Deep Learning based Aesthetic Evaluation of State-Of-The-Art 3D Reconstruction Techniques

Gernot STUEBL, Christoph HEINDL, Harald BAUER, Andreas PICHLER
PROFACTOR GmbH, Steyr-Gleink, Austria

DOI: 10.15221/17.306 http://dx.doi.org/10.15221/17.306

Abstract

In the field of 3D human body scanning, due to different scanning technologies different reconstruction approaches have emerged. The two main ones are based either on pure 2D information, like photogrammetry, or 2D plus depth, as used with RGBD active structured light sensors.

Reconstruction results of these technologies differ in geometric as well as aesthetic quality. Whereas the judgement of geometric quality is straight forward, a judgement of the aesthetics aspects (e.g. proper texture mapping, etc.) strongly depends on the subjective perception of the human viewer.

Recent advances in image aesthetics assessment, demonstrate that machine-learning algorithms, specifically deep neural networks, are able to model human aesthetic perception in a reasonable way. Especially if they are trained with a huge number of data.

This work presents research towards an unbiased aesthetic judgement of 3D reconstructions by utilizing a deep neural network. In detail, two state-of-the-art software suites as representatives for 2D and 2D plus depth reconstruction approaches are compared according to the aesthetics of their results. The models of a publicly available dataset are virtually scanned with a sensor simulator, which produces the necessary 2D and depth information. This data serves as input to the mentioned software suites. The resulting 3D reconstructions are aligned and a deep neural net aesthetically compares frontal views of the models. To ensure a fair comparison between different models a normalized aesthetic value is introduced.

Keywords: 3D bust scanning, deep neural network, aesthetic, quality, RGBD, photogrammetry

1. Introduction

Geometric comparison of 3D models is a solved problem. However when it comes to aesthetic assessment, human subjective perception make a fair comparison difficult. In this paper, a deep neural network trained on 500000 aesthetic judgements is the base for aesthetic evaluation. It is assumed that the huge number of training data can model a mean human aesthetic perception.

The neural net is used to compare the reconstruction results of two different software suites namely Agisoft PhotoScan [1] as representative of photogrammetry with 2D data only and ReconstructMe [2] as representative for 2D plus depth reconstruction. This is of further interest, since both technologies are heavily used in 3D body scanning systems and can be seen as competitors.

This work is outlined as followed: in Section 2, related work to reconstruction techniques as well as image assessment is given. In Section 3 the used dataset and the experimental setup is explained in detail. A discussion of the results can be found in Section 4, where also the measure of normalized aesthetic value is introduced. Finally, Section 5 provides a conclusion and an outlook to further research questions.

2. Related Work

The related work part is divided into two subsections, the first concerning reconstruction techniques and the second concerning aesthetic assessment.

2.1. Reconstruction techniques

In this paper, two different reconstruction techniques are compared. The first one is based on photogrammetry, which is the science to make measurements out of photographs. A classical approach in this area bases on Structure from Motion (SfM) [3], with the software suite Bundler [4] as representative. Bundler constructs a 3D model through Scale-Invariant Features (SIFT) [5] feature point matches which are verified by a Random Sampling Consensus (RANSAC) [6] based algorithm.

A different approach, as followed by Labatut [7], is graph cut based. The work of Heller [8] extends this idea to further account "weakly-supported surfaces" which allows also constructing semi-transparent or poorly textured surfaces. This is a technique, which is the base of current state-of-the-art software [9]. Reconstruction by photogrammetry needs only image data to generate a 3D model. If there is also additional depth information available, as e.g. with RGBD cameras, other reconstruction techniques can be applied. The most famous one in this field is the KinectFusion approach by Izadi et al. [10]. This was a break-through publication, which inspired lot of follow-up work. The KinectFusion algorithm builds up a 3D volume, which is afterwards filled by processing the RGBD data. This leads to a point cloud representation of an object. Further extensions as described in Gusev [11] and Heindl et al. [12] combine this approach with triangulation of the points and texturing to have a full watertight 3D model. The ReconstructMe software suite an implementation of [12].

