An Effective One-shot Body Part Multi-View Reconstruction Device with Self-calibration Capabilities

Matteo Bonotto[1][2], Daniele Evangelista[2], Marco Imperoli[2], Alberto Pretto[1]

[1] Department of Information Engineering, University of Padova, Padova, Italy -
matteo.bonotto.2@phd.unipd.it, alberto.pretto@dei.unipd.it
[2] FlexSight Srl, Padova, Italy -
(matteo.bonotto, daniele.evangelista, marco.imperoli)@flexsight.eu

Technical Commission II

Keywords: Auto-Calibration, Multi-View Stereo, Active Stereo, Point Cloud Processing

Abstract

This paper introduces a custom-built low-cost camera ring device designed for automatic cast synthesis, able to accurately and instantly scan body parts. The scanned mesh will be used as a backbone model for the cast design and 3D printing. The system is based on the multi-view active stereo principle and it is composed of a circular array of 16 synchronized cameras (Fig. 1) and 4 equally distributed IR pseudo-random laser pattern projectors. We employ a custom multi-view stereo reconstruction pipeline based on (Schönberger et al., 2016), which guarantees optimal results without the downsides of the supervised data-driven multi-view stereo algorithms, i.e. data collection and ground truth labeling. Additionally, inspired by (Duda and Frese, 2018), we propose a novel, automated calibration system to extract intrinsic and extrinsic camera parameters which are required to perform robust multi-view stereo reconstructions.

1. Introduction

The traditional process of making customized casts is a time-consuming process that requires, in addition to the collection of functional data itself, the presence of a professional figure in charge of designing and making the final product. Custom 3D printed casts based on a scan of the body part to be treated can overcome these limitations. The developed system aims to obtain these scans by a one-shot acquisition processed by a multi-view stereo reconstruction pipeline. An instant acquisition allows to reduce the stress of the acquisition process on the patient while avoiding the presence of artifacts on the reconstruction due to the temporal desynchronization of the images. Extracting 3D models of human limbs, however, represents a challenging task for classical MVS approaches based on Patch Match Stereo (Bleyer et al., 2011). Due to the repetitiveness of the skin texture, measuring the visual similarity using a photometric measurement like normalized cross correlation (NCC) could lead to wrong estimations. In order to alleviate the problem we make use of 4 equally distributed IR pseudo-random laser pattern projectors and propose simple yet effective changes on COLMAP framework (Schönberger et al., 2016) to address the challenge represented by low textured areas. To successfully perform the 3D reconstruction the cameras must be calibrated, i.e. both intrinsic and extrinsic parameters have to be estimated. Traditional calibration systems require a human operator to move around in the 3D space an object with a well-known geometrical pattern, e.g. a checkerboard. Our system allows to perform calibration in an automated way, minimizing the errors introduced by human intervention. After the acquisition phase, the computation is performed by a consumer-grade external compute unit equipped with an Intel Core i7-12700KF, 16 GB of RAM and an Nvidia 2060 capable of extracting a low-resolution mesh within 15 seconds and the final high-resolution output within 40 seconds with sub-millimeter accuracy.

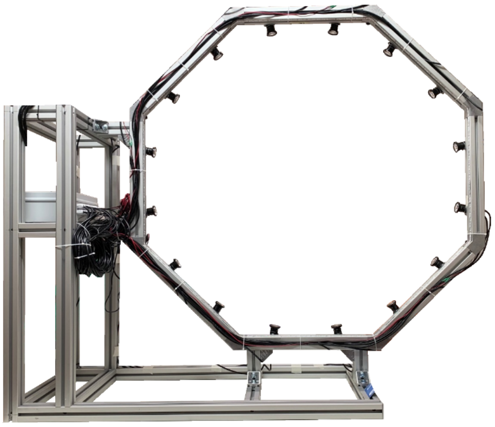

Figure 1. The custom-built low cost 3D camera ring device.

2. Related Work

The proposed system is related to prior work in multi-view stereo 3D reconstruction and camera calibration.

