EVALUATION OF 3D RECONSTRCUTION USING MULTIVIEW BACKPROJECTION

K. Mueller[1], X. Zabulis[2], A. Smolic[1], and T. Wiegand[1]

[1]Fraunhofer Institute for Telecommunications,
Heinrich-Hertz-Institut,
Germany

[2]Informatics and Telematics Institute,
Centre for Research and Technology,
Greece

ABSTRACT

This paper evaluates the final reconstruction quality of 3D objects from different reconstruction methods by comparing rendered views of a 3D model to the original views, initially taken from 2D cameras. The paper uses pixel-by-pixel error measures, like pixelwise reconstruction error for non-textured objects and PSNR values for colored or textured objects. Concurrently, the limitations of such measures in connection with 3D reconstruction evaluation are highlighted and a reconstruction measurement based on differential values is investigated, where deviations from reference values are analyzed instead of absolute PSNR-values.

1. INTRODUCTION

Reconstructed 3D models are often presented in special rendering tools to provide free navigation for the user while watching the scene. The 3D models often stem from different reconstruction approaches, e.g. colored voxel reconstruction or view-dependent multi-texture synthesis with synthetic geometry. However, in all approaches, considered in this paper, a real-world scene is initially captured by a number of cameras. To analyze the reconstruction quality in a specific approach, only these original images or sequences are available for reconstruction evaluation. Therefore, these images need to be considered in reconstruction error analysis.

In the reconstruction chain, a number of different errors are induced, starting from erroneous calibration and segmentation information. Furthermore, reconstruction errors are caused by approximation steps in the 3D geometry creation process. Finally, the 3D content is projected on the 2D render plane. The renderer itself applies special filters for object scaling during visualization. Overall, a number of different error sources influence the final 3D object that is displayed to the user. These sources have to be considered when reconstruction evaluation is to be applied.

Therefore, the paper starts by reviewing common reconstruction methods in chapter 2, discusses on the applied multi-view back projection evaluation methods in section 3 and shows some results for 3D geometry reconstruction and textured object evaluation in section 4, and finally concluding the paper.

This work is supported by EC within FP6 under Grant 511568 with the acronym 3DTV.

2. 3D RECONSTRUCTION METHODS

In contrast to pure virtually generated objects, the creation of 3D video objects from natural scenes starts with a number of more or less overlapping camera images or sequences. The number of cameras is highly specific to the appropriate scene, but strongly influences the type of 3D scene reconstruction that can be applied. For our evaluation process we consider a tradeoff between reconstruction quality and cost by limiting the number of cameras. In exchange we need to apply 3D geometry reconstruction to compensate for the sparser setting.

To obtain a relation between real 3D world and 2D images, a classical calibration procedure has to be applied, which delivers deriving intrinsic and extrinsic parameters for all cameras [10]. The second preprocessing step is the object segmentation to separate the 3D video objects from background information for the geometry reconstruction process. Segmentation in multi-camera scenarios again highly depends on the type of scene content. One possibility is to use Kalman filter formalism, which track the desired objects over time in all views [5]. Other segmentation methods use objects relations from the different views by applying structure-from-multiview approaches. Such methods are related to structure-from-motion methods, but use disparity vectors or depth cues instead of motion vector information to obtain the required segmentation of 3D scene objects [13].

2.1. Textured Voxel Reconstruction

Space-carving [3] and voxel coloring [7] approaches provide reconstructions only in terms of determining if a voxel is occupied or not. They do not provide surface normal information, which can boost the rendering quality. Shortcomings of such approaches, are that radiometric calibration is difficult to achieve and retain and that correlation metrics are more powerful matching operators, because they account for local order of pixels. Besides the traditional, epipolar and correlation based approach to stereo a number of works compensate for the

projective distortion in the matching process (or otherwise, match the textures in 3D), to obtain more matches and accuracy [1][2][6][12]. Treating the imaged surfaces as locally planar allows the back projections of images at hypothetical planar patches and, in turn, the prediction that back projections should match if the patch coincides with the surface. A match then provides the estimations of locus and orientation of the surface. Such correlation-based volumetric approaches are reviewed in more depth below along with some details on the specific approach that was adopted for the voxel-based reconstructions, in this paper.

