A robust pipeline for rapid feature-based pre-alignment of dense range scans

Anonymous ICCV submission

Paper ID 1435

Abstract

Aiming at reaching an interactive and simplified usage of high-resolution 3D acquisition systems, this paper presents a fast and automated technique for pre-alignment of dense range images. Starting from a multi-scale feature point extraction and description, a processing chain composed by feature matching and correspondence searching, ranking grouping and skimming is performed to select the most reliable correspondences over which the correct alignment is estimated. Pre-alignment is obtained in few seconds per million point images on a off-the-shelf PC architecture. The experimental setup aimed to demonstrate the system behavior with respect to a set of concomitant requirements and the obtained performance are significant in the perspective of a fast, robust and unconstrained 3D object reconstruction.

1. Introduction

Acquisition of multiple scans from different viewpoints is the first step of a wide class of 3D object modelling pipelines. At some early stage, after the acquisition, each dataset (e.g. range image or point cloud) generated by a 3D scanning device (e.g. a laser or structured light optical scanner) should be accurately aligned (or coregistered) in a common coordinate system. The quality of this alignment strongly influences the subsequent object modelling steps in which the aligned dataset is fed to a surface reconstruction technique (see for example [2, 12, 14]).

Multiple scan alignment can be conceptually split in two different problems: 1) independent scans must be roto-translated into a common reference system, and 2) they should be accurately coregistered. These two problems, which are usually referred to as *coarse* and *fine* alignment, are different in nature and require distinct solving approaches. In this work, we focus on the coarse alignment problem (that is, to find a common reference system). Being the first step of a modelling chain, its performance are the most critical from the point of view of error propagation throughout the 3D modelling chain. In particular, even if called coarse, a certain degree of accuracy is strongly required for the success of the subsequent fine alignment. In fact, it is well known that, for fine alignments, classic solutions (either pairwise, e.g. ICP [3] and its variants [19], or global, e.g. [18],[13]) are based on optimization routines which often suffer from local minima problems or position ambiguities which should be maximally reduced by proper initialization.

Regarding datasets, state-of-art optical scanning devices have increased in the last years their spatial resolution as well as other acquisition performance (accuracy, acquisition time,...), and their usage is expected to be more and more unconstrained, toward devices that can be easily used like a digital camera. This would be suitable in response to an increasing demand of "3D" either in today professional applications (industry, biomedicine, cultural heritage,...) as well as for the expected increment of 3D contents of future web applications. Now, despite the coarse alignment problem has been long studied and several solutions have been proposed (some representative works are cited in Sec. 1.1), the reference applications are more and more demanding and require solutions that satisfy at the same time all these emerging requirements (in Sec. 1.2 we will better define the multiobjective problem we want to tackle). We therefore observe and believe that "high-performance coarse alignment is an open problem and is stille demanding for research efforts and effective solutions.

1.1. Related work

The coarse registration problem has been extensively studied, and several methods can be found in literature. Many of them can be reconducted to one of the two main philosophies that have emerged during these years, i.e. with or without the exploitation of feature descriptors.

The first approach exploits the ever-increasing computational capabilities of modern calculators to find, within a large solution space, the affine transform that best aligns two views. The main advantage of the techniques which fall into this category is that they are independent from the data given as an input and more robust to noisy data. On the other hand, they are usually computationally expensive. The progenitor of this family is considered to be the

RANSAC, devised by Fischler and Bolles [8]. During the years, improvements to this algorithm have been proposed in order to reduce the computation time, also by exploiting point neighborhood descriptors [6],[1]. A second approach for coarse registration relies on the extraction and subsequent matching of global (e.g. spin images [11]) or local shape descriptors. Advantages with respect to brute-force approaches are mainly related to computational gain achieved through a selective choice and skim of descriptive features. On the other hand, they usually fail in describing featureless (at some scale) surfaces, and are quite sensitive to noise. Multi-scale feature based approaches (also used in this work) allow a better adaptation to different kind (and dimesion) of object features. Related works are those presented by Li and Guskov [16] and Lee *et al*. [15] which introduced extensions of Lowe's 2D SIFT [17] to 3D datasets. Their approach has subsequently been exploited by Castellani *et al*. [5]. Thomas and Sugimoto [20] proposed to use the reflectance properties for images registration to better work with featureless images.

An important choice within the described approach regards the feature descriptor to be employed. An ideal descriptor should associate an unambiguous signature for each feature, fast to compute, robust to rotation of viewpoint and to variations of point density for the image. For range images, Li and Guskov [16] proposed a descriptor based on a combination of Discrete Fourier transform and Discrete Cosine transform to describe the neighborhood of each feature point. Gelfand *et al*. in [9] proposed the use of volumetric descriptors, that is the estimation of the volume portion inscribed by a sphere centered at some points belonging to the surface. Castellani *et al*. [5] proposed a statistical descriptor based on hidden Markov chain that is trained through its neighborhood.