2.2. Aesthetic assessment

One distinguishes three types to assess the aesthetics of images: The simplest one is the type of objective image quality assessment, where different distortions like blocking, ringing, mosaic patterns, blur, noise, ghosting, etc. are measured. Objective image quality assessment is mainly used to assure the quality during transmission or compression of images. Recently Stübl et al. [13] showed how a simple image distance function could be used as quality assurance measure, which correlates to human aesthetic perception. A more complicated quality assessment type is to use handcrafted features, often combined with machine-learning techniques. This is a huge research area with various publications, see [14, 15]. The authors will not go into detail, since this is not the focus of the paper. The last type of image quality assessment is named deep image aesthetic quality assessment. In this type, deep neural nets are used for aesthetic quality assessment. Systems build on deep neural nets nowadays outperform classical approaches in many fields of computer vision, e.g. in classification [16]. Also in the field of aesthetic assessment, deep neural nets beat traditional systems, as recent publications like ILGnet [17] show. ILGnet is trained on nearly 500000 aesthetic image ratings and one of the first systems, which models human aesthetic perception in a reasonable way.

3. Aesthetic Comparison of 3D Reconstructions

In this section, the used dataset and the experimental setup are described.

3.1. Dataset

The dataset consists of 12 publicly available 3D bust scans, which are hosted on the website of Sketchfab[1]. The model references names on the webpage are: `Alexander`, `Alexey`, `Dmitry`, `Georgy`, `Ilya_b`, `Ilya_l`, `Ilya-ustinov`, `Nick`, `Sonya`, `Supportitseez3d`, `Tanya`, and `Victor` provided by the contributor itSeez3D. All models are under the Creative Commons Attribution-Noncommercial license and can be freely used within the scientific community. Special care was taken that none of the tested software suites was involved during creation of the model.

3.2. Experimental setup

Fig. 1 illustrates the experimental setup. A 3D model from the dataset is scanned with a virtual software RGBD scanner, which simulates the behavior of a PrimeSense Carmine 1.08 sensor.

[1] https://sketchfab.com/itseez3d

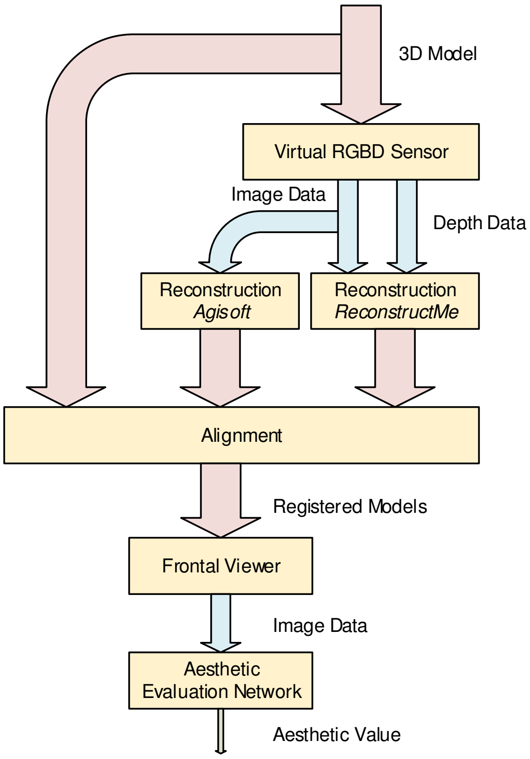
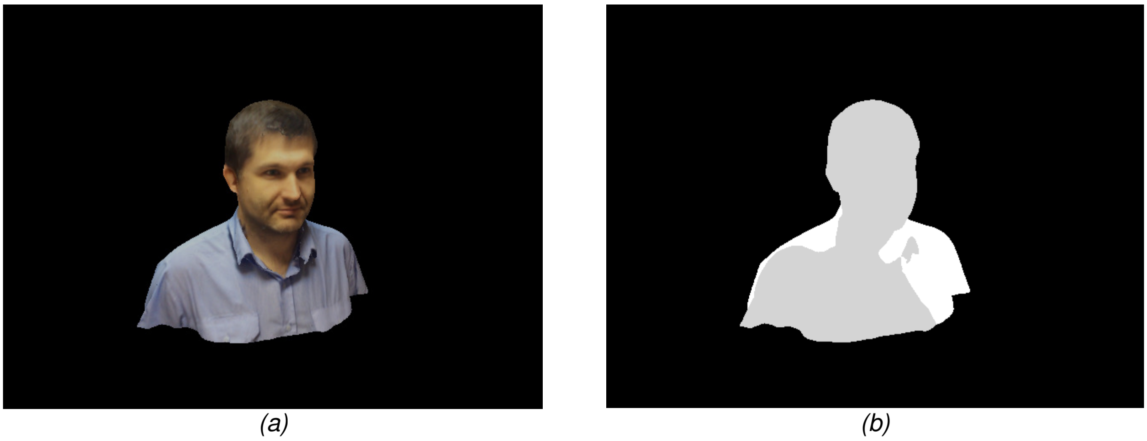


