Handheld scanning with ToF sensors and cameras

Enrico Cappelletto, Pietro Zanuttigh, Guido Maria Cortelazzo
Dept. of Information Engineering, University of Padova
enrico.cappelletto,zanuttigh,corte@dei.unipd.it

Abstract

Novel real-time depth acquisition devices, like Time-Of-Flight (ToF) cameras or Microsoft Kinect, allow a very fast and simple acquisition of 3D views. These sensors have been used in many applications but their employment for 3D scanning purposes is a very challenging task due to the limited accuracy and reliability of their data. In this paper we present a 3D acquisition pipeline explicitly targeted to ToF cameras. The proposed scheme aims at obtaining a reliable reconstruction that is not affected by the limiting issues of these cameras and is at the same time simple and fast in order to allow to use the ToF sensor as an handheld scanner. In order to achieve these targets an ad-hoc pre-processing stage is used together with a modified version of the ICP algorithm that is able to recognize the most reliable and relevant points and to use only them for registration purposes. Furthermore, ToF cameras have also been combined with standard color cameras in order to acquire colored 3D models. Experimental results show how the proposed approach is able to produce reliable 3D reconstructions from the ToF data.

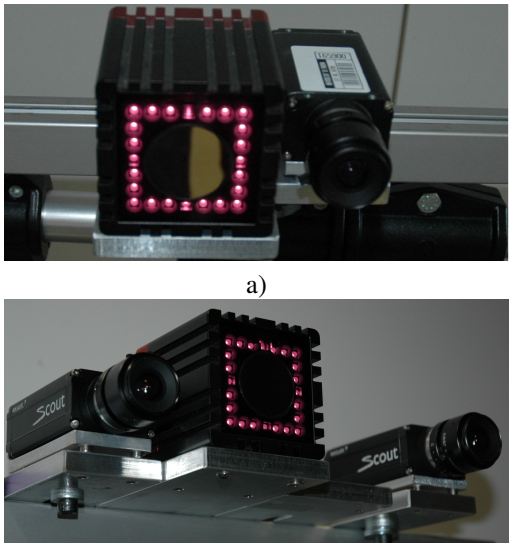
1. Introduction

The reconstruction of the three-dimensional shape of complex objects has always been a very challenging task. Most approaches rely on the acquisition of a set of different 3D views of the object and then fuse such views into a complete 3D shape representation. Both tasks are very challenging. The acquisition of the 3D views has been traditionally solved by 3D structured light scanners but these devices are very expensive, slow and cumbersome to use. Passive methods, most notably stereo vision approaches [10], have also been employed but they are not very robust. The recent introduction of real-time 3D acquisition devices, like ToF cameras and the Kinect, has made the acquisition of 3D views much faster and simpler than before. Unfortunately these devices have also several limiting issues, like high noise level, limited resolution or artifacts in proximity of edges, that the algorithms employed for the reconstruction of 3D shapes from their data must take into account.

While these devices have been widely used in setups with a fixed camera acquiring moving people or objects for dynamic 3D acquisition and human motion tracking [11], their employment for the reconstruction of static 3D scenes is a novel research field. Among the research projects which have investigated this task, Microsoft's KinectFusion [7] is probably the most relevant. In this project each frame acquired by the Kinect is registered in real-time using the ICP algorithm [1] over the complete 3D scene description reconstructed by using a variation of the volumetric truncated signed distance function (TSDF) [3]. Similar results can also be obtained using Time-Of-Flight (ToF) cameras in place of the Kinect. For example in [2] the data acquired by a MESA SR4000 ToF camera are firstly improved with a super-resolution algorithm and then the different 3D views are aligned and combined together. The latter task uses a probabilistic approach based on Expectation Maximization (EM). Color cameras can also be employed together with the depth sensors in order to improve the reconstruction accuracy. In [6] the data acquired by the ToF sensor is firstly used to reconstruct a coarse 3D representation of the scene by a volumetric approach. Then the data coming from multiple color cameras are used in order to improve the reconstruction by enforcing a photoconsistency measure and silhouette constraints. Furthermore the first commercial applications exploiting the Kinect for 3D reconstruction, like *Reconstructme* [8] and *Kinect@home* [9] are starting to appear.