1.2. Problem definition and requirements

In this paper we wish to address the problems related to an unconstrained usage of modern, highly resoluted acquisition devices, capable of granting superior accuracy performances. With the term unconstrained usage we intend that the operator is given the liberty to choose the acquisition path he prefers to follow during the scanning phase, thus free from any constraint such as positioning the scanning device at predetermined positions or angles. The only requirement that we still need to maintain is that each image has a certain degree of overlap with respect to the rest (at least with one of the other scans). In practice, however, this constraint is always fulfilled, since whenever multiple views are required to acquire the area of interest, the operator is implicitly required to plan a suitable acquisition path. This is also a prerequisite for the subsequent steps of the modelling chain, such as fine alignment and surface extraction. We can therefore assume that we are given a set of scans that follow an acquisition path for which each image presents an overlap area with respect to the previous one.

With our work we would like to fulfill a number of requirements, which we now briefly describe. First of all, we would like our pipeline to be equally effective regardless to the nature of the acquired object (industrial, artistic, and so on) as well as its size. The developed solution should also be fast: ideally it should allow an interactive usage, which means that the alignment is performed while the operator varies the scanner (or object) position in order to acquire the next scan. Accuracy would of course be a desirable property as well, however since we are addressing a coarse alignment procedure, care must be put when defining what accuracy means in this context. In fact, the objective of coarse alignment is to approximately register a couple of views, so that a subsequent procedure of fine alignment (such as ICP, for example) is capable to convergence to the optimal alignment, without getting stuck in a local minima of the error function.

We focus our attention on pairwise alignment since it constitutes the fundamental block of any progressive approach (that is, align one scan with respect to the ones that have been already successfully aligned), as well as for registering images that do not belong to the acquisition path.Our solution consists in a pre-alignment pipeline (described in detail in Sec.2) that has been specifically designed to fulfill all the requirements previously stated. Main contributions of this paper are:

- A complete and fully functional pipeline for range images alignment;
- A lightweight feature signature devised in order to quickly reduce the matches space;
- A matching chain developed to progressively skim the correspondence space.

1.3. Notation

A range image can be conceived as the projection of a 2D image grid on a 3D target object surface and the acquisition of depth related information from that surface. The resulting dataset is a "structured" point cloud, that is a number of points lying in a 3D space, and associated to a pixel of acquisition grid. We define a range image as a map $I \in \mathbb{Z}^2 \rightarrow R \in \mathbb{R}^3$, where the domain I is a rectangular grid (usually corresponding to the CCD matrix), and the co-domain R corresponds to the set of 3D points representing the acquired surface. Because of the acquisition's nature (measure range limitations, occlusions due to the object shape, etc.), not all pixel positions $i \in I$ may have a valid corresponding point $\boldsymbol{p}_i \in R$, therefore only a subset $I_V \subseteq I$ of valid points is acquired for each image. We take advantage of range images data structure in order to speed up the processing: in particular, by exploiting the image domain I, neighborhood information can be retrieved quickly

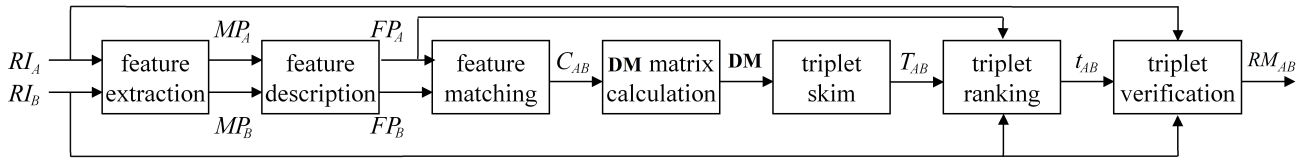

Figure 1: Block diagram representing the proposed system

and efficiently, while data processing is performed over 3D target space R.

2. The proposed pairwise alignment pipeline

We describe in detail all blocks of Fig.1 which contribute to the automatic alignment of two given range images RI_A and RI_B. Aiming at a substantial reduction of the problem dimensionality, solutions based on the exploitation of distinctive features detected through automatic analysis of the acquired views appear particularly interesting. However, irregularities and "holes" that may be present over the scan (due to out-of-range measures, borders and line-of-sight occlusions) have a critical impact on the repeatability of features detected over scans taken from different viewpoints, thus potentially severely degrading the performance of such approaches. Notwithstanding, we found the multiscale feature extraction method of Bonarrigo *et al.* [4] particularly suited to our objectives (resilience to the above degradations and computational efficiency), and therefore we implemented it as a first step of our pipeline (Sec.2.1). However, [4] doesn't suggest any feature description, so from this point on we proceed with our original contribution. Following the pipeline of Fig.1, in Sec.2.2 we introduce a feature descriptor that is at the same time representative and cheap to compute, specifically conceived to be invariant with respect to any Euclidean transformation that may be applied to the scans. The matching process between two of these signatures is described in Sec.2.3. Next, a computationally effective search for reliable correspondences between features is described in Sec.2.4 and is articulated in several substeps with the objective to progressively skim entries that are considered unlikely or incoherent. At first this is done on single correspondences, next triplets of correspondences are considered and classified in order to select a small set of them over which the pre-alignment transformation is estimated.

