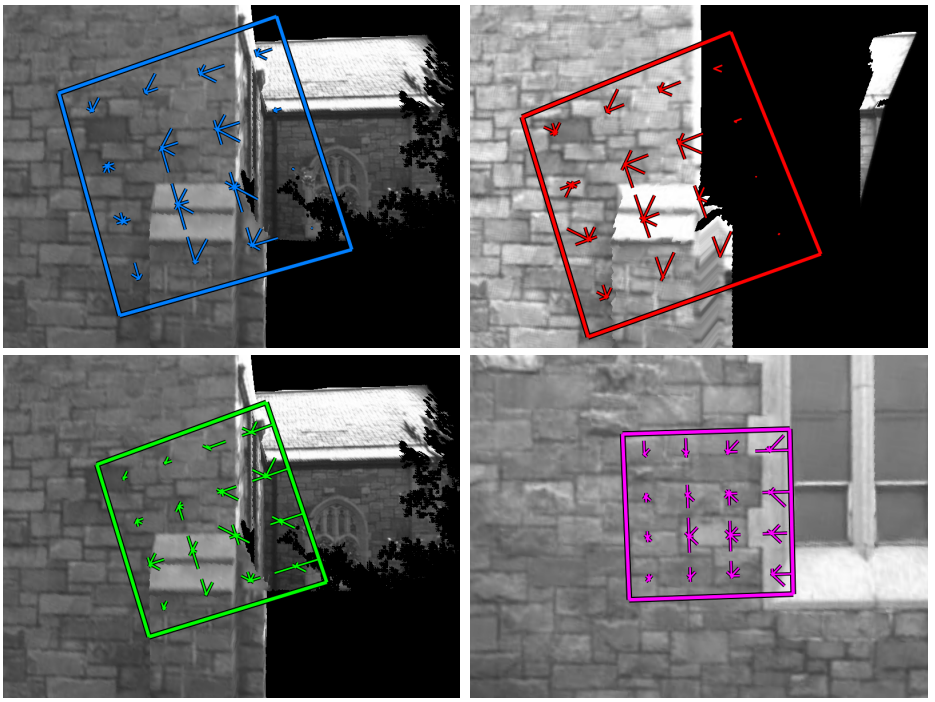

Fig. 1 The top row shows a correct match between two automatically-detected physical scale keypoints from two different scans computed near a depth discontinuity. The bottom row shows a back-projected SIFT keypoint (green) computed at the same location, and its (incorrect) match in purple.

We also develop a novel match ranking system that combines a familiar descriptor based ratio metric [12] with a metric that measures the confidence in the matching of the actual underlying range data. We also show the integration of this keypoint technique into a previously-published range data registration algorithm [19]. This algorithm generates initial rigid transformation estimates from rank-ordered keypoint matches, it refines these using both range and intensity constraints in a region-growing iterative closest point (ICP) formulation, and then applies decision criteria to select the correct registration (or reject them all). We show how to use our new physical-scale keypoints in both the initialization and decision steps of this algorithm. Our new match ranking system, integration of physical-scale keypoints with the aforementioned registration algorithm, new experiments, and new comparisons are the main contributions over our earlier paper [20].

The remainder of this paper is organized as follows. Section 2 summarizes related work. Section 3 describes the physical-scale keypoint detection process. Section 4 presents the keypoint descriptor computation and matching algorithms. Section 5 shows the integration of physical-scale keypoints into the range registration process. Section 6 presents experimental results on several large-scale datasets. Section 7 offers concluding remarks.

2 Related Work

The focus of our related-work discussion is on methods for detection and matching of keypoints and other shape descriptors in range data, sometimes with associated intensity information.

Shape descriptors that collect points into a histogram have been used extensively. Johnson and Hebert's spin images [9] collect points into a 2D histogram that is rotated about the surface normal at a central point. Mian et al. [13] extended spin images to a tensor representation vector and developed a method for matching them. Frome et al. [7] extended 2D shape contexts to 3D, using an oriented basis point to construct a 3D histogram. 3D shape context descriptors have an extra degree of freedom in rotation about their normal, which is accounted for by creating additional descriptors at sampled rotations. Recently, Zhong [25] developed Intrinsic Shape Signatures, which histograms points in a spherical coordinate system around a basis point. The coordinate frame is based on the eigenvectors of the scatter matrix of the neighborhood. Geometry-only keypoints have been quite successful in many applications; however, it can be difficult to establish a highly repeatable coordinate frame for matching. This is often addressed by using additional keypoint matches to constrain the surface matching. Thus surface matching is difficult on noisy or smooth/near-planar surfaces. We can mitigate this problem by using the coincident intensity images produced by our range scanner, thus we seek techniques to exploit this information.

For 2D images, scale space was studied thoroughly by Lindeberg [11] and used by Lowe [12] in the development of the well-known SIFT scale invariant keypoints. These were extended to affine invariance by Mikolajczyk et al. [16, 15]. SIFT keypoints have been combined with 3D data in several ways. In particular, our previous work in range registration [19] used 3D geometry to provide a measure of affine invariance in the keypoint descriptors by backprojecting image gradient vectors onto planes in the range data. We also used the 3D information to filter keypoints near discontinuities and to eliminate matches with widely differing scales. Wu et al. [22] extended SIFT using a framework called viewpoint-invariant patches (VIP) by mapping the image intensities onto a plane fit in 3D, and then detecting and describing the keypoint in this plane. Pose is obtained using a one-point RANSAC algorithm.

Assuming full, uniformly sampled meshes, Zaharescu et al. [24] extended Wu's viewpoint invariant patches to use full 3D gradients and descriptors. They built a scale space of intensities on a mesh using repeated convolutions of a fixed-width 3D Gaussian. Extrema are detected at one-ring neighborhoods of the difference-of-Gaussian meshes. Gradients are computed using weighted directional derivatives around a vertex, which are binned on the three axes of the keypoint to form a descriptor. While our approach to detection and description builds on concepts from this work, our focus is on single viewpoint range data and the challenges it introduces.

A strictly geometric scale space was developed by Novatnack and Nishino [17] in which a mesh is embedded into a 2D normal map and smoothed in this space using geodesic Gaussians. Interest points are detected based in this normal map scale space. Akagunduz and Ulusoy used a scale space of mean and Gaussian curvatures to detect interest points on parameterized 3D surfaces [1]. Other approaches to scale space based on geodesics have been proposed. For example, Hua et al. [8] proposed a conformal mapping of a mesh, detected interest points on this mesh using geodesic diffusion, and used 2D SIFT descriptors to characterize the interest points. Zou et al. [26] developed an intrinsic geometric scale space using Ricci flow on 3D surfaces. Starck and Hilton proposed a geodesic intensity histogram for

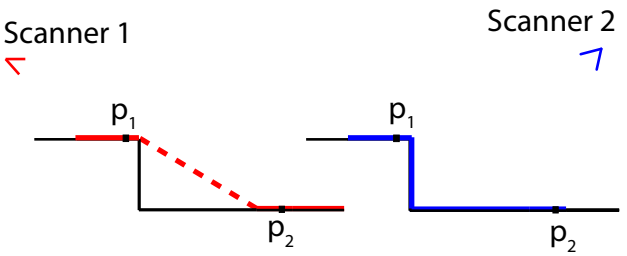

Fig. 2 The red lines indicate the mesh formed by scanner 1, and the blue by scanner 2. Since scanner 1 did not see the corner, it believes the shortest path between p_1 and p_2 is over the dotted line, which disagrees with scanner 2's view. Because of this property we do not use geodesics as a distance measure.

manifold matching [21]. A primary limitation of using geodesics is the difficult of defining them properly in the presence of strong discontinuities, as illustrated in Figure 2.

Our decision to use geometry-augmented photometric features instead of strictly geometric features is largely due to the nature of the data. Many datasets of outdoor scenes, especially involving buildings, lack distinguishable geometry except on very large scales. King et al [10] presented results on an earlier version of our dataset and showed that use of spin image descriptors allowed matching on only 6 of 11 datasets whereas intensity-based methods matched 10 out of 11. The spin image failures were due to extended flat regions, discontinuities, and low overlaps. Flat regions are indistinguishable using spin images, while discontinuities and low over both produce substantialy difference in the data sets used to form the spin images. Because of these results on an overlapping data set, our focus here is on combining intensity and range data.

