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

Francesco Bonarrigo, Alberto Signoroni and Riccardo Leonardi
University of Brescia
Information Engineering Dept., via Branze, 38, I25123 Brescia, ITALY
firstname.lastname@ing.unibs.it

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 concurrent 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 reference 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, 13, 15]).

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 fine alignment approaches, either pairwise (usually ICP-based [3],[20]) or global (e.g. [19],[14],[4]), are based on optimization processes which are vulnerable to local minima problems or position ambiguities, thus requiring proper initialization.

Regarding datasets, state-of-art optical scanning devices have increased in the last years their spatial resolution and improved other key features (accuracy, distortion, acquisition time,...), and it is expected to see them used in unconstrained environments with lightweight devices like digital cameras. This well fits the increasing demand of “3D” either in today professional applications (industry, biomedicine, cultural heritage,...) as well as the expected increment of 3D contents of future web applications. Despite the coarse alignment problem has been long studied and several solutions have been proposed (some representative works are quoted in Sec.1.1), the reference applications are more and more demanding and require solutions which could satisfy multiojective problems. Therefore, high-performance coarse alignment remains an open investigation field.

1.1. Related work

The coarse registration problem has been extensively studied, and several methods can be found in the literature. Two main philosophies have emerged during these years, i.e. with or without the exploitation of feature descriptors. The first approach exploits the ever-increasing computational capabilities to identify, 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 input data. In addition, such solutions are usually more robust with respect to noise. On the other hand, they are usually computationally intensive. The progenitor of this family is considered to be the RANSAC, devised by Fischler and Bolles [9]. During the years, improvements to this algorithm have been proposed in order to reduce the computation time, also by exploiting
point neighborhood descriptors [7],[1]. A second approach for coarse registration relies on the extraction and subsequent matching of global 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, this second family of approaches usually fails in describing featureless (at some scale) surfaces, and is quite sensitive to noise. The ‘spin-image’ approach introduced by John-son and Hebert [12] produces viewpoint invariant representations of 3D shapes which can be used for multiple view alignment. Multi-scale feature based approaches (also used in this work) are suitable for a better adaptation to different kind (and dimension) of object features. Related works are those presented by Li and Guskov [17] and Lee et al. [16] which introduced extensions of Lowe’s 2D SIFT [18] to 3D datasets. Their approach has subsequently been exploited by Castellani et al. [6] for mesh alignment. In this context, a key choice is related to the feature descriptor to be used. An ideal descriptor should provide an unambiguous signature for each feature, it should be fast to compute, robust to viewpoint changes and to variations of point density in the image. For range images, Li and Guskov [17] proposed a descriptor based on a combination of local Discrete Fourier Transform and Discrete Cosine Transform to describe the neighborhood of each feature point. Gelfand et al. in [10] 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. [6] proposed a statistical descriptor based on hidden Markov chain trained through the neighbouring points, visited following the order given by a spiral originated at the feature point.

1.2. Problem definition and requirements

In this paper we wish to address the problems related to an unconstrained usage of modern, high resolution acquisition devices, capable of providing superior accuracy performances. With the term unconstrained usage we intend that the operator is given the freedom to choose the acquisition path he prefers for the scanning process. As such, no positional constraints exist such as placing the scanning device at predetermined positions or angles. The only requirement that still has to be preserved is that each image should have a certain degree of overlap with respect to the other scans. In practice, however, this constraint is commonly fulfilled, since whenever multiple views are required to acquire the area of interest, the operator is implicitly required to plan a suitable acquisition path. Note that 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 provided with a set of scans that follow an acquisition path for which each image is partially overlapping the previous ones. Following, we describe the objectives we are aiming at. First of all, the processing pipeline should be suited and equally effective for different kinds of acquired objects (industrial, artistic, ...) of different sizes. The developed solution should also be fast: ideally it should allow some level of interactivity for the operator, i.e. the alignment is computed while the operator varies the scanner (or object) position in order to perform the next scan. Accuracy should of course be a desirable property as well. However, robustness is the prioritary issue. In fact, the objective of coarse alignment is to approximately register a pair of views, so that the subsequent processing stage of fine alignment (such as ICP, for example) is capable to achieve an optimal alignment, without getting stuck in local minima of the error function. In this context, precision in feature localization and feature repeatability among overlapping views are the most critical issues which determine the robustness and accordingly the accuracy of the alignment. In this work, we focus our attention on pairwise alignment since it constitutes the fundamental unit of any progressive approach (that is, align one scan with respect to the ones that have been already successfully aligned). The proposed solution consists in a pre-alignment pipeline (detailed in the next section) that has been specifically designed to fulfill all the previously stated requirements. Accordingly the main contributions of this work are: a complete and fully functional pipeline for pairwise range images alignment; a lightweight feature descriptor devised in order to quickly reduce the matches set dimension; and a chain of operations developed to progressively skim the set of correspondence pairs.