2.1. Feature extraction

As stated, our feature extraction technique builds on [4], which we briefly resume for the sake of completeness. Their approach can be thought as an extension of the Lowe's SIFT approach [17] to 3D point data according to the following steps: a) given a range image RI, M filtered images $G(r)$, at scales $r \in [1, M]$, are derived by applying Gaussian kernels of growing dimension; b) a set of $R-1$ saliency maps $S(r)$ are derived from pairs of $G(r)$ at consecutive scales, from which they identify a set of feature points that are provided with the information of scale at which each feature point has been detected. We now assume that we are given these set of feature points, and propose to characterize each feature through a signature W_f (as described in Sec.2.2) computed exploiting the feature point's neighborhood.
To produce the filtered images $G(r)$ at various scales $r \in [1, M]$, a first unconstrained geometric Gaussian filtering on valid points $\boldsymbol{p}_i$ of the RI is done, obtaining $\boldsymbol{p}_i^g(r)$:

$$\boldsymbol{p}_i^g(r) = \frac{\sum_{\boldsymbol{p}_j \in B_{2\sigma_r}(\boldsymbol{p}_i)} \boldsymbol{p}_j \cdot e^{-\frac{\|\boldsymbol{p}_i - \boldsymbol{p}_j\|^2}{2\cdot\sigma_r^2}}}{\sum_{\boldsymbol{p}_j \in B_{2\sigma_r}(\boldsymbol{p}_i)} e^{-\frac{\|\boldsymbol{p}_i - \boldsymbol{p}_j\|^2}{2\cdot\sigma_r^2}}} \quad (1)$$

where $B_{2\sigma_r}(\boldsymbol{p}_i)$ identifies the points within a distance $2\sigma_r$ from $\boldsymbol{p}_i$. The effect of the geometric processing (1) is well balanced only if one can assume that the position of the points $\boldsymbol{p}_j \in B_{2\sigma_r}(\boldsymbol{p}_i)$ is regularly distributed over the object surface. However, despite the regularity of the acquisition domain I, this assumption is in general not true, as Fig.2 illustrates. Therefore, when $B_{2\sigma_r}(\boldsymbol{p}_i)$ contains non uniform point distributions with respect to the surface, 1 tends to generate a positional bias of the filtered points $\boldsymbol{p}_i^g(r)$. Points close to borders and holes are similarly affected. To introduce resilience to the above distortions, $\boldsymbol{p}_i^g(r)$ are only allowed to move along the normal direction $\hat{n}_i$ associated to the original point $\boldsymbol{p}_j$ by the following projection (see again Fig.2):

$$g_i(r) = \boldsymbol{p}_i + \langle \boldsymbol{p}_i^g(r) - \boldsymbol{p}_i, \hat{n}_i \rangle \hat{n}_i \quad (2)$$

Now $G(r)$ is defined as the set of points $g_i(r)$, $i \in [1, |I_V|]$. As the kernel radius σ_r increases, details which size is smaller than σ_r are smoothed out from $G(r)$ and, when the kernel size doubles, computations are made on factor two subsampled images. Moreover, as strongly suggested in [4], both neighborhood scanning and subsampling are performed on the regular grid I, thus in a very fast way.

Once the filtered versions $G(r)$ have been calculated, unitary length normal vectors $\hat{n}_i(r)$ are recomputed, and $R-1$ saliency maps are derived. A saliency map is a 2D

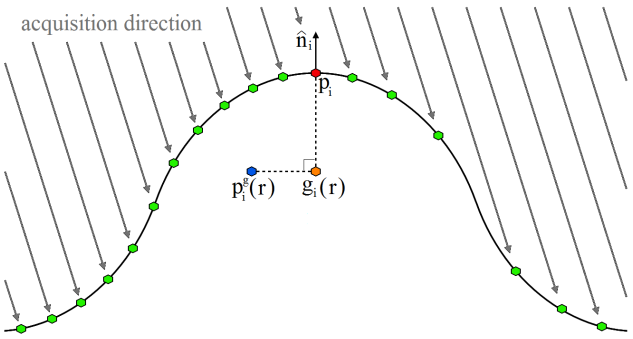

Figure 2: Original point $\boldsymbol{p}_i$ (red) is Gaussian filtered to get $\boldsymbol{p}_i^g(r)$ (blue), which is projected over $\hat{n}_i$ direction to get $g_i(r)$ (orange).

array of scalar values, obtained by pairwise subtraction of $G(r)$ at adjacent scales. This retains only the details comprised between the two bounding scales r and $r+1$, in other words it highlights features which dimension is comprised between two kernel sizes σ_r and σ_{r+1}. Saliency maps $S(r) = \{s_i(r)\}$ are actually calculated as follows:

$$s_i(r) = |g_i(r) - g_i(r+1)| \cdot \langle \hat{n}_i(r), \hat{n}_i(r+1) \rangle \quad (3)$$

where the correction factor $\langle \hat{n}_i(r), \hat{n}_i(r+1) \rangle$ have been introduced to better concentrate saliency over stable image points, i.e. points for which the normal direction $\hat{n}_i(r)$ doesn't vary too much across the scales. Subsequently, for each saliency map $S(r)$, its maximum values are located by an iterative search where, once the greatest valid saliency value for $S(r)$ is found, no other maximum can be selected within an invalidation neighborhood region $B_{2\sigma_{r+1}}(\boldsymbol{p}_i)$. This prevents from finding redundant overlapping feature points, as the greatest detail size that can be detected within $S(r)$ is σ_{r+1}. Each maximum is further tested in order to make sure that 1) its neighborhood is well defined (that is, it is not close to a border or hole, otherwise the associated feature descriptor would result incomplete); 2) it does not lie over a saliency ridge, because in such cases small variations in saliency estimation may cause great variations of feature position. Points $\boldsymbol{f}_{k,r,h}$ associated to the above maxima of the saliency map at scale r and associated to the kth range image RI_k, form the feature point set $F_{k,r}$, with $h \in [1, |FP_{k,r}|]$. This concludes our summary of what we implemented from [4]. Hereinafter, for a neater and more compact notation we will omit unnecessary indexes when things have general validity. For example, if we need to address a feature point, we will refer to it as a generic feature point $\boldsymbol{f}$.

