Noname manuscript No.
(will be inserted by the editor)

3D Scanning of Cultural Heritage with Consumer Depth Cameras

Enrico Cappelletto · Pietro Zanuttigh · Guido M. Cortelazzo

Received: date / Accepted: date

Abstract Three dimensional reconstruction of cultural heritage objects is an expensive and time-consuming process. Recent consumer real-time depth acquisition devices, like Microsoft Kinect, allow very fast and simple acquisition of 3D views. However 3D scanning with such devices is a challenging task due to the limited accuracy and reliability of the acquired data. This paper introduces a 3D reconstruction pipeline suited to use consumer depth cameras as hand-held scanners for cultural heritage objects. Several new contributions have been made to achieve this result. They include an ad-hoc filtering scheme that exploits the model of the error on the acquired data and a novel algorithm for the extraction of salient points exploiting both depth and color data. Then the salient points are used within a modified version of the ICP algorithm that exploits both geometry and color distances to precisely align the views even when geometry information is not sufficient to constrain the registration. The proposed method, although applicable to generic scenes, has been tuned to the acquisition of sculptures and in this connection its performance is rather interesting as the experimental results indicate.

Keywords 3D Reconstruction · Kinect · ICP · Depth Map

1 Introduction

Three dimensional descriptions of cultural heritage objects play useful roles in many different applications such as restoration, evaluation of the conservation state and virtual visits. Unfortunately most approaches for building 3D representations require expensive hardware and a large amount of highly

All authors
Dept. of Information Engineering
Tel.: +39-049-8277642
Fax: +39-049-8277699
E-mail: enrico.cappelletto,zanuttigh,corte@dei.unipd.it

skilled labor and are not affordable for many cultural heritage institutions typically fighting with budget issues. In particular active methods, most notably implemented by laser and structured light scanners for small objects and by time-of-flight scanners for larger architectural structures allow to obtain very accurate and detailed reconstructions. On the other side this type of scanning equipment is very expensive, the acquisition process is rather time consuming and the registration and the fusion of the various acquired views into a single 3D object is a challenging technical problem usually requiring a lot of manual interaction. The main alternative is to use passive methods, like stereo vision approaches, shape-from-silhouette or structure from motion. These methods allow to reconstruct the 3D shape of an object from a collection of pictures, thus avoiding the use of expensive hardware but most of such methods have several limitations and are typically neither as robust nor as accurate as the active methods.

The recent introduction of real-time consumer depth acquisition devices, like Microsoft Kinect or Time-Of-Flight cameras, has made the acquisition of 3D views much faster and simpler than before. Unfortunately these devices suffer several issues, like high noise level, limited resolution and artifacts in proximity of edges, that reconstruction algorithms must take into account. On the other hand their ability to acquire data at interactive frame-rates makes possible to acquire large numbers of views in very short times. The availability of views much closer than the ones used in typical 3D registration pipelines with data from laser or structured light scanners at the same time is a major advantage with respect to registration since it is possible to use short-baseline approaches much more reliable and faster than their long-baseline counterparts and a critical issue for computation and memory requirements.

Nevertheless consumer depth acquisition devices can make 3D acquisition much simpler and cheaper, thus really opening the way to the usage of 3D data in fields like cultural heritage. This work follows this rationale and present a 3D reconstruction pipeline explicitly targeted to consumer depth cameras. The proposed scheme, which extends the approach of [5] and applies it to cultural heritage, allows to automatically obtain accurate textured 3D models from the data acquired by the Kinect camera or similar devices (e.g., Asus Xtion or Time-Of-Flight cameras). A very commonly used approach for the registration of multiple views is the ICP algorithm [4], that computes the rototranslation between couples of views by assuming that each point in one view correspond to the closest one in the second view, then iterates the process by computing a new set of correspondences on the aligned views and continues until the optimal registration is obtained. As [21] and other approaches, the proposed method uses the ICP algorithm but, with respect to previous works it introduces new elements in order to adapt the reconstruction pipeline to the characteristics of the Kinect data and to exploit the color information acquired by the device not only for texturing but also for registration purposes. Firstly an ad-hoc filtering scheme is used in order to reduce the noise level and to remove the most common artifacts typically affecting the data acquired by the Kinect (note how this step has been largely improved with respect to [5]).

Then the salient points are extracted and used in a modified version of the ICP algorithm.

A new important element, not present in previous works, is the use of color information acquired by the video-camera not only for texture reconstruction but also to improve the extraction of salient points and the geometry reconstruction. This allows to obtain reliable reconstructions also on flat smooth surfaces lacking relevant features for 3D registration.

The paper is organized as follows: the related work is presented in Section 2, the proposed reconstruction pipeline is described in Section 3, the experimental results are given in Section 4 and Section 5 draws the conclusions.

2 Related works

3D scanning of cultural heritage has long been an active research field and a huge number of different approaches have been presented. Extensive reviews can be found in [22] and [26]. As previously noted, the available methods can be divided between active and passive methods. Active methods have been applied to a large number of hardware devices based on laser scanners, structured light systems or time-of-flight lidar systems for larger scenes. From the Digital Michelangelo project [19] on , several research projects have been devoted to the construction of software and hardware solution suitable to the acquisition of complex heritage works. In particular a great effort has been devoted to the development of effective solutions for the so-called 3D modeling pipeline, i.e. the fusion of the views and pictures acquired by the various sensors into a single textured 3D object. A detailed description of this process and a review of the various available methods can be found in [3]. However 3D modeling of cultural heritage objects typically requires considerable manual interaction even if some automatic registration schemes for cultural heritage objects have been proposed [1]. On the other side passive methods allow to perform 3D reconstruction using only pictures of the scene. Stereo vision approaches [28] are one of the most commonly used solution, see [24] for their usage in the cultural heritage field. Other commonly used schemes are shape-from-silhouette [18] for small objects and structure from motion for large scenes [20]. The latter approach has attracted a lot of interest in recent years due to projects like Microsoft's Photosynth [31] aiming at the usage of large collection of pictures available on the web for 3D reconstruction purposes. There are also hybrid methods combining image-based passive methods and active light scanners, e.g. the approach proposed by Hakim et al. [10].

As previously noted the introduction of consumer depth cameras has recently opened the way to a new research field trying to exploit their data for 3D scanning purposes. The Kinect and other similar devices have been widely used in setups with a fixed camera acquiring moving people or objects for dynamic 3D acquisition and human motion tracking [29]. On the other side their employment for the reconstruction of static 3D scenes is still an open issue since, as pointed out in recent studies on these devices, e.g. [15] or [9],

they have several impairing issues including high noise level, limited spatial resolution and edge artefacts.

Among the research projects which have investigated this task, Microsoft's KinectFusion [21] is probably the most relevant. In this project each frame acquired by the Kinect is registered in real-time using the ICP algorithm [4] over the complete 3D scene description reconstructed by a variation of the volumetric truncated signed distance function (TSDF). The approach of KinectFusion allows accurate reconstructions, but the large amount of memory needed limits its application to small scenes. E.g., "KinFu", the implementation contained in the PCL library [23] is limited to a $3 \times 3[m]$ area. Kinect Fusion has also been extended in the Kintinuous project [36]. In a recent work by Henry et al. [12] both geometric and visual features are used for the reconstruction of indoor environments from the Kinect data. A super-resolution scheme and a probabilistic approach are used for 3D reconstruction from Kinect and Time-Of-Flight data in [7]. Another research project [34] aims at capturing full 3D human body models using 3 Kinects. This approach is able to register the various body parts under non-rigid deformations. Furthermore commercial applications exploiting the Kinect for 3D reconstruction are now appearing, e.g., ReconstructMe [25] or Skanect [30]. These applications are typically able to capture full color 3d models of objects, people or rooms. However reconstruction accuracy is not always satisfying.