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, which can be associated to a pixel of the acquisition grid. We define a range image as a map \( I : \mathbb{Z}^2 \rightarrow \mathbb{R}^3 \), 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 \( p_i \in R \), 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 and efficiently, while data processing is performed over the 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. The multiscale feature extraction method we proposed in [5] constitutes the first step of the proposed pipeline (Sec.2.1), since it proved to be robust with respect to the degradations described above, as well as computationally efficient. However, in [5] we did not suggest any feature description, so from this point on we proceed with original contribution. Following the pipeline of Fig.1, in Sec.2.2 we introduce a feature descriptor that is both representative and cheap to compute, as well as invariant with respect to viewpoint changes. The matching process between two of these signatures is described in Sec.2.3. Next, a computationally effective search for reliable correspondences among features is described in Sec.2.4 and is articulated in several subssteps, with the objective to progressively skim entries that are considered unlikely or incoherent. At first, this is done on single correspondences. Subsequently, triplets of correspondences are considered and classified in order to select a subset over which the pre-alignment transformation can be estimated.

2.1. Feature extraction

As stated, our feature extraction technique builds on [5], which we briefly resume for the sake of completeness. The approach can be thought as an extension of the Lowe’s SIFT approach [18] to 3D range data according to the following steps: 1) given a range image \( RI \), \( M \) filtered images \( G(r) \), at scales \( r \in [1, R] \), are derived by applying Gaussian kernels of increasing dimension; 2) 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 as well as the scale at which each feature point has been detected. To produce the filtered images \( G(r) \) at various scales \( r \in [1, R] \), a first unconstrained geometric Gaussian filtering on valid points \( p_i \) of the RI is done, obtaining \( p_i^g(r) \):

\[
p_i^g(r) = \frac{\sum_{p_j \in B_{2\sigma_r}(p_i)} p_j \cdot e^{-\frac{||p_i-p_j||^2}{2\cdot\sigma^2}}}{\sum_{p_j \in B_{2\sigma_r}(p_i)} e^{-\frac{||p_i-p_j||^2}{2\cdot\sigma^2}}}
\]

where \( B_{2\sigma_r}(p_i) \) identifies the points within a distance \( 2\sigma_r \) from \( p_i \). The effect of the geometric processing (1) is well balanced only if one can assume that the position of the points \( p_j \in B_{2\sigma_r}(p_i) \) 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 \( B_{2\sigma_r}(p_i) \) contains non uniform point distributions with respect to the surface, (1) tends to generate a positional bias of the filtered points \( p_i^g(r) \). Points close to borders and holes are similarly affected from this problem. To introduce resilience to the above distortions, \( p_i^g(r) \) are only allowed to move along the normal direction \( \hat{n}_i \) associated to the original point \( p_i \) by the following projection (see again Fig.2):

\[
g_i(r) = p_i + (p_i^g(r) - p_i) \cdot \hat{n}_i \hat{n}_i
\]

so that \( G(r) \) is defined as the set of points \( g_i(r) \), \( i \in [1, |R|] \). As the kernel radius \( \sigma_r \) increases, details which size is smaller than \( \sigma_r \) are smoothed out from \( G(r) \), while, whenever the kernel size doubles, the input range image is subsampled of a factor two to reduce the computational burden. As suggested in [5], both neighborhood search 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 array of scalar values, obtained by pairwise subtraction of \( G(r) \) at adjacent scales. This operation only retains the details comprised between the bounding scales \( r \) and \( r+1 \), in other words it highlights features which dimension is comprised between two kernel sizes \( \sigma_r \) and \( \sigma_{r+1} \). Saliency maps \( S(r) = \{s_i(r)\} \) are calculated as follows:

\[
s_i(r) = |g_i(r) - g_i(r+1)| \cdot (\hat{n}_i(r) \cdot \hat{n}_i(r+1))
\]
where the correction factor \( \langle \hat{n}_i (r), \hat{n}_i (r + 1) \rangle \) has been introduced to better concentrate saliency over stable points, i.e., points for which the normal direction \( \hat{n}_i (r) \) does not vary too much across the scales. Subsequently, for each saliency map \( S (r) \), maxima values are identified 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 \( B_{2\sigma + 1} (p_i) \). This prevents the identification of redundant, overlapping feature points, as the greatest detail size that can be detected within \( S (r) \) is \( \sigma_{r + 1} \). Each maximum is further tested in order to make sure that 1) its neighborhood is well defined (that is, it is not close to borders nor holes, 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 \( f_{i,k,r,h} \) associated to the above maxima of the saliency map at scale \( r \) and associated to the \( k \)th range image \( RI_k \), form the feature point set \( F_{i,k} \), with \( h \in [1, |EF_{i,k} |] \). This concludes our summary of what we implemented from [5]. 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 \( f \).

