Published in the proceedings of ICPRAM 2015.
The final publication is available at scitepress.org

Interactive 3D modeling
A survey-based perspective on interactive 3D reconstruction

Julius Schöning and Gunther Heidemann
Institute of Cognitive Science, University of Osnabrück, Germany
{*juschoening, gheidema*}*@uos.de*

Keywords: Image-based Modelling, 3D Reconstruction, Interactive, User Centered, CAD-ready, Survey

Abstract: 3D reconstruction and modeling techniques based on computer vision show a significant improvement in recent decades. Despite the great variety, a majority of these techniques depend on specific photographic collections or video footage. For example, most are designed for large data collections, overlapping photos, captures from turntables or photos with lots of detectable features such as edges. If the input, however, does not fit the particular specification, most techniques can no longer create reasonable 3D reconstructions. We review the work in the research area of 3D reconstruction and 3D modeling with a focus on the specific capabilities of these methods and possible drawbacks. Within this literature review, the practical usability with the focus on the input data — the collections of photographs or videos — and on the resulting models are discussed. Upon this basis, we introduce our position of interactive 3D reconstruction and modeling as a possible opportunity of lifting current restrictions from these techniques, which leads to the possibility of creating CAD-ready models in the future.

1 INTRODUCTION

A multitude of computer vision based techniques reconstruct and model 3D objects or scenes from photographs or video footage captured in 2D (monocular) or 3D (stereo). Many of these techniques and approaches reconstruct objects or scenes with automatic algorithms from image sequences (Tanskanen et al., 2013; Pan et al., 2009; Pollefeys et al., 2008; Pollefeys et al., 2004; Zollhöfer et al., 2014; Snavely et al., 2006). Alongside to these academical approaches a growing number of commercial software products have been presented in the last years (Agisoft, 2014; Autodesk, Inc., 2014; Microsoft Corporation, 2014; Trimble Navigation Limited, 2014). But would an architect or an engineer call the resulting 3D reconstruction a CAD-model? Is a non-expert user able to apply these tools? Can the user apply the method without special hardware like a rotating plate, laser, stereo camera or even a main frame computer? Are users able to reconstruct models of real world objects in such a way that they will be able to translate them back to real world replicas? The answer to these questions is mostly – *No* – but why? Therefore, we provide a literature review of automatic and interactive 3D reconstruction and modeling techniques, compare them and analyse weaknesses of existing methods. Under these perspectives, we provide a discussion of how an interactive computer vision application could be employed to overcome existing weaknesses. Further, we develop an idea for an application for the interactive creation of non-monolithic, functioning 3D models — models that architects or engineers would call a model, and CAD-ready models that can be applied for simulation tasks, reverse engineering, replication and many more purposes.

2 VARIETY OF TECHNIQUES

For 3D reconstruction techniques various research areas can be identified. In order to provide a simple taxonomy, the input data type is used as the identifying header. Monocular and stereo inputs are not separate, because the stereo advantage is not as significant as commonly expected due to the limited stereo distance of typically less than $5m$, the noisy depth estimation and field of view of $\sim 60^\circ$ (Henry et al., 2014). The 3D reconstruction techniques and applications discussed in the following do not necessarily focus on reconstructing scenes or objects for modeling purposes only. Instead, other applications like, e.g., robot navigation, might be considered. We do

not claim that the list is complete but we argue that it contains the relevant work which should be considered in this context.

2.1 Collections of photographs

A well known technique that uses a collection of photographs as input data is *Photo tourism* by Snavely et al. (Snavely et al., 2006). This tool uses an unstructured collection of photographs of a scene, e.g., acquired from the internet, and converts them to a sparse 3D model of the scene. Thus the user can browse, explore and organize large photo collections in a 3D model of the scene. Camera resections, the reconstruction of viewpoints from where the photos are taken, are also possible. *Photo tourism* reconstructs the model by computing correspondence features of each image using SIFT and RANSAC, which are based on descriptors that are robust with respect to variations in pose, scale, and lighting. An optimization algorithm is used to recover the camera parameters and the 3D position of the feature points. A robust structure from motion algorithm for 3D structure estimation is the backbone of *Photo tourism*. Some weaknesses of this system are that it works only with huge collections. Also, textureless or repeating structures cannot be reconstructed, only continuous scenes. Also, the resulting model is quite sparse.