2.1 Multi-view Stereo

In Multi-View Stereo (MVS) the desired 3D model is computed starting from a set of images and their corresponding camera extrinsics and intrinsics. Differently from Structure from Motion (SfM) (Schönberger and Frahm, 2016, Lindenberger et al., 2021) where camera instrinsics and extrinsics estimation is performed along with 3D reconstruction, MVS can achieve better accuracy as the calibration process, performed beforehand, usually exploits regular structures to perform a better estimation of the necessary parameters. In (Seitz et al., 2006) MVS methods are grouped into four main categories. The first class of algorithms (Seitz and Dyer, 1997, Treuille et al., 2004, Ji et

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-2-W7-2024-33-2024 | © Author(s) 2024. CC BY 4.0 License.

al., 2017) first compute a cost function on a 3D volume to extract the desired surface from it. The second class (Fromherz and Bichsel, 1995, Kutulakos and Seitz, 1999) iteratively evolve the hypothesis surface. The third class (Szeliski, 1999, Zheng et al., 2014, Yao et al., 2018) outputs for each for each image on the dataset its corresponding depth map, usually exploiting consistency constraints between output depth maps. Finally in the last group (Faugeras et al., 1990, Taylor, 2003) are inserted algorithms which try to fit a surface over a set of feature previously extracted and matched. Modern classical MVS methods typically use patch match with photometric consistency (Zheng et al., 2014, Schönberger et al., 2016, Romanoni and Matteucci, 2019, Xu and Tao, 2019, Wang et al., 2023). Belonging to the third class, input images consists of one reference image for which the depth map is estimated and two or more source images.

One of the first data-driven MVS method exploiting an end-to-end convolutional neural network was presented in (Ji et al., 2017). Each input image is converted to a special 3D volume representation denoted as colored voxel cube (CVC) and used as input for the network. More recent methods usually extract dense image features from multiple views and backproject them into 3D volumes. In MVSNet (Yao et al., 2018) features are extracted to produce a matching cost volume. Following methods (Gu et al., 2020, Yao et al., 2019, Sayed et al., 2022) based on MVSNet reduced the GPU memory consumption due the employment of 3D convolutional layers while retaining the same level of performance or improving over it.

2.2 Camera Calibration

The calibration process refers to the estimation of the intrinsic (i.e. camera-specific) and extrinsic parameters of the rig. Intrinsic (or internal) parameters are represented by the focal length, the coordinates of the principal point, distortion parameters of each camera while extrinsic parameters identify the pose (i.e. rotation and translation) of the camera w.r.t. a fixed coordinate system. In Direct Linear Transform (DLT) methods (Abdel-Aziz and Karara, 1971, Heikkila and Silven, 1997) given at least 6 known control points in the 3D space and their relative projection the camera matrix P is computed by means of SVD decomposition. Tsai (Tsai, 1987) proposed a new algorithm for camera calibration named two stage method, in the first the rotation, position of the pattern and scale factor are computed. In the second stage, instead the internal parameters of the camera are estimated. In the seminal paper (Zhang, 2000) Zhang discovered that a camera could be calibrated by just showing a planar pattern, at a few different orientations to the camera. Previous available calibration techniques consisted of using precisely fabricated 3D objects with painted patterns which were expensive to make and not very practical to use.

3. Self Calibration Procedure

We estimate the intrinsic and extrinsic parameters in two consecutive steps.

3.1 Intrinsic parameters calibration

As most of the current calibration methods our calibration system is based on the following procedure:

1. Collect k images, each one framing a checkerboard with m internal corners from different point of views
2. List the m 3D corner positions $P_0, P_1, P_2, ..., P_{m-1}$ in the checkerboard reference frame
3. From each image i, extract the m 2D corners projections
4. Associate P_0 with $p_{i,0}$, P_1 with $p_{i,1}$, …, P_m with $p_{m,1}$
5. Initialize the calibration parameters by solving a linear system with k constraints given by the k estimated homographies projective mappings
6. Initialize the k checkerboard positions with the rotations and translations extracted from the (normalized) homographies
7. Find the intrinsic parameters set along with the camera positions with respect to the checkerboard that minimize the squared distances in the image space between extracted corners and corners projections, using conventional least square methods.

