Generating Super-Resolved Depth Maps Using Low-Cost Sensors and RGB Images

Leandro Tavares Aragão dos Santos, Manuel Eduardo Loaiza Fernandez, and Alberto Barbosa Raposo

Dept. of Informatics, Pontifical Catholic University of Rio de Janeiro, Brazil

Abstract. There are a lot of three-dimensional reconstruction applications of real scenes. The rise of low-cost sensors, like the Microsoft Kinect, suggests the development of systems cheaper than the existing ones. Nevertheless, data provided by this device are worse than that provided by more sophisticated sensors. In the academic and commercial world, some initiatives try to solve that problem. Studying that attempts, this work suggests the modification of super-resolution algorithm described by Mitzel et al. [1] in order to consider in its calculations colored images provided by Kinect. This change improved the super-resolved depth maps provided, mitigating interference caused by sudden changes of captured scenes. The tests showed the improvement of generated maps and analysed the impact of CPU and GPU algorithms implementation in the super-resolution step.

1 Introduction

Real scene 3D reconstruction can be applied to several areas: engineering surveys, like those provided by private companies technologies as Leica Geosystems [2] and LFM [3], ergonomic studies, robots navigation [4], customers' body capture for clothing stores [5] and historical monuments models creation, as proposed by Chen et al. [6]. Focusing on applications related to the human body, there are several technologies that offer high quality reconstructions, however, their costs are too high. Considering this scenario, an approach that allows depuration of Kinect data to a quality level closer to that of the devices mentioned above, would allow this type of use on a broader scale.

If, in addition to reconstructing high qualtity human body 3D models, such a system could be manipulated by the movements of the joints detected by the Kinect, a range of possibilities would open up. Grouping and processing depth maps, RGB images and joint movements detected by the Kinect, this system would enable real time controlled avatars generation. Primarily, a reconstruction process applied to depth maps and RGB images would provide 3D meshes related to each captured user body. Later, through Rigging and Skinning methods [7], each 3D user model would be integrated to captured joints. Thus, each user movement would distort the respective meshes in real time.

Scrutinizing the rebuilding process, we have three basic stages: the super-resolution, global rigid registration and non-rigid registration. Super-resolution

G. Bebis et al. (Eds.): ISVC 2014, Part II, LNCS 8888, pp. 632–641, 2014.

generates high resolution from N low resolution depth maps. The global rigid registration originates a 3D mesh reconstructed from the high-resolution maps. However, this reconstruction contains many artifacts as a result of user movements during capture. To eliminate such problems, the preliminary mesh is subjected to non-rigid registration.

To the best of our knowledge, there isn't an open library that covers all steps needed by those goals. In this context, this paper focuses on the first fundamental step of the process: super-resolution. When analyzing data provided by low-cost sensors, we note that available resolution is insufficient for a series of high quality 3D reconstruction applications. From this simple observation, it is possible to understand why reconstruction strategies, based on these sensors, depend fundamentally on super-resolution algorithms.

Our research highlights the work of Cui et al. [8], due to the use of RGB images information in the super-resolution calculation. Our work, based on open technologies and on ideas present in commercial solutions, adapts the super-resolution approach proposed by Mitzel et al. [1] and incorporates the concept of using RGB information in the reconstruction process.

During the capture of several low resolution depth maps required for the super-resolution process, a person's clothing, for example, can move. This movement creates a difference between input frames that culminates in artifacts in final reconstruction. Using OpenCV library, this study looked into incorporating RGB images provided by the Kinect to super-resolved depth maps generation, in order to minimize mentioned problems. This approach has shown promising results regarding removal of interference present in depth maps captured by Kinect.

This paper is organized as follows. In Section 2 we present related work. In Section 3 we discuss the incorporation of colors consideration in the super-resolution calculation. In Section 4 we present tests and results. Finally, Section 5 discusses our main conclusions and points to future work.

2 Related Work

A full human body 3D scan system using the Kinect would need two basic elements. First, for a particular person, it should be able to capture the minutiae of the body with reasonable quality. Additionally, another module would associate the mesh generated by Kinect and, in each movement, would distort the model.