Similar results can also be obtained from Time-Of-Flight (ToF) cameras in place of the Kinect, e.g., in [6] a probabilistic approach based on Expectation Maximization (EM) is used to combine and align the acquired views. Color cameras can also be employed together with the depth sensors in order to improve the reconstruction accuracy. In [37] the data acquired by the ToF sensor are firstly used to reconstruct a coarse 3D representation of the scene by a volumetric approach. Then the data from multiple color cameras are used in order to improve the reconstruction by enforcing a photoconsistency measure and silhouette constraints.

Two approaches for the registration of 3D views are described and evaluated in [17]. The first approach is based on RGB images and estimates a sensor pose using image features, while the second uses only geometrical information. The results show that image-based registration method is particularly suitable to scenes with texture, while the object space-based method, although able to work on scenes without texture requires an adequate amount of geometric information in the scene. Such a complementary behaviour suggests that geometry and texture should be combined in order to provide a highly reliable method, as proposed in this paper.

3 Geometry reconstruction pipeline

The proposed 3D reconstruction pipeline, shown in Fig. 1, is made by 4 basic steps: in the first step each depth map acquired by the Kinect is registered together with color data in order to associate a color value to each 3D point.

The depth map is then filtered and polished in order to remove typical acquisition artifacts of the employed sensor. In the next step the salient points are extracted from the point cloud on the basis of both depth and color data; the saliency information is used in the following step where the acquired views are aligned together by a modified ICP algorithm which also exploits color and geometry together; in the final post-processing stage the surface is simplified and polished and color data from the various views are fused to get the final color representation.

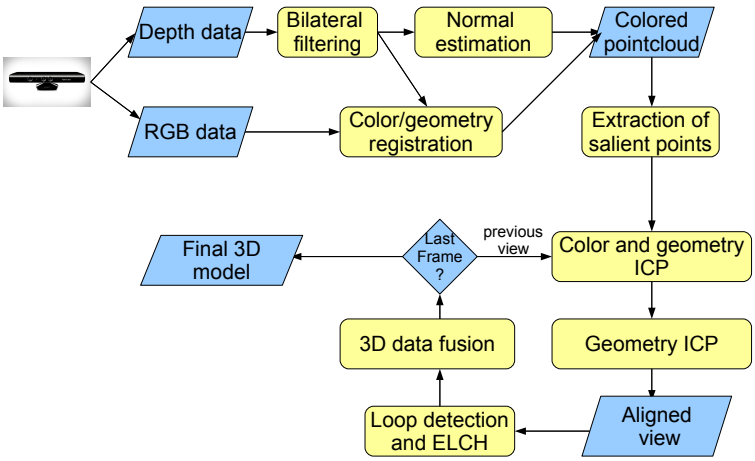


Fig. 1 Architecture of the proposed system

3.1 Pre-processing of depth information

In this work we assume that a sensor capable to acquire both color and depth data is available (devices like the Kinect or Asus' Xtion include both a depth and a color camera). Such data can also be obtained by combining a Time-Of-Flight sensor with a standard color camera. Before any acquisition the depth camera (e.g., the IR camera in the case of the Kinect) and the color camera must be calibrated. In the case of the Kinect both the intrinsic parameters of the two cameras and the extrinsic parameters relating the depth and color sensor can be computed by the approach of Herrera et al. [13]. For Time-Of-Flight and camera setups ad-hoc approaches exist, e.g., [8]. The calibration parameters of the depth and color cameras and the relative position between them are used to compute the samples position in 3D space and to reproject color data over the depth information. Color data are also converted to the CIELAB color space. In this way the acquisition produces a set of colored 3D points $p_i = (X_{p_i}, Y_{p_i}, Z_{p_i}, L_{p_i}, a_{p_i}, b_{p_i}), i = 1, .., N$, where $(X_{p_i}, Y_{p_i}, Z_{p_i})$ are

the spatial 3D coordinates of p_i and $(L_{p_i}, a_{p_i}, b_{p_i})$ its color components in the CIELAB space.

It is worth noting that the 3D views acquired by the Kinect with respect to those acquired by standard 3D structured light or laser scanners are characterized by limited accuracy, high noise levels and by the presence of many erroneous depth samples. For all these reasons it is necessary to pre-process the acquired data before using them in the 3D reconstruction process. The proposed method encompasses 2 basic steps. In the first the depth information acquired by the Kinect is filtered in order to reduce the amount of noise. Bilateral filtering [33] is very effective for smoothing the acquired surfaces while preserving the edges. The filter behaviour in the spatial and range domains is controlled by two standard deviations, denoted as σ_d and σ_r respectively. The bilateral filter preserves edges larger than σ_r and tends to average across smaller discontinuities. It also averages structures thinner than $\sim 2\sigma_d$. The two parameters thus allow to trade-off between noise removal and the preservation of the structures in the depth map. Unfortunately setting them in a way that is optimal for the various image regions, specially if the image (or the depth map in our case) has different noise levels in different regions, is a challenging problem.

In the case of Kinect data, the resolution in the z direction (i.e., the direction corresponding to the optical axis) decreases quadratically with the distance from the sensor since the Kinect working principle is based on disparity estimation [9] and, as well known, depth is inversely proportional to disparity. This observation holds also for stereo vision systems and many other structured light depth cameras exploiting similar principles. The z quantization issue can be easily appreciated by acquiring some sample surfaces by the Kinect [16]: at $1[m]$ the spacing between the points in the z direction is quite small (about $2[mm]$) but it rapidly increases and at $3[m]$ the quantization of z values is quite evident (the spacing is $2.5[cm]$). At $5[m]$ the situation is even worse with a z spacing of $7[cm]$. As reported in [11] the quantization relationship between the distance and the quantization step is the following:

$$q(z) = \frac{2.73z^2 + 0.74z - 0.58}{1000} \tag{1}$$

Hence also the error random component increases with the distance. Therefore it makes sense to use different filtering parameters at different distances. We developed an adaptive bilateral filtering scheme where the size of the window and the standard deviation of the two components of the bilateral filter dynamically change with the distance of the considered points. The basic scheme is the same of the standard bilateral filter, except that the filtering parameters depend on the depth of the considered point, i.e.:

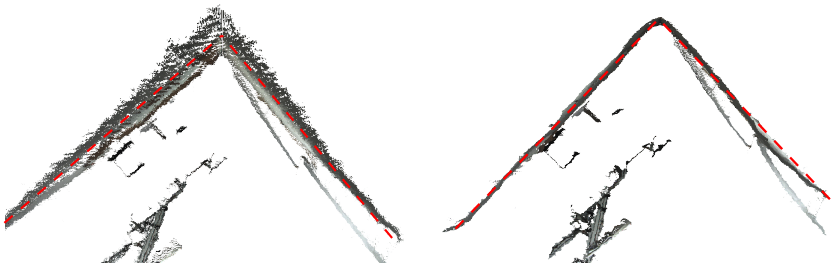
$$\hat{z}_j = \frac{\sum_{i \in W_j(z_j)} \left(e^{-\frac{1}{2}\left(\frac{-d_d(i,j)}{\sigma_d(z_j)}\right)^2} e^{-\frac{1}{2}\left(\frac{-d_r(i,j)}{\sigma_r(z_j)}\right)^2} \right)}{N_f} \tag{2}$$

where $W_j(z_j)$ is the square window used for the filter computation, $d_d(i,j)$ and $d_r(i,j)$ are the distances between sample (i,j) and the window center in the spatial and range domain respectively and N_f is a proper normalization factor computed as in [33]. Differently from the standard bilateral filtering approach the two standard deviations σ_d and σ_r in Eq. (2) are not constant values but depend on the distance from the sensor according to:

$$\sigma_d = K_d z_j^2 \tag{3}$$

$$\sigma_r = K_r z_j^2 \tag{4}$$

For simplicity we considered only the second order term in (1) while the two parameters K_d and K_r have been set according to the error model presented in [16]. This filter has a stronger behaviour at larger distances where Kinect data are less accurate and a milder one at closer distances. In particular it removes the quantization noise on farther surfaces without affecting too much the closer ones, are already quite accurate. Fig. 2 shows a top view of the acquisition of a room corner before and after the application of the proposed modified bilateral filter. The proposed filter can clearly restore the proper shape of the planar surfaces affected by quantization artifacts and sensor noise. At the same time the sharp edge between the two surfaces is preserved.



a) b)

Fig. 2 View of a room corner from the top: a) before applying the proposed filter; b) after filtering by the proposed approach. The red dotted line shows the actual profile of the acquired wall.