3.2 Extrinsic Parameters Calibration

Given the intrinsic camera parameters we fix the reference frame of the camera ring device so that it is coincident with the reference frame of a chosen camera, say the camera with ID 0. Therefore, we estimate the $n-1$ rigid body transformations T_i that relate the other $n-1$ cameras to the camera 0. To estimate $T_i, i = 1, \ldots, n-1$ we designed the following procedure:

1. For each camera i and for each view of the checkerboard, we solve a perspective-n-point (PNP) problem to estimate the relative rigid body transformation that relates the checkerboard with such a camera. We use here the intrinsic parameters estimated in the previous step
2. If two nearby cameras frame the checkerboard at the same time, given the transformations computed in point 1, it is possible to compute the relative rigid body transformation that relates such a couple of cameras, e.g., $T_{i,j}$
3. From the transformations $T_{i,j}$ it is possible to compute in a closed form an initial guess for $T_i = T_{0,1} \cdot T_{0,1} \cdot \ldots \cdot T_{i-1,i}$
4. Find the extrinsic parameters T_i that minimize the squared distances in the image space between extracted corners and corners projections, using conventional least square methods. We use here the intrinsic parameters estimated in the previous step

The method presented above allows for consistent and accurate calibrations. Unfortunately, the procedure for acquiring the calibration dataset is manual, time-consuming, and must be performed by an experienced operator who moves the checkerboard to a set of suitable and fixed positions To perform automated calibration we employ a multi-board pattern (Fig. 2) that allows maximizing the distribution of corners over the image and that, at the same time, can be moved in a convenient and automatic way to obtain an adequate distribution of checkerboards positions. It allows to retain all the benefits of the methods employing planar surfaces in contrast to 3D calibration objects, i.e a relatively cheaper calibration apparatus and a simple setup. Additionally the conformation of the multi-board pattern, the number of rotations and the axis of rotation can be easily adjusted to better fit other camera configurations. Our multi-board model is made up of two groups of 3 opposite boards. In each group, two of the three boards are tilted vertically (Fig. 2).

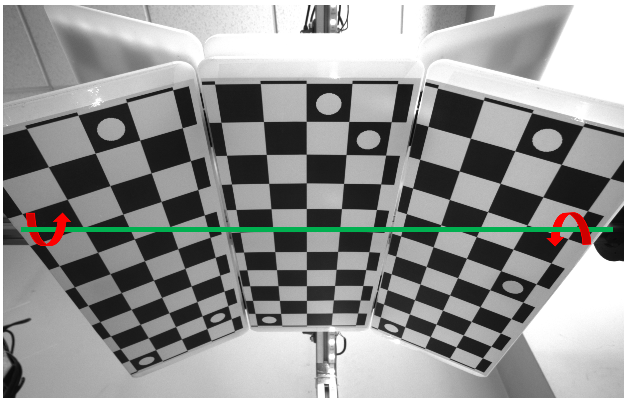

Figure 2. Calibration pattern. The pattern rotates along the highlighted axis.

The 6 tables are installed on a rotating pin, so as to generate different horizontal inclinations by rotating the pin. In order to distinguish the checkerboards, each board is characterized by a unique pattern of white circles drawn inside specific black squares. We commanded the system to perform a sequence of synchronized acquisitions rotating the movable pattern of a 12 degrees steps, until a total rotation of 360 degrees was obtained. The corners extraction algorithms have been modified accordingly to detect more than one checkerboard per image and to recognize the unique pattern of white circles. To minimize the changes to the default procedure and hence to keep unchanged the backbone of the method the new dataset has been loaded in such a way as to simulate the presence of a single checkerboard. Multiple checkerboards in single images were simply considered as a single checkerboard framed in different positions (and times).