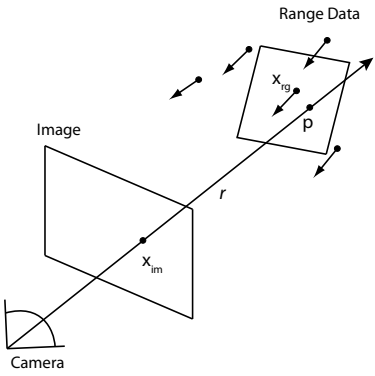

Fig. 3 This figure demonstrates the backprojection of image pixels into the range space. The points **p** form the image mesh.

3 Keypoint Detection

We assume that a range scanner with an associated, calibrated camera acquires the datasets, producing point measurements in 3D and grey-scale intensity image taken essentially at the same time (seconds or minutes apart). Calibration parameters are available to map range measurements to image coordinates and to backproject image points into 3D. We also assume that unit normal vectors are available for each range data point. These are computed using a prioritized iteratively-reweighted least-squares (IRLS) plane fitting algorithm, but other algorithms may be used.

Intensities are backprojected onto 3D and formed into a mesh using the following technique, illustrated in Figure 3. Let $\mathbf{x}_{im}$ be a pixel location mapped using the intrinsic calibration parameters into the coordinate system of the range data, and let $\mathbf{r}$ be the line from the camera center through $\mathbf{x}_{im}$. Next, we find the closest (within a threshold) range data point to $\mathbf{r}$; let this point be $\mathbf{x}_{rg}$. We compute the 3D location $\mathbf{p}$ corresponding to $\mathbf{x}_{im}$ as the intersection of the ray r with the plane formed by $\mathbf{x}_{rg}$ and the computed normal. Location $\mathbf{p}$ inherits its normal from range point $\mathbf{x}_{rg}$, and its intensity value from the original image pixel. The set of data formed by backprojecting each image pixel in this way is triangulated to form a mesh, which we refer to as the *image mesh*. We also compute a 2D Delaunay triangulation on the image mesh, which allows triangles to be created across small holes that arise when a pixel has no range point sufficiently near $\mathbf{r}$. (Such pixels can arise from glass windows or specular surfaces, for example.) The image mesh may be thought of as a piecewise 2D manifold embedded in 3D. Our detection algorithm extends 2D SIFT operators to work on this image mesh. The first problem we address is how to develop a physical scale space.

3.1 Physical Scale Space

Since the image mesh is embedded in 3D and formed from the physically-meaningful measurements of the range scanner, there is a physical meaning to the distances between adjacent pixels. We use this as the basis for forming our "physical scale space", choosing beforehand a sequence of scales, $S = (s_1, \ldots, s_k)$, measured in centimeters, at which to detect keypoints for any image mesh. In order to build a physical scale space, we must define (1) a smoothing

kernel to be applied to the image mesh, (2) a discrete convolution method for applying this kernel to the mesh, and (3) a downsampling method to create the image meshes corresponding to the different physical scales.

The smoothing kernel that we use is similar to a mesh bilateral filter [6]. We choose this filter because it does not use geodesic distances (as our data contains discontinuities), yet at the same time allows for more repeatability than a simple 3D Gaussian by reducing contribution between neighboring points taken from different surfaces. This in turn increases our algorithm's insensitivity to viewpoint. If the smoothed intensity is being computed at current mesh location $\mathbf{p}_0$ with normal η, then the bilateral weight applied to the intensity of nearby image mesh point $\mathbf{p}$ with normal $\mathbf{n}$ is

$$B(\mathbf{p},\mathbf{n};\mathbf{p}_0,\eta,\sigma,\nu) = \exp(-\|\mathbf{p}-\mathbf{p}_0\|^2/2\sigma^2)\ \exp(-(1-\mathbf{n}\cdot\eta)^2/2\nu^2). \quad (1)$$

The first factor is the standard Gaussian kernel using the 3D Euclidean distance between two points. The second factor the lowers the contribution of points whose normal $\mathbf{n}$ differs significantly from η. We used a value of 0.4 for ν throughout the paper, giving near-zero weight to points whose normals differ from η by 90° or more.

Figure 4 illustrates the bilateral weighting near the corner of a building. Notice the drop in weight both far from the point and across the normal orientation discontinuity. The latter is important because changes in viewpoint can convert this structural corner into a depth discontinuity; if these points had significant influence on the formation or the description of the keypoint, the keypoint would no longer be repeatable in such viewpoints. Moreover, this bilateral weighting also increases robustness to changes in light source position, as it will mitigate contributions from the adjacent surface, which might be illuminated or even shadowed differently from the surface on which the keypoint lies. Our surface normal comparison approach differs from the analogous "intensity" difference term used in [6], which used a point-to-plane distance, as illustrated in Figure 5. Finally, we note our filter will fail to produce repeatible results on very porous surfaces (e.g. treetops) where normal calculation is difficult. However, such locations are poor choices for keypoints in the first place as they can look completely different from a relatively small viewpoint difference.

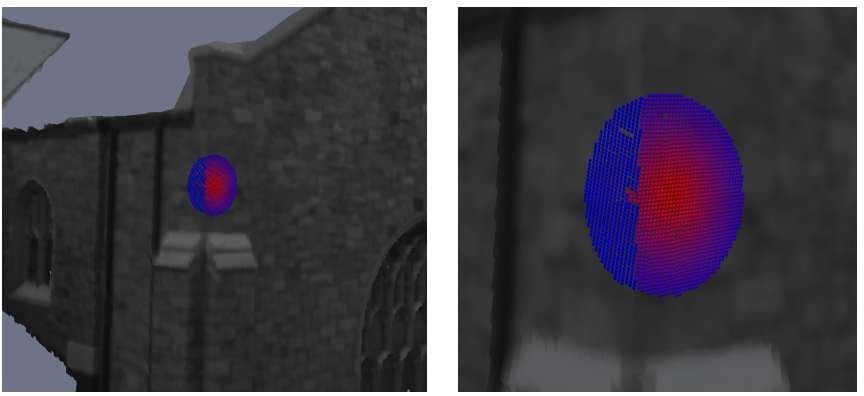

Fig. 4 Points are colored by their bilateral filter weight from red being the highest to blue being the lowest. A clear drop-off in the weighting can be seen at the corner of the building. The right image is a zoomed-in view of the left image.

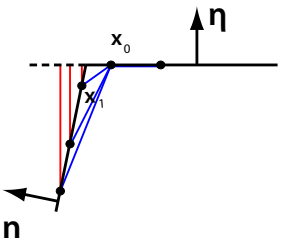

Fig. 5 The blue lines represent the distances used for the first (Gaussian) term in the bilateral filter. Fleishman et al. [6] used the red distances for their second term. Using these distances, $\mathbf{x}_1$, which may not exist in a different scan will make a sizable contribution to $\mathbf{x}_0$. Our technique mitigates this problem using the dot products of the normals $\mathbf{n}$ and η.

In the smoothing process, the foregoing bilateral filter weighting is applied to both the intensities and the normals on the image mesh. Bilateral filter weights are computed for points $\mathbf{x}$ within 2σ of the mesh point $\mathbf{x}_0$. The weights are normalized to unit sum, while the weighted average normal is converted to a unit vector.

Given this background on smoothing, we are now ready to discuss the formation of the physical scale space. Importantly, the formation of our physical scale space does not use the relatively-dense sampling of scales needed in SIFT. Such sampling is needed in 2D SIFT to localize keypoints in scale and to facilitate comparison across different scales. With known physical scales, we can simply select our scales in advance, form the image mesh at each scale, detect keypoints at each scale, and match between keypoints within each scale between two (or more) different image sets. However, before we begin computing the physical scale space we calculate the minimum scale appropriate for the image mesh from our set of scales. This is done by selecting the scale that is closest to the median edge length of the image mesh, let this scale be s_{min}.

We compute the first smoothed image mesh from the original image mesh and bilateral filter using $\sigma = s_{min}$ in Equation 1. In order to compute subsequent layers of scale space, we downsample the points used for smoothing. We do this by labeling a subset of the points (starting with all) of the image mesh as control points, and using only control points as support for smoothing in (1). After each layer has been smoothed, we remove points from the control point subset such that no two points within the subset are closer than $s_i/2$ to each other. At each layer $i > min$, the smoothing scale applied to the previous layer (image mesh) is $\sigma = \sqrt{s_i^2 - s_{i-1}^2}$. Figure 6 demonstrates four layers of the physical scale space.

