This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Multi-view Non-rigid Refinement and Normal Selection for High Quality 3D Reconstruction

Sk. Mohammadul Haque, Venu Madhav Govindu
Indian Institute of Science
Bengaluru, India
{smhaque, venu}@ee.iisc.ernet.in

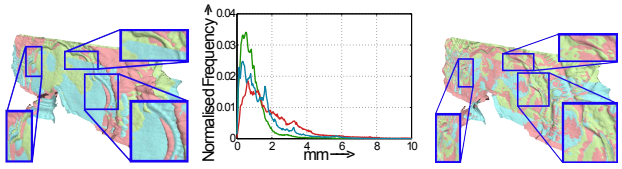
Abstract

In recent years, there have been a variety of proposals for high quality 3D reconstruction by fusion of depth and normal maps that contain good low and high frequency information respectively. Typically, these methods create an initial mesh representation of the complete object or scene being scanned. Subsequently, normal estimates are assigned to each mesh vertex and a mesh-normal fusion step is carried out. In this paper, we present a complete pipeline for such depth-normal fusion. The key innovations in our pipeline are twofold. Firstly, we introduce a global multi-view non-rigid refinement step that corrects for the non-rigid misalignment present in the depth and normal maps. We demonstrate that such a correction is crucial for preserving fine-scale 3D features in the final reconstruction. Secondly, despite adequate care, the averaging of multiple normals invariably results in blurring of 3D detail. To mitigate this problem, we propose an approach that selects one out of many available normals. Our global cost for normal selection incorporates a variety of desirable properties and can be efficiently solved using graph cuts. We demonstrate the efficacy of our approach in generating high quality 3D reconstructions of both synthetic and real 3D models and compare with existing methods in the literature.

1. Introduction and Relevant Work

The recent availability of consumer-grade depth camera technology has lead to very significant advances in dense 3D scene reconstruction techniques. Of the numerous approaches, the methods of RGBD Mapping [10], Kinect Fusion [11] and DynamicFusion [15] are representative. Despite such advances, there is a fundamental limit to the 3D reconstruction quality achievable due to the inherent low quality of individual depth maps obtained from devices such as the Kinect.

One class of approaches enhance 3D reconstruction quality by using additional information obtained through radiometric methods. Of these methods, some implicitly use shading information [6, 8, 12, 17] whereas others explicitly solve for surface normals using photometric stereo. While depth maps are noisy in nature, their low frequency component is of good quality. Conversely photometric normals are good at preserving high frequency or fine scale 3D surface details. Therefore, many methods fuse 3D depth representations with their corresponding photometric normals [2,5,9,14,16,18]. While the approach presented in this paper falls into this category and assumes inputs of depth and normal estimates, our formulation is in principle not limited to photometric stereo for obtaining normals.



(a) W/O non-rigid refinement (b) Misalignment Histograms (c) Our non-rigid refinement

Figure 1: Non-rigid misalignments: a) shows the presence of significant non-rigid misalignments in 3 rigidly registered input scans, b) shows the histograms of non-rigid misalignments present in the 3 scans, c) shows that these misalignments are corrected by our method. Please view this figure in colour.

Of the methods mentioned above, our approach is closest to those of [2, 9]. In [2], the authors obtain 2D normal maps from multiple viewpoints and combine them in a weighted fashion before passing on to a depth-normal fusion step. In [9], instead of averaging the normals the authors select one of them. Both [2] and [9] carry out their mesh-normal fusion using approaches similar to [14] to provide a high quality refined scan. Additionally, in [9] the photometric normals are obtained using the IR camera of the depth cam-

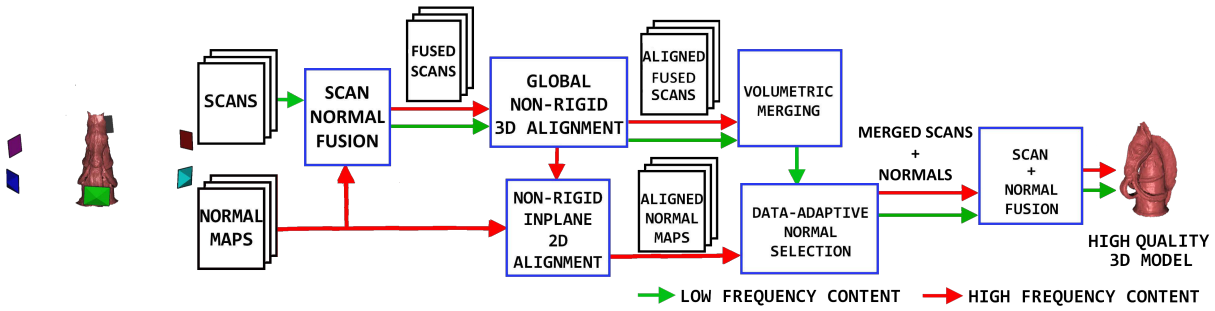


Figure 2: A schematic representation of our full pipeline. Please view this figure in colour and see text for details.

era itself, with the resulting advantage of automatically co-registered depth and normal maps.

