Saliency-aware Real-time Volumetric Fusion for Object Reconstruction

Sheng Yang\(^1\), Kang Chen\(^1\), Minghua Liu\(^1\), Hongbo Fu\(^2\) and Shi-Min Hu\(^1\)

\(^1\)Tsinghua University
\(^2\)City University of Hong Kong

Abstract
We present a real-time approach for acquiring 3D objects with high fidelity using hand-held consumer-level RGB-D scanning devices. Existing real-time reconstruction methods typically do not take the point of interest into account, and thus might fail to produce clean reconstruction results of desired objects due to distracting objects or backgrounds. In addition, any changes in background during scanning, which can often occur in real scenarios, can easily break up the whole reconstruction process. To address these issues, we incorporate visual saliency into a traditional real-time volumetric fusion pipeline. Salient regions detected from RGB-D frames suggest user-intended objects, and by understanding user intentions our approach can put more emphasis on important targets, and meanwhile, eliminate disturbance of non-important objects. Experimental results on real-world scans demonstrate that our system is capable of effectively acquiring geometric information of salient objects in cluttered real-world scenes, even if the backgrounds are changing.

CCS Concepts
- Computing methodologies → Reconstruction; Object detection;

1. Introduction
As a milestone in real-time depth fusion, KinectFusion [IKH\(^*\)11] has aroused great research interests among the vision and graphics communities. Various works have been proposed to improve KinectFusion in different aspects [RV12, WKF\(^*\)12, NZIS13, KPR\(^*\)15]. However, robust and high-precision real-time reconstruction of real-world scenes with consumer-level depth sensors is still a hard problem. The difficulties come from two-folds: first, depth information acquired by cheap sensors (e.g., Microsoft Kinect) is noisy, distorted and incomplete; second, real-world environments are typically complex and cluttered, where scanner-unfriendly surface materials, featureless areas and dynamic objects frequently occur. Even the state-of-art fusion techniques [KPR\(^*\)15, DNZ\(^*\)16] cannot easily produce clean, complete and accurate results. Actually, reconstruction qualities are affected by many factors, including the conditions of target scenes, strategies adopted by fusion algorithms, parameter configurations, experiences of users, etc. Thus, different application scenarios favor different kinds of reconstruction solutions. For instance, placing target objects on a rotating platform under fixed camera(s) would always be an ideal condition, if possible.

In this work, we aim to improve real-time reconstruction performances in a specific scenario, i.e., reconstructing objects in complex real-world backgrounds with a hand-held Kinect-style RGB-D sensor. Although KinectFusion and its follow-up works already have the ability to solve this problem to some extent, they do not distinguish between the background environment and the target objects therein. As a consequence, sensor noise and depth distortion will cause equal influences to each voxel, which is reasonable when reconstructing the whole scene since each RGB-D pixel is of equal importance. However, in the scenario of object reconstruction, registration precision of regions belonging to target objects is clearly much more important. With the presence of noise and distortion, such differences between object and scene reconstruction would result in very clear changes. We kindly refer readers to [CZMK16] to find the state-of-the-art object reconstruction techniques, which however do not always produce satisfactory reconstruction results. In addition, existing methods are sensitive to dynamic background objects, which often occur in real-world scenes, e.g., walking pedestrians or windblown curtains. Any changes in background during scanning, may easily break up the registration process, which leads to a strict restriction upon existing reconstruction systems.

We observed that if a reconstruction framework is smart enough to understand user intentions, the aforementioned issues can all be addressed. We can thus put more emphasis on important targets during registration, and meanwhile, eliminate disturbance of non-important objects, even if they are moving in the background. The idea of protecting local geometry around points of interest in scans was first introduced and successfully used in an off-line dense reconstruction pipeline [ZK13]. However, their global cam-

\(\copyright\) 2017 The Author(s)
Computer Graphics Forum \(\copyright\) 2017 The Eurographics Association and John Wiley & Sons Ltd. Published by John Wiley & Sons Ltd.

DOI: 10.1111/cgf.13282
era tracking algorithm for POI (point of interest) cannot be adapted to meet our real-time requirements. Thus, we resort to inferring user-intended objects by detecting visual saliencies from RGB-D frames, which are then incorporated into a traditional real-time volumetric fusion pipeline. Experimental results on real-world scans demonstrate that our system is capable to effectively acquire detailed geometric information of salient objects in cluttered real-world scenes, and outperforms the status quo.