3.2 Extrema Detection

Since we treat each scale independently, we compute the Laplacian of the image intensities on the mesh and then find maxima and minima. We compute the Laplacian using the Laplace-Beltrami Operator (LBO), a generalization of the Laplacian to functions on manifolds. In our case, the function is the image intensity and the manifold is the 3D mesh formed from backprojected image points. We use the indirect discretization by Xu [23], which first requires a gradient to be computed for each vertex. We use Xu's discrete linear approximation to the gradient. This technique works by computing the intensity gradient on each triangle adjacent to vertex $\mathbf{p}_i$, and then using an area-weighted average of the adjacent tri-

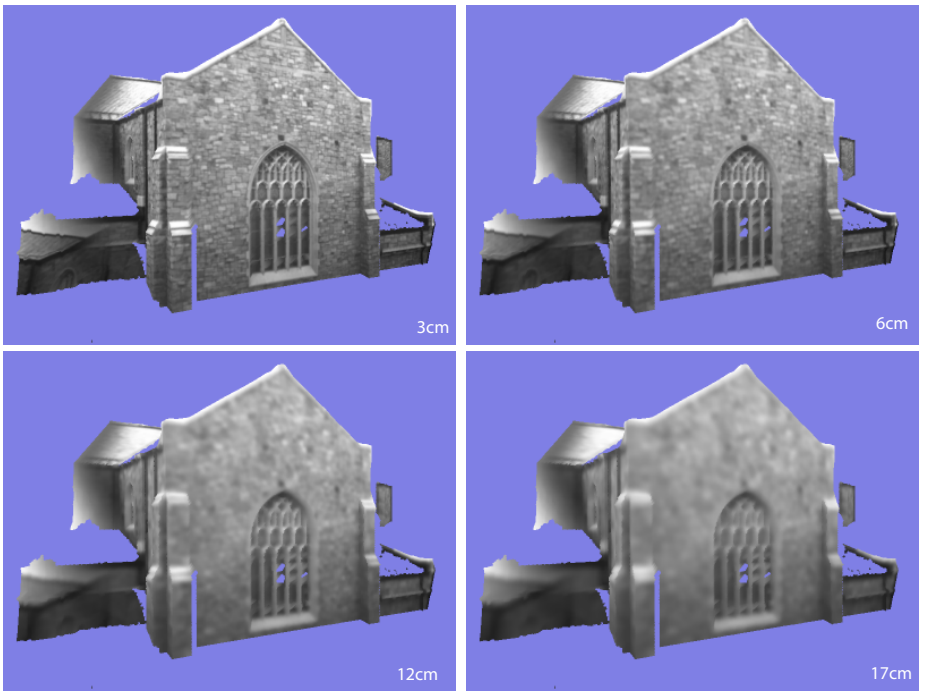

Fig. 6 This figure shows the effects of the bilateral filter at different scales.

angle gradients to approximate the gradient at $\mathbf{p}_i$. Figure 7 illustrates the process. Adopting the notation from [23], the gradient on a triangle is

$$\nabla_{T_j} I = \frac{1}{2A_j}[I(\mathbf{p}_i)\mathbf{v}_i + I(\mathbf{p}_j)\mathbf{v}_j + I(\mathbf{p}_{j+})\mathbf{v}_{j+}], \tag{2}$$

where A_j is the area of triangle $[\mathbf{p}_i, \mathbf{p}_j, \mathbf{p}_{j+}]$. $I(\mathbf{p_i})$ is the intensity at point $\mathbf{p}_i$, and $\mathbf{v}_i$ is the inward normal of the edge opposite $\mathbf{p}_i$ (i.e. edge $[\mathbf{p}_j, \mathbf{p}_{j+}]$) scaled by the length of that edge. It is computed by

$$\begin{aligned}\mathbf{v}_i = [&((\mathbf{p}_i - \mathbf{p}_j)\cdot(\mathbf{p}_j - \mathbf{p}_{j+}))(\mathbf{p}_{j+} - \mathbf{p}_i) + \\ &((\mathbf{p}_i - \mathbf{p}_{j+})\cdot(\mathbf{p}_{j+} - \mathbf{p}_j))(\mathbf{p}_j - \mathbf{p}_i)]/2A_j,\end{aligned}$$

and similarly for the other edge normals. Thus the intensity gradient at a vertex with respect to the mesh is simply approximated by

$$\nabla_M I(x_i) = \frac{1}{A(\mathbf{p}_i)} \sum_{j \in N_1(i)} A_j \nabla_{T_j} I, \tag{3}$$

where $A(\mathbf{p}_i)$ is the sum of areas of the 1 ring of triangles connected to $\mathbf{p}_i$ ($N_1(i)$, see Figure 7). The LBO can now be approximated as

$$\Delta_M I(x_i) = \frac{1}{2A(\mathbf{p}_i)} \sum_{j \in N_1(i)} -\mathbf{v}_i^T [\nabla_M I(\mathbf{p}_j) + \nabla_M I(\mathbf{p}_{j+})]. \tag{4}$$

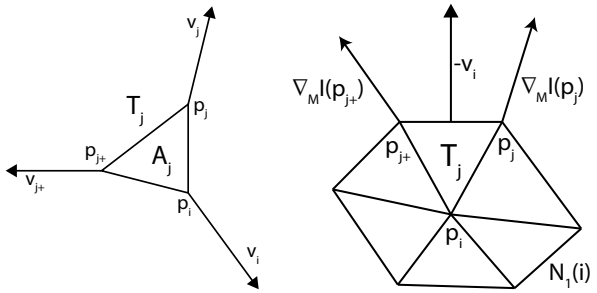

Fig. 7 Left: the terms in the gradient's calculation. Right: a 1-ring neighborhood and the vectors used to compute the contribution of triangle T_j to the Laplacian.

Extrema are detected as the maxima and minima of the scale-normalized LBO by comparing a vertex with its 1-ring neighbors. Extrema are filtered by thresholding the strength of the Laplacian response to be within the top 10% of all responses. Finally, extrema are further filtered by a non-maximal suppression [4] with a suppression radius of $3s_i$.

As mentioned earlier, range data often has holes due to surfaces that scatter the return away from the scanner; we only compute values for the LBO when a full ring of vertices surrounds a point. We require at least 5 neighboring vertices with Laplacians computed within distance s_i to detect extrema. Since we do not require a point to be an extremum across scales, the number of neighbors that a vertex is compared against is significantly fewer than in regular SIFT. This results in more candidate keypoint locations than regular SIFT before spatial filtering.

After the extrema have been detected, the gradients computed above are projected into the tangent space of their point for later use in computing the keypoint's orientation and descriptor. These gradients are made more repeatable by computation in the physical scale space.

3.3 Keypoint Coordinate Frame

A 3D coordinate frame can now be computed for each keypoint [19,24,22]. One axis is taken to be the normal of the extremal vertex. The second axis is defined to be the dominant gradient direction of the intensities at the keypoint, computed as follows. We project the gradients of nearby vertices (e.g., within $3s_i$) into the tangent plane of the keypoint, weight them using the same bilateral filter as before with $\sigma = s_i$, and enter them into a histogram (we use 36 bins in our experiments). The maximum of this histogram is found and a parabolic fit to the values around this maximum determines the dominant gradient direction [12]. The third axis of the keypoint is merely the cross product of the first two.

This coordinate frame is an improvement over previous work [19] because the normals in the image mesh are smoothed as the scale increases, giving more support to the keypoint center's normal as scale increases. The computation of the dominant gradient direction is improved by the physical scale space's invariance to strong gradients in the image caused by structural discontinuities. It is also improved by down-weighting the contribution of gradients from locations that are more sensitive to viewpoint variation. Finally, the size of the contribution region is fixed by the physical scale. Through the fixed size of the support

region and the application of the bilateral filter we increase the repeatability of gradient contributions, and in turn the repeatibility of the dominant gradient direction.