This work presents a novel 3D reconstruction pipeline to obtain textured 3D models in real-time from the data acquired by the combination of a ToF camera with a standard color camera. As [7] and other schemes, the proposed approach exploits the ICP algorithm but, with respect to the previous approaches, we introduce new elements in order to adapt the reconstruction pipeline to ToF data. Mainly an ad-hoc pre-processing step is used in order to reduce the noise level and to remove the most common artifacts typically affecting in the data acquired by ToF cameras. The paper is organized as follows: Section 2 introduces the proposed acquistion system, then the proposed 3D reconstruc-

tion pipeline is described in detail in Section 3. Section 4 presents the experimental results and finally Section 5 draws the conclusions.



a)

b)

Figure 1. Employed acquisition setup. a) ToF camera and standard color camera; b) extended version with 2 color cameras.

2. Proposed acquisition system

Since a Time-Of-Flight camera can not acquire color information (except for the reflectivity in the employed IR band) we combined it with a standard color camera in order to have a system capable of acquiring both geometry and color information. In particular the used acquisition setup, depicted in Figure 1, is made by a MESA SR4000 Time-Of-Flight camera together with a Basler Scout A1000 color camera. A second color camera can also be added to the setup in order to improve the calibration accuracy and to reduce occlusion issues. In order to use this combined setup the two devices need to be precisely calibrated together. For this purpose we used a variation of the method proposed in [4]. Furthermore if a second camera is available it is also possible to build two different 3D point clouds, one from the checkerboard corners acquired by the ToF sensor and the other obtained from the two color cameras considered as a stereo system. An iterative approach can then used to find the extrinsic parameters that correspond to the best alignment between the two point clouds. After calibrating the two devices, for each depth frame each 3D sample p_i acquired by the ToF camera can be reprojected to a location $p_{cam} = (u, v)$ in the color frame. In this way a color value can be associated to each 3D point by using bilinear interpolation on the color samples surrounding (u, v). Finally to avoid issues on samples visible from the ToF viewpoint but not from the camera one, the Z-buffer algorithm is used in order to check for occlusions (i.e., two 3D points can be associated to the same 2D location and the color value refers only to the 3D point closer to the camera).

3. Geometry reconstruction pipeline

The proposed 3D reconstruction pipeline, shown in Figure 2, is made of 4 basic steps: the pre-processing of depth information acquired by the ToF; the extraction of the salient points that will be used for the registration process; the alignment (registration) of the views; and finally the surface simplification and polishing.

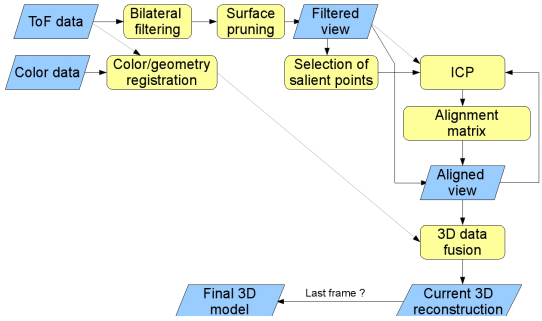


Figure 2. Architecture of the proposed system

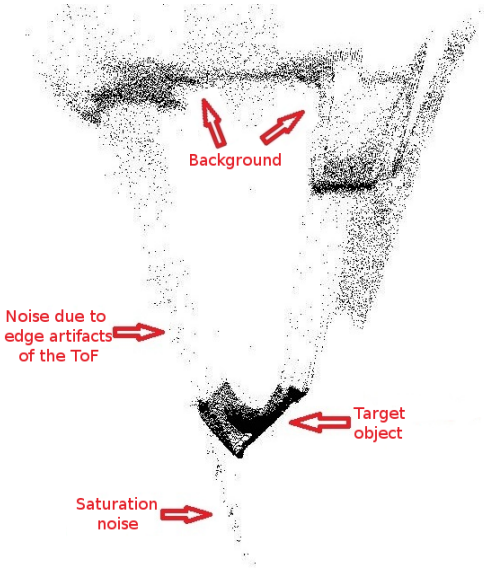


Figure 3. Sample view of the *ball* scene from the experimental results showing some examples of the artifacts in the data acquired by the ToF sensor.

3.1. Pre-processing of depth information

If compared with the data acquired by standard structured light or 3D laser scanners the 3D views acquired by

ToF sensors are characterized by limited accuracy, high noise level and by the presence of many erroneous depth samples. Fig. 3 shows an example of point cloud acquired by the ToF sensor: note the large amount of erroneous points. On both sides of the object there are many dangling points due to edge artifacts (basically, due to the limited resolution of the ToF, many edge pixels capture the reflected light from both foreground and background regions and the measured depth is a weighted average of the two distances). For these reasons they can not be directly sent to the reconstruction pipeline, but need a cleaning and refinement stage in order to remove unreliable data that can then affect the registration and reconstruction stages.