4. COLMAP framework overview

We'll briefly review in this section the COLMAP framework (Schönberger et al., 2016) which our method is based on. Given as input a reference image X^{ref} and a set of source images $X^{src} = \{X^m | m = 1...M\}$, the framework estimates for each pixel l the depth θ_l, the normal n_l and binary variables $Z_l^m \in \{0, 1\}$ which indicates if the pixel l is occluded in image m. The joint likelihood function $P(X, Z, \Theta, N)$ is accordingly defined as:

$$\begin{aligned}\prod_{l=1}^{L}\prod_{m=1}^{M} & [P(Z_{l,t}^m | Z_{l-1,t}^m, Z_{l,t-1}^m) \\ & \cdot P(X_l^m | Z_l^m, \theta_l, n_l) \\ & \cdot P(\theta_l, n_l | \theta_l^m, n_l^m)]\end{aligned} \tag{1}$$

The smooth term $P(Z_{l,t}^m | Z_{l-1,t}^m, Z_{l,t-1}^m)$ enforces spatially smooth output occlusion maps. The second term represents the photometric consistency of the patch X^m, expressed as: where the color similarity ρ_l^m is computed using a bilateral NCC and a slanted plane-induced homography The last term $P(\theta_l, n_l | \theta_l^m, n_l^m)$ encourages consistency among the output depth and normal maps of different views. The equation 1, being intractable is approximated and plugged inside a variant (GEM) of the generalized Expectation-maximization (EM) algorithm. The real posterior $P(Z, \Theta, N | X)$ is factorized in its approximation $q(Z, \Theta, N) = q(Z)q(\Theta, N)$. During the iteration t of the E step of GEM the term $q(Z^m, l, t)$ is inferred, while the values (Θ, N) are kept fixed. In the M step, instead, PatchMatch propagation and sampling are used to choose the optimal depth and normal. From the set of hypothesis:

$$\begin{aligned}\{&(\theta_l, n_l), (\theta_{l-1}^{prp}, n_{l-1}), (\theta_l^{rnd}, n_l), (\theta_l, n_l^{rnd}), \\ &(\theta^{rnd}, n^{rnd}), (\theta_l^{prt}, n_l), (\theta_l, n_l^{prt})\}\end{aligned} \tag{2}$$

where θ_l^{rnd} is a randomly generated depth value and θ_l^{prt} is the perturbated depth value, the tuple (θ, n) satisfying:

$$(\theta^{opt}, n^{opt}) = \operatorname*{argmin}_{\theta_l^*, n_l^*} \frac{1}{|S|} \sum_{m \in S} (1 - \rho_l^m (\theta_l^*, n_l^*)) \tag{3}$$

is selected for the pixel l.

5. Multi-view Stereo 3D Reconstruction Pipeline

Figure 3 shows the main elements of the proposed multi-view reconstruction pipeline. The point cloud estimation is the main step of the proposed 3D reconstruction pipeline. The core algorithm is based on the multi-view stereo (MVS) method proposed in (Schönberger et al., 2016). A depth map and a normal map are computed for each camera view by using the multi-view patch matching approach (Bleyer et al., 2011). The estimated depth and normal maps are then fused together for final point cloud computation using the graph-based technique used in (Schönberger et al., 2016). As depicted in Fig. 3, we extend the COLMAP framework by performing several MVS estimations considering input scaled at multiple resolutions. We define N as the total number of scales, then the spatial resolution of the image at the k^{th} stage is defined as $\frac{W}{2^{N-k}} \times \frac{H}{2^{N-k}}$. The idea is that from lower-resolution images we can estimate the rough 3D structure while being robust to noise; from higher-resolution images, instead, we can focus on details improving the overall quality of the reconstruction. We run the COLMAP framework starting from the low resolution images. The output depth and normal maps are used to initialize the depth and normal hypothesis for the higher resolution images. Additionally this procedure allow to make the method converge faster. While the number of processed images increases, depth and normal maps from lower resolution samples are generated faster while upsampling the results of the previous scale k^{i-1} to initialize the hypothesis of the current scale k makes the method converge much faster achieving an overall speedup. Exploiting the fact that object must be inside the operative region of the camera ring, to further reduce the computational time required to obtain the final outpoint point cloud, we use the meshes obtained from the low resolution scale to generate a bounding box around the object for each image. After generating the mesh at scale k^0 we backproject its silhouette for each input image. The image is cropped to the bounding box containing the generated 2D mask and the corresponding camera parameters are modified accordingly. The generated 3D point clouds (one for each resolution) are then fused together by using a custom Clustering Filter, which removes redundant points on lower resolution point clouds.