2.2. Feature description

In order to search for correspondences between feature points belonging to different views, we need to define and use a viewpoint invariant signature. For each feature point $\boldsymbol{f}$, at some scale r dimension σ_r, we propose a novel descriptor computed exploiting both normal vectors and saliency data of its neighbor points $\boldsymbol{p}_j \in B_{\sigma_{r+1}}(\boldsymbol{f})$. To generate the descriptor, at first a reference system $\hat{x}_{\boldsymbol{f}}, \hat{y}_{\boldsymbol{f}}, \hat{z}_{\boldsymbol{f}}$, centered over the feature point $\boldsymbol{f}$, is constructed. $\hat{z}_{\boldsymbol{f}}$ is set toward the direction of $\hat{n}_{\boldsymbol{f}}$, while $\boldsymbol{P}_{\boldsymbol{f}} = span\{\hat{x}_{\boldsymbol{f}}, \hat{y}_{\boldsymbol{f}}\}$ is the tangent plane to $\hat{n}_{\boldsymbol{f}}$. Orientation of axis $\hat{x}_{\boldsymbol{f}}$ is irrelevant, since we will later introduce a rotation invariant matching process. On the plane $\boldsymbol{P}_{\boldsymbol{f}}$ we define a polar grid of radius σ_{r+1} subdivided into M radial sectors and L angular sectors, as shown in Fig.3. We've empirically found that $M = 3$ and $L = 32$ generate a discriminative signature, while allowing fast computation. Each point $\boldsymbol{p}_j$ belonging to $B_{\sigma_{r+1}}(\boldsymbol{f})$ is projected to $\tilde{\boldsymbol{p}}_j$ which lies onto the plane $\boldsymbol{P}_{\boldsymbol{f}}$, and associated to the respective index (m_j, l_j) of the polar grid. Given the feature point $\boldsymbol{f}$ and a point $\boldsymbol{p}_j$, and defining a vector $\vec{v} = \boldsymbol{p}_j - \boldsymbol{f}$, the computation of (m_j, l_j) is performed as follows:

$$m_j = \left\lfloor \|\boldsymbol{p}_j\| \frac{\sigma_{i+1}}{M} + 0.5 \right\rfloor \qquad l_j = \left\lfloor \theta_j \frac{2\pi}{L} + 0.5 \right\rfloor \quad (4)$$

where

$$\begin{aligned} &\boldsymbol{p}_j = \boldsymbol{f} + \|\vec{v}\| \cdot \hat{v}_{xy} \qquad \hat{v}_{xy} = \frac{\vec{v} - \|v\| \, cos(\varphi_j) \cdot \hat{z}_{\boldsymbol{f}}}{\|\vec{v} - \|v\| \, cos(\varphi_j) \cdot \hat{z}_{\boldsymbol{f}}\|} \\ &\varphi_j = \arccos(\langle \hat{v}, \hat{z}_{\boldsymbol{f}} \rangle) \\ &\theta_j = \begin{cases} \arccos(\langle \hat{v}_{xy}, \hat{x}_{\boldsymbol{f}} \rangle) & \langle \hat{v}_{xy}, \hat{y}_{\boldsymbol{f}} \rangle \geq 0 \\ 2\pi - \arccos(\langle \hat{v}_{xy}, \hat{x}_{\boldsymbol{f}} \rangle) & \langle \hat{v}_{xy}, \hat{y}_{\boldsymbol{f}} \rangle < 0 \end{cases} \end{aligned} \quad (5)$$

Once each point $\boldsymbol{p}_j \in B_{\sigma_{r+1}}(\boldsymbol{f})$ has been associated to a sector, it is possible to compute $\boldsymbol{w}_{\boldsymbol{f}}$, the descriptor associated to feature point $\boldsymbol{f}$. At first, for each sector (m, l) the average normal vector $\hat{n}(m, l)$ and saliency $s(m, l)$ are computed (if a sector does not contain any point, it is considered not valid). Then, given $\hat{n}_{\boldsymbol{f}}$ and $s_{\boldsymbol{f}}$ respectively the normal vector and saliency value associated to the feature point $\boldsymbol{f}$, the sector descriptor $\boldsymbol{w}_{\boldsymbol{f}}(m, l)$ is computed as follows:

$$\begin{aligned} &\boldsymbol{w}_{\boldsymbol{f}}(m, l) = [\Delta n(m, l), \Delta s(m, l)] \\ &\Delta n(m, l) = 1.0 - |\langle \hat{n}(m, l), \hat{n}_{\boldsymbol{f}} \rangle| \\ &\Delta s(m, l) = 1.0 - \frac{s(m, l)}{s_{\boldsymbol{f}}} \end{aligned} \quad (6)$$

The proposed descriptor is fast to compute, since both normals and saliency information are already available once feature points have been identified. Moreover, it is moderately light as it only requires $2 \times M \times L$ floating values. Nevertheless, we will see that it can still provide enough selectivity to skim the correspondence space to a more treatable dimension. With respect to other known approaches, which don't exploit the informative content associated to saliency variations, we chose to exploit it in our signature.