The proposed pre-processing algorithm is made by 2 basic steps. In the first a bilateral filter [12] is applied to the depth information acquired by the ToF sensor in order to reduce the noise but at the same time preserve the edges. The edge smoothing behavior property of standard low-pass filtering would create severe artifacts at the objects' boundaries, while the bilateral filtering scheme ensures that the shape boundaries are correctly preserved.

In the next step from the depth map and the calibration information we build the corresponding set of 3D points $p_i = (X_i, Y_i, Z_i), i = 1, .., N$. Then we consider the window W_{p_i} of size $k \times k$ surrounding each sample p_i in the depth map (for the experimental results we set $k = 3$). We compute 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. If the number of samples in S_{p_i} is large enough ($|S_{p_i}| > 0.8|W_{p_i}|$) the point p_i is considered valid, otherwise the point is on a too slanted surface or an isolated point and it is discarded. Note how this quite strict thresholding is necessary due to the high unreliability of ToF data, specially in proximity of edges, where the sensor pixel captures light coming from different surfaces at different distances thus providing unreliable depth values as described in [5] and shown in Fig. 3.

3.2. Extraction of salient points

The Iterative Closest Points (ICP) algorithm [1] requires to select a subset of the acquired points to compute the rototranslation matrix between a pair of views. This step is particularly critical in the proposed setup since the data acquired from the ToF sensor have many unreliable points and 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 which is both reliable and meaningful for registration purposes.

To achieve this target we used a saliency measure [13] that computes the usefulness of each point for registration purposes by taking into account geometry information in a neighborhood of the considered point. The idea is that the more distinctive points (i.e. the ones in regions of high curvature) are the most salient ones.

In particular we compute the normal n_p to the surface at each point p and then we associate to each sample p the set

$$A_p = \{(p' \in W_p) \wedge (\mathbf{n}_{p'} \cdot \mathbf{n}_p > T_g)\} \quad (1)$$

of the points for which the surface normals $\mathbf{n}_{p'}$ form an angle smaller than $\arccos(T_g)$ with the normal $\mathbf{n}_p$ of point p. The area of A_p 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 $|A_p| \geq 1$. The geometric distinctivity measure is computed as the inverse of the cardinality of A_p:

$$D_g(p) = \frac{1}{|A_p|} \quad (2)$$

Note that 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 $1/k^2 \leq D_g(p) \leq 1$, i.e. $D_g(p) = 1$ corresponds to the most salient points and $D_g(p) = 1/k^2$ to quite flat regions. The idea is that points corresponding to high curvature regions can be considered more distinctive since they force tighter bounds on the surface alignment. Fig. 4 shows an example of the computation of geometric distinctivity on a sample 3D view for different values of the threshold T_g.

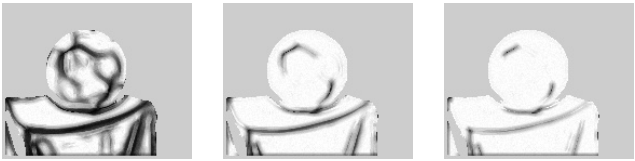


Figure 4. Geometric saliency corresponding to different values of T_g. Darker points correspond to larger values of $D_g(p)$

In order to build the set of relevant points $\mathcal{P}$ that will be used to register the considered view we selected the N_d more distinctive points (for the experimental results we set $N_d = 500$).

3.3. Real-time 3D geometry registration

The ToF 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. Let us denote with $\mathcal{V}_i$ the set of points relative to view V_i. 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 are then used as input for the registration algorithm. The proposed algorithm is based on the ICP method and works in the following way:

1. The relevant points $\mathcal{P}_i$ of view V_i are extracted by the method of Section 3.2.

2. The ICP algorithm is used to register the relevant points $\mathcal{P}_i$ over the previously aligned view V^r_{i-1} of the scene. Finally, once the ICP algorithm has converged to a rotation R_i and a translation $\mathbf{t}_i$, view V_i is rototranslated according to R_i and $\mathbf{t}_i$ and the rototranslated set of points $\mathcal{V}^r_i$ is added to the current scene reconstruction S'_i.

3. The resulting point cloud is polished using the method of Section 3.4 in order to produce the 3D scene description S_i.

4. The procedure is iterated until all the acquired views are processed.

It is also interesting to notice that the use of salient points only for the new view that is added at each step in the registration process does not only allow to drastically reduce the computation time but also to improve the registration accuracy (specially if compared with random subsampling approaches) since the points used in the registration process are the most relevant ones.