4 Keypoint Description and Matching

The remaining stages of the keypoint algorithm are description and matching. We first describe computation of the descriptor and then consider matching. The latter goes beyond current techniques by using both the descriptor and range information to rank-order keypoint matches.

4.1 Description

The SIFT-like descriptor that we compute lies in the tangent space of the keypoint — i.e., tangent to the image mesh surface in 3D on which the keypoint is located. The x axis of the descriptor is aligned with the dominant direction of the projected intensity gradients. The support for the descriptor consists of the image mesh points that lie within a sphere around the keypoint with radius $8\sqrt{2}s_i$ (i.e., half the length of the descriptor's diagonal). The descriptor is a 4×4 spatial grid with 8 orientation bins per spatial bin, resulting in a standard 128-dimensional descriptor. Gradients of vertices from the support region are projected onto the descriptor's plane, then weighted with the bilateral filter (with σ as the descriptor's radius) and placed in their appropriate bins using partial volume interpolation. The descriptor is normalized and thresholded as in the original SIFT [12]. Example descriptors can be seen in the top row of Figure 1.

We choose to project intensity gradients onto the single tangent plane in order to be robust to changes in viewpoint. Consider a keypoint near a corner, such as the end of a building. If the keypoint were to wrap around the corner, then the keypoint could only be matched when both planes are seen. By only using one plane, any view that sees that plane will create a similar descriptor at that location. An alternate 3D descriptor built using all three axes of the keypoint [24] would also suffer from this problem.

Our descriptor has advantages over previous algorithms (e.g., [19,22]) by being less sensitive to structural discontinuities. Since the support of a descriptor is larger than the scale at which the keypoint is detected, descriptor bins frequently extend across depth discontinuities into the scan's free space. In our algorithm, distant pixels from such structural discontinuities are either not in the support region or are downweighted due to the bilateral filter. Such bins will have nearly empty orientation histograms, which is a repeatable property when the empty bins lie in the free space of the scan for any view of the keypoint.

4.2 Matching

Keypoints are only matched against other keypoints of the same physical smoothing scale. A separate k-d tree is built on the descriptors for each scale used by the image mesh being matched. As in SIFT [12], for each keypoint in one scan, we search the scale-specific k-d tree from the other scan to find the two closest (in the sense of Euclidean distance) descriptors and their associated keypoints. We then compute the ratio, r, of the best descriptor distance to the second best distance.

Matching keypoints only against other keypoints of the same physical scale significantly culls the necessary search space. The reduced search space improves the power of the ratio metric by ensuring that the second best keypoint is of the same scale and is not an implausible match. Also, it allows a set of physical scale keypoints to be matched more quickly than a similarly-sized set of mixed scale keypoints. Finally, since keypoints at one scale can be

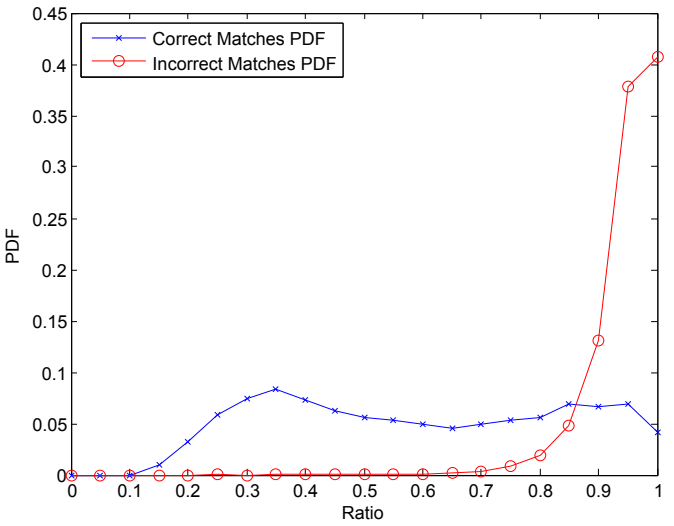

Fig. 8 PDFs for ratio measure on the experimental dataset (presented in section 6) for both correct and incorrect matches.

computed and matched independently of keypoints at another scale, increasing the number of computed scales cannot decrease the total number of correct matches.

The ratio, r, for a keypoint match provides a strong indication of the distinctiveness and therefore the quality of the match as demonstrated by figure 8. Lowe [12] and others [4] have applied an empirically-determined threshold of 0.8 to cull the set of matches and then used the surviving matches in recognition and registration. In our earlier work on image registration [19], and here, we instead use the ratios to generate a ranked set of initial transformations to test. The rest of this section describes a method to improve the ranking by exploiting the prospective alignment of the range data suggested by a keypoint match.

The first step for creating a range match measure is to create correspondences between the underlying range data of one keypoint with the underlying range data of its match. This first part of the algorithm creates a set, M, of corresponding range points between the scans in the vicinity of the keypoints. We compute this set by first recalling that each keypoint match can be used to form a rigid transformation between two scans simply by aligning the origins and axes of these coordinate systems. Specifically, let S_p and S_q be the two range scans (i.e. *not* the image meshes), let k_p and k_q be descriptor-matched keypoints from these two scans whose 3D locations are $\mathbf{c}_p$ and $\mathbf{c}_q$ in their respective scans. Let $\mathbf{T}$ be the transformation from S_p to S_q formed as just described. We gather all range points within a fixed distance D [1] of $\mathbf{c}_p$ in S_p into a set P and map both the point locations and surface normals into scan S_q using the computed transformation, $\mathbf{T}$. Let the mapped set be P'. Next we need to compute which points in P' were actually visible by S_q so that we do not try to match points that were not seen by both scanners. We approximate the visible set of mapped points by culling all points from P' whose normal points away from S_q's scanner. For each remaining mapped point location, $\mathbf{p}'$, we use projection-based matching [18] to find the approximate closest

[1] We use .75m in our experiments, which is about 25 times the sample spacing of our scans.

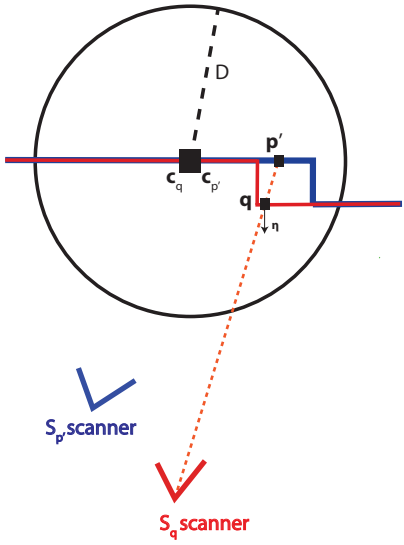

Fig. 9 Illustrating how correspondences are found for computing the range measure. This figure shows S_p (blue) transformed onto S_q (red) using $\mathbf{T}$. We see the keypoint center $\mathbf{c}_q$ and transformed keypoint center, $\mathbf{c}_{p'}$, aligned resulting in a misregistration of the 1-dimensional scans. The correspondence of $\mathbf{p}'$ (a mapped point) is found via projection matching (orange dotted line). The corresponding point found is $\mathbf{q}$ with normal vector η.

point $\mathbf{q}$ and its associated normal η in S_q. Figure 9 illustrates this correspondence matching technique. We likewise ensure that this point is within D from $\mathbf{c}_q$ and does not face away from S_p's scanner. If $\mathbf{q}$ passes these tests we can then compute the point-to-plane distance

$$d(\mathbf{p}',\mathbf{q}) = |(\mathbf{p}'-\mathbf{q})^\top \eta|, \tag{5}$$

which gives an indication of how well the transformation aligns $\mathbf{p}$ with scan S_q. Let M be the resulting set of $(\mathbf{p}',\mathbf{q})$ correspondences. In figure 10 we demonstrate the computed correspondences for both a correct and an incorrect match on real data.