In the second step a moving window W_{p_i} of size $k \times k$ is first centered on each sample p_i of the acquired depth map (the experimental results of Section 4 use $k = 3$). The set $S_{p_i} = \{p' \in W_{p_i} \wedge |Z_{p'} - Z_{p_i}| < T_z\}$ of the samples in the window with a depth value similar to the one of the considered point p_i is then computed. If the number of samples in S_{p_i} is large enough (i.e., $|S_{p_i}| > 0.8|W_{p_i}|$) the point p_i is considered valid, otherwise it is discarded since the point is either on a too slanted surface or it is an isolated point. This thresholding is used to remove unreliable depth values, specially in proximity of edges where the Kinect data is less reliable.

Finally the surface normals $\mathbf{n}_{p_i}$ are estimated for each point p_i using the robust and efficient border- and depth-dependent smoothing scheme of [14]

(the normals will play an important role for salient point extraction). Since the acquired data are not reliable on slanted surfaces, they are also thresholded based on the angle between the surface normal and the viewing direction. In order to exclude points corresponding to surfaces too slanted with respect to the viewing direction a point $\mathbf{p}$ is kept only if $\mathbf{n}_p \cdot (-\mathbf{v}) > T$.

3.2 Extraction of salient points

The employed registration algorithm requires to select a subset of the acquired points to compute the roto-translation matrix between each couple of consecutive views. This step is particularly critical in the proposed setup since the data acquired from the Kinect have many unreliable points that can impact on the computed registration parameters. Furthermore it is not possible to process in real-time too large amounts of samples. In order to obtain an accurate real-time reconstruction it is necessary to extract a small subset of the original points both reliable and meaningful for registration purposes.

To achieve this target a saliency metric measuring the usefulness of each point for registration purposes is proposed. The idea is that the more distinctive points (i.e., the ones either in regions of articulated geometry or of high color variance) are the most salient ones.

For what concerns geometry information, the curvature of the local surface was used as the distinctivity measure (as suggested by [35]). The idea is that points corresponding to high curvature regions can be considered more distinctive since they force tighter bounds on the surface alignment. In particular the normal n_{p_i} to the surface at each point p_i is compared to the normals of the close samples (i.e., the samples in a window W_{p_i} surrounding p_i) in order to compute the set

$$A_{p_i} = \{(p' \in W_{p_i}) \wedge (\mathbf{n}_{p'} \cdot \mathbf{n}_{p_i} > T_g)\}. \tag{5}$$

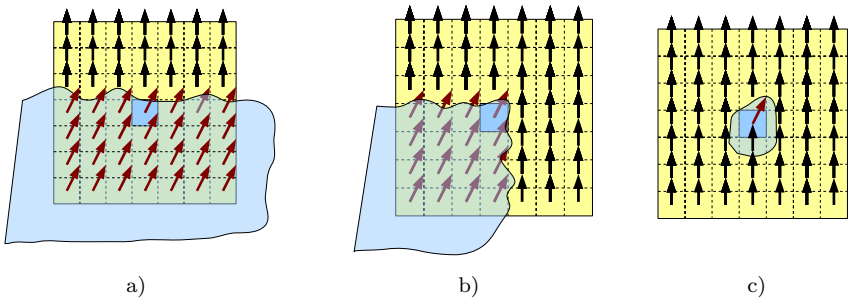
A_{p_i} is the set of the points for which the surface normals $\mathbf{n}_{p'}$ form an angle smaller than $\arccos(T_g)$ with the normal $\mathbf{n}_{p_i}$ of point p_i. The cardinality of A_{p_i} is therefore inversely proportional to the local curvature of the surface surrounding the selected point. Note that the point p itself is included in the computation in order to ensure that $|A_{p_i}| \geq 1$. Note how a large value of $|A_{p_i}|$ corresponds to samples in flat regions, not very informative for registration purposes. Samples with small $|A_{p_i}|$ are typically associated to edges, corners and high curvature regions. These points represent tighter bounds for the surface alignment. On the other side, they also have the risk of being less reliable since the acquired data can be less reliable close to edge or corners, but notice how edge points have been processed by the filtering scheme of Section 3.1. In order to avoid to use samples in the middle of smooth regions, samples with $|A_{p_i}| > |W_{p_i}|/2 + \sqrt{W_{p_i}}$ are excluded from the salient point set and their distinctivity is set to 0. Note how, as shown in Fig. 3 $|A_{p_i}| = W_{p_i}/2$ is the typical value for edge points, the rationale for the threshold value is to keep edge samples or samples with a comparable saliency. On the other side a low

value of $|A_{p_i}|$ is usually associated to isolated points typically unreliable or to artifacts due to noise. We decided to exclude points for which $|A_{p_i}| < |W_{p_i}|/4$ (note how a quarter of the window size is the region covered by the considered surface in the case of a typical corner, as shown in Fig.3). The geometric distinctivity measure is therefore computed as the inverse of the cardinality of A_{p_i}, i.e.:

$$D_g(p_i) = \begin{cases} 0 & \text{if } |A_{p_i}| \leq |W_{p_i}|/4 \\ 1/|A_{p_i}| & \text{if } |W_{p_i}|/4 \leq |A_{p_i}| \leq |W_{p_i}|/2+\sqrt{|W_{p_i}|} \\ 0 & \text{if } |A_{p_i}| \geq |W_{p_i}|/2+\sqrt{|W_{p_i}|} \end{cases} \quad (6)$$

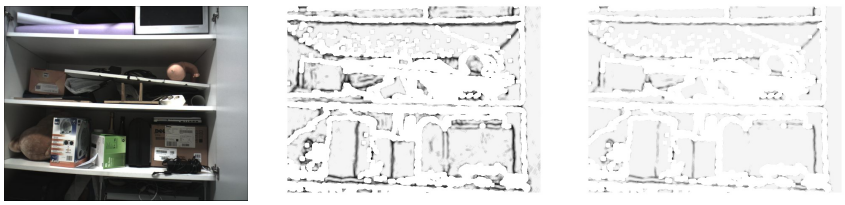
since $1 \leq A_p \leq k^2$ (where k is the size of the window W_p), $D_g(p)$ is included in the range $1/k^2 \leq D_g(p) \leq 4/|W_{p_i}|$, i.e., larger $D_g(p) = 4/|W_{p_i}|$ corresponds to the most salient points and $D_g(p) = 1/k^2$ to quite flat regions.

Fig. 4 shows an example of the computation of geometric distinctivity on a sample 3D view for different values of the threshold T_g.



a) b) c)

Fig. 3 $|A_{p_i}|$ in different situations: a) On edge samples $|A_{p_i}| \simeq |W_{p_i}|/2$; b) On corner samples $|A_{p_i}| \simeq |W_{p_i}|/4$; c) On isolated samples $|A_{p_i}|$ is very small.



a) b) c)

Fig. 4 Geometric saliency corresponding to different values of T_g. a) Color view; b) Geometric saliency for $T_g = 0.22$; c) Geometric saliency for $T_g = 0.44$; Darker points correspond to larger values of $D_g(p_i)$.