The system described by Tong et al. [9] proposes the use of three Kinects. At the beginning, it records a very rough template of human body. This template is used to deform sucessive frames. To distribute resulting errors deformations, global registry geometries is used, treating problems like occlusion. Successive iterations, alternating between paired and global registry, occur until the algorithm converges. Finally, model reconstruction is made using Poisson reconstruction method as described by Kazhdan et al. [10].

Cui et al. [8] use a different approach to allow the use of a single Kinect to obtain meshes with a higher detail level when compared to those presented by Tong et al. [9]. Depending only of Kinect provided data and not demand-

ing a previously captured model, they reproduce in detail all face and clothes geometry.

Low resolution and high noise level of data provided by Kinect demand a smoothing step of surfaces generated from map provided by this device. Newcombe et al. [11] apply a bilateral filter in order to obtain a map with reduced noise and preserved discontinuities. Schuon et al. [12] develop a super-resolution algorithm capable of increasing depth resolution and quality of the data generated by a scanner based on structured light. Later, Cui et al. [13] improve this method. Cui et al. [8] devise a new algorithm for processing the color and depth data of Kinect. This algorithm presents a higher resolution, reduces noise and preserves original surfaces details.

Cui et al. [13] approached the global rigid registration through a scan alignment probabilistic model that takes into account noise sensor characteristics. However, this strategy solves local alignment. The same authors proposed a global probabilistic alignment algorithm [8].

Non-rigid deformation of human body during scan makes the previous step result not ideal. Taking advantage of the fact that the human body is highly articulated, Cui et al. [8] improve the algorithm presented by Chang and Zwicker [14] using a scan alignment probabilistic model which is robust to noise device.

Cui et al. [8], computing appropriate transformations from a total energy function, solve the already known closed-loop problem. At the end, a texture mapping is applied to attach to the model the scanned body appearance.

Those works have proven the feasibility of acquiring input data to human body models generation using the Kinect. More than that, our paper suggests the incorporation of colors consideration proposed by Cui et al. [8] to the data term of Mitzel et al. [1] super-resolution approach. Basically, we have modified the equation 1 proposed by Mitzel et al. [1].

$$I_H^{n+1} = I_H^n + \tau \left(\sum_{k=1}^{N} W_k^T B^T D^T (I_L^{(k)} - DBW_k I_H^{(n)}) + \lambda \triangle I_H^{(n)} \right) \qquad (1)$$

In equation 1, τ describes time in each iteration and $\triangle I_H^{(n)}$ corresponds to the regularization term. After warping (W_k), blurring (B_k) and down-sampling (D_k), the two equation terms induces super-resolved image I_H, to all low resolution frames, while imposing a linear diffusion of the intensities weighted by λ. Changes made in equation 1 are described in the following section.

3 Incorporating Colors to Super-Resolution

This work focuses on the development of a super-resolution algorithm efficient to be applied to depth maps provided by Kinect. During research, super-resolution class provided by OpenCV library has proved to be the most promising as the core of this work.

Well documented and widely used by the community, this library is available for academic and commercial use under BSD licensing. The respective super-resolution class is based on the strategy presented by Mitzel et al. [1]. However,

Mitzel et al. [1] focus on video sequence images. This scenario does not consider environments where we have two sensors information, that provide different data of the same scene, as Kinect does, providing depth data and RGB images. Considering this, the present work unified the algorithm already implemented in OpenCV [15] with the benefits of Cui et al [8], leveraging RGB images to depth maps reconstruction process.

Both the use of the proposal of Cui et al. [8], as the adaptation of Mitzel et al. [1] approach, were only viable because of negligible rotation of the object between frames captured from a certain images group. In the case of adaptation of Mitzel et al. [1] approach, for example, such characteristic between frames allow an interpretation of the depth maps as images represented by levels of intensity without impacting in any way the super-resolution algorithm originally proposed for sequences of video frames.