Inspired by *Photo tourism*, the software *Microsoft Photosynth* (Microsoft Corporation, 2014) is a tool for capturing, viewing and sharing photos in 3D. It works in the same way as *Photo tourism* and automatically reconstructs a collection of overlapping photographs into a 3D model. A scene of 200 photographs is computed in five to ten minutes on an average laptop. To overcome the drawbacks of *Photo tourism*, a photography guide is provided which helps the user with how to take photos such that *Photosynth* can be used to best advantage. The guide provides hints such as to not shoot repetitions, complex occlusions, or shiny objects. For well made collections, acceptable models can be obtained. Moving or dynamic objects cannot handled by *Photosynth*.

The main purpose of the first two techniques is browsing, exploring and organizing photographic collections in a 3D model. The purpose of *Agisoft PhotoScan* (Agisoft, 2014) is the automatic creation of textured 3D models from photographs. For a successful and complete reconstruction, photographs shown in a 360° view of the object are required. It depends on the viewpoints whether the software works automatically or interactively, because the user is required to outline the object of interest manually in the reconstruction process. By this means, irrelevant elements such as the background or the turntable are excluded. Further processing steps like aligning photographs, building a dense point cloud, meshing point clouds and creating textured surfaces work automatically in *PhotoScan*. Like *Photosynth*, *Agisoft PhotoScan* comes with a guideline, also, its drawbacks are similar.

Like *PhotoScan*, *Autodesk 123D Catch* (Autodesk, Inc., 2014) creates 3D models from a collection of photographs and like *PhotoScan* it builds a monolithic 3D model. But in *123D Catch*, the user can manipulate the model in a post-processing phase. To increase speed, the overall process is outsourced to cloud computing. Like the other systems, *123D Catch* can not handle moving objects, reflections, under- or overexposure, blurred photographs, occlusions etc.

Kowdle et al. (Kowdle et al., 2014) addressed these issues of automatic reconstruction for object creation by a semi-automatic approach. In this approach, the user is in the loop of computational reconstruction, which allows the user to intuitively interact with the algorithm. The user guides the process of image-based modeling to find a model of the object of interest by interacting with the nodes of the graph. Thus, the 3D reconstruction achieves a much higher quality in an acceptable time span, especially for scenes with textureless surfaces and structural cues. This semi-automatic approach works with multiple images of a scene, captured from different viewpoints. In the pre-processing steps a structure from motion algorithm creates a dense 3D point cloud. This cloud is used for growing 3D superpixels. With user interaction the superpixels are labeled to segments and finally to the object of interest. Using this knowledge, a RANSAC based plan-fitting on the labeled 3D points estimates the 3D Model.

2.2 Single photograph

Photo collections acquired according to guidelines for 3D reconstruction are rare. Single photos or very small collections of photos are more common. Debevec et al. (Debevec et al., 1996) presented an early approach to hybrid modeling, that can model and render existing architectural scenes from a sparse set of still photographs. But to extract a model from a small collection requires addition information. The *Façade* software combines an interactive image-guided geometric modeling tool with model-based stereo matching and a view-dependent texture mapping. In the interactive modeling phase the user selects block elements and aligns their edges with visible edges in the input images. The system then automatically computes the dimensions and locations of the blocks

along with the camera resection. Based on the assumption that man-made architecture relies on geometric elements, this approach reconstructs more reliable models from geometric primitives.

Upon this principle the plug-in *Match Photo* for the Trimble Navigation Limited tool *SketchUp* provides a manual method for creating a 3D model from one photograph or match an existing CAD-model into a photograph. After calibrating the photo to the coordination system of the *SketchUp* workspace, the user draws the edges of the object over the photograph. When all visual and occluded edges are drawn, the 3D model is built and can be texturized (Trimble Navigation Limited, 2014).