Our next step is to combine the distances for all correspondences to determine the local quality of $\mathbf{T}$. This quality measure must account for poor correspondences without allowing them to swamp the measure. We do this using a robust weighting scheme. Specifically, for each $(\mathbf{p}',\mathbf{q}) \in M$ we compute a Cauchy weight

$$w_{p'q} = w(d(\mathbf{p}',\mathbf{q})) = \frac{1}{1+\frac{d(\mathbf{p}',\mathbf{q})}{C^2})}. \tag{6}$$

where C is a tuning constant that accounts for sensor noise. (We use 3 times the estimated sensor noise standard deviation.) We combine these values to obtain a range alignment accuracy score as

$$R(k_p,k_q) = \frac{1}{|M|} \sum_{(\mathbf{p}',\mathbf{q})\in M} w_{p'q} \tag{7}$$

A value of $R(k_p,k_q)$ close to 1 means that all points in the local neighborhood of $\mathbf{p}_c$ are well-matched with S_q. Therefore, the measure $R(k_p,k_q)$ indicates how well the keypoint

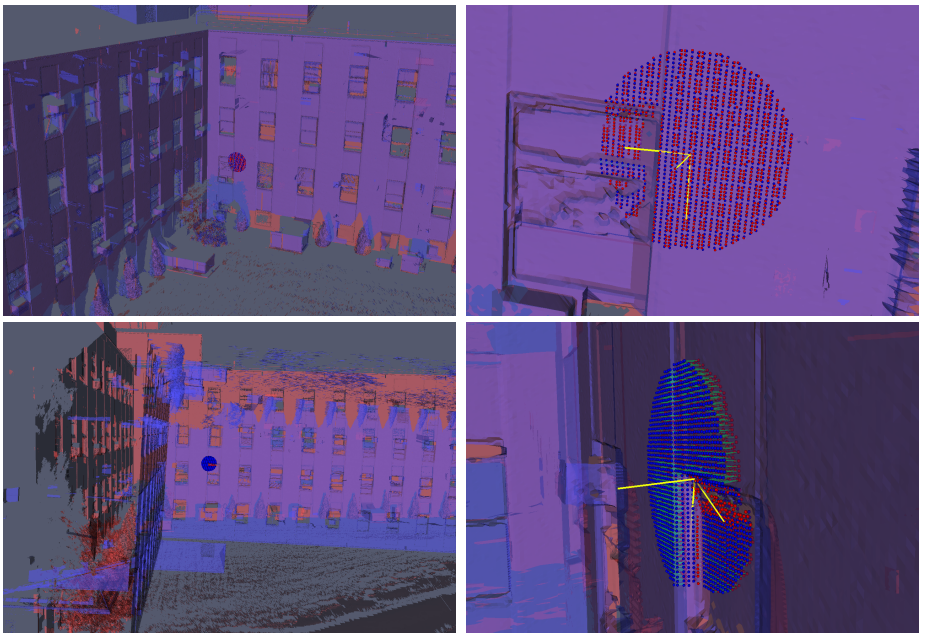

Fig. 10 The top row of this figure demonstrates the correspondences for the computation of the range measure in a correctly matched keypoint. The bottom row demonstrates it for an incorrect match. In the bottom row, the window is not only matched incorrectly but is also upside-down (note the distances between the correspondences around the window). The right column is a zoomed in view of the left column. In this figure S_p is blue and S_q is red.

match locally aligns the range data. We are now ready to combine the match's ratio r with the range accuracy measure into a single score for ranking.

We combine the ratio and the range accuracy using a simple linear discriminant. Training our discriminant begins by computing r for each match in the training set and discarding matches whose ratios are greater than 0.8. As these matches tend to be either completely wrong or repeated structure matches that would weaken the range accuracy measure if trained on. We then compute $R(k_p, k_q)$ for the remaining matches. We train a simple linear discriminant on the ratio and the range accuracy measure. Figure 11 shows the trained discriminant over the experimental data. Matches are sorted by their signed distance from the discriminant. By using the range match accuracy in the ranking we aim to not only rank additional correct matches higher, but to also rank the matches by accuracy of their transformation.

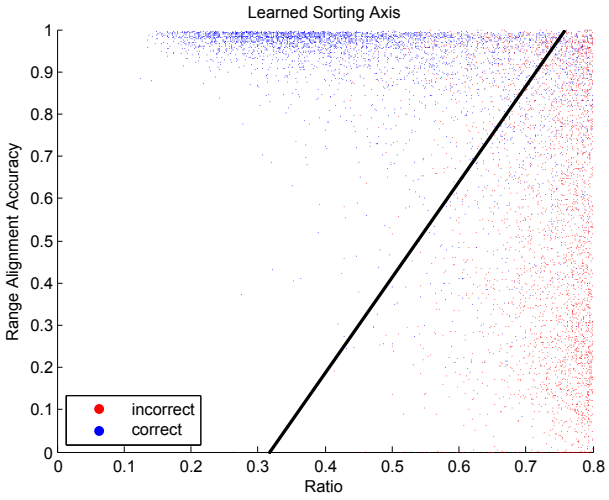

Fig. 11 The discriminant is shown over matches plotted by their ratio vs range accuracy. The discriminant is perpendicular to the axis the matches are sorted on.

5 Application to Range Scan Registration

We demonstrate the application of the physical scale keypoint algorithm to registering pairs of range scans. The objective is to compute a rigid transformation that aligns one scan, designated as the moving scan, with the other, designated as the fixed scan. While many algorithms have been proposed in the literature [2, 5] and any algorithm that uses keypoints on combined range and intensity data can also use our physical-scale keypoints, we focus on improvements to our own earlier algorithm [19].

This algorithm works by rank-ordering keypoint matches, generating initial rigid transformations from each keypoint match, iteratively refining each transformation, and applying a decision criteria to select the correct registration (or to decide that all tested transformations are incorrect). The refinement technique is a combination of the well-known ICP algorithm and a region growing algorithm. The initial region is a small volume surrounding the keypoint match, and growth of this region alternates with ICP refinement until the region grows to encompass the entire region of overlap between the scans.

The single keypoint match initialization fits well with region growing ICP. A major advantage of the initialization technique is that even when there is very low overlap or significant changes between the scan pairs, we can often create a reasonable initialization since only a single correct keypoint match is necessary. This correct keypoint match will produce a rigid transformation that produces a close but imperfect alignment of the range scans in the regions that immediately surround the keypoints. (Globally, the result can be far off, especially in the rotation parameters.) ICP in this small region quickly converges and the resulting rigid transformation brings more of the range scans into accurate alignment, allowing for expansion of the scan regions over which ICP is applied. Exeriments in [19] should that iterating ICP refinement and region growth produces accurate global alignment for more than 98% of the correct initial matches.

The change to the initialization stage is simple. Physical scale keypoints are matched and sorted as described above to produce initial transformations. If the use of physical scale keypoints produces more correct initializations higher in the list, then fewer matches will typically have to be tested before the algorithm succeeds, and conversely, fewer matches will have to be tested rejected before deciding that the scans do not overlap. We demonstrate this experimentally below and use this to reduce the overall number of initial estimates that the algorithm tries.

No changes are needed in the region-growing ICP refinement step.

5.1 Application to Registration Verification

The change to the final decision criteria is more involved, but straight forward. In fact, a complex algorithm is replaced by a simpler one. The criteria presented in [19] combine measures of stability, positional accuracy, angular consistency and boundary alignment into a seven-component vector which is then mapped to a single decision value using a linear classifier. In this paper, this classifier is replaced with a single measure that counts the number of highly-ranked matches that are consistent with the estimated transformation and then applies an empirically-determined threshold to separate good and bad transformations.

Given an estimated transformation and a set of keypoint matches, we apply the transformation to each moving scan keypoint. If the mapped position and orientation in fixed scan are consistent with the position and orientation of the matched fixed scan keypoint, then this keypoint match is considered to be consistent with the transformation. We then simply count

the number of consistent matches and compare against a threshold. This is quite similar to an earlier algorithm used for intensity-image registration [3].

Experimentally, we have found that considering PSK matches with ratio test values below 0.75, and then using a location distance threshold of $5s_0$ and an orientation different threshold of at most 5° on the matches yields good results. These aggressive thresholds reflect the purpose of the verification. Using a value below 0.75 eliminates some correct matches, but a lot more incorrect ones, focusing the test on only the most likely matches. Similarly, the distance and orientation thresholds require that only the matches most consistent with the transformations be used. As we will see, this produces substantial separation between the correct and incorrect transforamtions.

Verification measures directly related to the refinement algorithm, such as range accuracy, can easily accept a repeated structure misregistration as correct. In our proposed algorithm, by only letting distinctive matches count towards the correctness of a transformation, relatively few matches will be consistent with a misregistration based on misaligning structures.