Following the proposal of Cui et al. [8], the adjustment term $A^{(k)}$ — a matrix which produces bigger values if original color frame is similar to the average of the other images — is

$$\frac{1}{C_k - \frac{1}{s}\sum_{i=1}^{s} C_i} \tag{2}$$

where C_k is the RGB frame corresponding to each depth map k, that is, the related low resolution image $I_L^{(k)}$. The i index varies between several RGB images used in the current reconstruction, being s the quantity of images used. We should adjust the equation which describes reconstruction process of super-resolved image from numerical method "Steepest Descendent", in order to behave the adjustment provided by the color map contribution. Such an adaptation leads to the following equation:

$$I_H^{n+1} = I_H^n + \tau \left(\sum_{k=1}^{N} A^{(k)} * \left(W_k^T B^T D^T (I_L^{(k)} - DBW_k I_H^{(n)}) \right) + \lambda \triangle I_H^{(n)} \right) \tag{3}$$

where $*$ is the term-by-term product of $A^{(k)}$ with the regularization term originally proposed by Mitzel et al. [1]. The array C_k dictates the dimensioning of adjustment array $A^{(k)}$. However, it is term-by-term multiplied by the resulting value of $(W_k^T B^T D^T (I_L^{(k)} - DBW_k I_H^{(n)})$, that is, it has to be $\beta n \times \beta m$ size, where β is the resolution increasing factor. In this paper, we use β equals to 2. The low resolution depth map provided by Kinect leads us to $n = 640$ and $m = 480$. Thus, our adjustment term size might be 1280×960. However, Kinect provides us RGB images with 1280×1024 size. So, the second step of colors incorporation was to adapt such images, so that each array C_k contains a image with 1280×960 resolution.

Later, after guaranteeing that, at instant k, low resolution map capture I_L and the respective already processed RGB image generation occurs at the same time, the arrays $I_L^{(k)}$ and C_k are processed according to the algorithm described in section 4.

In summary, prior to processing the new algorithm, there are the preliminary steps described in the algorithm below.

Preliminary processing

1: Adapt the low resolution depth map to the intensity levels representation.
2: Adapt RGB images resolution from 1280×1024 to 1280×960.
3: Ensure the concomitance between low resolution map capture I_L and the respective already processed RGB image generation C.

The new algorithm guarantees that, if colored image corresponding to the instant k, presents some difference when compared to other images of the group, the respective map $I_L^{(k)}$, will contribute less than others to the super-resolution result. This is due to the incorporated term-by-term multiplication, where the adjustment term $A^{(k)}$ has lower values for the pixels that are very different from their corresponding pixels in other frames.

4 New Algorithm

The goal of the algorithm is: given a sequence of N low-resolution images $\{I_L^k\}_{k=1}^N$, to estimate the movement between frames and infer high resolution image I_H of the current scene. If some disparity in the colors map $C^{(k)}$ corresponding to each frame $I_L^{(k)}$ is detected, such low resolution image may not contribute to the final result.

New algorithm

1: Choose an image of the sequence as reference.
2: Estimate for each pair of consecutive frames the shift between the current frame and the next one using optical flow algorithm.
3: Use motion fields u_i^f and v_i^f to compute motion fields u_i^r, v_i^r related to the reference image, observing that f indexes (movement between mutual frames) and r (movement between reference frames and individual images) must indicate differences between motion maps.
4: Interpolate motion fields u_i^r e v_i^r in order to reach the same dimension of I_H image.
5: Initialize I_H, setting values to 0.
6: $\mathbf{C_{sum}} := \mathbf{0}$;
7: from $\mathbf{k = 1}$ to $\mathbf{N}$ do
8: $\mathbf{C_{sum} = C_{sum} + C^{(k)}}$
9: end for
10: from $\mathbf{k = 1}$ to $\mathbf{N}$ do
11: $\mathbf{A^{(k)} = \frac{1}{C^{(k)} - \frac{1}{s} C_{sum}}}$
12: end for
13: from $t = 1$ to T do
14: $sum := 0$;

```
15:         from k = 1 to N do
16:               b := W_k I_H^t ("backward" warping);
17:               c := h(x, y) * b (convolution with Gaussian Kernel);
18:               c := D_c (down-sampling to I_L dimensioning);
19:               d := (I_L^k − c);
20:               b := D^T d (up-sampling without interpolation);
21:               c := h(x, y) * b
22:               d := W_k^T c ("forward" warping);
23:               sum := sum + A^(k) d (colors map consideration);
24:         end for
25:         I_H^{t+1} = I_H^t + τ(sum − λ △ I_H^t);
26: end for
```

With these changes, a given frame $I_L^{(k)}$ contribution to the final result is lower in the case that corresponding C_k presents a color pattern different when compared to the average color of the other.