Handling Non-rigid Deformations: In practice in a multi-view setup, even after accurate rigid 3D registration of the scans, there is a significant non-rigid misalignment among the depth and normal maps. This may arise due to a variety of reasons including residual alignment discrepancies, errors in normal estimates, camera non-linearities etc. Figure 1 (a) shows misaligned 3D surface features in a set of 3 rigidly registered scans. Figure 1 (b) shows a histogram of the magnitudes of non-rigid misalignments present among the three scans. This histogram is obtained by running our non-rigid refinement step explained in Section 2, but for now we can easily note that the amount of such misalignment is significant enough to result in either blurring of the finer details if averaged or misaligned details at the seam lines. Figure 1 (c) shows that our approach successfully corrects for the non-rigid misalignments across different scans.

1.1. Our Approach and Contributions

In this paper, we develop global, multi-view methods for the non-rigid refinement of depth and normal maps and adaptive selection of normals that are fused to result in high quality 3D representations that preserve very fine-scale surface details and minimise non-rigid misalignments. Figure 2 shows the full pipeline of our method. We obtain depth and normal maps from different viewpoints that cover the entire surface of the object to be scanned. In our approach, following that of [9], the depth and corresponding normal maps are obtained in the same co-ordinate system, *i.e.* they are automatically co-registered. For normal maps that are obtained using an external RGB camera, an additional registration step will be required. The rigid alignment of the input depth data is solved using a multi-view ICP algorithm. The salient steps of our approach are:

1. Global, multi-view non-rigid refinement of 3D scans as well as corresponding 2D normal maps

2. Graph-cut based global, multi-view normal selection.

We recall that a key objective of our method is to correct for non-rigid misalignments between normal maps that cover the same surface region. To this end, we first fuse the individual depth-normal maps obtained from each individual viewpoint. This ensures that the warping of the normal maps is done in a manner consistent with both the depth and normal information. The fused high quality scans are then aligned by a global, multi-view non-rigid refinement method described in Section 2. The resulting non-rigid motion estimates are used to warp the individual normal maps as described in Section 2.1. We also note that the multi-view nature of this refinement step ensures that the individual errors and deformations are properly accounted for by being distributed over all the scans. At this point we have obtained sets of scans and corresponding normal maps that are both aligned according to the estimated global non-rigid alignment. The aligned scans are merged using a volumetric method [7] giving us a mesh representation of the scanned object or scene. This mesh gives a complete representation that is accurate in the low frequency sense.

We note that each 3D surface point has a number of candidate corresponding normal estimates in the warped normal maps from different viewpoints from which it is visible. However, despite adequate care, the intuitive idea of averaging these candidate normal estimates will invariably result in blurring of the finest scale details available in the individual normal maps. Therefore, as described in Section 3, we use a graph-cut based multi-view adaptive approach to select only *one* of the normals to be associated with each mesh vertex. Our graph-cut approach carefully accounts for the relative reliability of the individual normals depending on the viewing direction and the amount of non-rigid displacement and thereby retaining the fine-scale details and minimising misalignment artefacts across different scans. Finally, the selected normals are fused with the 3D mesh obtained from the aligned scans using the method described in [9].

2. Global Non-rigid Refinement of 3D Scans

A key step in our pipeline is the non-rigid refinement of the 3D scans or meshes. Since we are interested in the correct registration of high frequency details in the form of normal maps, we fuse the corresponding depth and normal maps from individual viewpoints using the method in [9]. The resulting scans are then used as inputs for our global, multi-view non-rigid refinement procedure. While there are a number of non-rigid refinement approaches in the literature [4, 13], we use a global method based on [1]. The use of an affine motion model for each individual scan vertex ensures that there are enough degrees of freedom that will allow the high frequencies components to align as well as possible. To make our method both robust and global, we use the normal information in weighting the pairwise alignments and an adaptive weighted averaging for the final non-rigid warping of the scans respectively.

Pairwise Non-rigid Alignment: We first describe the non-rigid alignment of a pair of scans, wherein a template scan is warped onto a target scan. Our global, multi-view approach builds on individual pairwise alignments. Here we closely follow the approach of [1] with appropriate modifications. For each vertex $\mathbf{v}_i$ in homogeneous form, we associate an affine transformation $\mathcal{T}(\mathbf{v}_i) = \mathbf{X}_i\mathbf{v}_i$ where $\mathbf{X}_i$ is a 3×4 matrix representing the affine transformation applied to vertex $\mathbf{v}_i$. Similar to [1], we develop a net cost to be minimised consisting of two terms. The first term C_{fit} in Equation 1 penalises the distance of each warped template vertex from its nearest neighbour in the target scan. The second term C_{stiff} in Equation 2 penalises the differences between the affine transformation at each vertex $\mathbf{X}_i$ and those of its neighbours. For N vertices in the template scan, we denote the C_{fit} term as

$$C_{\text{fit}}(\mathbb{X}) = \left\| (\mathbf{W} \otimes \mathbf{I}_3) \left(\begin{bmatrix} \mathbf{X}_1 & & \\ & \ddots & \\ & & \mathbf{X}_N \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \vdots \\ \mathbf{v}_N \end{bmatrix} - \begin{bmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_1 \end{bmatrix} \right) \right\|_2^2 \quad (1)$$

where $\mathbb{X} = [\mathbf{X}_1, \cdots, \mathbf{X}_N]^T$ is the set of all affine models to be estimated, $\mathbf{u}_i$ is the nearest neighbour in the target scan for the warped template vertex $\mathcal{T}(\mathbf{v}_i)$ and $\mathbf{W} = \text{diag}(\mathbf{w})$ is a diagonal weight matrix. Similarly, we denote the net stiffness term as

$$C_{\text{stiff}}(\mathbb{X}) = \|(\mathbf{Q} \otimes \mathbf{G})\,\mathbb{X}\|_F^2 \quad (2)$$