Fig. 1. Experimental setup illustrated as pipeline. The 3D model is scanned with a virtual sensor, which produces the input for the reconstructions suites. Their output is than aligned and frontal views of the models are generated. These serve as input of the deep neural network, which does the aesthetic judgement.

The virtual scan process imitates a rotary plate scan that takes a photo every 10° in a distance of 1.6m. The distance and angle values are chosen to have a good coverage of all models in the dataset. The virtual RGBD sensor produces 2D images as well as depth data, see Fig. 2 for an example.



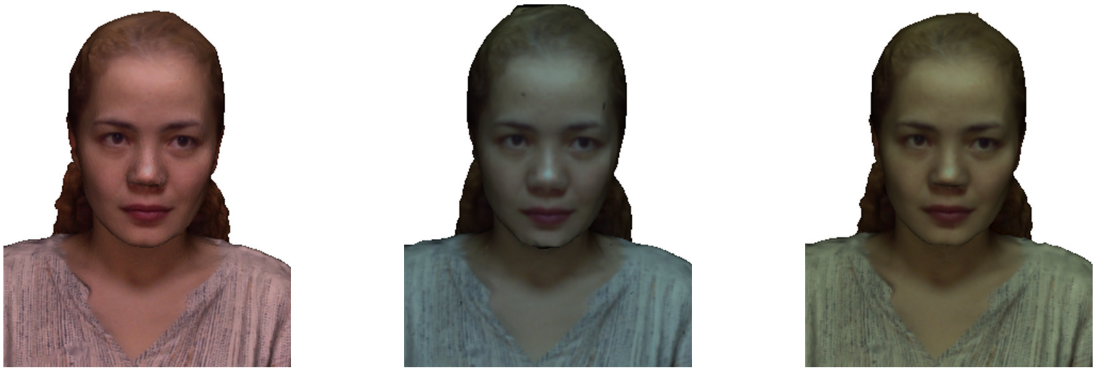
(a) *(b)*

Fig. 2. Example of virtual scan of Model `Ilya-ustinov`: (a) the 2D information, (b) depth information where brighter is farer away. The black background indicates no depth information.

These two data types are fed into the ReconstructMe reconstruction suite, which produces again a 3D model. The PhotoScan reconstruction suite on the other hand needs only the 2D data for reconstructing a 3D model. For both software suites, the standard settings for high precision reconstruction have been used.

In a next step, the two reconstructed models are aligned to the base model by using the Iterative Closest Point Algorithm [18] with a previous manual rough alignment. The alignment assures that the models have the same size and orientation in space, which is important for the further process, in which frontal images of the models are taken. Previous work [13] showed that aesthetic judgement of busts is dominated by the view on the face; this is why a single frontal image is assumed to be enough to compare aesthetics of a model.