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 \( f \), at some scale dimension \( \sigma_r \), we propose a novel descriptor computed exploiting both normal vectors and saliency data of its neighbor points \( p_j \in B_{\sigma_{r + 1}} (f) \). To generate the descriptor, at first a reference system \( \hat{x}_f, \hat{y}_f, \hat{z}_f \), centered over the feature point \( f \), is defined. \( \hat{z}_f \) is set toward the direction of \( \hat{n}_f \), while \( P_f = \text{span} \{ \hat{x}_f, \hat{y}_f \} \) is the related tangent plane. Orientation of axis \( \hat{x}_f \) is irrelevant in the present context, since we will later introduce a rotation invariant matching process. On the plane \( P_f \) we define a polar grid of radius \( \sigma_{r + 1} \), subdivided into \( M \) radial sectors and \( L \) angular sectors, as shown in Fig. 3. We’ve empirically found that \( M = 3 \) and \( L = 32 \) represents an adequate tradeoff between discriminative capability and computation speed. Each point \( p_j \) belonging to \( B_{\sigma_{r + 1}} (f) \) is projected to \( \hat{p}_j \) which lies onto the plane \( P_f \), and associated to the respective index \( (m_j, l_j) \) of the polar grid. Given the feature point \( f \) and a point \( p_j \), defining the vector \( \hat{v} = p_j - f \), the computation of \( (m_j, l_j) \) is performed as follows:

\[
m_j = \left\lfloor \frac{\| p_j \|}{M} \sigma_{i + 1} + 0.5 \right\rfloor \quad l_j = \left\lfloor \frac{2\pi}{L} + 0.5 \right\rfloor \quad (4)
\]

where

\[
p_j = f + \| \hat{v} \| \cdot \hat{v}_{xy} \quad \hat{v}_{xy} = \frac{\hat{v} - \| \hat{v} \| \cos (\varphi_j) \cdot \hat{z}_f}{\| \hat{v} - \| \hat{v} \| \cos (\varphi_j) \cdot \hat{z}_f \|}
\]

\[
\varphi_j = \arccos (\langle \hat{v}, \hat{z}_f \rangle) \quad \cos (\varphi_j) \cdot \hat{z}_f
\]

\[
\theta_j = \left\{\begin{array}{ll}
\arccos (\langle \hat{v}_{xy}, \hat{x}_f \rangle) & \langle \hat{v}_{xy}, \hat{y}_f \rangle \geq 0 \\
2\pi - \arccos (\langle \hat{v}_{xy}, \hat{x}_f \rangle) & \langle \hat{v}_{xy}, \hat{y}_f \rangle < 0
\end{array}\right.
\]

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

\[
w_f (m, l) = [\Delta n (m, l), \Delta s (m, l)]
\]

\[
\Delta n (m, l) = 1.0 - |\langle \hat{n} (m, l), \hat{n}_f \rangle|
\]

\[
\Delta s (m, l) = 1.0 - \frac{s (m, l)}{s_f}
\]

The proposed descriptor is fast to compute, since both normal and saliency information are already available once
feature points have been identified. Moreover it is moderately lightweight, since 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. Contrarily to other known approaches, which do not exploit the information content associated to saliency variations, we found it useful and discriminative its exploitation within our signature.

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 $f_s \in F_k$, $f_d \in F_{k+1}$ is a potential correspondence $c$. Fig.4 gives visual insight of how signatures actually look like and how feature similarities can define good 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 $c_{sd}^{\text{score}}$ is computed by matching their signatures. To render the matching invariant to viewpoint rotations, as well as agnostic with respect to $\hat{x}_f$ orientation, 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:

$$c_{sd}^{\text{score}} = \max_{l \in [1, L]} \{ c_{sd}^{\text{score}} (\vec{l}) \}$$

(7)

with

$$c_{sd}^{\text{score}} (\vec{l}) = \sum_{m=1}^{M} \sum_{l=1}^{L} n_{sd}^{\text{score}} (m, l, \vec{l}) \cdot s_{sd}^{\text{score}} (m, l, \vec{l})$$

$$n_{sd}^{\text{score}} (m, l, \vec{l}) = (1 - |\Delta n (m, l) - \Delta n (m, \vec{l})|)$$

$$s_{sd}^{\text{score}} (m, l, \vec{l}) = (1 - |\Delta s (m, l) - \Delta s (m, \vec{l})|)$$

Whenever a sector is marked as not valid, its contribution to $c_{sd}^{\text{score}}$ 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 $c_d$ found between $RI_k$ and $RI_{k+1}$ which possess the highest score. In our implementation we have experimentally set $Q$ to 150. This choice is justified by the fact that setting an hard threshold on the score is not a viable option, since the distribution of score values is not constant with respect to view pairs, therefore a threshold adequate for a given pair may result insufficient, or excessive, for other pairs.