In summary, we present a real-time framework for acquiring 3D objects from complex real-world environments using hand-held consumer-level RGB-D scanning devices. Compared with existing works, our framework has several advantages: (i) geometric features of target objects are better preserved; (ii) users can conveniently tap to change target objects if multiple salient objects are present; (iii) tracking and registration will not be disturbed by changes in the background. To achieve them, we present a novel spatial-temporal visual saliency detection method and successfully incorporate visual saliency into real-time depth fusion.

2. Related Work

Our framework involves two key components: saliency detection and volumetric fusion. Thus, in this section, we give a brief review of the most related works in these two areas.

Saliency Detection. Visual saliency has long been a fundamental problem in neuroscience, psychology, and vision perception. Although there have not yet been consensus on how visual saliency should be defined and evaluated in the computer vision community, various saliency models have been proposed for different applications, for example, predicting human fixation or extracting salient regions [BCJL15]. In our framework, visual saliency is adopted as cues for identifying POI of users, and shares similar definition to the saliency recognized in image and video segmentation works [AEWS08, FXL17].

Classical saliency detection methods typically extract features at pixel level [CMH*15] or super-pixel level [JWY*13, QCB*15], which are then propagated into spatially-coherent meaningful regions based on different kinds of saliency metrics [PLX*14, GRB16]. Image segmentation techniques like GrabCut [RKB04] are often used at the final stage to decide whether a region is salient or not [FPCC16]. For RGBD images, the depth channel is handled isolately [GRB16], or used for extracting additional features for comparison [PLX*14]. Recent deep learning methods have also addressed the problem [KWW16, HCH*16] but relatively heavy (cost hundreds of milliseconds per frame) for real-time applications. Spatio-temporal coherence has also been addressed when determining salient regions in video sequences [FWL14, KKS15, ZLR*13], where optical flow is incorporated in these methods for motion estimation. However, their methods are not suitable for identifying foreground objects of interests or determining which one is more important when multiple foreground objects exist. In fact, such methods require high-quality RGB sensors and good lighting conditions, while extra depth information from sensors used in our scenario offers stronger cues about focus and motions of objects than traditional 2D flow estimation. To the best of our knowledge, none of the previous video saliency detection methods have coupled depth fusion to consider such coherence in 3D space.

In our framework, saliency detection serves a very specific purpose, i.e. real-time depth fusion, which makes our saliency detection component addressed different from those typically tackled in computer vision in three aspects: (i) input RGB-D streams are normally captured around a target which the user intends to reconstruct, and thus salient regions are much clearer and more recognizable; (ii) salient objects in the scene are strictly static rigid bodies; (iii) salient model must be computed very quickly to achieve real-time depth fusion. Despite various saliency detection algorithms studied in computer vision, none of them suits our application scenario. Thus, we introduce a novel spatio-temporal visual saliency detection method by customizing a salient model for single-frame RGB-D images based on [PLX*14], and integrate it into the frame-to-model registration process in the traditional depth fusion workflow. Experimental results have demonstrated that the proposed reconstruction-orientated saliency model and frame-to-model registration are mutually complementary in terms of precision.

Volumetric Fusion. Volumetric fusion is the most classic 3D reconstruction method studied in computer graphics [CL96]. Traditional volumetric fusion pipelines involve two key phases: tracking and integration, where ICP (Iterative Closest Point) [BM92] and SDF (Signed Distance Function) techniques are typically used in each phase respectively. In 2011, Iazadi et al. introduced KinectFusion [IKH*11] which demonstrated that real-time volumetric fusion systems could be achieved with the parallel processing abilities of modern graphical hardwares. Ever since, real-time 3D reconstruction systems based-on low-end depth sensors have become a research hotspot [CLH16].