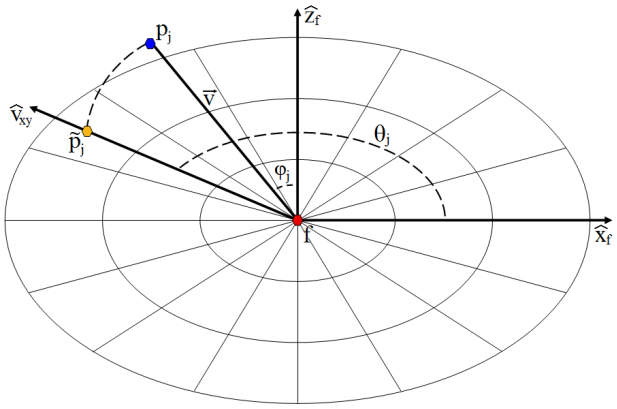

Figure 3: Signature grid description.

2.3. Feature matching

Given a pair of range images (RI_k, RI_{k+1}) and the related feature sets F_k and F_{k+1}, each couple of features $\boldsymbol{f}_s \in F_k$, $\boldsymbol{f}_d \in F_{k+1}$ is a potential correspondence c. Fig.7 gives visual insight of how signatures actually look like and how feature similarities/dissimilarities can define good/bad feature matches. In order to quantitatively assess which ones are more likely to be correct, each couple of feature points detected at same scale level are examined and a correspondence score c_{sd}^{score} is computed by matching their signatures. To render the matching invariant to viewpoint rotations, as well as agnostic with respect to the direction of $\hat{x}_{\boldsymbol{f}}$, one of the descriptors is allowed to rotate around its normal axis L times, one for each possible circular direction, and the maximum score is determined as follows:

$$c_{sd}^{score} = \max_{\bar{l}\in[1,L]} \left\{ c_{sd}^{score}\left(\bar{l}\right) \right\} \tag{7}$$

with

$$c_{sd}^{score}\left(\bar{l}\right) = \sum_{m=1}^{M}\sum_{l=1}^{L}\left[n_{sd}^{score}\left(m,l,\bar{l}\right) \cdot s_{sd}^{score}\left(m,l,\bar{l}\right)\right]$$
$$n_{sd}^{score}\left(m,l,\bar{l}\right) = \left(1 - \left|\Delta n_s\left(m,l\right) - \Delta n_d\left(m,\bar{l}\right)\right|\right)$$
$$s_{sd}^{score}\left(m,l,\bar{l}\right) = \left(1 - \left|\Delta s_s\left(m,l\right) - \Delta s_d\left(m,\bar{l}\right)\right|\right)$$

Whenever a sector is marked as not valid, its contribution to c_{sd}^{score} is set to zero. The score value is used to skim the correspondence space from its original size of $|F_k| \cdot |F_{k+1}|$ to a more treatable dimension. We define the correspondence set C_k of size Q as the list of correspondences c_q found between RI_k and RI_{k+1} which possess the highest score. In our implementation we've experimentally set Q to 150, this choice is justified by the fact that setting an hard threshold on the score is not an option, since the distribution of score values is not constant with k.
This correspondence selection is far from guaranteeing that C_k does not contain false correspondences due to incidental signatures similarity. However, experiments with pre-aligned datasets have shown that correct matches are concentrated in the highest positions of the score ranking, along with several false matches. It is therefore necessary to introduce a robust selection step in order to ascertain the reliable correspondences that are present in C_k.

2.4. Correspondence test and selection

In order to determine a roto-traslation matrix that references the current range image RI_k to the next one, we need to locate at least 3 correct correspondences (a triplet) within the set C_k. Each triplet t is defined as follows:

$$t = \{c_g, c_h, c_j\}, \text{with} \quad \begin{cases} c_g, c_h, c_j \in C_k \\ g, h, j \in [1, Q] \\ g \neq h \neq j \end{cases} \tag{8}$$

Given the correspondence set C_k of size Q, the number of non-repeating triplets corresponds to $Q^3 - 3Q^2 + 2Q/6$. Determining which (if any) of the triplets is correct is computationally expensive; for Q equal to 150 we would obtain more than half million triplets, therefore brute-force approaches such as directly test each of the possible roto-translations is not a viable option. Hence we have devised another selection procedure which dramatically decreases the computational cost related to the test. Our procedure consisting into three progressive steps: 1) every correspondence belonging to C_k is validated against each other and a distance score is calculated for each couple of correspondences; 2) for each triplet of correspondences, a score is assigned based on the three pairwise scores previously computed, and a subset T_k of triplets is retained; 3) for each triplet in T_k, a roto-traslation matrix RM is estimated and applied to the image feature set F_{k+1}, corresponding points are searched within image RI_k. The triplet which collects the highest number of such correspondences is considered as the more reliable estimate.