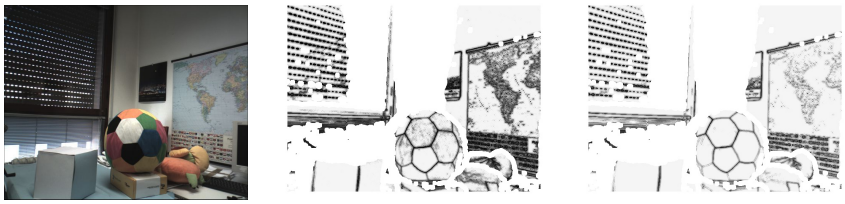
With respect to color information let us recall that a uniform color space, such as CIELab, ensures the consistency of the distance measurements between

the different color components, i.e., it ensures that the euclidean distance in the color space corresponds to the perceived color difference. Furthermore, since the L component of the CIELab color vector (i.e., the luminance), is strongly affected by the viewing direction, specially on reflective surfaces and in presence of non-uniform illumination, it has not been considered in the proposed algorithm and only the a and b components are used.
Similarly to the approach used for geometric information the window W_{p_i} around point p_i is considered and the points with color properties similar as those of p_i are computed, i.e. we compute the set:

$$C_{p_i} = \{(p' \in W_{p_i}) \wedge (\sqrt{(a_{p_i} - a_{p'})^2 + (b_{p_i} - b_{p'})^2} < T_c)\} \tag{7}$$

where a'_p and b'_p are the a and b color components of point p' in the CIELab color space. The color saliency is given by the set of the points of W_{p_i} with color components $(a_{p'}, b_{p'})$ similar to those of p_i (also in this case p_i is included in the computation). If it belongs to a uniform color region the cardinality of C_{p_i} will be large. If p_i belongs to regions with a complex texture pattern (more suitable for registration purposes since color data can be used to properly align the surfaces) $|C_{p_i}|$ will assume lower values. As for the case of geometry we threshold the values in order to avoid points in uniform regions or in too noisy areas. The color relevance of point p_i (an example is shown in Fig. 5) is computed as

$$D_c(p_i) = \begin{cases} 0 & \text{if } |C_{p_i}| \leq |W_{p_i}|/4 \\ 1/|C_{p_i}| & \text{if } |W_{p_i}|/4 \leq |C_{p_i}| \leq |W_{p_i}|/2 + \sqrt{|W_{p_i}|} \\ 0 & \text{if } |C_{p_i}| \geq |W_{p_i}|/2 + \sqrt{|W_{p_i}|} \end{cases} \tag{8}$$



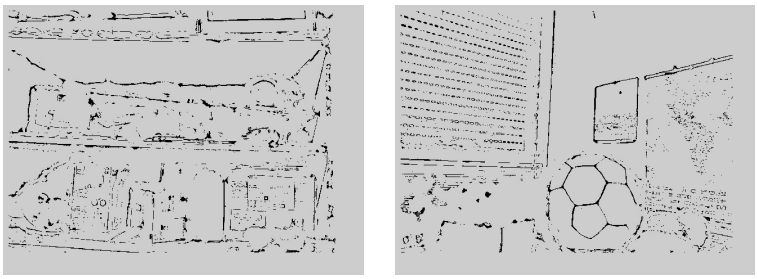
a) b) c)

Fig. 5 Color saliency corresponding to different values of T_c. a) Color view; b) Color saliency for $T_c = 7$; c) Color saliency for $T_c = 15$; Darker points correspond to larger values of $D_c(p_i)$.

Finally geometry and color distinctivity are combined together. According to the idea that a point is useful for registration purposes if it is able to bound the registration either by using geometry or color constraints, the distinctivity of a point is computed as the maximum of the color and geometric distinctivity:

$$D_p(p_i) = \max(D_g(p_i), D_c(p_i)) \tag{9}$$

Fig. 6 shows a couple of examples of the computed saliency values on two sample scenes. In order to build the set $\mathcal{P}_i$ of the relevant points of view V_i that will be used for the registration, a further constraint has been added in order to force a reasonably uniform spatial distribution (i.e., in order to avoid to have all the salient points concentrated in a spatial region of the scene). Namely each acquired view is divided into quadrants of $40x40$ pixels by a regular grid on the Kinect depth map, i.e., the $640x480$ depth map of the Kinect is sub-divided into $16x12 = 192$ quadrants. For each quadrant we search for the N_q highest saliency points which will become the salient points for the corresponding regions. If a quadrant contains $N_i < N_q$ salient points (e.g., because it corresponds to a flat and untextured region), the N_i salient points are selected and the missing $N_q - N_i$ points are taken from the other quadrants by increasing their number of salient points (i.e., for the Kinect each quadrant gets $(N_q - N_i)/(16x12-1)$ extra salient points). For the experimental results we used a total of $N_d = 2880$ salient points, i.e. $N_q = 2880/192 = 15$ points for each quadrant (less than 1% of the acquired samples).



a) b)

Fig. 6 Maximum of the color and geometric distinctivity $D_p(p_i)$ for two sample scenes framing a set of shelves (a) and a *ball* object in our lab (b). Darker points correspond to larger values of $D_p(p_i)$.

3.3 3D geometry registration with color-aware ICP

The Kinect sensor is used as an hand-held scanner and is moved around the scene in order to acquire I frames each corresponding to a 3D view $V_i, i = 1, .., N$ of the scene. The approach presented in the previous section can be used to extract the set $\mathcal{P}_i$ of the relevant points in view V_i that will be used as input for the registration algorithm. For the registration step we extended the Iterative Closest Points (ICP) algorithm [4]. Notice that a well-known drawback of this algorithm is the risk of falling into local minima, but since the employed depth cameras can acquire at high frame rates, the views are very close together and ICP can be applied directly to the acquired data without the need of a preliminary coarse registration.

The proposed approach is outlined in Algorithm 1. Firstly the relevant points $\mathcal{P}_i$ of view V_i are extracted by the method of Section 3.2. Then a 5-dimensional KD-tree is built where each sample has 5 dimensions, the 3 spatial coordinates (x, y, z) and the two color components a and b of the corresponding color value in the CIELAB color space. In order to allow the nearest neighbour search on the 5 dimensional representation that includes two completely different measurement spaces both the geometry and the color are normalized by their standard deviations σ_g and σ_c. The color is then further multiplied by a weighting factor (we experimentally set it to $k_{cg} = 1/7$), i.e., each sample is represented by the vector (x', y', z', a', b'), where:

$$x' = \frac{x}{\sigma_g} \quad y' = \frac{y}{\sigma_g} \quad z' = \frac{z}{\sigma_g} \quad a' = k_{cg}\frac{a}{\sigma_c} \quad b' = k_{cg}\frac{b}{\sigma_c} \tag{10}$$

A modified version of the ICP algorithm is then used to register the relevant points $\mathcal{P}_i$ over the previously aligned view V^r_{i-1} of the scene. Various extensions and variations of the ICP algorithm have been proposed in the literature [27], in this work we used as the distance between corresponding point the distance in the (x', y', z', a', b') space, i.e., the distance depends on both the geometrical distance in the (x, y, z) space and on the difference between the color of the two samples. The search algorithm used to find the correspondence has also been modified in order to use the 5-dimensional KD-tree and the improved distance measure. After the ICP algorithm reaches the convergence the set of relevant points $\mathcal{P}_i$ is analyzed and a new set $\mathcal{P}'_i$ is built by removing from the $\mathcal{P}_i$ the samples for which a good correspondence has not been found, i.e., only correspondences with a distance smaller than a threshold T_{icp} in the (x', y', z', a', b') space are preserved.