Many subsequent works aimed to improve KinectFusion from various aspects. Moving Volume KinectFusion [RV12] and KinectFusion [WKF*12] eliminated the graphical memory limitation by straight-forwardly shifting volumes maintained in memory according to camera trajectories. This idea was further extended by deploying better data structures and exploiting smarter memory-swapping strategies [NZS13, KPR*15], which successfully improved the efficiency of graphical memory usage and has become a prevalent choice in practical volumetric fusion systems. With improvements of RGB cameras equipped by low-end RGB-D scanners, visual correspondences estimated from color images have also been considered to improve local registration precision [KSC13, WJKN*15] or to reduce drifting by loop-closure detection [CZK15, DNZ*16]. Thanks to advances in graphical hardwares, these time-consuming mechanisms used to appear in non-real time SLAM systems can now be integrated into real-time applications. Our framework follows this trend, i.e. we incorporate visual saliency into the classic volumetric fusion pipeline. Specifically, our application poses two major demands which have rarely been addressed before: (i) preserving detailed geometric information for target objects; (ii) reducing disturbance from non-important backgrounds.

3. Methodology

Our framework consists of two interdependent modules. The saliency detection module (Sec. 3.1) estimates salient image regions based on features extracted from each RGB-D frame and spatio-temporal information maintained by the volumetric fusion module (Sec. 3.2), which iteratively registers and fuses each incom-
Before detailed discussions, we first give the mathematical noting frame, with an aligned saliency map, into a global volumetric put frame at time \( t \) which is defined as a set of voxels \( C \) for other notes, the induced vertex map from \( \sigma \) and enable voxel hashing for unbounded reconstruction. Specifically, we compute saliency values considering current frame, previous frames, camera trajectory, and potential user interactions. For each super-pixel based on multi-contextual contrast (local, global, pseudo-background, and the optional focus contrast), which are later propagated into larger salient image regions. The whole procedure is parallelized and accelerated by GPU to ensure real-time performance.

**Feature Extraction.** For each incoming pair of RGB and depth frame at time \( t \), we first extract super-pixels via a GPU-accelerated SLIC algorithm [ASS12] named gSLICr [RPR15] to construct nearly 200 super-pixels on each frame. Afterwards, we calculate the feature vector \( S_u = [S^l_u, S^m_u, S^b_u, S^g_u]^T \) for each super-pixel \( u \subseteq \Omega \). where \( S^l_u \) represents the centroid, \( S^m_u \) the PCA normal of the points in the super-pixel, \( S^b_u \) the average color in CIE-Lab color space, \( S^g_u \) the average depth, and \( S^m_u \) the number of pixels in \( u \). All distance and location units are given in meters, and the three channels of CIE-Lab color are normalized in \([0, 1]\).

**Low-level Saliency Value.** At time \( t \), our low-level saliency value \( S^l_u \) of each super-pixel \( u \) is defined as the multiplication of the following three terms:

\[
S^l_u = C(S_u) \times R(u, S^l_{t-1}) \times U(S_u),
\]

where \( C(S_u) \) is a weighted combination of local, global, and pseudo-background contrasts defined in [PLX14]. In our implementation, this term was calculated with the same parameter settings suggested in [PLX14].

\[
R(u, S^l_{t-1}) = \text{weighted average saliency value from a subset of all pixel-wise correspondences as } \bigcup_{\nu \in \mathbb{N}} \{ \nu, \nu' + 1 \}. \text{ In order to reduce the influence of dynamic regions or outliers, the weight } \chi \text{ for each correspondence is calculated through their location as:}
\]

\[
\chi_j = \exp\left(- \frac{\|V_j(p) - V_{j-1}(p_{j-1})\|^2}{\sigma_{\chi}^2}\right),
\]

where \( \sigma_{\chi} = 0.2 \) in our implementation. In the frame-to-model registration spirit [IKH11], here we use the saliency map \( S^l_{t-1} \) ray-casted from fused volumetric data rather than directly adopt \( S_{t-1} \), as \( S^m_{t-1} \) contains the saliency information of all previous frames in order to coherently focus on the target. For the first frame, \( R(\cdot) \) is set to one.

\( U(S_u) \) depicts the influence of user-specified POI hints. Specifically, we construct a focus contextual set which consists of super-pixels having its centroid \( S^m_u \) within the user-specified focal point \( F_t \) and radius \( F_r \), and then follow [PLX14] to compute the feature contrast. If no specification has been given, our system will take the centroid of the super-pixels which get the highest saliency as \( F_t \), and the distance to the furthest pixels among those with saliency values higher than their average as \( F_r \) for subsequent inputs.