3.4. 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 the two points are collapsed together into a single 3D point.

Finally, since the aim of the proposed reconstruction technique is to build colored 3D models, it is also necessary to add color data to the acquired geometry. A color value is associated to each sample in each acquired view using the method described in Section 2. In the fusion step it is necessary to assign a color value to the samples obtained by merging the points coming from different views. For this task we consider the normals $\mathbf{n}_{p_i}$ corresponding to the 3D samples that are going to be merged together with the viewing direction $\mathbf{v}_j$ corresponding to the view V_j in which each point p_i has been acquired and we assign to the merged sample the color value corresponding to the sample for which $-(\mathbf{n}_{p_i} \cdot \mathbf{v}_j)$ is maximum.

4. Experimental results

In order to evaluate the effectiveness of the proposed approach we acquired several different scenes by the proposed acquisition setup, made a by a MESA SR4000 ToF camera (with a resolution of 176×144) together with a Basler Scout A1000 color camera (with a 1034×779 resolution) as shown in Fig. 1. We acquired about 150 frames at $30 fps$ for each considered scene. Note that this means that each scene has been acquired in just about $5 sec$.

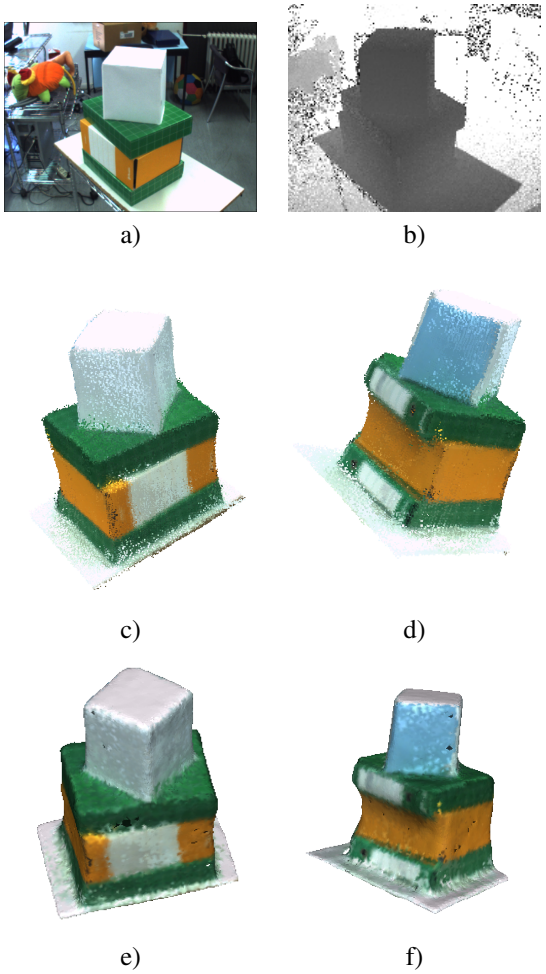


Figure 6. Reconstruction of the *boxes* scene: a) Color view of the scene; b) Depth map representing one view acquired by the ToF sensor (*contrast-stretched for visualization purposes*); c) Snapshot of the obtained point cloud; d) Snapshot of the obtained point cloud from another viewpoint; e) Snapshot of the 3D model obtained by simplifying and triangulating the point cloud; f) Snapshot of the 3D model from another viewpoint.

Fig. 5 shows a first example relative to the reconstruction of a ball. The ball has been acquired by moving the acquisition setup on a circle around it (for this reason the top and bottom area are missing since they have not been acquired). The geometry reconstruction of the ball in Fig. 5 is not trivial, in spite of the simplicity of the object's shape, as its regularity and symmetry properties may lead to alignment errors when the scene points are randomly sampled. By using relevant points along the table edge as the input of Section 3.2 algorithm, instead, the scene geometry in Fig. 5 is correctly reconstructed. The basic shape of the ball is recognized and the texture is correctly aligned on the ge-

a) b) c)

Figure 5. Reconstruction of a *ball* scene. a) Color view of the scene; b) Snapshot of the obtained 3D model; c) Another snapshot of the model rendered from a different viewpoint more to the top showing the alignment of the different views.

ometry, even if, as expected, the precision is below the one of laser or structured light scanners. However consider that the ToF camera has an accuracy of the order of $1cm$ on flat surfaces and quite lower in many practical situations and that the acquisition time is just $5sec$ for the whole scene. By comparing Fig. 3 with the reconstruction results of Figs. 5b) and Figs. 5c) it is possible to appreciate that the proposed method can cope well with the acquired data reliability and noise problems previously mentioned. Another issue is that there is a small error accumulation due to the lack of a final global alignment step. In this work this step has not been included in order to favor fast on-line reconstruction, however a further off-line post-processing stage including global alignment (inevitably much slower than the current version) is under development.

a) b)

Figure 7. Reconstruction of a *teddy bear* object. a) Color view of the teddy-bear; b) Obtained 3D reconstruction.