The original ILGnet deep neural network, was modified to give an aesthetic rating between 0 and 1 for an input image and is used to perform aesthetic judgement of the three individual frontal images, see Fig. 3 for an example. The described procedure is repeated for all 12 models in the database.



(a) Original, 0.5294   *(b) PhotoScan, 0.1754*   *(c) ReconstructMe, 0.4520*

Fig. 3. Frontal views of different 3D representations of Model `Supportitseez3d` and their aesthetic values (higher is better). The reconstruction with ReconstructMe is juged to be more aesthetic than the one with PhotoScan.

4. Results

To allow a comparison of different models, one has to normalize the aesthetic values of the reconstructions. Therefore, we introduce the normalized aesthetic value

$$\widehat{\boldsymbol{A}}(\boldsymbol{M_x}) = \frac{\boldsymbol{A}(\boldsymbol{M_x})}{\boldsymbol{A}(\boldsymbol{M_o})} \qquad (1)$$

defined as the ratio between the aesthetic value of the reconstructed model $\boldsymbol{M_x} \in \{\boldsymbol{M_P}, \boldsymbol{M_R}\}$, with superscript $\boldsymbol{P}$ for PhotoScan and $\boldsymbol{R}$ for ReconstructMe, and the aesthetic value of the original model $\boldsymbol{M_o}$**.** Values near 1 indicate similar aesthetic values of the reconstruction and the original model, where values higher than 1 refer to an improved aesthetic value.

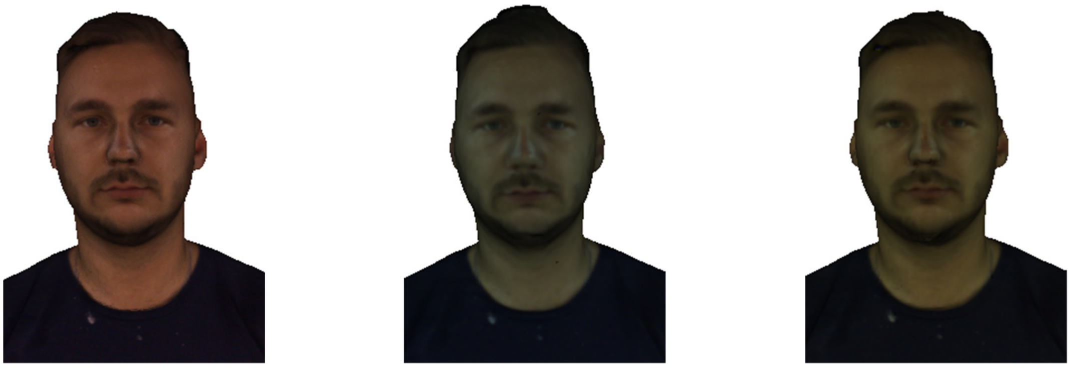
Table 1. Models and their normalized aesthetic values (higher is better). The reconstructions of ReconstructMe have been judged to be more aesthetic than the ones of PhotoScan in 9 out of 12 cases. In 6 cases (marked bold) the aesthetic value is close the value of the base model.

Model Name	**Normalized Aesthetic Value**	
	$\widehat{\boldsymbol{A}}(\boldsymbol{M_P})$ (PhotoScan)	$\widehat{\boldsymbol{A}}(\boldsymbol{M_R})$ (ReconstructMe)
`Alexander`	0.928	1.437
`Alexey`	0.694	0.846
`Dmitry`	0.883	**1.019**
`Georgy`	1.503	1.295
`Ilya_b`	0.745	**1.060**
`Ilya_l`	0.821	**1.017**
`Ilya-ustinov`	0.891	0.700
`Nick`	0.892	**1.057**
`Sonya`	0.546	0.733
`Supportitseez3d`	0.331	0.854
`Tanya`	**0.986**	**1.075**
`Victor`	0.912	0.761

Table 1: lists the normalized aesthetic values of all models in the dataset. Values near one are marked bold, since they are close to the value of the original model. It showed up that the reconstructions of ReconstructMe have been judged more aesthetic than the ones of PhotoScan in 9 out of 12 cases.