**Propagation.** To remove outliers and obtain more meaningful salient image regions, we need to group super-pixels into larger regions. To further reduce effects from uninterested areas through user hints based on the strategy from [PLX14], we similarly construct an undirected graph of super-pixels in \( C \) based on their adjacency in the image space, and to generate a set of salient seed s, each of which is then greedily propagated into a spanning tree. On account of speed, super-pixels with low-level saliency value \( S^l_u \) greater than the average are all considered as seeds by default and
the propagation of each seed is computed parallelly on GPU. However, if \( F_c \) and \( F_l \) have been specified, the following modifications would be adopted.

First, a super-pixel \( u \) with \( ||S_u^c - F_c|| < F_l \) and its empirical weighting \( S_u^M \) above their average will be chosen as seeds, which is calculated as:

\[
S_u^M = \exp\left(-\frac{||S_u^c - F_c||^2}{F_l^2}\right) \times S_u^f.
\]

By multiplying this exponential term, we emphasize those seeds closer to the focal point for propagation, and also raise their saliency (see the comparison between \( S_u^f \) and \( S_u^M \) in Figure 2). Second, a new terminating condition is applied to the spanning tree algorithm, i.e. the edge \((S_u^f, S_u)\) taking \( S_u^f \) as their seed will be added to the tree if

\[
||S_u^f - S_u|| < F_f \cdot S_u^f.
\]

since the propagation through one seed should be limited in a certain extent with regard to the object size.

The final saliency value of \( S_u^f \) is then calculated as:

\[
S_u^f = Q(S_u) \times S_u^M.
\]

where \( Q(\cdot) \) is the normalized \((0, 1)\) frequency of \( S_u \) appeared in all the spanning trees. After propagation, the saliency map for current frame \( S_f \) is obtained and passed into the volumetric fusion step for tracking in the volumetric fusion module (Sec. 3.2).

Some results of our saliency detection module are illustrated in Figure 3, from which we can see that our method does produce more spatially and temporally consistent saliency maps compared with [PLX∗14], and thus suits our application better. Please refer to our supplementary video for additional examples.

3.2. Volumetric Fusion

Compared with the previous fusion pipelines (e.g., [IKH∗11] and [KPR∗15]), our framework involves two additional channels in the tracking and integration process: the calculated saliency map for current frame \( S_f \) and the ray-casted saliency map from the camera pose of last frame \( S_f^{t-1} \). Specifically, in the tracking stage, both \( S_f \) and the estimated pose \( T \) are iteratively refined, and in the fusion stage, saliency information is integrated into volumes along with the geometric and color information.

Saliency-aware Pose Estimation. Our goal is to estimate the local transformation \( T \), which aligns the current frame to the previous one. In our implementation, \( T \) is obtained by iteratively solving the following minimization problem:

\[
\arg\min_{\xi} E(\xi) = E_{wicp}(\xi) + \sigma_R E_{wrgb}(\xi),
\]

where \( \sigma_R = 0.1 \) in our implementation. \( E_{wicp} \) and \( E_{wrgb} \) measure the saliency-aware geometric cost and the saliency-aware photometric costs, respectively. Following [WKJ∗15], they are mainly defined based on the point-to-plane error in ICP registration \( E_{icp} \) and the intensity differences between RGB pixels \( E_{rgb} \), respectively, which can both be formalized as least-squares problems after some derivations:

\[
E_{icp} = ||J_{wicp}\xi + r_{icp}||^2,
\]

\[
E_{rgb} = ||J_{rgb}\xi + r_{rgb}||^2.
\]

We refer readers to [WKJ∗15] for detailed definitions and derivations, here we focus on how saliency information is integrated into this cost function. Specifically, we define a saliency weight for each pixel-wise correspondence \((p_i^t, p_i^{t-1})\) as:

\[
w_i = \exp(\sigma_W \cdot \frac{S_{p_i}^w (p_i^{t-1}) \cdot W_{p_i}^w (p_i^{t-1}) + S_f (p_i)}{W_{p_i}^w (p_i^{t-1}) + 1}),
\]

where \( \sigma_W = 4 \) in our implementation is the parameter adjusting the weight of salient regions through the tracking process. We follow the combination strategy in TSDF fusion to achieve a smooth tracking. Detailed evaluation of this parameter is given in Sec. 4.3.