5 Tests

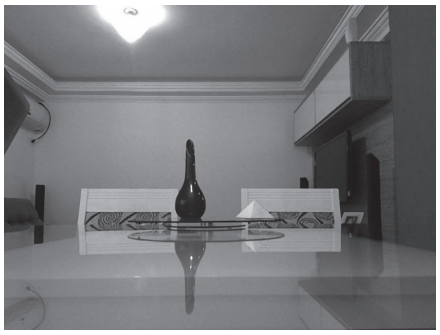
Aiming at finding some improvement arising from colors consideration included in the proposed algorithm, a turntable test was designed.

A turntable was positioned two meters from the Kinect. It was marked so as to be divided into eight identical pieces. At each frame generated, the turntable was rotated 45 °. Furthermore, two marks have been made on turntable surface to indicate the position of two interference objects over tests. Test sequence is as follows:

1. With turntable empty, depth maps are captured until the first super-resolved frame is generated;
2. Interference objects are positioned on the respective marks;
3. Each super-resolved frame generated, the turntable is rotated 45 °.
4. Previous steps are repeated 25 times.

This procedure was executed once for the original Mitzel el al [1] algorithm implementation, without colors consideration, and again for the proposed implementation that incorporates RGB images to the process. Objects added after first super-resolved frame generation act as interferences between captured RGB images. Due to turntable rotation, these differences move through pixels at each frame. So, colors consideration inclusion might avoid that depth of interference objects has some influence upon the super-resolved frames.

Using the original Mitzel et al. [1] algorithm, at the twelfth high resolution image generated, interference objects which must not appear, were visualized, as shown in figure 1b. Executing the algorithm proposed in this paper, after 25 repetitions of the third step, no signal of interference objects emerged, as shown in figures 1c and 1d. Thus, this result indicates that colors consideration



(a) Captured scene with interference objects

(b) Interferences detected when RBG images are not considered

(c) First scene considering RGB information

(d) After several repetitions frame does not show interferences

Fig. 1. Turntable test

avoids that abrupt changes in the captured elements have some influence on the distances presented in super-resolved map.

In addition, we wanted to show that the proposed algorithm has mitigated the interference caused by abrupt modifications in captured scene, human being clothes for example, during the super-resolved images generation process.

For such an assessment, a human being was positioned in the centre of captured scene. We capture this environment continuously. On average, every 10 s, capture was frozen and the person put on a suit. After three new frames capture, the process was frozen again and the person undressed the suit. This procedure has been repeated for 120s.

If the proposed algorithm is capable of eliminating artifacts caused by punctual changes between captured frames, as the result reached by Cui et al. [8], when super-resolved maps were generated considering RGB images, the suit will not be detected.

The results of the original algorithm described by Mitzel et al. [1] suffered interference of the frames which contain the suit, as shown in figure 2b. When using the algorithm proposed at this work, the suit does not appear, as shown in figures 2c and 2d.

This test has shown that, within a 3D digitalization system dedicated to human body, the use of RGB images would avoid clothes movement of captured body, for example, harm depth levels provided by final mesh reconstruction.

These tests indicate that, with color consideration, super-resolved depth maps will be less prone to interferences during capture process.

(a) Captured body

(b) Captured body with interference

(c) First scene considering RGB information

(d) After several repetitions frame does not show any interference of the suit

Fig. 2. Test with human body

Finally, to analyse the impact of CPU and GPU algorithms we use a computer with Intel Core i7 with a 1.73GHz Processor, 6 GB of RAM and GEFORCE GT 425M Video Card. Runtime related to the ten first frames has been computed for the following distinct combination:

- Using CPU and without colors consideration;
- Using GPU and without colors consideration;
- Using CPU and with colors consideration;
- Using GPU and with colors consideration.

The captured resolution of depth maps was 640 x 480 and of the colored images was 1280 x 1024. Since scaling factor β was 2, super-resolved resolution was 1280 x 960.

Tests have showed that CPU runtime is sevenfold bigger than GPU runtime. Analising color consideration impact, CPU runtime increased about 9% — to *frame0* — and 18% — to the other ones. GPU runtime increased about 42% — to *frame0* — and 41% - to the other frames.