6 Results

We demonstrate the quality of our Physical Scale Keypoint (PSK) algorithm using pairwise range scan matching on real-world outdoor scans. The datasets were collected using a Leica HDS-3000 scanner. For each scan, we built the image mesh from the image that views the greatest portion of the range data. These images were acquired by the scanner at roughly the same time as the scan and are 1024×1024 in resolution. We selected a set of scales fixed for all experiments beforehand as $S = \{s_i \mid s_i = 3\text{cm} \times 2^{i/2}, i = 0, \ldots, 5\}$.

6.1 Data Set

Our dataset, illustrated in Figure 12, consists of eight outdoor range scan pairs, which include large intensity differences, low overlap, repeated structure, large viewpoint differences, and numerous discontinuities. Two of the scan pairs, VCC North and Parkinglot, are considered easy, since the difference in viewpoint between the scans is low for both pairs. The VCC South scan pairs are more difficult since the two images from VCC South1 were taken under substantially different illumination conditions, while VCC South2 has low overlap. The remaining four datasets are all very difficult; we include them to demonstrate our improvements in keypoint matching and registration, but also to show the limits of our algorithm. Among these, the DCC scan pair has numerous occluding trees and a large viewpoint difference. The JEC, JROWL, and Biotech pairs all have extensive repeated structure. The JROWL pair has one scan taken during the day and the other at dusk. The Biotech pair is the most difficult, in part because of its repeated structure, in part because one scan is taken from an oblique view of the building resulting in severely non-isotropic range point sampling, and in part because the other scan has significant glare in the images.

We have established a manually-verified "ground truth" transformation between each scan pair, sufficient to extract quantitative results on the effectiveness of our physical scale keypoint techniques. A match is considered to be correct if the ground-truth-transformed moving keypoint's orientation vector is within 10° of the fixed keypoint's, and if its location is within $15cm$ of its matched keypoint.

6.2 Comparison Techniques

The algorithm is compared to a technique we call back-projected SIFT (bpSIFT), used in our earlier work. This technique is closely related to VIP patches [22]. In bpSIFT, keypoints are detected in the images, and back-projected into a plane fit on the range data. The 2D gradients from the images are then mapped onto the plane in the range data to give the descriptor affine invariance [19]. Correct matches for bpSIFT are measured in the same way as for PSK. For completeness we also present results using the original SIFT descriptor [12].

6.3 Illustrative Examples

We begin our analysis with some illustrative examples of the effectiveness of our keypoints on difficult examples. Figure 1 already demonstrated one such example from the VCC South pair. Three more are shown in Figure 13. The first of these shows a keypoint taken from an ashtray in front of a door entrance, the second is taken from the support pillar of a building,

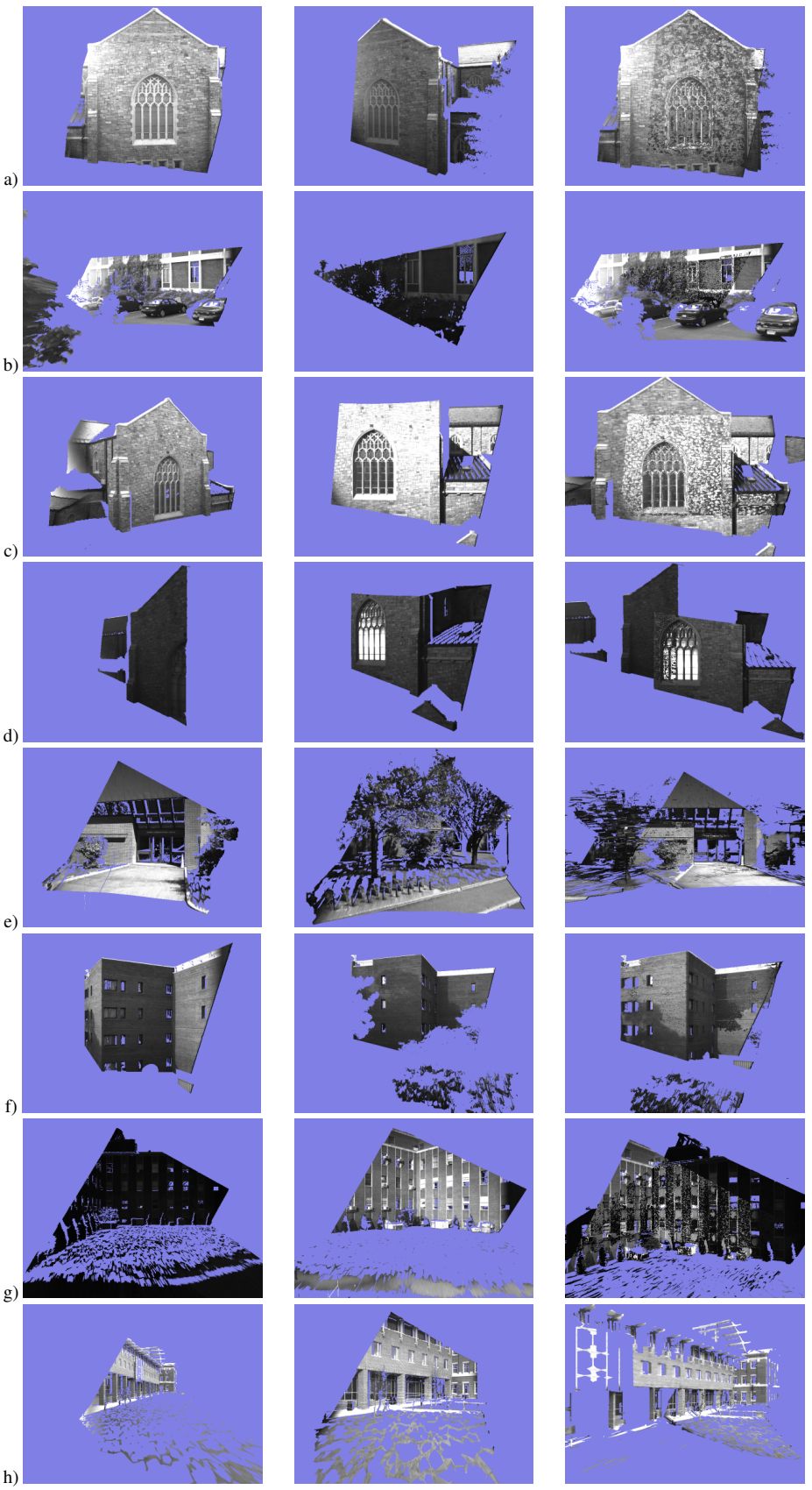

Fig. 12 The experimental dataset. Columns 1 and 2 show the individual image meshes, and Column 3 shows their ground-truth alignment. The scans are: a) VCC North, b) Parkinglot, c) VCC South1, d)VCC South2, e) DCC, f) JEC, g) JROWL, and h) Biotech.

and the third is taken from the side face of an air-conditioner unit. In each case, when keypoints are detected and described on a plane, as in bpSIFT or VIP patches, the descriptor region will include projections from points on different surfaces. Since our bilateral weighting technique substantially downgrades the influence of points from different surfaces in both keypoint detection and keypoint description, the PSK technique effectively matches these keypoints. The descriptor illustrated for each figure shows that the vast majority of the values are from a single surface.

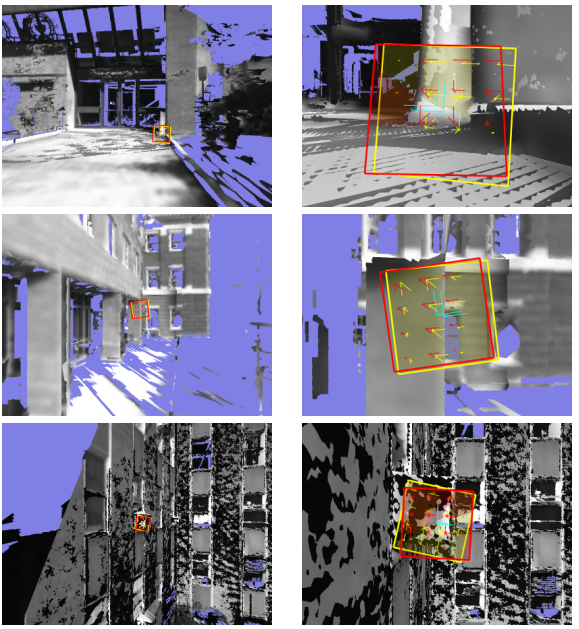

Fig. 13 Example PSK matches near discontinuities. The left column is a wider view of the area being matched; the keypoint is shown close up in the right column. These examples are taken from the DCC, Biotech, and JROWL scans respectively. The moving keypoint is shown in yellow with its matched fixed keypoint in red.