We then organize all computed \( w_i \)s of the current frame as a vector \( \mathbf{W} \), based on which, the saliency-aware costs \( E_{wicp} \) and \( E_{wrgb} \) referred in equ 8 are finally defined as:

\[
E_{wicp} = ||\mathbf{W} \cdot J_{wicp}\xi + \mathbf{W} \cdot r_{icp}||^2,
\]

\[
E_{wrgb} = ||\mathbf{W} \cdot J_{rgb}\xi + \mathbf{W} \cdot r_{rgb}||^2.
\]
4. Evaluation

We have evaluated our system in various aspects, including reconstruction quality, saliency coherence, and computational performance. Our testing data came from two sources: the large dataset of object scans constructed by Choi et al. [CZMK16] and a small dataset we built with 14 scenes scanned from 4 different places (i.e., a lobby, two seminar rooms, and an office) by an ASUS Xtion Pro Live sensor. The large dataset is mainly used for regression testing. The small dataset is established for evaluating focus switching and dynamic background objects during scanning, and thus either contain multiple targets or dynamic objects in the background (e.g., walking pedestrians, windblown curtains, etc.). We also made two 3D printed crafts with known geometry for quantitative evaluation.

4.1. Model Quality

Some representative results of both [CZMK16] and ours are shown in Figure 6, including various locations such as outdoors, office, and seminar room. We have compared our system with some of the state-of-art works, where KF, DV, CI, KT, EF stand for KinectFusion [IKH11], Dense RGB SLAM [SSC11], Combined-ICP [WJK13], Kintinuous [WKJ15], Elastic-Fusion [WLSM15], respectively. Here we briefly classify these methods into three categories based on their cost functions used in pose estimation: KF only considers $E_{rgb}$, DV only considers $E_{rgb}$, while CI, KT, and EF consider both $E_{rgb}$ and $E_{rgb}$. KT and EF are different from CI because they also involve the idea of loop-closure detection. All results are produced using the suggested default settings in their papers.

As illustrated in the top two rows of Figure 6, scans from [CZMK16] typically contain a single target in a static scene. Hence, all systems in our test produced similar results. However, once dynamic objects present in the background (Figure 6-sc1,sc2,sc3), the reconstruction processes of KF, DV, and CI are broken and severe drifting starts to appear. Although KT and EF can benefit from correctly detected loop closures, geometric details of their constructed target objects in large scenes are not well preserved with respect to ours (Figure 6-sc2,sc4). Generally, our method successfully protects local geometry by reducing registration errors of salient regions. Although misalignments may be amplified in other areas, such trade-off is reasonable and valuable in

3.3. Interactive Focus Switching

Our system has further functionality allowing users to provide additional hints for focusing while reconstructing, since automatic inferred salient regions may be unexpected when multiple objects appear in front of the camera. In such a case, users can specify hints through a simple click-and-drag interaction, similar to the tap-to-focus operations in modern camera applications. A click on the input frame or ray-casted frame specifies a focal point $F_1$ in the local or global coordinate. A drag from the focal point to a destination point tells a length $F_2$ for the approximate size of the target. Focus switching on different objects can affect the tracking process and better capture geometric details of the target objects, as shown in Figure 4.

In many crowd or cluttered scenes, the target might be obstructed sometimes in the scanning process, e.g., due to a person walking across in front of the target (Figure 5). To reduce the impact of severe occlusion, we generated another ray-casted image as $\hat{S}_t$, whose rays only intersect with salient TSDF volumes. Comparisons between $S_t$ and $\hat{S}_t$ enables our system to filter out the undesired regions before the tracking and fusion process.

Figure 5: An example of occlusion cases. Top: the color image and ray-casted images when the person was blocking the target. Bottom: final reconstructed models. Heavy occluded regions are filtered out through comparisons between $S_t$ and $\hat{S}_{t-1}$ in our implementation.

Figure 4: Examples of focus switching. Left: user specifications (top) and saliency (green) of reconstructed meshes (bottom). Right: quality of the target has been improved through focus switching.

our application scenario. Since ground truth geometry of models in [CZMK16] are unknown, for quantitative evaluation, we scanned two 3D printed crafts shown in Figure 7, twice of each in both static and dynamic scenes separately. Both Combined-ICP [WJK∗13] and our method were performed for each scan. The reconstructed results were manually segmented and registered to their ground truth mesh models, with their average point-to-model distances reported in Table 1. As a result, models reconstructed from our
pipeline outperform those from Combined-ICP, especially in dynamic cases where the fusion process is influenced by dynamic backgrounds.