An operator is utilized that yields a measure of the confidence about the occupancy of a voxel in 3D space given a strongly calibrated image pair (I_1, I_2). It is applied at world point p and outputs a confidence score $s(p)$ and a 3D unit normal $\boldsymbol{\kappa}(p)$.

Let a planar surface patch S, which size is $\alpha \times \alpha$ units of length (*mm*), centered at p, with unit normal $\boldsymbol{n}$. Back-projecting I_1 and I_2 onto S yields two collineation images $w_1(p, \boldsymbol{n})$ and $w_2(p, \boldsymbol{n})$. Their formation rule is:

$$I_i(P_i \cdot (p + R(\boldsymbol{n}) \cdot [\Delta x\ \Delta y\ 0]^T)),$$

where P_i is the projection matrix of camera i, $R(\boldsymbol{n})$ is a 3×3 rotation matrix that maps $[0\ 0\ 1]^T$ to $\boldsymbol{n}$, and Δx, Δy *are in* $[-\alpha/2, \alpha/2]$ are local horizontal and vertical coordinates of a point on the patch. To obtain $s(p)$ and $\boldsymbol{\kappa}(p)$, an $r\times r$ point lattice is assumed on S and the correlation of $w_1(p, \boldsymbol{n})$ and $w_2(p, \boldsymbol{n})$, $Corr(w_1(p, \boldsymbol{n}), w_2(p, \boldsymbol{n}))$ is optimized as:

$$s(p) = \max_n (Corr(w_1(p, \boldsymbol{n}), w_2(p, \boldsymbol{n}))),$$
$$\boldsymbol{\kappa}(p) = argmax_{\boldsymbol{n}} (Corr(w_1(p, \boldsymbol{n}), w_2(p, \boldsymbol{n}))) .$$

Scalar $s(p)$ and vector $\boldsymbol{\kappa}(p)$ are then combined into vector $\boldsymbol{v}(p) = s(p)\_\kappa(p)$. It is assumed that I_1 and I_2 image Lambertian and textured surfaces as well as that world surfaces are locally planar.

Reconstruction of surfaces is performed by detecting the strong local maxima of $|\boldsymbol{V}(p)|$. The corresponding voxels are considered occupied, with normals: $\boldsymbol{\kappa}_{\beta(p)}(p)$.

2.2. Textured 3D Wireframe

If mask or silhouette information in provided, shape-from-silhouette approaches can be used to reconstruct 3D geometry information. Here, a hierarchical octree-based approach is used to reconstruct a high-resolution voxel model [8]. Than a wireframe transformation is carried out, starting with a marching cubes algorithm [4], followed by first order neighborhood smoothing [9] and surface primitive reduction to obtain the low resolution wireframe model. The associated geometry evaluation of these stages is shown in chapter 3.1.

For photo-realistic rendering, texture information is mapped onto the reconstructed geometry. Depending on the refection properties and lighting conditions, natural surfaces may change appearance, when navigating through the scene. Therefore, view-dependent multi-texturing is used, where a number of original images or videos are mapped onto the 3D geometry and weighted according to the currently selected viewpoint. The initial parameter for a specific texture weight calculation is the cosine term between viewing vector and associated camera direction vector. The cosine term is than transformed into a normalized texture weight to provide original textures only close to original viewpoints, while guaranteeing smooth interpolation weight transition while navigating through intermediate viewpoints [11].

3. EVALUATION METHODOLOGY

For evaluation, view back projection of the 3D model towards each original view was utilized in this paper. Then a 2D pixel-based error measure is used for comparison, depending on the type of 3D objects under investigation: recontruction error and PSNR for non-textured and textured models respectively.