This correspondence selection is far from guaranteeing that $C_k$ does not contain false correspondences due to incidental signature similarities. 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 within $C_k$.

![Image](image-url)

Figure 4: 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.

2.4. Correspondence test and selection

In order to determine a roto-translation 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:

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

(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 would be excessively expensive: in our case, for $Q$ equal to 150, more than half million triplets should be tested, therefore brute-force approaches such as direct test of 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 consists into three progressive steps: 1) every correspondence belonging to \( C_k \) is validated against the others, 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-translation matrix \( R M \) is estimated and applied to the image feature set \( F_{k+1} \), and 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. Following, each step will be described in detail.

1) Correspondence distance 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.5, and defined as follows:

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

Due to the normalization term at the denominator, relative distance is bound between 0 (same distance) and 1 (maximum distance). This allows to perform a more reliable correspondence ranking, since the error is evaluated proportionally to the absolute distance between the correspondences. Once the relative distances have been estimated, they are organized into a \( Q \times Q \) matrix \( DM \):

\[
DM = \begin{bmatrix}
0 & d_{12} & d_{13} & \cdots & d_{1Q} \\
\vdots & 0 & d_{23} & \cdots & d_{2Q} \\
& 0 & \ddots & \ddots & \ddots \\
& & 0 & \ddots & \ddots \\
& & & 0 & \ddots \\
& & & & 0
\end{bmatrix} \tag{10}
\]

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.6. 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) Triplet scoring and skimming 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:

\[
t_{\text{score}} = 1 - \frac{d_{gh} + d_{hj} + d_{jg}}{3} \quad \begin{cases} g, h, j \in [1, Q] \\
g \neq h \neq j
\end{cases} \tag{11}
\]

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 usually appear at the highest positions of the ranking.

3) Triplet verification and selection 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 [11];
- 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 the one that is most likely to be correct. Its associated roto-translation matrix \( \bar{R}M \) 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 $R_{I_{k+1}}$, we roto-translate them through $RM$, and verify that at least a given percentage of points find a correspondence in $R_{I_k}$. In our implementation, such threshold is set to 20%. If the number of matches is above that threshold, image $R_{I_{k+1}}$ 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. In order to demonstrate the fulfillment of the target application requirements (Sec.1.2), successful alignment rate and computation time measurements have been experimentally obtained on a realistic, well assorted (in terms of object features) number of test datasets. 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), we collected different objects and acquired them with a commercial high-res structured-light scanner (1280x1024 CCD, i.e. max 1.3Mpoints/RI) according to common usage procedures, i.e. following a suitable and freely chosen 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 [8]), 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 pair 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. The datasets are shown in Fig.7, while their characteristics are detailed in the first three columns of Tab.1. A total of 276 range images coupled in 267 RI pairs underwent the proposed alignment pipeline configured as follows: preemptive factor 2 subsampling (except for low-res Bunny), three octaves, 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 successful 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 very 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 were due to either an insufficient overlap area (that is, close to the lower bound of 20%), or particularly featureless areas. For comparison purposes, an updated selection of the datasets used in this paper (those of which we have copyright permission) and others, with the related alignment performance, can be downloaded from the webpage [http://www.ing.unibs.it/~signoron/](http://www.ing.unibs.it/~signoron/).

**Table 1: Experimental results summary**

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

**Figure 7: The test datasets:** Venus, Denture, Bunny, Hurricane, Platelet, Decoration, Angels, Dolphin, Capital.
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. Computational speed is somehow difficult to compare to the literature, 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 further improved the computation time for feature extraction obtained in [5] and, as also observed in [5], we confirm to be, at least, one order of magnitude (comprising HW obsolescence compensation) under the times declared in the related works [16, 17, 6].

4. Conclusions

We have presented an automatic processing pipeline for 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 [5]. 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: one can conceive to use the proposed technique during the acquisition phase, where a fast alignment (coarse+fine) can take place as new images arrive, with evident benefits in terms of scanner usability (better user orientation, visual feedbacks, immediate object covering check) and acquisition 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 respect to the previous one. In the future we wish to address this limitation, so that the constraint may be relaxed in demanding for an overlap with at least one of the previously aligned scans.

References