where $\mathbf{Q}$ is the node-arc incidence matrix for the template scan and $\mathbf{G} := \text{diag}(1, 1, 1, \gamma)$ where γ is a parameter used to weight the linear and the translational components of the transformation differently. The solution for the non-rigid alignment $\mathbb{X}$ is obtained by the minimisation of the total cost, *i.e.*

$$\arg\min_{\mathbb{X}} C_{\text{fit}}(\mathbb{X}) + \alpha C_{\text{stiff}}(\mathbb{X}) \quad (3)$$

which is equivalent to solving a sparse linear system of equations. Much like the standard ICP algorithm, upon solving for the motion model $\mathbb{X}$ using Equation 3, we can update the correspondences $\mathbf{u}_i$. These steps are iterated till convergence. In our experiments, we also progressively anneal $\mathbb{X}$ by sequentially decreasing the stiffness weight α with values $\alpha = \{2000, 800\}$. The reader may refer to [1] for a detailed explanation of all of these terms.

Adaptive Weighting: In Equation 1 we provide a per-vertex weight value w_i that reflects our confidence in the alignment of vertex $\mathbf{v}_i$. Since we are interested in correcting for misalignments in high frequency details, we utilise the normal information available at all vertices. Specifically

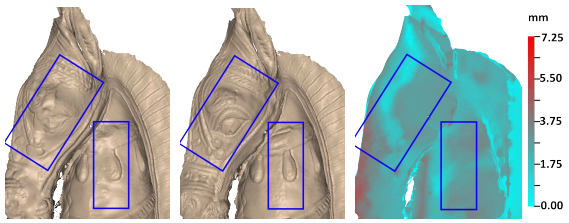
$$w_i = \begin{cases} \mathbf{n}(\mathbf{v}_i) \cdot \mathbf{n}(\mathbf{u}_i) & \text{if } \mathbf{n}(\mathbf{v}_i) \cdot \mathbf{n}(\mathbf{u}_i) \geq \delta \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

where $\mathbf{n}(\mathbf{v}_i)$ and $\mathbf{n}(\mathbf{u}_i)$ are the normals co-registered to the template vertex $\mathbf{v}_i$ and its nearest neighbour correspondence $\mathbf{u}_i$ in the target scan respectively. In Equation 4, while we provide a higher weightage to vertices with similar normals, we also ensure that vertices with dissimilar normals do not influence the non-rigid warping of the template scan. Such a weighting function ensures that our final non-rigid alignment has good agreement with both the low and high frequency components of the scans.

Global, multi-view non-rigid alignment: While the solution of Equation 3 allows us to non-rigidly align a pair of high quality scans, for our purposes we need to jointly align a set of such scans. Since we have already registered the scans using a rigid Euclidean motion model, we know the overlap relations between the entire set of scans. For global, multi-view non-rigid alignment, we carry out the following steps. Consider a template scan k and let its overlapping set of target scans be $O(k)$. Now each target scan $l \in O(k)$ will induce a warp on vertex $\mathbf{v}_i$ in scan k as $\mathcal{T}^l(\mathbf{v}_i^k)$ where $\mathcal{T}^l$ denotes that the warp is into the target scan l and $\mathbf{v}_i^k$ denotes the i-th vertex in the template scan k. However, since each target scan $l \in O(k)$ will "pull" vertex $\mathbf{v}_i^k$ towards itself, we need to warp $\mathbf{v}_i^k$ in a manner most consistent with all the individual warps. In our approach, we use a weighted average of all the individual warped points $\mathcal{T}^l(\mathbf{v}_i^k)$, *i.e.* the updated vertex $\mathbf{v}_i^{'k}$ is given as

$$\mathbf{v}_i^{'k} = \left(\sum_{l \in O(k)} P_i^l \; \mathcal{T}^l(\mathbf{v}_i^k) \right) / \left(\sum_{l \in O(k)} P_i^l \right) \quad (5)$$

where the adaptive weights $P_i^l = \frac{1}{||\mathcal{T}^l(\mathbf{v}_i^k) - \mathbf{v}_i^k||_2 + \eta}$ for some positive η such that when a particular scan l pulls the vertex $\mathbf{v}_i^k$ by a large amount towards itself, we reduce its influence so as to keep the warped vertex as consistent with the remaining scans in $O(k)$ that have a better fit. We note



(a) W/O non-rigid alignment (b) With our non-rigid alignment (c) Non-rigid misalignment map

Figure 3: Zoomed-in views of final reconstructions of a clay horse model: (a) without non-rigid alignment and (b) with our global, multi-view, non-rigid alignment. (c) Colour map of globally estimated mean non-rigid misalignments across all views. It is evident that the input data has significant non-rigid misalignments that are accurately corrected for by our approach. Please view this figure in colour.

here that one may prefer that the global, multi-view non-rigid alignment solution be such that all warped scans are fully consistent among themselves, *i.e.* the warped individual vertices are co-incident. While that requires an expensive iterative process, we find that for all our experiments a single iteration multi-view warp of Equation 5 is adequate to correct for the non-rigid misalignments present in the full set of scans.