5. Discussion

At a first glance, the results of Section 4 suggest that reconstructions with ReconstructMe are more aesthetic than the ones produced by PhotoScan. However, one has to have in mind that the judgement is done via a deep neural net, whose working principle is mainly a black box. It is not sure on which features the judgement is based, neither if it truly reflects human judgement. For example for the Model Georgy $\widehat{\boldsymbol{A}}(\boldsymbol{M_P}) > \widehat{\boldsymbol{A}}(\boldsymbol{M_R})$, although for a human viewer there is only a minor difference visible, see Fig. 4. Nevertheless, the results show a trend in the direction of ReconstructMe.



(a) Original, 0.2213 *(b) PhotoScan, 0.3328* *(c) ReconstructMe, 0.2866*

Fig. 4. Frontal views of different 3D representations of Model Georgy and their aesthetic values (higher is better). The reconstruction with PhotoScan is judged to be more aesthetic than the one of ReconstructMe. Visually there is only minor difference.

A further detail, which came up during the inspection of the results, is, that models reconstructed with PhotoScan are more elongated that the ones created with ReconstructMe, see Fig. 3b and 4b. Having in mind that PhotoScan does not have depth information this behavior is expectable. For ReconstructMe it is easier to have better geometric accuracy through the use of depth information.

An interesting result of the experiment is that for ReconstructMe a high number of normalized aesthetic values near one appear (5 out of 12). For all of them the aesthetic assessment is better than with PhotoScan. However, a value near one implies that the aesthetic value does not change significantly through the scan and reconstruction procedure. This was a surprise, since the virtual scan procedure combined with the reconstruction principle is known to be not lossless. A possible explanation is that the post-processing of ReconstructMe, which bases on the work of Heindl et al. [12], tries to texture the face out of a single image only. This single image inherently might be similar to the one taken by the FrontalViewer. From another point of view, it also means that in nearly half of the cases this feature works quite well. To texture the face out of one image only could also be a potential improvement for the PhotoScan software.

6. Conclusion and Future Work

This paper presents a first step towards an unbiased aesthetic assessment of 3D reconstructions. For this, 3D models of a publicly available dataset have been virtually scanned and the image and depth data was used as input for two 3D reconstruction software suites, namely PhotoScan and ReconstructMe. To do a fair comparison between two different models the concept of normalized aesthetic value has been introduced. The results show that for ReconstructMe 9 out of 12 aesthetic quality assessments are better than the reconstructions with PhotoScan. However since the deep neural network is a black box it is not known which features led to this decision. A possible future work might be to do a detailed analysis of the neural net's behavior. By this, one should show that the assessment is independent from local image properties (e.g. smoothness), which could simply be a property of the reconstruction principle.

An additional finding is that in 5 out of 12 cases the aesthetic value does not alter significantly from the original model to the reconstruction with ReconstructMe. Since in all these cases the aesthetic value is higher with ReconstructMe than with PhotoScan, this does not change previous results. However, for further experiments one might create more than the frontal view and use the mean of aesthetic values. In addition, the influence of different kind of noise would be interesting.

7. Acknowledgment

This research is carried out within the ”FTI-Project ProTechLab” project funded by the State of Upper Austria through the Strategic Economic and Research Program ”Innovatives OÖ 2020”.