1) In order to validate each correspondence through the others we rely on the rigidity constraint which states that the distance between two points subject to an Euclidean transformation remains constant. We introduce the concept of relative distance between a pair of correspondences, illustrated in Fig.4, and defined as follows:

$$d_{gh} \equiv d\left(c_g, c_h\right) = \frac{\left|\left\|\boldsymbol{p_g^A} - \boldsymbol{p_h^A}\right\| - \left\|\boldsymbol{p_g^B} - \boldsymbol{p_h^B}\right\|\right|}{\max\left(\left\|\boldsymbol{p_g^A} - \boldsymbol{p_h^A}\right\|, \left\|\boldsymbol{p_g^B} - \boldsymbol{p_h^B}\right\|\right)} \tag{9}$$

Due to the normalization term at the denominator, relative distance is bound between 0 (equal distance) and 1 (maximum distance). This allows to perform a more reliable correspondence ranking, since the error is evaluated in proportion to the absolute distance between the correspondences.

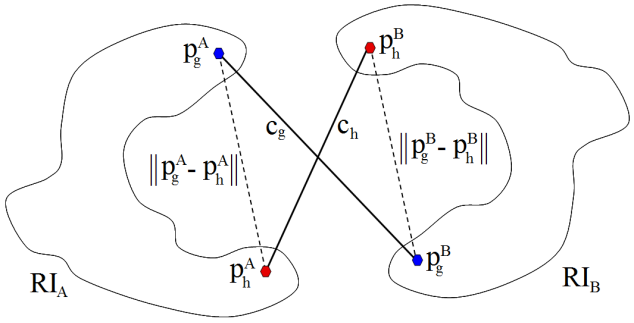

Figure 4: Exemplification of correspondences distance d_{gh} of (9).

Once the relative distances have been estimated, they are organized into a $Q \times Q$ matrix **DM**:

$$\mathbf{DM} = \begin{bmatrix} 0 & d_{12} & d_{13} & \cdots & d_{1Q} \\ d_{21} & 0 & d_{23} & & d_{2Q} \\ d_{31} & d_{32} & 0 & & d_{3Q} \\ \vdots & & & \ddots & \vdots \\ d_{Q1} & d_{Q2} & d_{Q3} & \cdots & 0 \end{bmatrix} \quad (10)$$

DM matrix is symmetric ($d_{hg} = d_{gh}$), and possesses zeros over its main diagonal ($d_{gg} = 0, \forall g \in [1, Q]$). An example of how such matrix looks like is presented in Fig.5.

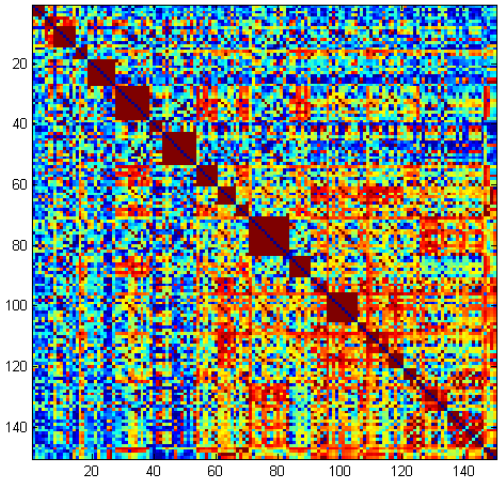

Figure 5: A distance matrix **DM**: blue dots represent low relative distance, while red ones identify distant matches. The red square clusters that can be seen along the diagonal are generated whenever evaluating pairs of correspondences that share one feature point (in such cases the relative distance is 1).

2) Once calculated **DM**, we can skim the triplet space by determining the set T_k of U triplets which present the maximum value of the following score:

$$t^{score} = 1 - \frac{d_{gh} + d_{hj} + d_{jg}}{3} \qquad \begin{cases} g, h, j \in [1, Q] \\ g \neq h \neq j \end{cases} \quad (11)$$

We have experimentally found that selecting the best U (which again has been set to 150) triplets ensure that the correct ones are retained, and appear as usual in the highest positions of the ranking.

3) In order to determine the most correct triplet within the set T_k, for each $t_u \in T_k, u \in [1, U]$ the following steps are performed:

- the roto-translation matrix RM_u associated to triplet t_u is estimated through Horn method [10];
- the feature set F_{k+1} is roto-translated through application of RM_u;
- corresponding points between F_{k+1} and RI_k are identified.

The triplet which is found to possess more corresponding points, labeled as $\bar{t}$, is considered as one that is most likely to be correct. Its associated roto-translation matrix $\bar{RM}$ is thus refined by taking into account the corresponding points just estimated. At last, we need to verify whether the obtained alignment has to be considered successful or not. To this end, we select a subset of points from RI_{k+1}, we roto-translate them through $\bar{RM}$, and verify that at least a given percentage of points find a correspondence in RI_k. In our implementation, such threshold is set to 20%. If the number of matches is above that threshold, image RI_{k+1} is considered as successfully aligned to the previous one. This last constraint implicitly imposes the requirement that each image couple possesses at least 20% of overlap, otherwise even if the correct roto-translation matrix is found, the alignment is likely to be considered wrong as the number of corresponding samples is below the threshold.