One important aspect for 3D reconstruction is the number and position of cameras that are used. For 3D geometry reconstruction, the importance of each camera can be evaluated by reconstructing the object without the respective camera and than back projecting and evaluating the object. Here, n-fold cross validation is the method of choice, which is often used in classification processes to determine the predictability of a data set, if a certain amount of test data is missing. Starting from a total dataset of M, the dataset is split into n subsets M_i, $i = 1\ldots n$. The common procedure in n-fold cross validation is than to define the training sets $T_i = M\text{-}M_i$, predict the missing subset M_i and evaluate the error between prediction and subset for all i. This process can also be applied vice versa by using only a certain subset M_i and predicting all other subsets $\{M_k \mid k=1\ldots n,\ k\neq i\}$. This approach is selected for textured model evaluation, where the 3D model is reconstructed with a subset of original camera views and than compared to all original views by back projection.

3. RESULTS

The processing chain in 3D reconstruction consists of a number of single steps, which are investigated in the following subsections by applying multi-view back projection in the form of the described n-fold cross validation with underlying 2D error measures, i.e. pixel-wise reconstruction errors for volume evaluation and differential PSNR analysis for textured models.

3.1. Geometry Reconstruction Evaluation

The first reconstruction step analyses the 3D geometry reconstruction from 2D silhouette information. As described above, a hierarchical octree model is reconstructed first and than transformed into a low-resolution wireframe. The intermediate 3D geometry models for the synthetic sequence are shown in Fig. 1. These models are back projected into the silhouette images to

evaluate the contour deviation between the 3D model and all its 2D views.

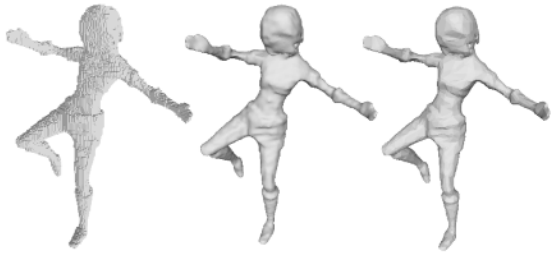

Fig. 1: Voxel model, high-resolution and reduced wireframes for “Lara”-sequence.

For the silhouette evaluation, the number of incorrectly projected pixel is compared to the number of correctly projected pixels in all views. Finally, the views are averaged to obtain the final value for the reconstruction error. The reconstruction error is shown in Fig. 2, together with the transformation error.

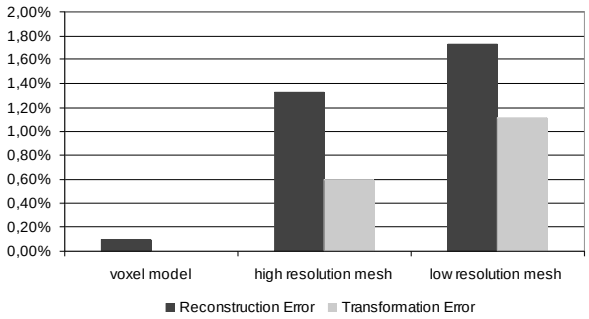

Fig. 2: Back projection errors for synthetic reconstruction models.

This second error measure analyses, how well the transformed models in Fig. 1 middle and right, approximate the high-resolution voxel model. The reconstruction error for the voxel model is nearly negligible at 0.09%, showing, how well the model approximates all silhouettes. For the high- and low-resolution wireframe, the reconstruction error increases to 1.33% and 1.73% respectively, which again is rather low. The similarity between wireframe models and voxel reconstruction is also represented by the transformation error. This error measures the deviation of a model in comparison to the voxel reconstruction. Therefore, the voxel model itself has no errors, while the transformation error increases from 0.60% to 1.11% for the high- and low-resolution wireframes. For the geometry reconstruction, the transformation steps from voxel model to a low-resolution wireframe model introduces only very small errors and therefore the 3D geometry represents very well the original 2D silhouette information.

The second example for voxel model transformation is shown in Fig. 3. This model originates from the real-word “Doo Young”-sequence.

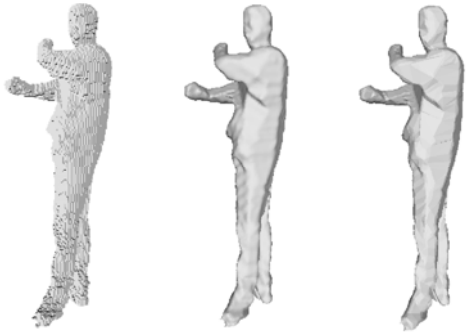

Fig. 3: Voxel model, high-resolution and redudeced wireframes for “Doo Young”-sequence.