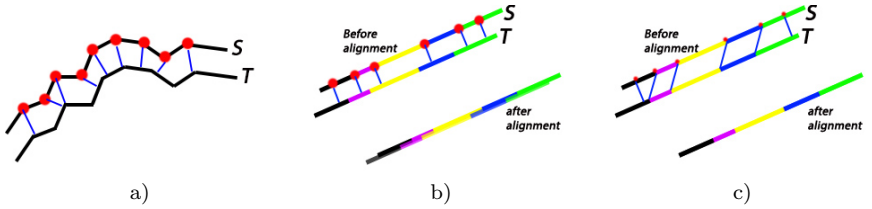
The remaining set of points $\mathcal{P}'_i$ is then used inside a second ICP procedure that performs a final refinement by using geometry information only. This allows to obtain an accurate geometry alignment and at the same time to avoid common ICP errors on regions or views with limited geometry details where geometry information alone is not sufficient to constrain the registration.

Algorithm 1 Color-aware ICP procedure

1: Extract salient points
2: Construct the 5 dimensional KD-tree
3: Run ICP with color and geometry based distances
4: Remove outliers
5: Run refinement ICP with pruned set of salient points and geometry information only.

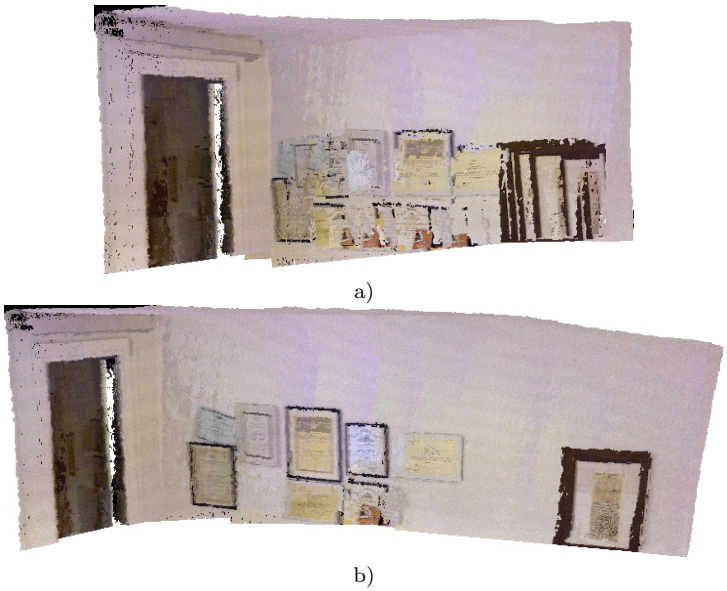
In order to understand how the proposed ICP algorithm with combined color and geometry distance can improve the registration performance a more detailed look to the registration process can be useful. If the two point clouds have a large amount of geometry details standard ICP can be applied with the euclidean distance in 3D space as shown in Fig. 7a. Unfortunately in the acquisition of large scenes (e.g., the room in the experimental results) it is quite

common to acquire large regions with just a large planar surface. Furthermore the low accuracy and high noise level of the Kinect and of the other consumer depth cameras makes difficult to constraint the registration on small objects or surface details as is done with 3D views from laser scanners. In this case, as shown in Fig. 7b, geometry information alone is not able to constraint the registration. In particular the planar surfaces can "slide" one over the other and the alignment is constrained only in the direction perpendicular to the plane while the alignment in the direction parallel to the plane surface is very unreliable. Fig. 8a shows an example of this problem on a real scene, note how the objects attached to the planar wall surfaces clearly shows the error accumulated in the registration process. In order to avoid the "sliding" effect the proposed approach considers also the color in the computation of the point distances in the registration algorithm. In the proposed algorithm the search for the correspondences in the euclidean 3D space has been modified by adding two further dimensions representing the color of the considered sample (luminance has not been used since it is not very stable across the different views, specially with reflective objects). In this way each point is related to the point on the target view that is spatially close but also have a similar color. Since on planar regions the salient points are typically selected on corner and edges of the texture information this approach allow to precisely align the edges of the objects in the color view and so to constraint the alignment even in the cases where geometry information is very limited or unreliable due to the low accuracy of the depth camera, as shown in Fig. 7c. Fig. 8b shows how by using also color data the scene of 8a can be correctly reconstructed avoiding the "sliding" effect. It is also interesting to notice that the use of salient points only for the new view that is added at each step allows to both drastically reduce the computation time and to improve the registration accuracy.



a) b) c)

Fig. 7 Alignment of a target point cloud T with an already registered source view S : a) alignment of two scenes containing enough geometry information to constraint the registration; b) alignment of two planar surfaces with geometry information alone; c) alignment of two planar surfaces with both color and geometry constraints.



a) b)

Fig. 8 Example of the reconstruction of a planar scene: a) Reconstruction with geometry-based distance; b) Reconstruction using color and geometry-based distance.

3.4 Global optimization of the registered views

The procedure is iterated until all the acquired views are processed. By registering each couple of views one after the other a complete 3D reconstruction is obtained but the registration error propagates and after registering several hundred frames it typically become quite large (consider that the employed consumer depth cameras are not very accurate). There exist many complex global optimization schemes for the registration of 3D views, but in order to handle very large number of views at the same time keeping very low memory and computation time requirements a simple technique based on the Explicit Loop Closing Heuristic (ELCH) method [32] has been used. During the registration process the algorithm stores in memory the viewpoint o_J, the viewing direction $\mathbf{v}_j$ and the centroid of the point cloud c_j corresponding to each view V_j. A graph of the connection between the different views is also built. After registering a new view V_j the algorithms compares the viewpoint and the viewing direction of V_j with the ones of the previously acquired views in order to estimate if any other view has more than 70% super-imposition with V_j. In order to make the computation very fast and to avoid to store in memory all the acquired views, the estimation is done by looking only at the viewing directions and at the viewpoint positions of the compared views. The set $\mathcal{N}(V_j)$ of the views connected to V_j is given by:

$$\mathcal{N}(V_j) = \{V_k : |o_k - o_j| < T_{pos}(|c_j - o_j|) \wedge |\mathbf{v}_k \cdot \mathbf{v}_j| < T_{angle}(|c_j - o_j|)\} \quad (11)$$

Views V_K and V_j are connected if the distance between the viewing positions $|o_k - o_j|$ is smaller than a threshold and if the angle between the two viewing directions is smaller than another threshold. Notice how both thresholds are not fixed but depend on the distance between the viewpoint and the object centroid, in order to account for the fact that the same distance between the viewpoint of the acquired views have a different impact depending on the distance of the object from the camera. E.g., a $10[cm]$ translation is quite relevant for the acquisition of an object at $50[cm]$ but much less for an object that is at $5[m]$ from the camera. Similar considerations hold also for the angular threshold T_{angle}. After computing which views are connected to V_j the graph is updated and the algorithms check if the introduction of V_j has created a new loop in the graph (i.e., in the chain of registered views) of length bigger than a threshold T_{loop}. Threshold T_{loop} avoids the construction of too short loops that may correspond just to roughly subsequent frames (for the results we set $T_{loop} = 50$). If a loop is detected the ELCH algorithm is used to refine the alignment of all the views involved in the loop by re-distributing the error in the position of the two views that close the loop on all the chain of registered views proportionally to the distance between the various views. This provision avoids the propagation of errors through the registrations of large amount of views.

3.5 Fusion of the geometry and color

After registering the new view over the previous acquired data it is necessary to fuse together the two point clouds in order to reduce the number of samples and to produce the final surface. For this task we firstly create a merged point cloud containing all the samples from both V_i and S_{i-1}. Then each point of the set $V_i \cup S_{i-1}$ is analyzed and if another point with a distance smaller than a threshold t_{res} (the threshold depends on the desired final model resolution) is found then a single 3D point is kept. This simple fusion algorithm allows reasonable performance within very limited computation time. Clearly offline accurate reconstructions can afford more complex fusion schemes.