The output from the point cloud estimation step is a high resolution dense point cloud that is typically affected by noise and outliers (Fig. 4-left). In order to improve the quality of the reconstruction and to reach sub-millimeter accuracy, in the post-processing step, the estimated point cloud is processed by exploiting a Statistical Outlier Removal filter (Fig. 4-middle).

Once most of the outliers are removed, the 3D mesh can be computed from the point cloud by using the Screened Poisson Surface Reconstruction algorithm (Kazhdan and Hoppe, 2013). In order to further reduce noise in the estimation, a Laplacian

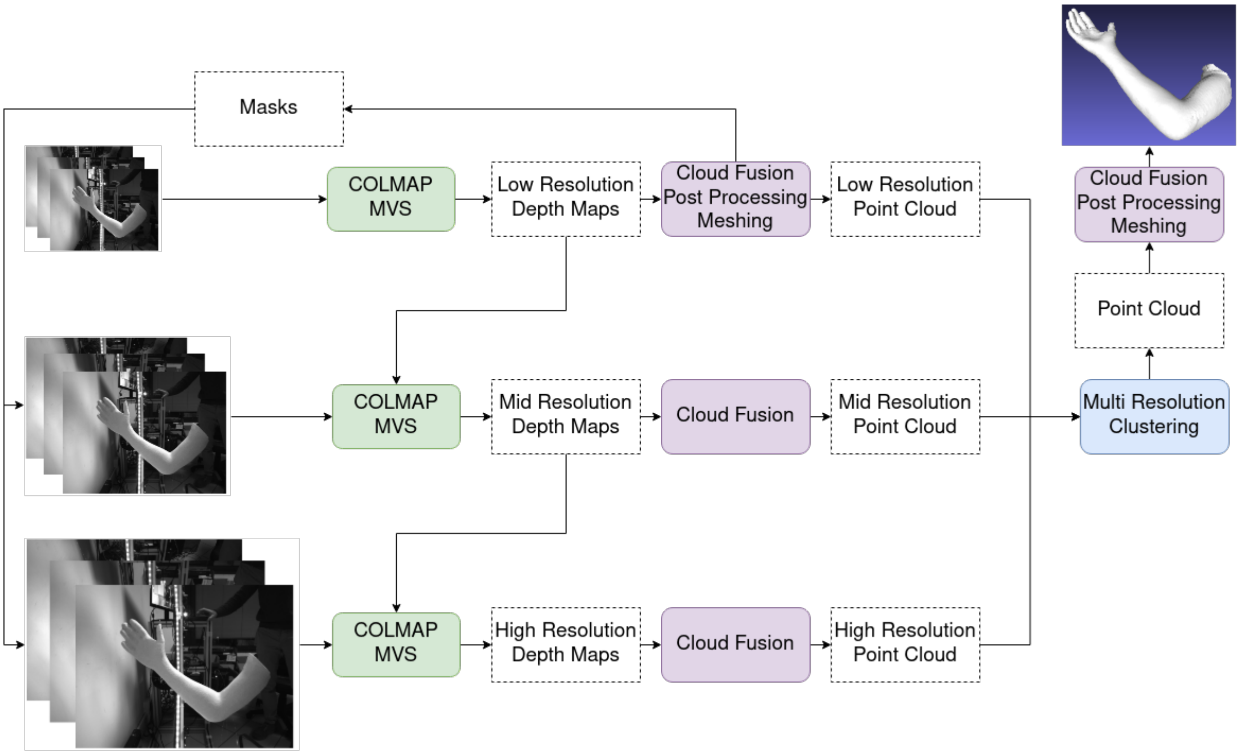
Masks
COLMAP MVS
Low Resolution Depth Maps
Cloud Fusion Post Processing Meshing
Low Resolution Point Cloud
COLMAP MVS
Mid Resolution Depth Maps
Cloud Fusion
Mid Resolution Point Cloud
COLMAP MVS
High Resolution Depth Maps
Cloud Fusion
High Resolution Point Cloud
Multi Resolution Clustering
Point Cloud
Cloud Fusion Post Processing Meshing

Figure 3. Multi-view stereo reconstruction pipeline.