Another approach we want to mention is the reconstruction of building interiors by Furukawa and Szeliski (Furukawa et al., 2009) because it is able to handle textureless planar surfaces such as uniformly-painted.

2.3 Video footage

An automatic 3D reconstruction method from video footage described by Pollefeys et al. (Pollefeys et al., 2008) produces a dense, geo-registered 3D model of an urban scene in real-time. The video footage is captured by a multi-camera system in conjunction with INS (Inertial Navigation System) and GPS measurements. To achieve real-time they decouple the problem into the reconstruction of depth maps from sets of images followed by the fusion of these depth maps, a simple and fast algorithm that can be implemented on GPUs. It yields a compact and geometrically consistent representation of the 3D scene.

ProFORMA (Probabilistic Feature-based Online Rapid Model Acquisition) (Pan et al., 2009) reconstructs freely rotated objects in front of a fixed-position video camera in near real-time. Due to the fact that the system guides the user with respect to the manipulation, i.e., the rotation of the object, this method might not be called an entirely automatic method. The user rotates the textured object in front of the camera. The final model is produced by Delaunay tetrahedralization of a point cloud obtained from online structure from motion estimation, followed by a probabilistic tetrahedron carving step to obtain a textured surface mesh of the object. A key-frame for the reconstruction is taken when a sufficiently large rotation is detected. During the initialization phase a photo of the background is captured for background removal.

VideoTrace (van den Hengel et al., 2007) is a system for interactive generation of realistic 3D models of objects from video. These models can be inserted into a simulation environment or another application. The user interacts with *VideoTrace* by tracing the shape of the object over one or more frames of the video. Immediate feedback mechanisms allow the user to rapidly model those parts of the scene which are of interest, up to the required level of detail. The combination of automated and manual reconstruction allows *VideoTrace* to model parts of the scene which are not visible, and to succeed in cases where purely automated approaches would fail. Before any interactive modeling takes place, structure and motion analysis is carried out on the video sequence to reconstruct a sparse set of 3D scene points. The resulting 3D point cloud is "drawn" over the frames of the input sequence. The interaction is done by modeling primitives. By default, these primitives (e.g., traced lines or curves) are automatically refined in 2D by fitting to local strong superpixel boundaries, so a user neither needs artistic abilities, nor is "pixel perfect" tracing required. The interaction process occurs in real time. This allows the user to switch between images from the original sequence naturally, selecting the most appropriate view of the scene at each stage. The model can be rendered using texture maps obtained from frames of the video.

Generating dense 3D maps of indoor environments using a RGB-D camera has been proposed by Henry et al. (Henry et al., 2014), despite the limited depth precision and field of view such cameras can provide. This technique effectively combines the visual and shape information of a RGB-D camera. Aligning the current frame to the previous frame with an enhanced iterative closest point algorithm combines the RGB and the D information. The resulting feature point cloud in 3D is visualized in surfels (surface patches).

Further reconstruction methods to be considered are automated reconstruction of buildings using a hand held video camera (Fulton and Fraser, 2009), live metric 3D reconstruction on mobile phones (Tanskanen et al., 2013), in situ image-based modeling (van den Hengel et al., 2009) and visual modeling with a hand-held camera (Pollefeys et al., 2004), but can not be described in this short paper.

3 DIRECT COMPARISON

By comparing all techniques mentioned in Section 2 yield, Table 1. It shows a comparison of most significant techniques (top half) and commercial software products (lower half).

Table 1: Direct comparison of significant techniques and commercial software products for 3D reconstruction and modeling.