<table>
<thead>
<tr>
<th></th>
<th>Bunny-S</th>
<th>Teapot-S</th>
<th>Bunny-D</th>
<th>Teapot-D</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined-ICP</td>
<td>2.81</td>
<td>2.11</td>
<td>3.41</td>
<td>3.99</td>
</tr>
<tr>
<td>Ours</td>
<td>2.48</td>
<td>1.72</td>
<td>2.61</td>
<td>2.01</td>
</tr>
</tbody>
</table>

Table 1: Average point-to-model distance of test cases, while S for static cases and D for dynamic cases. All given in millimeters.

4.2. Running-time

We deployed our system on a desktop with Intel Core i7 CPU, 32GB memory, and GeForce GTX 980Ti graphical card (2,816 CUDA cores). Average time costs for each step of our pipeline are reported in Table 2 with comparison to Combined-ICP. The saliency detection step includes super-pixel construction, feature extraction and two-level saliency calculation. Statistic results show that both the weighted tracking and the saliency integration bring insignificant burdens to our system. Finally, the overall speed meets the real-time requirement (less than 33ms per frame for 30Hz streams).

<table>
<thead>
<tr>
<th></th>
<th>Saliency</th>
<th>Tracking</th>
<th>Fusion</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Combined-ICP</td>
<td>-</td>
<td>11.23</td>
<td>1.40</td>
<td>12.63</td>
</tr>
<tr>
<td>Ours</td>
<td>12.87</td>
<td>12.09</td>
<td>1.50</td>
<td>26.46</td>
</tr>
</tbody>
</table>

Table 2: Average computational performance of the test scenes. All timings are given in milliseconds.

4.3. Parameter Study

The key parameter of our framework is the weight value $\sigma_W$ in Equation 11, which controls the influence of saliency to volumetric fusion. Intuitively, high values of $\sigma_W$ will significantly increase the importance of salient regions while decrease non-salient regions. If we set $\sigma_W = 0$, our framework is degenerated to Combined-ICP [WJK*13]. However, the quality of the reconstructed models is not monotonically increasing with $\sigma_W$, because the convergence speed of the Combined-ICP algorithm will be slowed down if the salient regions are planar surfaces and/or with featureless textures. In such cases, registration errors will be raised if we do not increase the number of iterations when solving Equation 8. In our experiments (see Figure 8), setting $\sigma_W$ to around 4 generally produces plausible results for all testing data.

4.4. Limitations

Our framework has two main limitations: (i) Our system is not suitable for full scene reconstruction applications. Since we put more emphasis on protecting local geometry of target objects, misalignments are forced into background areas. Thus, the quality of non-target objects in the reconstructed results is typically poorer compared with existing methods. (ii) Our system is not friendly to unpredictable users. Apparently, our system benefits from visual saliency detection to understand user intentions. However, if such intention were vague or obscure, our system might get confused and thus fail to produce high-fidelity results.

5. Conclusions and Future works

We have presented a real-time system for acquiring 3D objects with high fidelity using hand-held consumer-level RGB-D scanning devices. A novel spatio-temporal visual saliency detection method is incorporated into the traditional real-time volumetric...
fusion pipeline, which successfully emphasizes the important targets and eliminates disturbance of non-important objects. We also present a simple user interface for focus changing. In the future, our system can be ported to mobile devices for free reconstruction, and the saliency information can be stored in the output mesh model to guide the object segmentation process.

Acknowledgements
We thank the reviewers for constructive comments. This work was supported by the Natural Science Foundation of China (Project Number 61521002) and Research Grant of Beijing Higher Institution Engineering Research Center. Hongbo Fu was partially supported by the Natural Science Foundation of China (Project 113513).

References
[ZK13] ZHOU Q.-Y., KOLTUN V.: Dense scene reconstruction with points of interest. ACM TOG 32, 4 (2013), 112. 1

© 2017 The Author(s)
Computer Graphics Forum © 2017 The Eurographics Association and John Wiley & Sons Ltd.