Smoothing (Sorkine et al., 2004) is applied to the 3D mesh. Finally, the final reconstructed 3D mesh is shown in Fig. 4-right.

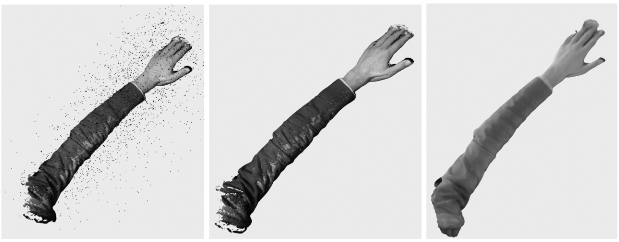

Figure 4. Example of result cloud (left) post-processing (center) and meshing (right).

6. Experiments

6.1 Calibration

We implemented the auto-calibration algorithms in C++, using the OpenCV library for image processing and the Ceres library for non-linear optimization. We validated the full auto-calibration submodule by using a set of six checkerboards each one composed of 12 x 5 internal corners, with a square size of 37 mm. We ran the system on the data collected following a total rotation of 360 degrees, finding that the system is able to correctly extract the corners of most framed checkerboards, distinguishing between different checkerboards. The conformation of the multi-board pattern combined with the movement around the axis allows to obtain an optimal corners distribution within the images of each camera (Fig. 5). In the intrinsic calibration procedure, we obtained a root mean square reprojection error of 0.32 pixels and a maximum reprojection error of 0.38 pixels. In the extrinsics calibration procedure, we obtained a root mean square reprojection error of 0.24 pixels and a maximum reprojection error of 0.33 pixels.

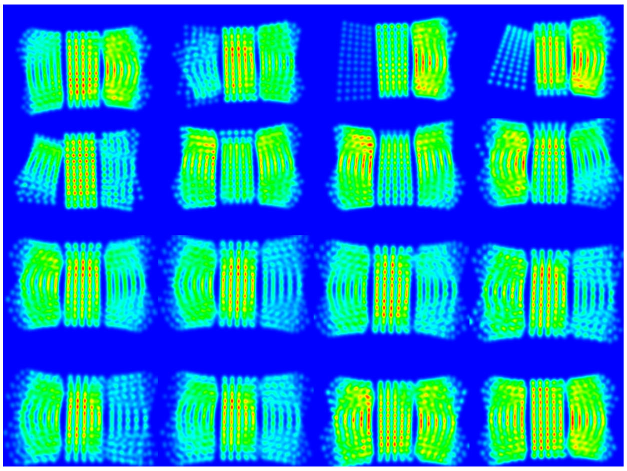

Figure 5. Corners distribution, obtained by means of kernel density estimation, i.e. a simplified version of the Parzen–Rosenblatt window method, using a Gaussian kernel with standard deviation of 30 pixels

6.2 Multi-view Stereo

The experimental evaluation has been performed on data acquired by using the custom-built camera ring device with 16 gray-scale cameras (2880x2160 px image size) with Intel Core i7-12700KF, 16 GB of RAM and a Nvidia 2060.

Some quantitative results are shown in Table 1. Moreover, in Table 1 a comparison with the original COLMAP implementation is provided, showing a sensible improvement in terms of