In contrast to the synthetic sequence, the calibration and segmentation information are distorted to a certain degree, which leads to higher overall reconstruction errors, as shown in Fig. 4.

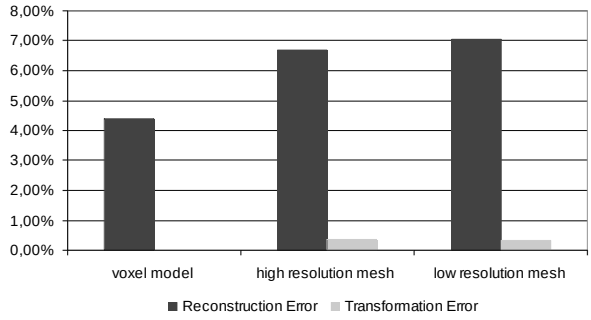

Fig. 4: Back projection errors for “Doo Young” reconstruction models.

Here, the reconstruction error for the voxel model starts at 4.37% and increases to 6.67% and 7.07% for the wireframe models. For this model, the multi-view back projection of each 3D model misses the silhouette information in some views For the transformation error, however, the values are 0.36% and 0.31% for both wireframe resolutions, indicating a very low distortion in comparison to the voxel model.

3.2. Evaluation of Textured Voxel Models

If a complete 3D wireframe reconstruction cannot be applied due to special camera settings, e.g. narrow baseline setups, a voxel model is reconstructed only for certain object parts and finally textured. In this case, the reconstruction quality strongly depends on the voxel size, used in the reconstruction. Thus, a larger voxel size also means a lower voxel resolution with less voxels. For the quality evaluation, different voxel sizes were selected for geometry reconstruction. Afterwards, texture information was taken from the first camera to reconstruct all views. Furthermore, the process was repeated for all cameras, as indicated by the *n*-fold cross validation.

The example in Fig. 5 shows the reconstructed and textured voxel models for view 1 of the trinocular sequence with no interpolation together with the associated amplified difference images at three different voxel sizes.

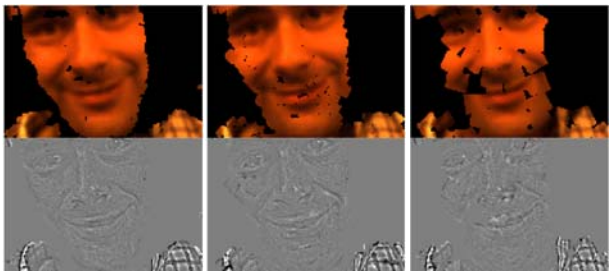

Fig. 5: Textured voxel model and associated difference images x4 with 5, 10 and 20 mm voxel resolution, trinocular sequence.

The original views in Fig. 5 indicate, that the completeness of the 3D model decreases for larger voxel sizes, leaving more holes for this type of reconstruction. In the difference images, the holes have been neglected to better analyze the differences on the voxel surfaces. Here, the difference images show more errors for larger voxel sizes due to worse approximation of the original 3D surfaces

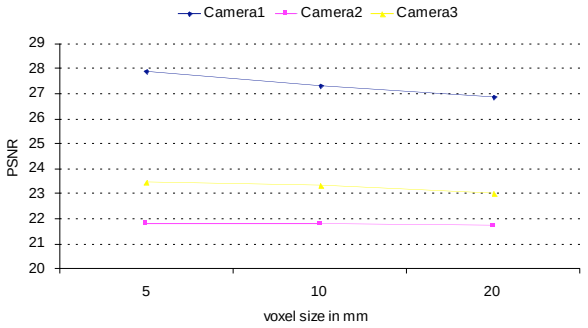

Fig. 6: Back projection PSNR for 3 camera views and different voxel sizes, trinocular sequence.