	Main purpose	**Input**	**Output**	**Mode**[1]	**Recording equipment**	**Handling of texture -less, shiny etc. objects**	**Initial-ization**[2]
Modelling architecture (Debevec et al., 1996)	model acquisition for man-made objects and scenes	single photograph or collection of photographs	monolithic 3D model of the object or scene	i	monocular camera	possible but should be prevented	n
Photo tourism (Snavely et al., 2006)	browsing, exploring and organizing photo collections in a sparse 3D the scene	huge collection of photographs ($>$ 100 photos)	sparse 3D model of the scene	a	monocular camera with or without GPS device	hard – photographs should not have repeating or textureless structures	n
VideoTrace (van den Hengel et al., 2007)	model acquisition for scenes	video sequence	monolithic 3D model of the scene or part of the scene	i	monocular video camera	not mention in the article	n
ProFORMA (Pan et al., 2009)	reconstruction of objects	uncut video sequence with static background	monolithic model of a object	a	monocular video camera on tripod and a rotatable object	hard – object must be sufficiently textured	r
Real-time urban 3D reconstruction (Pollefeys et al., 2008)	reconstruction of urban scenes	uncut video sequence with a fix orientation	dense 3D model of the scene	a	monocular multi-camera system	not mention in the article	r
User in the Loop for Image-Based Modeling (Kowdle et al., 2014)	model acquisition for objects	single photograph or collection of photograph	monolithic 3D model of the object	i	monocular camera	handled by the user	n
RGB-D modeling of indoor environments (Henry et al., 2014)	3D indoor mapping	uncut video sequence in RGB-D	3D map of the scene for, e.g. robot navigation	a	active stereo camera	handled by the D-channel of the active camera	*
123D CATCH (Autodesk, Inc., 2014)	model acquisition for objects	collection of photographs ($>$ 20 photos and $<$ 70 photos)	monolithic 3D model of the object	a	monocular camera, cloud computing	hard – photographs should not have occlusion, repetition, shininess etc.	n
Agisoft PhotoScan (Agisoft, 2014)	model acquisition for objects	collection of photographs ($>$ 10 photos and $<$ 70 photos)	monolithic 3D model of the object	a	monocular camera	hard – photographs should not have occlusion, repetition, shininess etc.	n
Photosynth (Microsoft Corporation, 2014)	browsing, exploring and organizing photo collections in a sparse 3D scene	collection of photographs ($>$ 10 photos)	sparse 3D model of the scene	a	monocular camera	hard – photographs should not have occlusion, repetition, shininess etc.	n
SketchUp Match Photo (Trimble Navigation Limited, 2014)	model acquisition for objects	single photograph or collection of photograph	3D model of the object	m	monocular camera	handled by the user	r

[1] execution mode: **a**utomatic, **i**nteractive, **m**anual; [2] Initialization: **r**equired, **n**ot required; *depends on the RGB-D camera

The first finding of the comparison is that there is a very strong correlation between the input and the execution mode of techniques. For a huge collection of photographs or video footage without many outliers, available automatic reconstruction seems to work quite well. In all other cases, interactive or manual methods perform better.

Almost all techniques require specific collections. The collections must be well planned photographs or videos that show the object of interest in a 360° view. Many other required properties, like an overlap by 50% of corresponding photographs, are described in the guidelines of each method. According to the guidelines, all automatic and almost all other techniques have problems with, reflection, occlusion and repeating structures. The output of all techniques is a monolithic model without any declaration of subparts. In most cases, the monolithic model is no more than a meshed dense point cloud. Such a model cannot be used in CAD applications, e.g., for simulation tasks, reverse engineering or creation of functioning replicas. Further information, e.g., if a technique needs special hardware equipment, can be found in Table 1.

4 PERSPECTIVE

After the direct comparison, drawbacks and missing elements of existing 3D modeling techniques can be spotted. It becomes obvious, that due to missing information like occlusions, automatic 3D reconstruction can only achieve acceptable results if enough photographs of the scene or knowledge of the underlying model are available. As pointed out in (Kowdle et al., 2014; van den Hengel et al., 2007; Debevec et al., 1996) interactive methods include real world model knowledge of the user into the reconstruction process. This enables a reconstruction from a few or even from a single photograph. Currently, the necessary real world knowledge cannot be explicitly integrated into existing algorithms, due to the sheer complexity it would create. Interactive modeling techniques, however, are the only way to reliably reconstruct 3D models from any collection of photographs or video footage. 3D reconstruction from arbitrary and unplanned collections will broaden the range of applications, examples are reverse engineering of mechanical parts, animal reconstruction for biology (e.g., for arachnology) and historical urban reconstruction.