The value of our non-rigid alignment procedure is illustrated in Figure 3 by comparing the reconstructions of a clay horse model. Figure 3(a) shows the result obtained by our pipeline without any non-rigid alignment. Figure 3(b) shows the result obtained when we apply our global, multi-view non-rigid alignment to the individual depth scans. Figure 3(c) shows the colour map of globally estimated non-rigid misalignments across all views. The mean and maximum warps for the 3D depth scans are 0.74 mm and 7.25 mm respectively. The equivalent 2D warp induced on the normal maps have a mean and maximum values of 0.27 pixels and 6.04 pixels respectively. It is evident that there is a significant amount of non-rigid misalignment present in the original data and that our approach accurately recovers the non-rigid alignments involved. As a result, our approach is able to remove the visible seam-line artefacts in Figure 3(a) and preserve fine-scale 3D details in the final reconstruction as shown in Figure 3(b).

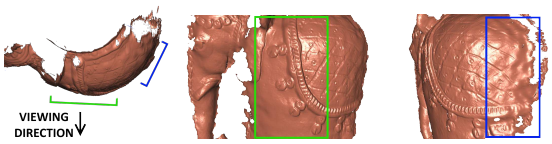
2.1. Non-rigid in-plane 2D warping of normal maps

Since we are applying a non-rigid warp to the depth scans, we have to also correspondingly warp the 2D normal maps so that they remain co-registered. Recall that we originally had co-registered depth and normal maps, *i.e.* in each scan, for every un-warped initial vertex $\mathbf{v}_i$, we assign a corresponding location $\mathbf{p}_i$ in the corresponding 2D normal map using the pin-hole camera equation $\mathbf{p}_i \equiv \mathbf{K}(\mathbf{R}\mathbf{v}_i + \mathbf{t})$, where $\mathbf{K}$ and $\{\mathbf{R}, \mathbf{t}\}$ are respectively the intrinsic and extrinsic calibration parameters of the Kinect's IR camera for a given scan. Note that although the depth and normals are obtained in the same co-ordinate system, for a global representation, we need to represent all scans and cameras in a common global frame of reference. Since the non-rigid warp on vertex $\mathbf{v}_i$ has moved it to $\mathbf{v}'_i = \mathbf{v}_i + \Delta\mathbf{v}_i$, the projected position of the corresponding normal map has also moved to a new location as $\mathbf{p}'_i \equiv \mathbf{K}(\mathbf{R}\mathbf{v}'_i + \mathbf{t})$, *i.e.* the 2D motion for the normal position is $\Delta\mathbf{p}_i = \mathbf{p}'_i - \mathbf{p}_i$. Using all the individual $\{\Delta\mathbf{p}_i\}$ shifts of normal positions, we can warp the normal map so as to account for the non-rigid 3D alignment. However instead of using a simple bilinear or cubic interpolation for warping, we use a bilateral-weighted kernel to ensure the accurate preservation of boundaries and edges in the warped normal map.

3. Graph-cut Based Adaptive Normal Selection

After applying the estimated global, multi-view non-rigid warps on the individual depth scans, we can obtain a single complete 3D mesh representation of the object. In our pipeline we use the popular volumetric merging approach using TSDF [7]. However, while our non-rigid warping removes the misalignments present in the individual depth maps, we still need to use the high quality normal information to add fine-scale 3D detail to the final mesh representation. In other words, we now need to associate a single normal vector to each vertex in the volumetric merged mesh. Since each vertex in the merged mesh is visible from multiple viewpoints or depth camera positions, we have multiple normals associated with it. An intuitive approach would be to take the average of these multiple normals and assign it to the mesh vertex. However, owing to the presence of residual positional errors as well as errors in the normal orientations themselves, a naive averaging is not desirable since it would result in blurring of the fine-scale details present in the individual normal maps.

In [2], the authors address the estimation of the final mesh vertex normals by combining the associated normals in a weighted fashion. Their cost function consists of a) a data term that measures the weighted difference between the estimated normal and that of the corresponding normal observations in the different normal maps and b) a smoothness penalty that compares the local gradient of the estimated normals with that in the corresponding locations in the 2D normal maps. This approach allows for a smooth estimation of normals, thereby reducing the effect of viewpoint transition artefacts. In [9], instead of averaging the available normals, the authors select one of the available normals and assign it to the mesh vertex. In their heuristic approach, the individual scans are ordered in a priority sequence and for each mesh vertex, amongst the available nor-



(a) Top view (b) Frontal view (c) Right view

Figure 4: Quality of photometric normals with varying viewing angle mapped to the corresponding 3D mesh. The top view in (a) shows the regions marked. The regions which are oblique to the viewing direction (marked in blue in (c)) have poorer quality of normals as compared to the regions in the direct front (marked in green in (c)). Please view this figure in colour.

mals, the one from the highest priority scan is selected without any smoothness constraint imposed on the adjacent normals. As we shall show in detail later, this approach results in poor normal selection since the fixity of the priority order results in unreliable normals that are almost orthogonal to the viewing direction being selected. More importantly, both approaches don't account for the non-rigid misalignments. Thus the uncorrected misaligned normals results in blurring of high frequency details in [2] and seam-line artefacts in [9].