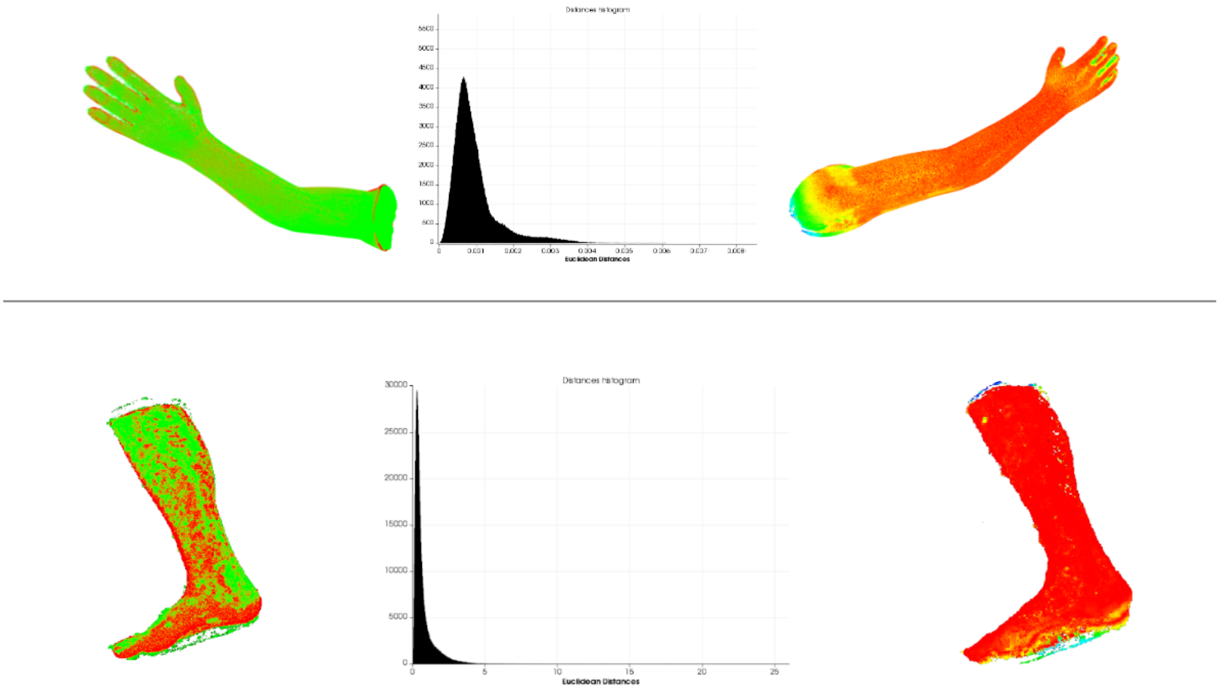

Figure 6. Left: Visualization of ground truth (red) and reconstructed (green) point clouds. Middle: Euclidean distances histogram. Right: Graphical visualization of distances on the reconstructed model (increasing distance values from red to blue).