6.4 Ranked Keypoints

Our quantitative analysis focuses on the rank-ordering of keypoints. We consider two versions of the PSK technique (one with the rank-ordering based on the intensity descriptor ratio and the other based on the combined intensity and range ranking produced by the linear discriminant, see section 4.2), comparing them to each other and to the earlier bpSIFT and SIFT techniques. These results are shown in Tables 1-3.[2]

Table 1 gives an idea of the density of the correct keypoint matches. Clearly, as the scans become harder to align, the advantages of the two versions of our new PSK technique over

[2] The correct keypoint criteria used here differ slightly from our earlier work, making the numbers for bpSIFT and SIFT different than what we presented in [19]. The scans also have a narrower field of view.

the earlier algorithms become dramatic, with bpSIFT and SIFT producing very few matches in the top 50 for the harder scans. The PSK techniques produce from two to ten times as many correct matches. The differences are less dramatic when considering only the top five or ten ranked matches, as shown in Table 2, or the first correct match, as shown in Table 3, but the advantages are still clear. Only the difficult Biotech scan pair shows no significant improvement.

Looking at rankings only, the differences between the two PSK techniques are much less significant than their advantages over bpSIFT. However, the version that combines the descriptor-ratio distance with the range alignment accuracy measure did rank twice as many matches for the tricky DCC and JROWL pairs in the top 50.

	PSK #t50 discriminant	PSK #t50 ratio	bpS #t50	SIFT #t50
VCC North	50	50	50	33
Parkinglot	46	36	23	18
VCC South1	50	49	18	7
VCC South2	50	47	10	13
DCC	23	12	3	3
JEC	16	15	0	0
JROWL	10	5	1	2
Biotech	2	2	1	1

Table 1 Number of correct matches in the set of 50 top-ranked matches for eight different scan pairs. The first column is the full PSK technique in which the ranking is computed based on the discriminant that combines the intensity descriptor ratio and the range alignment accuracy. The second column is the PSK technique using only the intensity descriptor ratio for ranking. The third column is the earlier bpSIFT technique. The fourth column is using the image-only SIFT descriptor.

	PSK #t5 discr	PSK #t10 discr	PSK #t5 ratio	PSK #t10 ratio	bpS #t5 ratio	bpS #t10 ratio	SIFT #t5 ratio	SIFT #t10 ratio
VCC North	5	10	5	10	5	10	5	6
Parkinglot	4	9	4	7	3	5	2	4
VCC South1	5	10	5	10	2	3	1	1
VCC South2	5	10	5	10	4	7	2	4
DCC	3	7	2	2	1	2	1	3
JEC	4	6	4	6	0	0	0	0
JROWL	1	2	0	0	0	0	1	2
Biotech	0	0	0	1	0	1	1	1

Table 2 Number of correct matches ranked in the top 5 / top 10 for eight scan pairs.

6.5 Registration Final Decisions

From the viewpoint of scan registration, these results have several important implications. First, building on our discussion in Section 5, within the context of our earlier algorithm, we need to investigate fewer initial keypoint matches. From Table 2 we see that testing at most 5 top-ranked PSK keypoints usually yields more than one correct match. Since our iterative

	PSK discriminant	PSK ratio	bpSIFT	SIFT
VCC North	1	1	1	1
Parkinglot	1	1	1	1
VCC South1	1	1	2	3
VCC South2	1	1	1	3
DCC	2	2	3	1
JEC	1	1	336	804
JROWL	4	16	12	3
Biotech	30	10	9	3

Table 3 Rank-order position of first correct keypoint for eight different image pairs.

refinement algorithm takes a correct initial keypoint match to a final correct transformation more than 98% of the time, we can safely use these small sets of keypoint-match candidates. While the other algorithms (bpSIFT and SIFT) also have a few highly ranked keypoints, PSK allows for more error in the verification process by providing more high-ranked correct matches.

Second, we demonstrate that our simpler decision criteria can be effective when we use physical scale keypoints.[3] We considered the top-ranked 50 keypoint matches in each scan, and refined each to a final transformation. We then tally how many of the large total set of keypoints in each scan was consistent with the final transformation as discussed in section 5. We found that the **minimum** number of consistent PSK matches for a correct registration is 15 (DCC), while the next two lowest counts are 20 and 43 for the JROWL and JEC scans respectively.[4] On the other hand, the **maximum** number of PSK matches consistent with an incorrect transformation is 6, which occurred only once, for the JROWL scan. For all other incorrect transformations, the maximum number was just 1. For more typical, and therefore easier, scans this separation will be even greater. This implies that we can safely set a threshold on the number of consistent matches as a verification criterion. The quality of this threshold could be improved with a much larger data set, at which point it would be worth adopting a more robust technique for calculating the threshold such as the one in [3].

6.6 Registration Example

For completeness we demonstrate use of the PSK algorithm in multiscan range registration. In the results above we used a single image for each scan. However, many scans require multiple images to cover the entirety of the range data. In order to use multiple images, we slightly modify the algorithm to perform the spatial filtering (described in section 3.2) after we have computed the full set of keypoints as some images may overlap.

We present a data set with five scans of the VCC building. The overlapping scan pairs are registered using PSK and the region growing refinement method described earlier. The non-overlapping pairs are automatically rejected by our decision criteria. The multiscan registration is finalized with a global minimization across the resulting correspondences of each pair-wise registration in order to place each scan into a single coordinate system. The resulting registration can be seen in figure 14. The first PSK match produced the final scan-pair estimate for all but the yellow/blue pair, which chose the alignment based on the 2nd

[3] We ignore for now the Biotech scan pair, which neither our current algorithm nor our previous algorithm can verify.

[4] Note that the number for DCC is lower than the number of consistent PSK matches ranked in the top 50 because the orientation acceptance criterion is tighter.

ranked keypoint. This pair is quite difficult because the viewpoints are far apart and the overlapping face is partially obscured in the yellow scan by the solar array.

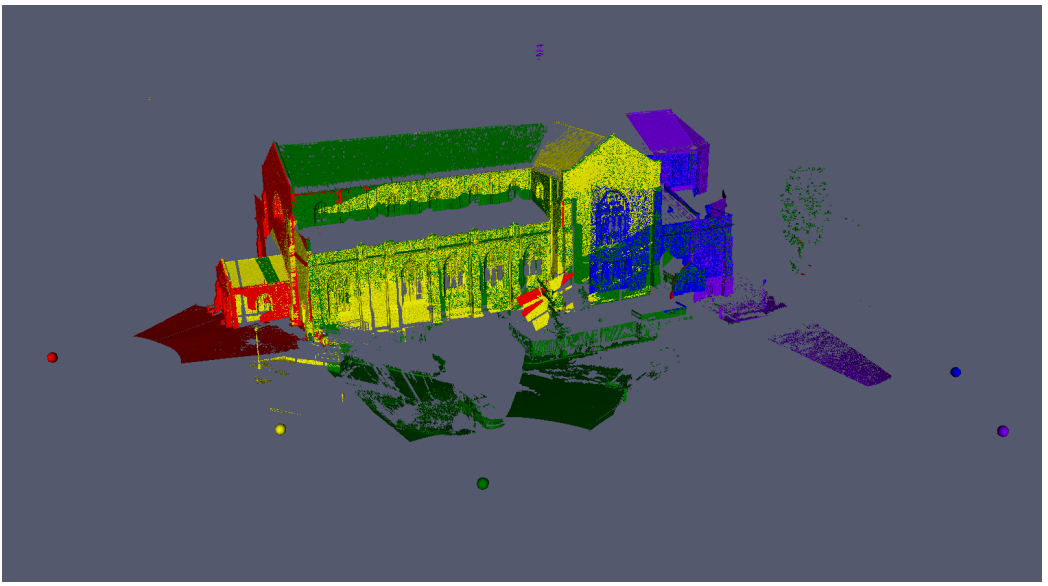

Fig. 14 This figure shows the resulting model after registering five VCC range scans. The spheres represent the viewpoint for the scan of the corresponding color.