Like the approach in [9] we also prefer to select a single normal from the available possibilities instead of averaging them. We choose to do so since there is always a certain amount of residual alignment error despite our global, multi-view non-rigid refinement step and even any intelligent averaging of normals will result in a loss of high frequency 3D detail. However, instead of a per-vertex normal selection as in [9], we jointly solve for the normal selection as a graph-cut based minimisation of a global cost function over the entire mesh. We desire our cost function to take into account: a) the amount of the non-rigid warp in our non-rigid refinement step of Section 2 and b) the viewing angles of the candidate normals in the different views. In other words, if a vertex has a higher non-rigid warp magnitude, its associated normal is deemed to be correspondingly less reliable. Additionally, if a vertex lies in a region which is highly oblique to the viewing angle, then the associated normal is deemed to be less accurate. We illustrate this scenario in Figure 4 which shows the quality of photometric normals mapped onto the corresponding single-view 3D mesh of a clay pot with an elephant motif. In the top view shown in Figure 4 (a) we have marked two different regions. The corresponding regions are also marked and shown in two convenient orientations in Figure 4 (b-c). It is clear that the quality of the normals obtained at the regions (marked in green and shown in Figure 4 (b)) which is frontal to the view is much higher than those on right side regions (marked in blue and shown in Figure 4 (c)).

Thus in our global cost function we incorporate a data term capturing the relative reliability of the normals and a smoothness term ensuring that the selection is smooth across regions. Our overall cost function is given as

$$C\left(\mathbf{L}\right) = \sum_{i=1}^{N} \left\| \left(\kappa - \mathbf{n}_i \cdot \mathbf{e}^{l_i}\right) \mathbf{f}_{i,l_i} \right\|_2^2 + \sum_{i=1}^{N} \sum_{j \in \mathcal{N}(i)} h\left(l_i, l_j\right) \tag{6}$$

where $\mathbf{L} = \begin{bmatrix} l_1 & l_2 & \cdots & l_N \end{bmatrix}^T, l_k \in \{1, 2, \cdots, K\}$ is the selection label vector that indicates the view from which a corresponding normal is chosen. N denotes the number of mesh vertices and K denotes the number of viewpoints from which the normal maps are available. κ is a positive constant in the range $(1, 2]$. Here, $\mathbf{n}_i$ is the normal at vertex $\mathbf{v}_i$ estimated using the mesh neighbourhood around $\mathbf{v}_i$. Further, $\mathbf{e}^{l_i}$ is the viewing direction for the l_i-th mesh and $\mathbf{f}^{l}_{i,l_i}$ is the magnitude of the non-rigid warp with respect to the l_i-th view. Since our solution is a label defining the selected viewpoint for each normal, we incur a smoothness penalty only when we make a transition from one viewpoint label to another. Therefore, we define the smoothness penalty $h\left(l_i, l_j\right) = \lambda \cdot \mathbf{1}\left(l_i \neq l_j\right)$ for $j \in \mathcal{N}(i)$ where $\mathbf{1}\left(\cdot\right)$ is the indicator function, $\mathcal{N}(i)$ is a neighbourhood of i and λ is a positive constant. Apart from its ability to preserve high frequency information in the normal map, our selection approach has the additional advantage that we can very efficiently solve for a global selection map using the graph-cut method proposed in [3].

We now demonstrate the advantages of our graph-cut based adaptive normal selection method by considering a simple synthetic example as shown in Figure 5 and comparing our result with the approach of **MERGE2-3D** [2] and the priority ordering approach of [9]. Three views ($40°$ apart) of deformed normals maps are synthetically generated from a sphere of radius 100mm with ridges of height 5mm and are fused on smoothed depth-maps of the sphere from the respective viewpoints as mentioned in [9]. The respective non-rigid 3D deformation magnitude maps are shown in the second, third and fourth columns of first row in Figure 5. We run the **MERGE2-3D** [2], Priority Selection (**PS**) [9] and our global, non-rigid refinement and adaptive normal selection steps for the final multi-view fusion and compare their performances. Specifically, we measure the mesh-to-mesh ℓ_2 distance in the overlapping region. The second and third rows in Figure 5 show the distance maps and the 1-d horizontal profiles along the highlighted blue lines for the initial mesh and outputs of the three methods. While clearly the maximum distance in the initial smooth mesh (first column) is at the sides of the ridges, the Priority Selection (second column) results in high amount of estimation error in the left half of the output. This is due to an assigned higher priority of the frontal view with high non-rigid deformation despite the availability of better nor-

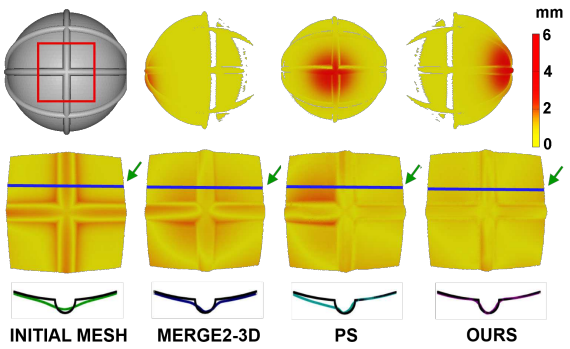
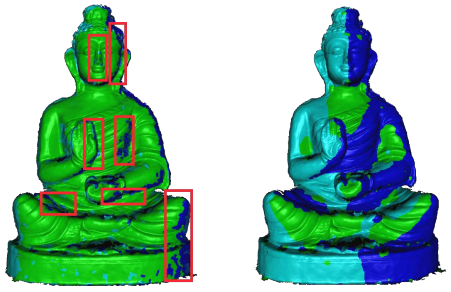


Figure 5: Synthetic sphere with ridges for 3 normal maps. Top row: Columns show the ground truth model and the initial error maps (in mm) of the single-view fused depth maps obtained from normal maps from 3 different viewpoints 40° apart respectively. Middle row: Columns show the error maps (in mm) of initial smooth reconstruction and the refined reconstructions obtained using **MERGE2-3D** [2], Priority Selection (**PS**) [9] and our method respectively. Bottom row: Shows the respective 1-d horizontal profiles (non-black coloured) along the highlighted blue lines against the ground truth (black coloured). Please view this figure in colour.

mals from the left view. The output (third column) from **MERGE2-3D** results in blurring of the ridges, hence high error in those regions. However, the error is minimum in the output of our method with more precise estimation of the ridges.