References

[1] *Agisoft Photoscan Professional v1.3.2 build 4205 (64bit)*, http://www.agisoft.com/, accessed 2017.
[2] *ReconstructMe v2.5.1034 (64bit)*, http://reconstructme.net/, accessed 2017.
[3] R. Hartley and A. Zisserman. “Multiple View Geometry in Computer Vision”, (2nd Ed.). Cambridge University Press, New York, NY, USA. 2003.
[4] N. Snavely, S.M. Seitz, and R. Szeliski. “Photo Tourism: Exploring Photo Collections in 3D”. In Proceedings of SIGGRAPH Conf., 2006.
[5] D.G. Lowe. “Object Recognition from Local Scale-Invariant Features.”, In Proceedings of the International Conference on Computer Vision (ICCV). Vol 2, pages 1150-1157, 1999.
[6] M.A. Fischler and R.C. Bolles. “Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography.”, Comm. ACM 24, 6, pages 381-395, June 1981, DOI:http://dx.doi.org/10.1145/358669.358692
[7] P. Labatut, J.-P. Pons, and R. Keriven. *“Efficient multi-view reconstruction of large-scale scenes using interest points, Delaunay triangulation and graph cuts.”,* In Conference on Computer Vision and Pattern Recognition (CVPR), pages 1-8, June 2007.
[8] J. Heller, M. Havlena, M. Jancosek, A. Torii, and T. Pajdla. “3D Reconstruction from Photographs by CMP SfM Web Service”, *In IAPR International Conference on Machine Vision Applications (MVA)*, pages 30-34, 2015.
[9] Capturing Reality s.r.o. Reality Capture, https://www.capturingreality.com/, accessed 2017.
[10] S. Izadi, D. Kim, O. Hilliges, D. Molyneaux, R. Newcombe, P. Kohli, J. Shotton, S. Hodges, D. Freeman, A. Davison, and A. Fitzgibbon. “KinectFusion: real-time 3D reconstruction and interaction using a moving depth camera.”, In Proceedings of the 24th annual ACM symposium on User interface software and technology (UIST '11). ACM, New York, NY, USA, pages 559-568, 2011. DOI: https://doi.org/10.1145/2047196.2047270
[11] G. Gusev*. “3D Self-Portraits. ACM Transactions on Graphics.”,* Proceedings of the 6th ACM SIGGRAPH Conference and Exhibition in Asia 2013, 11/2013
[12] C. Heindl, S.C. Akkaladevi, and H. Bauer. *“Capturing Photorealistic and Printable 3D Models Using Low-Cost Hardware.”,* Springer International Publishing, Cham, 2016, pages 507-518.
[13] G. Stübl, C. Heindl, H. Bauer, and A Pichler. „On Quality Assurance of 3D Bust Reconstructions.”, Proceedings of the 2nd OAGM-ARW Joint Workshop Vision, Automation and Robotics, 2017
[14] Y. Niu and F. Liu, "What Makes a Professional Video? A Computational Aesthetics Approach," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 22, no. 7, pp. 1037-1049, July 2012. DOI: 10.1109/TCSVT.2012.2189689
[15] H.-H. Su, T.-W. Chen, C.-C. Kao, W.H. Hsu, and S.-Yi Chien. “Scenic photo quality assessment with bag of aesthetics-preserving features”. In *Proceedings of the 19th ACM international conference on Multimedia* (MM '11). Pages 1213-1216. 2011. DOI: http://dx.doi.org/10.1145/2072298.2071977
[16] A. Krizhevsky, I. Sutskever und G. E. Hinton, „Imagenet classification with deep convolutional neural networks,“ in Advances in neural information processing systems, 2012.
[17] H. Li, E. Vouga, A. Gudym, J.T. Barron, L. Luo, and X. Jin, J. Chi, S. Peng, Y.Tian, C. Ye, and X. Li, “Deep Image Aesthetics Classification using Inception Modules and Fine-tuning Connected Layer”, in Proc. of 8th Int. Conf. on Wireless Communications and Signal Processing (WCSP), Yangzhou, China, 13-15 October, 2016. https://arxiv.org/abs/1610.02256/
[18] T. Jost and H. Hügli. 2002. Fast ICP Algorithms for Shape Registration. In *Proceedings of the 24th DAGM Symposium on Pattern Recognition*, Luc J. Van Gool (Ed.). Springer-Verlag, London, UK, UK, 91-99.