6 Conclusion

This paper aimed at making a super-resolution system capable of increasing resolution of depth maps provided by low-cost sensors. That system would serve as a basis for human bodies 3D scanning. To mitigate problems caused by possible movements of captured bodies, we used color images of the same scene that provided the depth maps.

Theoretical contribution of this work consists in incorporating Cui et al. [8] approach, to Mitzel et al. [1] algorithm, through equation 3. The two concepts junction, here proposed, has allowed use of Mitzel et al. [1] algorithm implementation provided in OpenCV library as a base for the presented strategy.

The classes, dedicated to super-resolution, provided by OpenCV library demanded some modifications to accomplish two aspects:

- depth maps and colored image concomitant acquisition;
- incorporation of color images consideration in the original super-resolution algorithm.

These changes, culminated in exposed results. Tests have indicated that colored images use improved super-resolved depth maps in the sense by avoiding interferences caused by abrupt changes in captured scene.

Addressing future work possibilities, it would be interesting to exploit two points:

- system implementation optimization;
- proposition of a new approach over colors consideration that could be computed faster.

Global and non-global registration modules development would continue low cost 3D digitalization system and would allow volumetric tests capable to provide new data regarding the efficiencies of colors consideration. More than that, that implementation would allow comparison with reconstruction results reached by Cui et al. [8].

Furthermore, future tests could analyse two aspects:

- What is the response of the proposed algorithm on scenes where the object has a small interference contrast from the rest of the captured scene?
- If we use a low-resolution image as input for the rigid and non-rigid registrations, which would be the impact on the reconstructed 3D mesh?

References

1. Mitzel, D., Pock, T., Schoenemann, T., Cremers, D.: Video super resolution using duality based TV-L^1 optical flow. In: Denzler, J., Notni, G., Süße, H. (eds.) Pattern Recognition. LNCS, vol. 5748, pp. 432–441. Springer, Heidelberg (2009)
2. Leica Geosystems: (Leica Geosystems), `http://www.leica-geosystems.com/en/index.htm`
3. LFM: (LFM), `http://www.lfm-software.com/lfm-products`
4. Cardon, D.L., Fife, W.S., Archibald, J.K., Lee, D.J.: Fast 3D reconstruction for small autonomous robots. In: Proceedings of the 31st Annual Conference of IEEE Industrial Electronics Society (IECON 2005) (2005)
5. BodyMetrics (BodyMetrics), `http://www.bodymetrics.com/`
6. Chen, J., Bautembach, D., Izadi, S.: Scalable real-time volumetric surface reconstruction. ACM Trans. Graph. 32, 113:1–113:16 (2013)
7. Baran, I., Popović, J.: Automatic rigging and animation of 3D characters. ACM Transactions on Graphics 26, 72:1–72:8 (2007)
8. Cui, Y., Chang, W., Nöll, T., Stricker, D.: Kinectavatar: Fully automatic body capture using a single kinect. In: AACV, Workshop on Color Depth Fusion in Computer Vision (2012)
9. Tong, J., Zhou, J., Liu, L., Pan, Z., Yan, H.: Scanning 3D full human bodies using kinects. Technical report, Hohai University (2012)
10. Kazhdan, M., Bolitho, M., Hoppe, H.: Poisson surface reconstruction. In: Proceedings of the Fourth Eurographics Symposium on Geometry Processing, SGP 2006, pp. 61–70. Eurographics Association, Aire-la-Ville (2006)
11. Newcombe, R., Kim, D., Kohli, P.: Kinectfusion: Real-time dense surface mapping and tracking. In: ISMAR (2011)
12. Schuon, S., Theobalt, C., Davis, J., Thrun, S.: Lidarboost: Depth superresolution for tof 3D shape scanning. In: CPVR 2009 (2009)
13. Cui, Y., Schuon, S., Chan, D., Thrun, S., Theobalt, C.: 3D shape scanning with a time-offlight camera. In: IEEE Proc. CVPR (2010)
14. Chang, W., Zwicker, M.: Global registration of dynamic range scans for articulated model reconstruction. ACM Trans. Graph. 30 (2011)
15. OpenCV: (OpenCV), `http://www.opencv.org`