We now demonstrate the smoothness of the solution from our normal selection method. In Figure 6 we show the normal selection index as a colour map for a Buddha figurine. In this example, we consider selecting a normal from a set of three views that are coloured as cyan, green and blue corresponding to the left, frontal and right views respectively. In Figure 6 (a) we can notice that the simple priority ordering heuristic results in labels that randomly alternate between adjacent mesh vertices which results in a diffusion of errors during the fusion step. As a result there is a corresponding loss of accuracy in preserving high frequency 3D detail. In contrast, as shown in Figure 6 (b) our solution promotes spatial continuity of the selection labels and we do not suffer from arbitrarily alternating labelling. It will also be noted in our solution that since the Buddha head and the base are nearly symmetric, our selection label is also almost symmetrically (vertically) split and each side of the head and the base are assigned to the left and right views. However, when the normal orientation is away from the viewing direction, then the label is switched to a better conditioned viewing direction. This is evident, for instance, if we consider the green patches on the left upper arm and hand of the Buddha.



(a) Priority Selection of [9] (b) Our Graph-cut Solution

Figure 6: Comparison of normal selection labels on a Buddha figurine for 3 normal maps from the left (cyan), central (green) and right (blue) views. a) shows the solution label map from [9]; b) shows our graph-cut based label map. Please view this figure in colour and see text for details.

Method	**Type**	**Minimisation Cost**	**Non-rigid Refinement**
Averaging (**AVG**)	EST	Squared norm	×
Weighted Average (**MERGE2-3D**) [2]	EST	Direction + Gradient Smoothness	×
Priority Selection (**PS**) [9]	SEL	×	×
OURS	SEL	Direction+Non-rigid + Selection Smoothness	✓

Table 1: Comparison of attributes of normal estimation in different approaches. (SEL - selection and EST - estimation)

Mesh-Normal Fusion: Once we obtain the high quality optimal normals from the previous step, we use the mesh normal fusion [14] to obtain the final results.

4. Results

4.1. Synthetic Datasets

In this Section, we evaluate the efficacy of our method compared to some other approaches. We consider the synthetic sphere with ridges and the Bunny and the Armadillo from the Stanford 3D Scanning Repository as our objects of interest. We resize the maximum dimension of each of them to 200mm to emulate a real world situation. We generate synthetic co-registered smooth depth maps and normal maps of the objects from 12 known viewpoints differed by 30° rotations about the Y-axis. However, each of the normal maps are synthetically obtained from non-rigid deformed versions of the true mesh. The deformations are restricted to 5% of the dimensions of the object. Thus, although the normals maps contain high quality details, they have non-rigid deformation. We then run our non-rigid refinement

method, adaptive normal selection and fusion steps on these input data and obtain the final multi-view 3D reconstructed models, denoted as **OURS**. We then compute the mesh-to-mesh ℓ_2 distances of the outputs from the ground truths. We compare our results with three approaches. The first approach is that of a simple averaging of all available normals for a given mesh vertex, denoted as **AVG**. The second method is of [2] that we denote as **MERGE2-3D**. The third approach we compare with is the priority scheme for normal selection described in [9], denoted as **PS**. Of these methods, **AVG** and **MERGE2-3D** estimate the mesh-vertex normals by averaging of available normals, whereas **PS** and **OURS** are normal selection schemes. Unlike our method, none of the other methods estimate any non-rigid alignment or correction step. While Table 1 summarises the comparative attributes of all these four methods, Table 2 shows the comparisons of the errors in the output from all these methods.

Model	**Reconstruction Errors (Mesh-to-mesh ℓ_2 distance)**			
	AVG	**MERGE2-3D** [2]	**PS** [9]	**OURS**
Sphere w/ridges	0.451	0.630	0.450	**0.418**
Bunny	0.498	0.710	0.531	**0.497**
Armadillo	0.325	0.374	0.387	**0.296**

Table 2: Comparison of performance of our method **OURS** with **AVG**, **MERGE2-3D** [2] and **PS** [9] on synthetic datasets with 12 views.

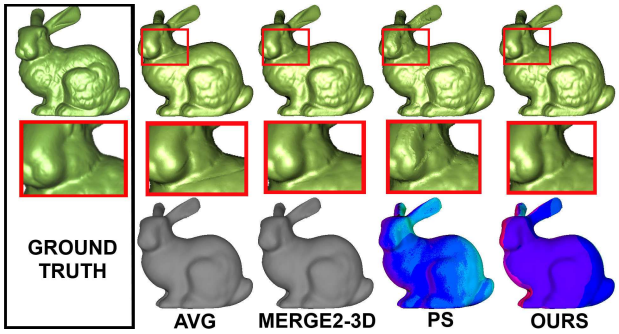


Figure 7: Comparative results on the Bunny model obtained from 12 views. The columns correspond to ground truth, **AVG**, **MERGE2-3D** [2], **PS** [9] and our method **OURS** respectively. The first and second rows show overall views and zoomed-in views of a part of the 3D reconstruction. The third row shows the normal selection labels for **PS** and **OURS**. **AVG** and **MERGE2-3D** are indicated in grey since they average normals. Please view this figure in colour and see text for details.