3. Experimental results

For the validation of our system we performed a series of quantitative tests. Successful alignment rate and computation time measurements have been experimentally obtained on a realistic and well assorted (in terms of object features) test dataset, in order to demonstrate the fulfillment of the target application requirements (Sec.1.2). Due to the lack of standard or widely-adopted high-resolution range image datasets (and a related difficulty in performing a fair comparison among different approaches of the literature, which moreover are often expressed only qualitatively) we

Dataset	RI pairs	Avg # points/RI	RI pairs aligned	Avg exec. time/RI [s]
Venus	60	835k	59	3.6
Capital	22	760k	22	3.4
Hurricane	31	690k	30	3.5
Decoration	47	510k	46	2.0
Platelet	11	80k	11	1.0
Angels	7	1M	7	4.1
Dolphin	19	410k	19	2.2
Teeth	7	410k	7	2.1
Bunny	63	37k	63	1.6

Table 1: Experimental results summary

collected different objects and we acquired them with a commercial high-res structured-light scanner (1280x1024 CCD, i.e. max 1.3Mpoints/RI) according common usage procedure, i.e. following a suitable and freely chosen multiple view acquisition path that cover the whole surface of each object. Each dataset represents a physical object containing features of different shape (such as grooves, bumps or small pits) at various dimensions. Objects sizes range from 50 mm up to 600 mm over their main dimension. Except for the Stanford Bunny dataset (from the Stuttgart repository [7]), which possess a low-resolution (400×400 I pixel grid), and presents an high overlap area between each scan couple, within the other 8 datasets each image couple has only a limited amount of overlap (usually above the 20% threshold), since the assumed acquisition policy was to minimize the number of scans while covering the entire surface of the object (see Fig.6). 3 datasets have been kindly provided by the authors of [4].

A total of 276 range images coupled in 267 RI pairs undergone the proposed alignment pipeline configured as follows: preemtive factor 2 subsampling (except for low-res Bunny), three ocataves, one saliency map for each octave and gaussian kernel size set to 4. Quantitative results are presented in Tab.1. In the fourth column, aligned RI pairs are counted, where the alignment is considered successfull only if both of the following tests give a positive result: 1) visual inspection check by the evaluation of geometry appearance and interpenetration patterns among the different point sets, 2) application of ICP fine alignment and verification of the alignment accuracy and of the absence of local minima trap occurrences. The technique demonstrated to be quite robust in that it correctly aligned 98.9% of the RI pairs. Further analysis performed over the few unaligned pairs concluded that main causes for failure was due to either an insufficient overlap area (that is, close to the lower bound of 20%), or particularly featureless areas.

Computational performance (in col.5) are related to a C++ implementation and run on a PC equipped with a processor Intel I5 M520 (2x2,4 GHz) and 4GB of RAM. It is important to note that the code has not yet been optimized for parallel execution, therefore time performances can be further improved. Computational performances show an average alignment time of 2600 milliseconds. This is distributed as follows: 58% for feature extraction, 11% for feature matching and the remaining 31% for correspondence skim and roto-traslation estimation. The two main factors that influence computation times are the number of points per range image to be processed, and the number of features detected over each image. In the "worst case" (that means, images close to 1 million of points and many features detected at all scales), alignment time reach a maximum of about 4 sec. Also computational speed are somehow difficult to infer and to compare from literature data because 1) not every work declare computational speed, 2) only subpart are usually considered (e.g. feature extraction) instead of the entire pipeline, 3) hardware obsolescence. However, we halved the computation time for feature extraction declared in [4] (on same datasets provided by the authors and on similar PC architecture) and, as also observed in [4], we confirm to be, at least, one order of magnitude (comprising HW obsolescence compensation) under the times declared in the related works [15, 16, 5].

4. Conclusions

We have presented a system for automatic pairwise pre-alignment of range images. The alignment is estimated from a selection of corresponding feature points on the scans, which are identified through a multi-scale analysis approach introduced in [4]. Correspondences are in turn created, ranked and skimmed by the matching of expressive feature descriptors. Computational complexity and problem dimensionality are kept low throughout the processing chain. The obtained performance satisfy all the application requirements about effectiveness, speed and accuracy. An interactive usage of high-resolution modern scanners is therefore possible: we can conceive to use the proposed technique during the acquisition phase, where a fast align-

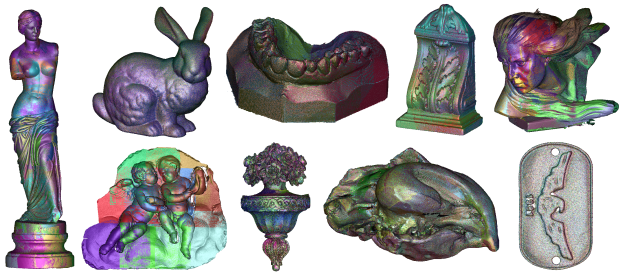

Figure 6: The test datasets: Venus, Bunny, Denture, Capital, Hurricane, Angels, Decoration, Dolphin, Platelet.