Fig. 6 shows the reconstruction of another scene made by a set of boxes of different size. Also in this case, as shown in Figs. 6c) and 6d) the objects in the scene are correctly reconstructed. Considering that the data provided by the ToF sensor to the reconstruction algorithm is a set of 3D views of the type shown in Fig. 6b), the algorithm appears capable to produce a 3D reconstruction that is not affected too much by the artifacts present in the ToF data. In particular it is worth noting how color information, that comes from a different sensor, is correctly aligned and fused on the geometry. Fig. 7 instead refers to the reconstruction of a teddy-bear. The shape of the object in this case is more complex and less regular but still the reconstruction is correct, most details of the teddy bear shape are recognized by the algorithm. Note also how the color alignment is correct. Probably the most evident artifact is the hole on the foot of the teddy-bear, due to a well-known drawback of the ToF cameras, i.e. the fact that they are not able to acquire very low reflective surfaces (such as black surfaces) since the reflected signal is too weak for a correct reconstruction. Another scene including different objects is shown in Fig. 8. Finally Fig. 9 shows the reconstruction of a person's shape.

Figure 8. Reconstruction of the *baby and boxes* scene.

5. Conclusions

In this paper we proposed a novel 3D reconstruction pipeline explicitly targeted to the data acquired by ToF cameras exploiting also the side information from color cameras. The proposed approach has demonstrated to be very robust against the noise and the other issues of the data acquired by these cameras. The reliable extraction of salient

Figure 9. Reconstruction of a person

points has allowed to use a smaller number of points in the registration process and so to speed-up the reconstruction algorithm. Experimental results have shown how the proposed approach allows to use the ToF camera as an hand-held scanner in order to produce accurate 3D reconstructions. Further research will be devoted to the development of a post-processing optional stage that is able to improve the reconstruction accuracy by the introduction of a global alignment step. The proposed approach is also very well-suited to be applied to the data produced by Microsoft's Kinect sensor and we are building a variation of the current 3D reconstruction system explicity targeted on this device. This is a very interesting research direction since the simplicity and speed of the proposed acquisition pipeline combined with the low cost of the Kinect sensor will allow unskilled users to build 3D models without complex and expensive ad-hoc hardware.

References

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

[2] Y. Cui, S. Schuon, C. Derek, S. Thrun, and C. Theobalt. 3d shape scanning with a time-of-flight camera. In *In Proc. of IEEE CVPR 2010*, 2010. 1

[3] B. Curless and M. Levoy. A volumetric method for building complex models from range images. *ACM Transactions on Graphics (SIGGRAPH)*, 1996. 1

[4] C. Dal Mutto, P. Zanuttigh, and G. Cortelazzo. A probabilistic approach to tof and stereo data fusion. In *3DPVT*, Paris, France, May 2010. 2

[5] C. Dal Mutto, P. Zanuttigh, and G. M. Cortelazzo. *Time-of-Flight Cameras and Microsoft Kinect*. SpringerBriefs in Electrical and Computer Engineering. Springer, 2012. 3

[6] Y. M. Kim, C. Theobalt, J. Diebel, J. Kosecka, B. Miscusik, and S. Thrun. Multi-view image and tof sensor fusion for dense 3d reconstruction. In *Proc. of ICCV Workshops*, pages 1542 –1549, 27 2009-oct. 4 2009. 1

[7] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. Kinectfusion: Real-time dense surface mapping and tracking. In *Proc. of IEEE ISMAR*, October 2011. 1

[8] http://reconstructme.net/. 1

[9] http://www.kinectathome.com/. 1

[10] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Proc. of CVPR*, volume 1, pages 519 – 528, june 2006. 1

[11] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake. Real-time human pose recognition in parts from a single depth image. In *Proceedings of CVPR*, June 2011. 1

[12] C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. In *Proc. ICCV*, 1998. 3

[13] A. Torsello, E. Rodol, and A. Albarelli. Sampling relevant points for surface registration. In *Proc. of 3DIMPVT 2011*, Hangzhou, China, May 2011. 3