From Table 2, it can be observed that the outputs of our method have the minimum errors. An interesting observation is that the **AVG** performs better than **PS** and **MERGE2-3D**. This is because **PS** prioritises the better views for selecting the normals but only on a per-view basis and **MERGE2-3D** blindly depends on the viewing direction without considering the reliability of the normals due to non-rigid deformations. On the other hand, **AVG** performs a blind uniform averaging of all available normals which for the synthetic datasets, enables in reducing the deformations to some extent. Figure 7 shows the results for the 3D reconstruction of the Bunny model for 12 views. The columns correspond to the methods **AVG**, **MERGE2-3D**, **PS** and **OURS** respectively. The first row shows the respective 3D reconstructions and the second row shows a label map for normal selection. Since both **AVG** and **MERGE2-3D** average the normals instead of selection, we indicate them as grey images, *i.e.* these two methods have no label selection. In the first row, it can be observed that the reconstructions of **AVG**, **MERGE2-3D** and **PS** contain non-rigid artefacts in the regions highlighted with red boxes while our method recovers the surface more accurately. It can be also noted from the second row of Figure 7 that the non-smooth nature of normal selection of **PS** at the viewpoint transitions leads to large amount of artefacts in its output. For more results on synthetic data, please refer to the supplementary material for this paper.

4.2. Real Datasets

In this Section, we present some results on real datasets. For our input depth maps, we use a version 1 Kinect (structured-light stereo) and also use the IR camera of the Kinect to obtain high quality photometric normals as described in [9], *i.e.* for each viewpoint we obtain co-registered depth and normal estimates. All the data from the multiple viewpoints are registered using a multi-view ICP method, resulting in the rigid Euclidean alignment of the depth and normal maps. We then run all the four methods as before, *i.e.* **OURS**, **AVG**, **PS** and **MERGE2-3D**.

In Figure 8, we present results for the 3D reconstruction of two objects, *i.e.* a clay horse using depth and normal observations from 6 viewpoints shown in the first four columns and a clay pot with an elephant motif viewed from 4 positions shown in last four columns. We also note that while the objects in our experiments used are of single albedo, for multi-albedo objects we could use the approach [5] to recover the photometric normals. The columns correspond to the methods **AVG**, **MERGE2-3D**, **PS** and **OURS** respectively. The first row shows the respective 3D reconstructions and the second row shows a zoomed-in detail of the reconstructions. The third row shows a label map for normal selection. Since both **AVG** and **MERGE2-3D** average the normals instead of selection, we indicate them as grey images, *i.e.* these two methods have no label selection. It will be immediately obvious that the naive averaging method (**AVG**) ends up blurring fine-

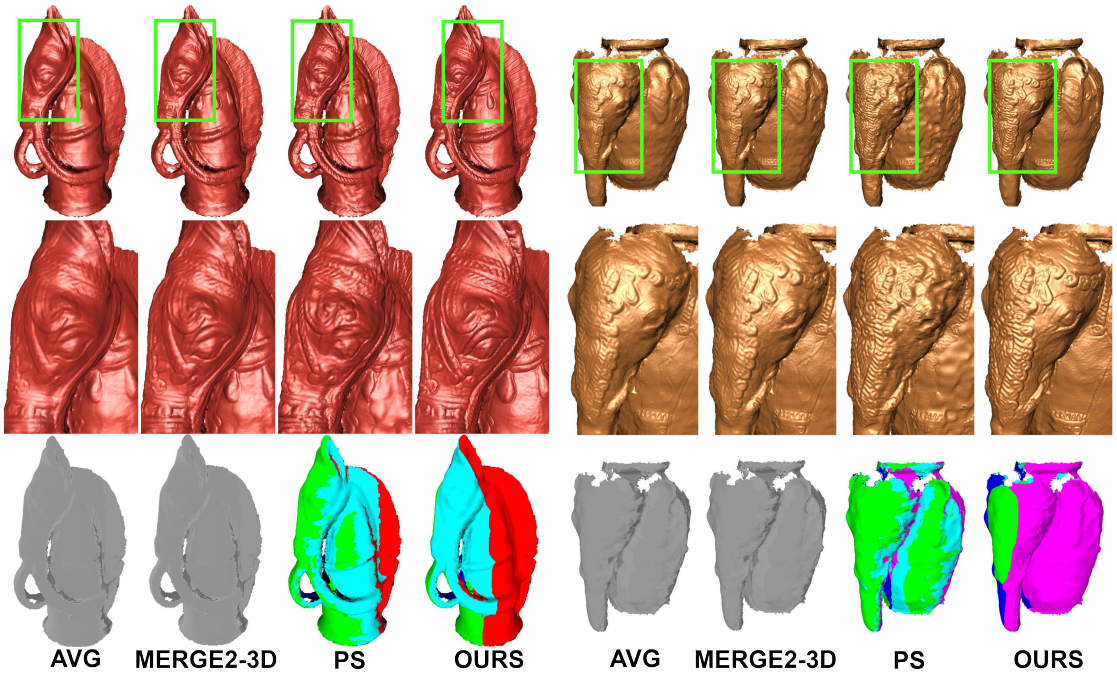


Figure 8: Comparative results on a horse model and a pot with an elephant motif obtained from 6 views and 4 views respectively. The columns correspond to the **AVG**, **MERGE2-3D** [2], **PS** [9] and our method **OURS** respectively. The first and the second rows show overall views and zoomed-in views of a part of the 3D reconstructions. The third row shows the normal selection labels for **PS** and **OURS**. **AVG** and **MERGE2-3D** are indicated in grey. Please view this figure in colour and see text for details.