The proposed approach is focused on the reconstruction of an accurate geometry, however color data also need to be added to the acquired geometry. Each acquired sample has the associated color information, but in the fusion step it is necessary to assign a color value to the samples obtained by merging points coming from different views. For each 3D sample p_i the corresponding color value must be computed from the different color values of the various points that have been merged into it. For this task a simple weighting function $W(p_i, V_j)$ that represent the reliability of the color of p_i in view V_j is built depending on two clues:

- The first is the angle between the the normal $\mathbf{n}_{p_i}$ of the surface at p_i and the viewing direction $\mathbf{v}_j$ corresponding to the view V_j. The idea is that the color data corresponding to the view in which the normal is better aligned

with the viewing direction (i.e., the one that maximizes $|\mathbf{n}_{p_i} \cdot \mathbf{v}_j|$) is more reliable.
– Since specular reflections are typically associated to very high luminance values, we underweight samples with a very high luminance value in presence of corresponding samples in other views with smaller luminance. More precisely for luminance values greater than 200 (a $[0, 255]$ luminance range is considered) we linearly decrease the reliability of the acquired data.

The employed relevance function is thus the following:

$$W(p_i, V_j) = \frac{1}{2}(|\mathbf{n}_{p_i} \cdot \mathbf{v}_j|) + \frac{1}{2} \min\left(1, \frac{(255 - L)}{55}\right) \tag{12}$$

Finally the function $W(p_i, V_j)$ is used to compute a weighted average in order to get the considered sample :

$$c = \frac{\sum_j R(p_i, V_j) c(i, j)}{\sum_j R(p_i, V_j)} \tag{13}$$

Notice how the proposed scheme is focused on the reconstruction of an accurate geometry, this simple euristic allows to obtain a reasonable color estimation but further research will be developed to the improvement of the color reconstruction module.

4 Experimental results

The effectiveness of the proposed approach was tested on several different objects and scenes. This section presents first some sample results on generic objects and scenes and then a complete evaluation of the performance on a set of statues in order to verify its applicability to cultural heritage data.

Experimental evaluation has been performed by acquiring several hundred frames for each considered scene and object using the Kinect camera. A frame rate of $10fps$ has been used and each acquired frame is made by a $640x480$ depth map and a $1280x1024$ color image (bilinear interpolation has been used to assign the color values to the lower resolution depth information). Note that this means that each scene or object has been acquired in just a couple of minutes, i.e., much faster than any standard 3D scanning techniques. Furthermore part of the cultural heritage objects used for the evaluation have also been acquired by a NextEngine 3D laser scanner and modeled by commercial 3D modeling tools in order to have a ground truth to validate the proposed approach. Unfortunately the NextEngine has a very high accuracy but a limited acquisition range, so it was not possible to get ground truth data for the rooms and for some of the larger objects.

Fig. 9 shows the reconstruction of a simple office scene. This scene has quite limited texture information but relevant geometrical features that can bound the registration. In this case the alignment process is robust both for the proposed approach and for standard geometry-based schemes. This example

Fig. 9 Reconstruction of a simple office scene (from 100 frames).

together with the one of Fig. 8b shows how the proposed approach is able to exploit the correct clues when only one of the two types of information (either color or geometry) is adequate.

Fig. 10 shows the reconstruction of a seated person obtained from about 800 frames acquired by moving the Kinect around the person. This case is slightly more difficult than the previous one since corner-less shape can lead to errors in alignment schemes based on geometry only. The proposed approach is instead able to correctly reconstruct the shape of the person exploiting both geometry and texture information. Note also that the scene around the acquisition place is not light-controlled and there are reflections and shadows affecting the color component of the distance measure employed by the ICP, but the choice of ignoring lightness makes the proposed approach robust with respect to this issue.

Fig. 11 shows our approach applied to a colored car. This scene is much more challenging since the acquisition was made outdoor where sun-light interferes with the IR pattern projected by the Kinect [9]. Furthermore the car's surface is very reflective and reflects the scene all around the car itself. The car is also a large object requiring a larger number of frames (around 1300). Finally the car shape has large uniform regions with very little features. The picture on the left is a view of the back of the car 3D model, while the right picture shows a lateral view of the 3D model with the roof purposely removed in order to show the interior. The latter allows to recognize several elements inside the car like the steering wheel and the seats. The feature extraction approach turns out to be fundamental to constrain the alignments in this scene, since many frames have poor geometric and color information and only the combined use of the two information sources permits a correct alignment. The proposed approach can be applied not only to closed objects but also to open scenes (a possibility not available for instance to many volumetric schemes that

a) b) c) d)

Fig. 10 Reconstruction of a seated person (from 800 frames).

a) b)

Fig. 11 Reconstruction of a Chrevolet Aveo (from 1300$frames$). a) View from the back; b) side view with the roof purposely removed in order to show the interior.

assume water-tight objects). Fig. 12 shows a section of the reconstructed surface of the interior of our research lab performed from 800 frames. This scene is particularly challenging since there are large flat surfaces (walls) without relevant geometry information but with texture due to the posted pictures and posters. Since the proposed approach exploits texture when geometry information is completely lacking it gives a complete and accurate 3D model of the scene. The left side is a good example of how color can bound the registration in areas with just a flat wall that would have been very difficult to reconstruct using geometry alone. Fig. 13 shows a larger section of the reconstructed 3D scene from which one can also appreciate how the global optimization of Sec. 3.4 avoids the propagation of the registration error in large scenes.

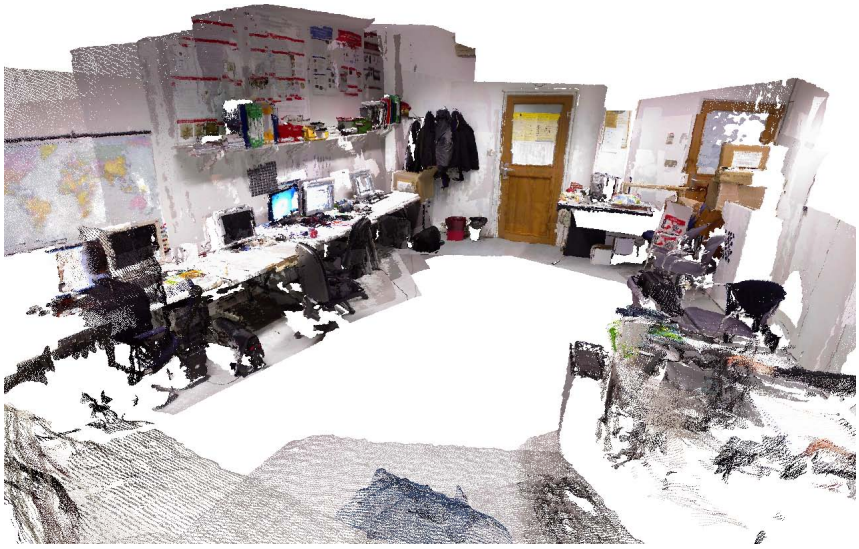


Fig. 12 Section of the reconstruction of the LTTM research lab.

In order to evaluate the effectiveness of the proposed approach in the cultural heritage field, we acquired a set of statues from the atelier of Gino Cortelazzo (1927-1985) [2], an Italian sculptor. The acquired artworks are of different sizes and of different colors and materials, thus they represent a good testbed of the objects of interest for the cultural heritage field. Some of them were covered by a quartz crystals varnish that causes the surface to be not only very irregular and so difficult to acquire but also very reflective making the acquisition very challenging by any scanning device and not only by consumer depth cameras. Table 1 shows the different characteristics of the various acquired artworks. The size of the *Clerk* and *Castle* statues is rather large and their acquisition by the commonly used active scanners require a great amount of views and very long scanning and processing times, while with the proposed method they were acquired and reconstructed in a few minutes.

Fig. 13 Overview of the reconstruction of the LTTM research lab (from 800 frames).