Figure 1 outlines our ideas for an interactive 3D reconstruction architecture. It consists of three main parts: i) the input data in the form of monocular or stereo photographs / video footage, as well as additional data like physical interrelationships, ii) the interactive reconstruction process and iii) the user as domain expert of the real world. The interactive reconstruction process should join the computational power of today's computers with the conceptual knowledge of the user to solve issues that are computationally unfeasible up to now. In contrast to most interactive 3D reconstruction methods, the computer remains the "work horse" of the process, while supervised and active learning algorithms shift the load from the user to the computer.

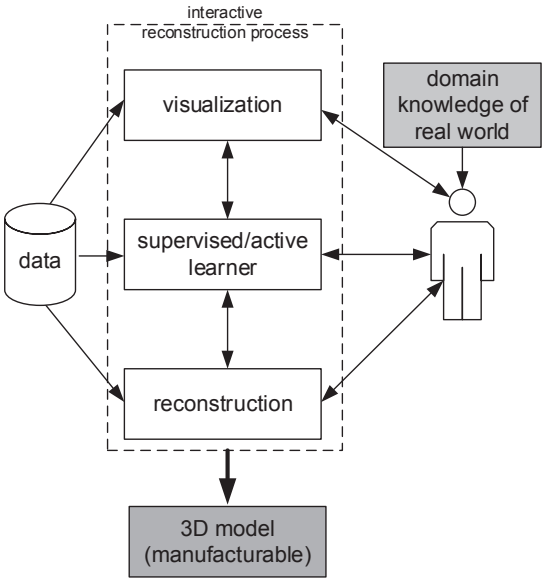

Figure 1: Interactive reconstruction architecture.

An application based on this architecture should be able to reconstruct from any collections of photographs and video footage or archive photos because missing information will be added by user knowledge. To achieve a better understanding of already existing techniques, algorithms, methods and processing pipelines, we describe a possible application based on the interactive architecture. Based on a photograph, a collection of photographs, a video or a collection of video, the user starts the interactive reconstruction process. In the first step, a monolithic model of the object or scene of interest should be created. Therefore, the user has to identify the object of interest with a marker. It should not be necessary to mark or outline the objects in every frame as it is common practice to date (e.g. (Agisoft, 2014)). The setting of a marker triggers the automatic calculation of a point cloud model in real-time. If the user recognizes problems in the point cloud, the user can exclude problematic items, such as reflections, with another marker from the automatic process, or directly modify the point cloud or the meshed point cloud. This direct modification allows the user to add occluded information to the 3D model or to delete projection errors

in the 3D model. Once a monolithic 3D model is reconstructed with a desired level of detail, the next step is to break down the model to its components or subparts until every subpart itself is monolithic. For breaking the monolithic 3D model into its subparts, the user roughly scribbles each subpart on the input data or uses common 3D breaking down techniques like cut-planes directly on the model. If all subparts have been identified, the user has to model the connections between all parts. In addition to pre-defined types of connections like glueing or screwing, this step has to account for moveable connections, such as a ball joint, where the user adds specific information like rotation axes, maximum angles etc. In the next step, the user assigns a material to each subpart, and an automatic consistency check should be included to ensure the compatibility of connection types and materials. The last step is the export of this model to common CAD format like *.dxf.

Such an application enables full CAD from reconstructed 3D models. This kind of CAD-ready 3D reconstruction can be used for simulation, reverse engineering, modeling, inverse modeling, testing, labeling and analysis tasks. The ultimate goal is that architects and engineers will accept and call the models out of 3D reconstruction — a model.