scale 3D surface features. Unlike the synthetic cases in Section 4.1, the **MERGE2-3D** method does better than **AVG** as in real datasets, the quality of normals at oblique angles from viewing direction is extremely poor and **MERGE2-3D** does a more intelligent averaging of the normals, but it can preserve surface details only upto an intermediate scale. This is largely due to the fact that **MERGE2-3D** does not account for the non-rigid misalignments and also carries out an averaging of misaligned normals resulting in a loss of detail. In contrast, since our approach corrects for the non-rigid misalignments and also selects a single normal out of many choices, it is able to preserve high frequency 3D surface detail at the finest scale. This is particularly evident if we compare the details of the eye and other parts of the horse and the elephant heads shown in the middle row of Figure 8. From the normal selection map shown in the third row of Figure 8, we notice that the priority selection method (**PS**) of [9] results in poor normal selection. In particular we note that green labels on the eye and the green patch on the sides of the horse neck and the clay pot. The surface normals in these regions are oriented away from the viewing direction selected, with the result that the selected normals are of poor quality. Consequently, there are strong distortions present in the final reconstruction. In contrast, our graph-cut based solution for the label map correctly selects the normals from the appropriate and desirable viewing directions, thereby ensuring that the normals are of good quality while also avoiding undesirable artefacts. For more results on real data, please refer to the supplementary material for this paper.

5. Conclusion

We have introduced a novel global, multi-view method for non-rigid refinement of 3D meshes with corresponding high quality 2D normal maps. We have also introduced a well-motivated normal selection scheme that can be efficiently solved using graph-cuts. Taken together, non-rigid alignment and normal selection result in high quality 3D reconstruction where both low and high frequency information is preserved.

Acknowledgement

This work is supported in part by an extramural research grant by the Science and Engineering Research Board, DST, Government of India. Sk. Mohammadul Haque is supported by a TCS Research Scholarship.

References

[1] B. Amberg, S. Romdhani, and T. Vetter. Optimal step nonrigid icp algorithms for surface registration. In Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on, pages 1–8. IEEE, 2007.

[2] S. Berkiten, X. Fan, and S. Rusinkiewicz. Merge2-3d: Combining multiple normal maps with 3d surfaces. In 3D Vision (3DV), 2014 2nd International Conference on, volume 1, pages 440–447, Dec 2014.

[3] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 23(11):1222–1239, 2001.

[4] B. J. Brown and S. Rusinkiewicz. Global non-rigid alignment of 3-d scans. ACM Transactions on Graphics (TOG), 26(3):21, 2007.

[5] A. Chatterjee and V. M. Govindu. Photometric refinement of depth maps for multi-albedo objects. In Computer Vision and Pattern Recognition (CVPR), 2015 IEEE Conference on, pages 933–941. IEEE, 2015.

[6] G. Choe, J. Park, Y.-W. Tai, and I. S. Kweon. Exploiting shading cues in kinect ir images for geometry refinement. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, pages 3922–3929, 2014.

[7] B. Curless and M. Levoy. A volumetric method for building complex models from range images. In Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, pages 303–312. ACM, 1996.

[8] Y. Han, J.-Y. Lee, and I. S. Kweon. High quality shape from a single rgb-d image under uncalibrated natural illumination. In Proceedings of IEEE International Conference on Computer Vision (ICCV), 2013.

[9] S. M. Haque, A. Chatterjee, and V. M. Govindu. High quality photometric reconstruction using a depth camera. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2275–2282, 2014.

[10] P. Henry, M. Krainin, E. Herbst, X. Ren, and D. Fox. Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments. International Journal of Robotics Research, 31(5):647–663, Apr. 2012.

[11] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, et al. Kinectfusion: real-time 3d reconstruction and interaction using a moving depth camera. In Proceedings of the 24th annual ACM symposium on User interface software and technology, pages 559–568. ACM, 2011.

[12] F. Langguth, K. Sunkavalli, S. Hadap, and M. Goesele. Shading-aware multi-view stereo. In European Conference on Computer Vision, pages 469–485. Springer International Publishing, 2016.

[13] H. Li, R. W. Sumner, and M. Pauly. Global correspondence optimization for non-rigid registration of depth scans. Computer graphics forum, 27(5):1421–1430, 2008.

[14] D. Nehab, S. Rusinkiewicz, J. E. Davis, and R. Ramamoorthi. Efficiently combining positions and normals for precise 3D geometry. ACM Transactions on Graphics (Proceedings of the ACM SIGGRAPH 2005), 24(3):536–543, 2005.

[15] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[16] R. Or El, G. Rosman, A. Wetzler, R. Kimmel, and A. M. Bruckstein. Rgbd-fusion: Real-time high precision depth recovery. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015.

[17] L. F. Yu, S. K. Yeung, Y. W. Tai, and S. Lin. Shading-based shape refinement of rgb-d images. In Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, pages 1415–1422, June 2013.

[18] Q. Zhang, M. Ye, R. Yang, Y. Matsushita, B. Wilburn, and H. Yu. Edge-preserving photometric stereo via depth fusion. In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pages 2472–2479. IEEE, 2012.