Artwork	Material	Size $[mm]$	Error ($\sigma[mm]$)	RMS ($[mm]$)
Impiegato (*Clerk*)	wood	1800x300x250	7.52	7.54
Civetta nera (*Black Owl*)	polystyrene, quartz	600x500x200	7.49	7.8
Civetta rosa (*Pink Owl*)	polystyrene, quartz	450x240x200	3.83	3.98
Donne e Gabbiano (*Women and Seagull*)	gypsum	550x500x500	5.13	5.21
Il Castello (*The Castle*)	iron, quartz	1300x600x450	6.67	6.7

Table 1 Accuracy of the proposed approach on some sample artworks. All the reported measures are in $[mm]$. *Quartz* denotes the covering by an epossidic quartz varnish.

The statues were acquired both by the proposed approach and by a NextEngine laser scanner in order to produce also a ground truth needed to validate the proposed method. Fig. 14 shows the reconstructions obtained by the proposed algorithm and the corresponding error maps. The proposed method produces accurate reconstructions of the complex geometries of the different artworks with accuracy ranging from around $3,8[mm]$ to $7,5[mm]$ depending on the specific object. If the obtained results are compared with the quantization error of the Kinect of Eq. 1, considering that the sensor was held at a distance of about a couple of meters from the acquired objects, it is possible to notice that the obtained error is even smaller than the quantization step, i.e., the proposed approach is able to improve the accuracy by combining multiple views. The *Pink Owl* is the most accurate object with a standard deviation on the acquired points of just $3,8[mm]$ in spite of the quartz varnish covering it. This is in accordance with the fact that *Pink Owl*, given its small size, was acquired with the sensor closer from it than the other objects. A very good accuracy (about $5[mm]$) has also been obtained on the *Women and Seagull* since the material is much smoother and less reflective. Even if the quartz crystals of the *Pink Owl*, *Black Owl* and *Castle* statues are very reflective and the surface is very rough, the proposed reconstruction pipeline still obtains a

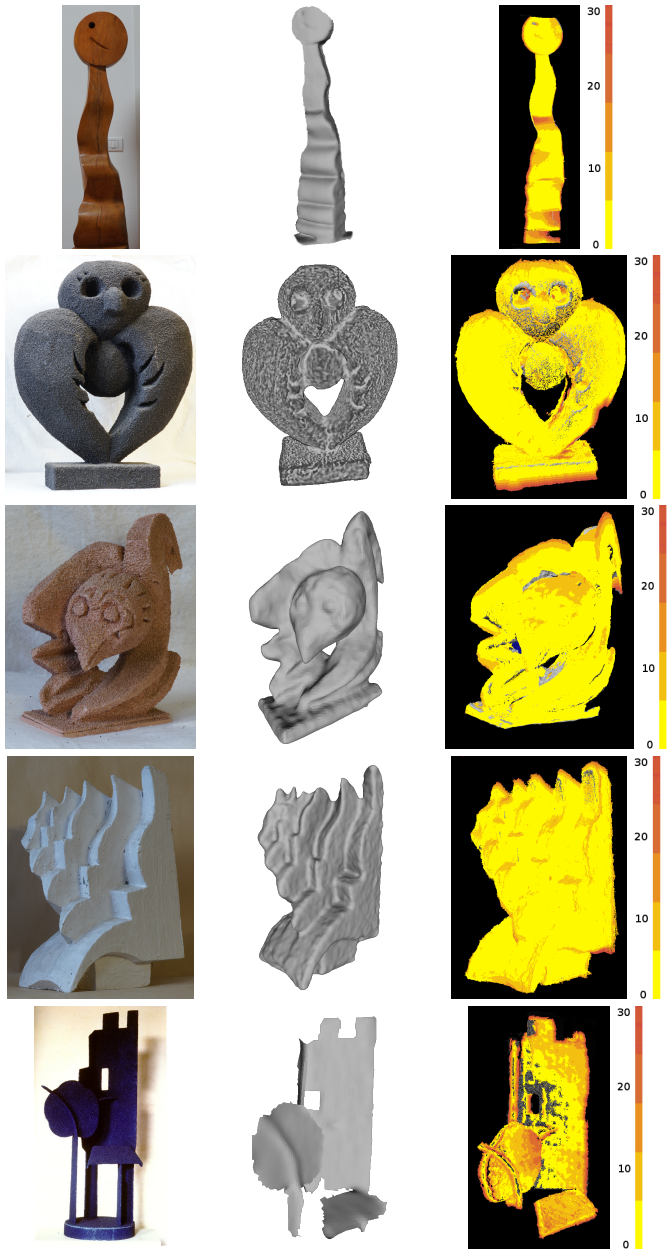


Fig. 14 Example of reconstructed artworks: a) Image of the artwork; b) Snapshot of a 3D mesh reconstructed from the computed point cloud; c) Error between the obtained point cloud and the ground truth (the colormap goes from 0[mm] in the yellow regions to 30[mm] in the red ones). Gray points are the ones for which no ground truth is available.

good accuracy on them (around $7[mm]$). Finally notice how even a quite large object like the *Clerk*, that is almost $2[m]$ tall, can be acquired with an accuracy comparable to the other ones (i.e., the obtained accuracy is always less than $0,4\%$ of the size of the object). Fig. 15 shows also some snapshots of colored point clouds. The proposed algorithm properly assigns color data to the reconstructed meshes, e.g., the *Clerk* in Fig. 15a. The *Pink Owl* (Fig. 15b) is completely covered of reflective quartz crystals that are very difficult to acquire with the low quality color camera of the Kinect. The results are reasonable but not very impressive, for a better reconstruction of this challenging material an high quality color camera should be associated to the Kinect. Finally notice how the proposed scheme was able to acquire the semi-transparent alabaster of the *Metamorphosis* statue of Fig. 15c that is very difficult to acquire with standard methods like laser scanners.

a) b) c)

Fig. 15 Colored snapshots of some reconstructed artwork: a) Clerk; b) Pink Owl; c) Metamorphosis.

Since the proposed reconstruction algorithm is also able to acquire larger scenes some rooms of the sculptor atelier were acquired. It was not possible to acquire the ground truth for these scenes since they are too large to be acquired with the available laser scanner. However Figs. 16 and 17 show some snapshots of the performed reconstructions for two different rooms. The shelves on the walls contain a lot of small objects and it is possible to notice how many small details of their complex geometries were captured.

Fig. 18 shows instead two different wood sculptures (i.e., the *Two Roosters* and *Character* artworks) in their current positions in the sculptor's atelier. The proposed approach is suited to perform the reconstruction of museums rooms, showing that depth camera used as handheld scanners can reconstruct museums environment very simply and rapidly thus obtaining valuable data for virtual reconstruction without all the burden associated to the 3D modeling of a large environment by standard approaches.

Finally some considerations on computation time and memory usage: for global optimization of the registration our approach exploits on-line loop detection together with the ELCH algorithm and is able to perform the optimization keeping in RAM only the last two frames and the computed position

Fig. 16 Snapshot of the reconstruction of the sculptor's lab

Fig. 17 Snapshot of another room of the sculptor's lab.

of the previously aligned frames, thus keeping the memory usage very low and allowing to use a very large number of frames. Computation time requirements are also very limited. The current implementation has not been fully optimized but it is able to process around 1 frame per second on a 2,4 Ghz Intel Q6600 processor. Notice that multi-threading has still to be included (i.e., the algorithm exploits a single core). By fully optimizing the code and employing multi-threading, besides eventually exploiting also the GPU, the proposed approach will be able to run in real-time. This will allow reconstructing the 3D objects while the acquisition goes on thus allowing the user to control the 3D reconstruction results during the acquisition process. This is very useful since the user can notice where holes and missing areas are and thus choose an optimal acquisition path for the various views.

a) b)

Fig. 18 Snapshots of a couple of artworks inside their setting: a) Due Galli (Two Roosters); b) Personaggio (Character).