5 CONCLUSION

The literature review of 3D reconstruction, comparison and the discussion of a possible application have shown that interactive 3D reconstruction is able to create CAD-ready model — not just dense or sparse models. We agree with (Kowdle et al., 2014; van den Hengel et al., 2007; Debevec et al., 1996) that a fully automatic reconstruction for high quality object creation is currently not feasible. To show the ability of interactive 3D reconstruction we started to implement the propsed methods. We expect the identification of even more weaknesses of current 3D reconstruction and computer vision methods in the course of research conducted in the proposed directions. Finally, we are optimistic that an interactive 3D reconstruction tool could create models of real world objects which can be "translated" back to the real world with 3D printers and CNC-machines or can be used for many CAD tasks.

REFERENCES

Agisoft (2014). Agisoft photoscan http://www.agisoft.ru/.

Autodesk, Inc. (2014). Autodesk 123d catch | 3d model from photos http://www.123dapp.com/catch/.

Debevec, P. E., Taylor, C. J., and Malik, J. (1996). Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques - SIGGRAPH '96*.

Fulton, J. and Fraser, C. (2009). Automated reconstruction of buildings using a hand held video camera. In *Innovations in Remote Sensing and Photogrammetry*, pages 393–404. Springer.

Furukawa, Y., Curless, B., Seitz, S. M., and Szeliski, R. (2009). Reconstructing building interiors from images. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 80–87. IEEE.

Henry, P., Krainin, M., Herbst, E., Ren, X., and Fox, D. (2014). Rgb-d mapping: Using depth cameras for dense 3d modeling of indoor environments. In *Experimental Robotics*, pages 477–491. Springer.

Kowdle, A., Chang, Y.-J., Gallagher, A., Batra, D., and Chen, T. (2014). Putting the user in the loop for image-based modeling. *International Journal of Computer Vision*, 108(1-2):30–48.

Microsoft Corporation (2014). Photosynth - capture your world in 3d https://photosynth.net/.

Pan, Q., Reitmayr, G., and Drummond, T. (2009). Proforma: Probabilistic feature-based on-line rapid model acquisition. *Procedings of the British Machine Vision Conference 2009*.

Pollefeys, M., Nistér, D., Frahm, J.-M., Akbarzadeh, A., Mordohai, P., Clipp, B., Engels, C., Gallup, D., Kim, S.-J., Merrell, P., and et al. (2008). Detailed real-time urban 3d reconstruction from video. *International Journal of Computer Vision*, 78(2-3):143–167.

Pollefeys, M., Van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., Tops, J., and Koch, R. (2004). Visual modeling with a hand-held camera. *International Journal of Computer Vision*, 59(3):207–232.

Snavely, N., Seitz, S. M., and Szeliski, R. (2006). Photo tourism: Exploring photo collections in 3d. *ACM Transactions on Graphics*, 25(3):835.

Tanskanen, P., Kolev, K., Meier, L., Camposeco, F., Saurer, O., and Pollefeys, M. (2013). Live metric 3d reconstruction on mobile phones. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 65–72. IEEE.

Trimble Navigation Limited (2014). Match photo: Modeling from photos | sketchup knowledge base http://help.sketchup.com/en/article/94920.

van den Hengel, A., Dick, A., Thormählen, T., Ward, B., and Torr, P. H. S. (2007). Videotrace:rapid interactive scene modelling from video. *ACM Transactions on Graphics*, 26(3):86.

van den Hengel, A., Hill, R., Ward, B., and Dick, A. (2009). In situ image-based modeling. *2009 IEEE International Symposium on Mixed and Augmented Reality*.

Zollhöfer, M., Theobalt, C., Stamminger, M., Niener, M., Izadi, S., Rehmann, C., Zach, C., Fisher, M., Wu, C., Fitzgibbon, A., and et al. (2014). Real-time non-rigid reconstruction using an rgb-d camera. *ACM Transactions on Graphics*, 33(4):1–12.