The corresponding PSNR values are shown in Fig. 6 for camera 1. With increasing voxel size from 5, 10 to 20mm, the PSNR value drops from 27.9 to 27.4 and 26.9. Again, for the PSNR calculation, background pixel and therefore object holes have been omitted to investigate the pure texture distortion of the object surface. In the diagram in Fig. 6, also the curves for camera 2 and 3 are shown, which were reconstructed with the texture information from camera 1. Therefore the total PSNR values are lower, since in addition to the surface approximation error caused by different voxel sizes, also texture reprojection errors occur. Plane 2D textures from one view are mapped into another view, causing a number of distortions, including stretching and occlusion artifacts in areas that were visible in the view of texture origin, but not in the destination view or vice versa.

However, considering the change in PSNR, the values drop from 21.8dB to 21.7dB for camera 2 and 23.5dB to 23.0dB for camera 3 by going from the higher voxel resolution with 5mm size to lowest with 20mm.

4. SUMMARY

In this paper, we have shown an evaluation procedure for two different 3D reconstruction methods from real world 2D camera view, using multi-view back projection. To evaluate the reconstruction quality, a back projection method has been introduced to compare the 3D model to 2D original camera views. To guarantee a complete evaluation, *n*-fold cross validation is applied, together with the underlying 2D error measures: reconstruction error and PSNR. The latter was analyzed in a differential way, since the absolute values don't consider the specific properties of projected textures in 3D. The results show, that hierarchical voxel reconstruction can be combined with wireframe transformation and surface primitive reduction, introducing only very small errors. For the textured voxel method, the voxel size and resolution not only influence the surface completeness, but also the texture approximation onto the original views.

REFERENCES

[1] A. Bowen et al., "Light field reconstruction using a planar patch model." in *SCIA*, 2005, pp. 85–94.

[2] O. Faugeras and R. Keriven, "Complete dense stereovision using level set methods," *Proc. ECCV 98*, 1998, vol. 1, pp. 379–393.

[3] K. N. Kutulakos and S.M. Seitz. A theory of shape by space carving. *IJCV*, vol. 38, no. 3, pp. 197–216, 2000.

[4] W. E. Lorensen, and H. E. Cline, "Marching Cubes: A high resolution 3D surface reconstruction algorithm," *Proc. SIGGRAPH*, vol. 21, no. 4, pp 163-169, 1987.

[5] K. Mueller et al., "Multi-Texture Modelling of 3D Traffic Scenes," Proc. ICME, Baltimore, MD, USA, July 6.-9. 2003.

[6] A. S. Ogale and Y. Aloimonos, "Stereo correspondence with slanted surfaces: critical implications of horizontal slant," in *Proc. CVPR04*, 2004, vol. 1, pp. 568–573.

[7] S. M. Seitz, C. R. Dyer, "Photorealistic Scene Reconstruction by Voxel Coloring", *Proc. Computer Vision and Pattern Recognition Conf.*, pp. 1067-1073, 1997.

[8] A. Smolic et al., "Free Viewpoint Video Extraction, Representation, Coding, and Rendering", *Proc. ICIP*, Oct. 24.-27. 2004.

[9] G. Taubin, „Curve and Surface Smoothing Without Shrinkage", *International Conference on Computer Vision (ICCV '95)*, pp. 852-857, 1995.

[10] R.Y. Tsai, "A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV camera and lenses," *IEEE Journal of Robotics and Automation*, Vol. RA-3, No. 4, August 1987.

[11] D. Vlasic et al., "Opacity Light Fields: Interactive Rendering of Surface Light Fields with View-dependent Opacity", *Proc. 2003 Symposium on Interactive 3D graphics*, pp. 65-74, 2003.

[12] X. Zabulis and K. Daniilidis, "Multi-camera reconstruction based on surface normal estimation and best viewpoint selection," in *Proc. of IEEE 3DPVT*, 2004, pp. 733–40.

[13] D. S. Zhang, G. Lu. "Segmentation of Moving Objects in Image Sequence: A Review", *Circuits, Systems and Signal Processing (Special Issue on Multimedia Communication Services)*, vol. 20, no. 2, pp. 143-183, 2001.