5 Conclusions

This paper proposes a novel solution for the 3D reconstruction of cultural heritage objects allowing accurate three-dimensional reconstructions without expensive hardware and time-consuming procedures. In particular the proposed approach allows to use the Kinect as an handheld scanner for very fast 3D reconstructions. The proposed pipeline has been explicitly targeted to the characteristics of the acquired data and is able to effectively exploit the side information coming from the color camera. Color information has been used both to extract salient points and to compute the distances between corresponding points in the ICP algorithm. Quite notably the reliable extraction of salient points allowed to use a smaller number of points in the registration process greatly speeding-up the 3D reconstruction algorithm. The use of salient points and color information in order to assist the registration process gives reliable 3D reconstructions also in situations where geometry information is not sufficient to constraint the registration. Experimental results proved the effectiveness of the proposed approach in the acquisition of artwork of different sizes and materials, often with accuracies even lower than the ones provided by the employed sensor.

Further research will be devoted to improve the final fusion step with respect to both geometry and color data. The final global alignment step will be improved in order to increase the reconstruction accuracy together with the realization of a more refined color fusion scheme. Finally the use of other consumer depth cameras, e.g., the second generation of the Kinect sensor and current consumer Time-Of-Flight sensor, will also be considered.

Acknowledgements We would like to thank Luca Palmieri for his contributions to the color fusion algorithm. Thanks also to Fabio Dominio and Francesco Michielin for their help in the acquisition of the experimental results data.

References

1. Andreetto M, Brusco N, Cortelazzo G (2004) Automatic 3d modeling of textured cultural heritage objects. Image Processing, IEEE Transactions on 13(3):354–369
2. Argan GC (2001) Il secondo Novecento - L'Arte Moderna. Sansoni
3. Bernardini F, Rushmeier H (2002) The 3d model acquisition pipeline. In: Computer Graphics Forum, vol 21, pp 149–172
4. Besl PJ, McKay ND (1992) A method for registration of 3-d shapes. IEEE Trans on PAMI 14(2):239–256
5. Cappelletto E, Zanuttigh P, Cortelazzo GM (2013) Handheld scanning with 3d cameras. In: Multimedia Signal Processing (MMSP), 2013 IEEE 15th International Workshop on, pp 367–372
6. Cui Y, Schuon S, Derek C, Thrun S, Theobalt C (2010) 3d shape scanning with a time-of-flight camera. In: In Proc. of CVPR 2010
7. Cui Y, Schuon S, Thrun S, Stricker D, Theobalt C (2013) Algorithms for 3d shape scanning with a depth camera. Pattern Analysis and Machine Intelligence, IEEE Transactions on 35(5):1039–1050
8. Dal Mutto C, Zanuttigh P, Cortelazzo G (2010) A probabilistic approach to tof and stereo data fusion. In: 3DPVT, Paris, France
9. Dal Mutto C, Zanuttigh P, Cortelazzo GM (2012) Time-of-Flight Cameras and Microsoft Kinect. SpringerBriefs in Electrical and Computer Engineering, Springer
10. El-Hakim S, Beraldin JA, Picard M, Godin G (2004) Detailed 3d reconstruction of large-scale heritage sites with integrated techniques. Computer Graphics and Applications, IEEE 24(3):21–29
11. Fossati A, Gall J, Grabner H, Ren X, Konolige K (eds) (2013) Consumer Depth Cameras for Computer Vision: Research Topics and Applications. Advances in Computer Vision and Pattern Recognition, Springer
12. Henry P, Krainin M, Herbst E, Ren X, Fox D (2012) Rgb-d mapping: Using kinect-style depth cameras for dense 3d modeling of indoor environments. Int Journal of Robotics Research 31(5):647–663
13. Herrera C D, Kannala J, Heikkil J (2012) Joint depth and color camera calibration with distortion correction. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34(10):2058–2064
14. Holzer S, Rusu RB, Dixon M, Gedikli S, Navab N (2012) Adaptive neighborhood selection for real-time surface normal estimation from organized point cloud data using integral images. In: Proc. of IROS, pp 2684–2689
15. Khoshelham K (2011) Accuracy analysis of kinect depth data. In: Proc. of ISPRS Workshop Laser Scanning, vol 38
16. Khoshelham K, Elberink SO (2012) Accuracy and resolution of kinect depth data for indoor mapping applications. Sensors 12(2):1437–1454

17. Khosravani A, Lingenfelder M, Wenzel K, Fritsch D (2012) Co-registration of kinect point clouds based on image and object space observations. In: Proceedings of LC3D Workshop
18. Laurentini A (1994) The visual hull concept for silhouette-based image understanding. Pattern Analysis and Machine Intelligence, IEEE Transactions on 16(2):150–162
19. Levoy M, Pulli K, Curless B, Rusinkiewicz S, Koller D, Pereira L, Ginzton M, Anderson S, Davis J, Ginsberg J, Shade J, Fulk D (2000) The digital michelangelo project: 3d scanning of large statues. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH '00, pp 131–144
20. Mara H, Kampel M, Niccolucci F, Sablatnig R (2007) Ancient coins and ceramics - 3d and 2d documentation for preservation and retrieval of lost heritage. In: 2nd ISPRS International Workshop 3D-ARCH 2007
21. Newcombe RA, Izadi S, Hilliges O, Molyneaux D, Kim D, Davison A, Kohli P, Shotton J, Hodges S, Fitzgibbon A (2011) Kinectfusion: Real-time dense surface mapping and tracking. In: Proc. of IEEE ISMAR
22. Pavlidis G, Koutsoudis A, Arnaoutoglou F, Tsioukas V, Chamzas C (2007) Methods for 3d digitization of cultural heritage. Journal of Cultural Heritage 8(1):93 – 98
23. PCL (2013) Point cloud library (pcl). http://pointclouds.org/
24. Pollefeys M, Proesmans M, Koch R, Vergauwen M, Van Gool L (2008) Detailed model acquisition for virtual reality,. In: Virtual Reality in Archaeology, ArcheoPress, pp 71–77
25. ReconstructMe (2013) Reconstructme. http://reconstructme.net/
26. Remondino F (2011) Heritage recording and 3d modeling with photogrammetry and 3d scanning. Remote Sensing 3(6):1104–1138
27. Rusinkiewicz S, Levoy M (2001) Efficient variants of the icp algorithm. In: Proc. of 3-D Digital Imaging and Modeling, pp 145–152
28. Seitz S, Curless B, Diebel J, Scharstein D, Szeliski R (2006) A comparison and evaluation of multi-view stereo reconstruction algorithms. In: Proc. of CVPR, vol 1, pp 519 – 528
29. Shotton J, Fitzgibbon A, Cook M, Sharp T, Finocchio M, Moore R, Kipman A, Blake A (2011) Real-time human pose recognition in parts from a single depth image. In: Proceedings of CVPR
30. Skanect (2013) Skanect. http://skanect.manctl.com/
31. Snavely N, Seitz SM, Szeliski R (2008) Modeling the world from internet photo collections. Int J Comput Vision 80(2):189–210
32. Sprickerhof J, Nüchter A, Lingemann K, Hertzberg J (2009) An explicit loop closing technique for 6d slam. In: Proc. of ECMR
33. Tomasi C, Manduchi R (1998) Bilateral filtering for gray and color images. In: Proc. ICCV
34. Tong J, Zhou J, Liu L, Pan Z, Yan H (2012) Scanning 3d full human bodies using kinects. Visualization and Computer Graphics, IEEE Transactions on 18(4):643–650

35. Torsello A, Rodol E, Albarelli A (2011) Sampling relevant points for surface registration. In: Proc. of 3DIMPVT 2011
36. Whelan T, Johannsson H, Kaess M, Leonard J, McDonald J (2012) Robust tracking for real-time dense RGB-D mapping with Kintinuous. Tech. rep., MIT
37. Young MK, Theobalt C, Diebel J, Kosecka J, Miscusik B, Thrun S (2009) Multi-view image and tof sensor fusion for dense 3d reconstruction. In: Proc. of ICCV Workshops, pp 1542 –1549