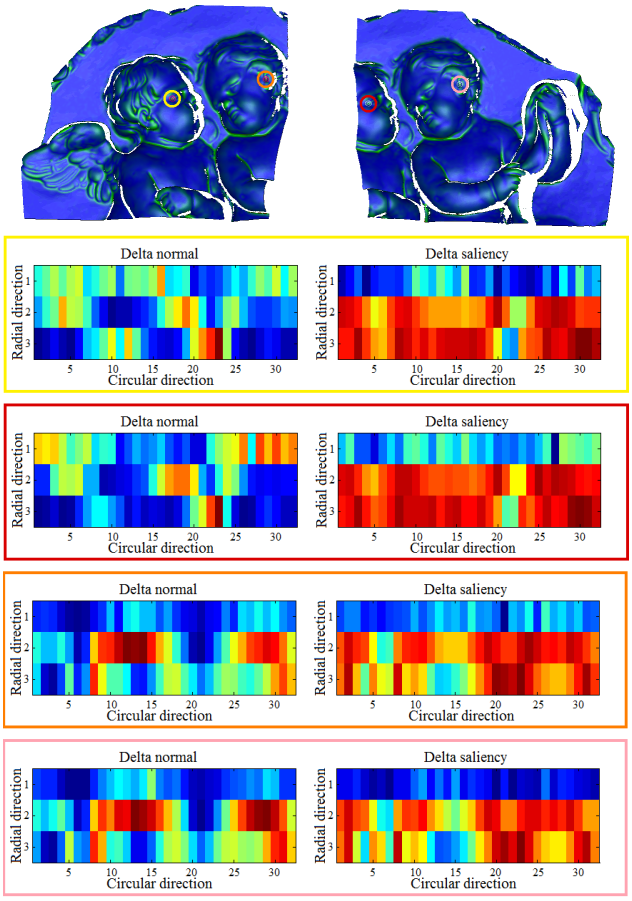

Figure 7: Feature signatures. Upper part: two range images on which some feature points are highlighted with different colors. Below, graphical visualization in a red-blue scale of the signatures, contoured with their corresponding colors.

ment (coarse+fine) can quickly take place as new images arrive, with evident benefits in terms of the scanner usage/usability (better user orientation, visual feedbacks, immediate object covering check) and acquisition (and modelling) speed-up. Since the alignment process is pairwise, the technique requires the adoption of an acquisition policy which guarantees that each image has an area of overlap with the previous one. In the future, we wish to address this (small) limitation, so that the constraint may be relaxed in demanding an overlap with any of the previously aligned scans.

References

[1] D. Aiger, N. J. Mitra, and D. Cohen-Or. 4-points congruent sets for robust pairwise surface registration. In *ACM SIGGRAPH*, 2008. 2

[2] F. Bernardini and H. Rushmeier. The 3D model acquisition pipeline. *Comp. Graph. Forum*, 21(2):149–172, 2002. 1

[3] P. J. Besl and N. D. McKay. A method for registration of 3-D shapes. *IEEE Trans. on Pattern Anal. Mach. Intell.*, 14(2):239–256, 1992. 1

[4] F. Bonarrigo, A. Signoroni, R. Leonardi, and M. Carocci. A multiscale feature extraction approach for 3D range images. In *International Workshop on Image Analysis for Multimedia Interactive Services*, 2010. 3, 4, 7

[5] U. Castellani, M. Cristani, S. Fantoni, and V. Murino. Sparse points matching by combining 3D mesh saliency with statistical descriptors. *Comp. Graph. Forum*, 27(2):643–652, 2008. 2, 7

[6] C. S. Chua and R. Jarvis. 3d free-form surface registration and object recognition. *International Journal of Computer Vision*, 17(1):77–99, 1996. 2

[7] Eisele. *Stuttgart Range Image Database*, 2001. http://range.informatik.uni-stuttgart.de/. 7

[8] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *ACM Communications*, 24(6):381–395, 1981. 2

[9] N. Gelfand, N. J. Mitra, L. Guibas, and H. Pottmann. Robust global registration. In *Symp. on Geom. Processing*, 2005. 2

[10] B. K. Horn. Closed-form solution of absolute orientation using unit quaternions. *Journal of the Optical Society of America*, 4:629–642, 1987. 6

[11] A. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3D scenes. *IEEE Trans. on Pattern Anal. Mach. Intell.*, 21(5):433–449, 1999. 2

[12] M. Kazhdan, M. Bolitho, and H. Hoppe. Poisson surface reconstruction. In *Symp. on Geom. Processing*, 2006. 1

[13] S. Krishnan, P. Y. Lee, J. B. Moore, and S. Venkatasubramanian. Optimisation-on-a-manifold for global registration of multiple 3D point sets. *International Journal of Intelligent Systems Technologies and Applications*, 3:319–340, 2007. 1

[14] P. Labatut, J. Pons, and R. Keriven. Robust and efficient surface reconstruction from range data. *Comp. Graph. Forum*, 28(8):2275–2290, 2009. 1

[15] C. H. Lee, A. Varshney, and D. W. Jacobs. Mesh saliency. In *ACM SIGGRAPH*, 2005. 2, 7

[16] X. Li and I. Guskov. Multi-scale features for approximate alignment of point-based surfaces. In *Symp. on Geom. Processing*, 2005. 2, 7

[17] D. Lowe. Distinctive image features form scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 2, 3

[18] K. Pulli. Multiview registration for large data sets. *3D Digital Imaging and Modeling*, 1999. 1

[19] S. Rusinkiewicz and M. Levoy. Efficient variants of the ICP algorithm. *3D Digital Imaging and Modeling*, 2001. 1

[20] D. Thomas and A. Sugimoto. Robust range image registration using local distribution of albedo. In *3D Digital Imaging and Modeling*, 2009. 2