6.7 Noisy Data

As a final test, we also performed experiments in which zero-mean Gaussian random noise was added to each range point measurement in the direction of the line of sight. After adding the noise, we rerun the entire algorithm, including the computation of surface normals. No parameters or thresholds were changed. Figure 15 plots the number of correct PSK matches ranked in the top 50 as a function of noise. In general, the fall-off in performance is gradual, as might be expected. The only catastrophic failure is the DCC scan with 8cm of noise. We note that this is an extreme case, especially since the noise in the original scanner is closer to 2mm and the inter-sample spacing is 2-5cm. Overall, since our algorithm is largely intensity based, the overly smoothed range data does not affect the distinctiveness of the keypoints as much as it would for a geometry based keypoint.

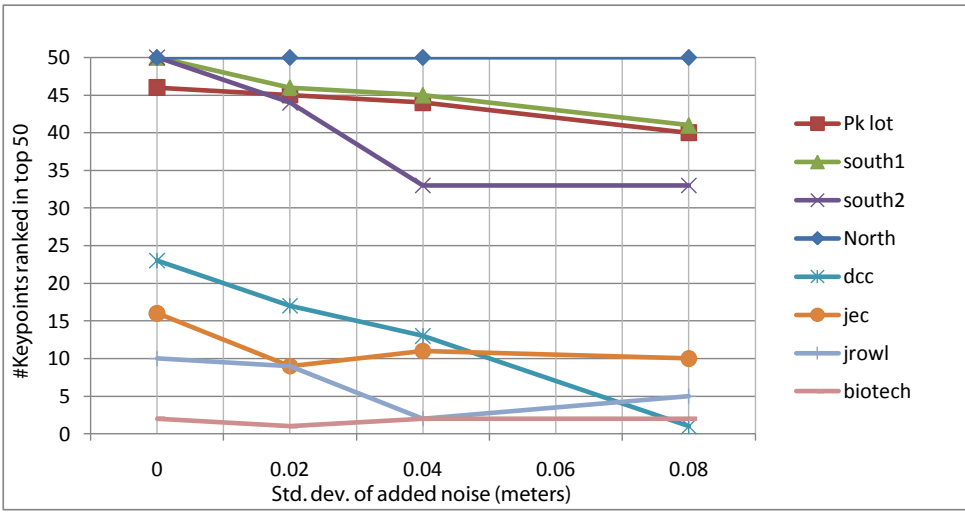

Fig. 15 This chart shows the effect of added noise to the range data on keypoint match rankings.

7 Conclusions

We proposed and tested a physical scale intensity-based keypoint (PSK) detection, description and matching algorithm that works with combined intensity and range data. The images are backprojected to form an image mesh in 3D, and all smoothing and differentiation steps are applied to this mesh. Detection and description of keypoints employs a novel bilateral filter to avoid the influence of points from across a depth or orientation discontinuity. Keypoints are detected and matched at multiple physical scales independently. We demonstrated the effectiveness of the PSK algorithm by matching pairs of challenging single-view range/intensity data sets taken from outdoor scenes. The results show it to be more effective than algorithms that detect keypoints through backprojection of intensity image regions onto single tangent planes. When integrated into a recent registration algorithm, physical scale keypoints reduce the number of initial matches required for registration and lead to a simpler decision criterion based simply on counting the number of PSK matches consistent with an estimated transformation.

The limitations of our PSK algorithm and its application to registration can be understood by examining the one scan pair — Biotech — for which it fails to generate sufficient matches for initializing and verifying a registration. The Biotech scans were taken from viewpoints that differ by about 45°, and one scan is taken at an extreme angle relative to the building. The resulting combination of viewpoint and sampling differences is hard to overcome. The second difficulty is that the strictly repetitive pattern of walls and windows make unique matching difficult. In other scan pairs that exhibit repetitive structures we typically find enough locally-distinctive features — such as air conditioner units — to produce distinctive keypoint matches. The PSK algorithm is able to accurately detect, describe and match such features.

On a more theoretical level, the primary limitation of our PSK descriptor is keypoints near discontinuities in which the area missing from the descriptor region lies in the free space of the scan. These regions are occluded by a foreground object in one scan but not in the other, and they are unlikely to be well-matched. A more sophisticated technique would be needed to construct and match descriptors for keypoints at such locations.

Despite these limitations, the physical scale keypoint technique has proven to be effective in matching between on all but the most extreme scan pairs. Combined with the refinement technique that requires only a single correct keypoint match to succeed, physical scale keypoints form the basis for efficient and effective registration of complicated, low-overlap scan pairs.

References

1. E. Akagndz and I. Ulusoy. Scale and orientation invariant 3d interest point extraction using hk curvatures. In *3dRR, ICCV Workshops*, 2009.
2. P. Besl and N. McKay. A method for registration of 3-d shapes. *PAMI*, 14(2):239–256, 1992.
3. M. Brown and D. G. Lowe. Recognising panoramas. In *Proceedings of the 9th International Conference on Computer Vision*, pages 1218–1225, 2003.
4. M. Brown, R. Szeliski, and S. Winder. Multi-image matching using multi-scale oriented patches. In *CVPR*, 2005.
5. Y. Chen and G. Medioni. Object modeling by registration of multiple range images. *IVC*, 10(3):145–155, 1992.
6. S. Fleishman, I. Drori, and D. Cohen-Or. Bilateral mesh denoising. *ACM Trans. Graph.*, 22(3):950–953, 2003.
7. A. Frome, D. Huber, R. Kolluri, T. Blow, and J. Malik. Recognizing objects in range data using regional point descriptors. In *ECCV*, 2004.
8. J. Hua, Z. Lai, M. Dong, X. Gu, and H. Qin. Geodesic distance-weighted shape vector image diffusion. *IEEE Trans. Vis. Comput. Graph.*, 14(6):1643–1650, 2008.
9. A. E. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 21(5):433–449, 1999.
10. 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*. IEEE Computer Society, 2005.
11. T. Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *Journal of Applied Statistics*, 21:224–270, 1994.
12. D. Lowe. Distinctive image features from scale-invariant key-points. *IJCV*, 60:91–110, 2004.
13. A. S. Mian, M. Bennamoun, and R. A. Owens. Matching tensors for automatic correspondence and registration. In *ECCV*, 2004.
14. K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *International Journal of Computer Vision*, 60(1):63–86, 2004.
15. K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 27(10):1615–1630, 2005.
16. K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool. A comparison of affine region detectors. *ijcv*, 65(1–2):43–72, 2005.
17. J. Novatnack and K. Nishino. Scale-dependent 3d geometric features. In *ICCV*. IEEE, 2007.
18. S. Rusinkiewicz and M. Levoy. Efficient variants of the ICP algorithm. In *3dim*, pages 224–231, 2001.
19. E. Smith, B. King, C. Stewart, and R. Radke. Registration of combined range-intensity scans: Initialization through verification. *Computer Vision and Image Understanding*, 110(2):226–244, 2007.
20. E. Smith, R. Radke, and C. Stewart. Physical scale intensity-based range keypoints. In *3DPVT*, 2010.
21. J. Starck and A. Hilton. Correspondence labelling for wide-timeframe free-form surface matching. In *ICCV*, 2007.
22. C. Wu, B. Clipp, X. Li, J.-M. Frahm, and M. Pollefeys. 3d model matching with viewpoint-invariant patches (vip). In *CVPR*, 2008.
23. G. Xu. Convergent discrete laplace-beltrami operators over triangular surfaces. In *GMP*, pages 195–204. IEEE Computer Society, 2004.
24. A. Zaharescu, E. Boyer, K. Varanasi, and R. Horaud. Surface feature detection and description with applications to mesh matching. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.
25. Y. Zhong. Intrinsic shape signatures: A shape descriptor for 3d object recognition. In *ICCV Workshops*. IEEE, 2009.
26. G. Zou, J. Hua, Z. Lai, X. Gu, and M. Dong. Intrinsic geometric scale space by shape diffusion. *IEEE Trans. Vis. Comput. Graph.*, 15(6):1193–1200, 2009.