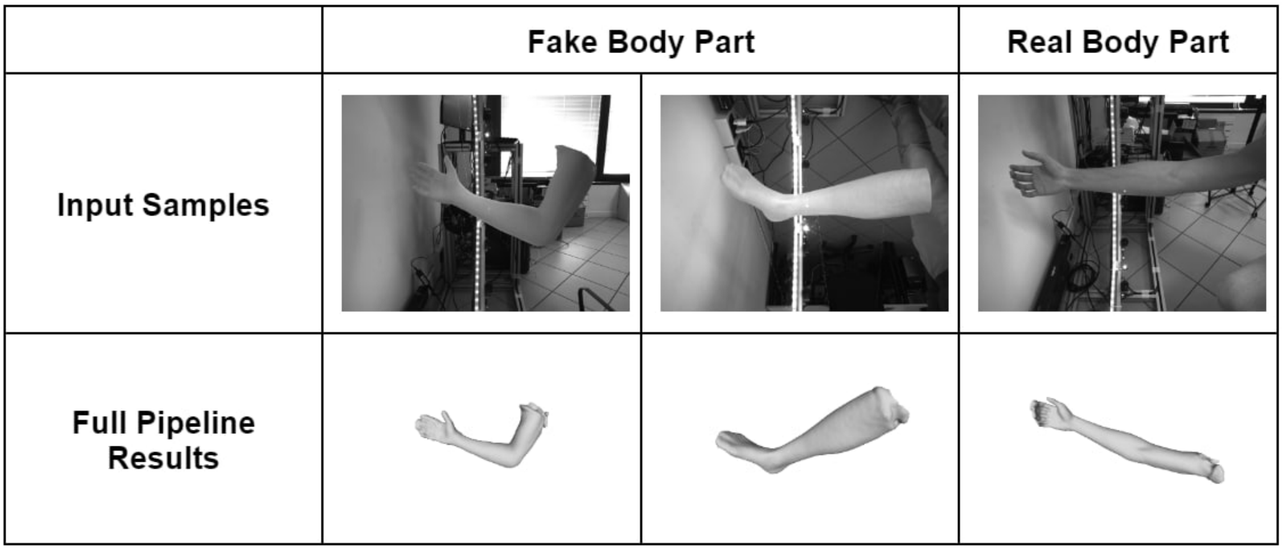

Figure 7. 3D reconstruction results (i.e., of fake and real human body parts) with the proposed multi-view stereo pipeline.

	Average Euclidean Distance	Hausdorff Distance	Chamfer Distance	Runtime	GPU memory
COLMAP	$1mm$	$11.3mm$	$(1.4mm)^2$	$\sim 115sec$	$\sim 1GB$
Ours	$0.9mm$	$9.6mm$	$(1.3mm)^2$	$\sim 40sec$	$\sim 1GB$

Table 1. 3D reconstruction's experimental evaluation.

accuracy of the reconstruction and runtime.
Preliminary tests show that our method can The quantitative evaluation has been performed on a test set of 30 different acquisitions of fake limbs (arm and leg) that have been appropriately 3D printed, of which we have the 3D CAD models (used as ground truth).
Some qualitative results of our approach are shown in Fig. 7.

7. Conclusions

In this paper, we introduced a custom-built low-cost 3D reconstruction system capable of extracting 3D models of small and medium objects. Human limbs, which are the main target for the 3D reconstruction of the proposed device, represent a critical challenge for classical MVS systems due to the repetitive

pattern of the skin. We proposed a pyramidal approach to improve over COLMAP, a robust MVS framework, and to reduce the time required to extract the mesh of the scanned object.
In order to reduce costs and human intervention we propose an automatic calibration system which can extract with good precision the camera parameters needed to perform an accurate MVS reconstruction. The layout of the checkerboards was chosen, through empirical tests, in order to maximize the number of corners projected in each camera and coverage. The proposed method, however, allow to perform calibration with different layouts which can be adjusted accordingly to the disposition of the cameras. Quantitative and qualitative performance evaluation for both automatic calibration and mvs approaches, that includes a comparison with a state of the art method, shows the effectiveness of the proposed framework.

References

Abdel-Aziz, Y. I., Karara, H. M., 1971. Direct Linear Transformation from Comparator Coordinates into Object Space Coordinates in Close-Range Photogrammetry. *Photogrammetric Engineering and Remote Sensing*, 81, 103–107.

Bleyer, M., Rhemann, C., Rother, C., 2011. Patchmatch stereo-stereo matching with slanted support windows. *British Machine Vision Conference*, 11, 1–11.

Duda, A., Frese, U., 2018. Accurate Detection and Localization of Checkerboard Corners for Calibration. *29th British Machine Vision Conference. British Machine Vision Conference (BMVC-29), September 3-6, Newcastle, United Kingdom.* http://bmvc2018.org/contents/papers/0508.pdf.

Faugeras, O., Bras-Mehlman, E., Boissonnat, J., 1990. Representing stereo data with the Delaunay triangulation. *Artificial Intelligence*, 44(1), 41–87.

Fromherz, T., Bichsel, M., 1995. Shape from multiple cues: Integrating local brightness information. *Proceedings of the Fourth International Conference for Young Computer Scientist*, 850–854.

Gu, X., Fan, Z., Zhu, S., Dai, Z., Tan, F., Tan, P., 2020. Cascade cost volume for high-resolution multi-view stereo and stereo matching. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2492–2501.

Heikkila, J., Silven, O., 1997. A four-step camera calibration procedure with implicit image correction. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1106–1112.

Ji, M., Gall, J., Zheng, H., Liu, Y., Fang, L., 2017. Surfacenet: An end-to-end 3d neural network for multiview stereopsis. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2307–2315.

Kazhdan, M., Hoppe, H., 2013. Screened poisson surface reconstruction. *ACM Trans. Graph.*, 32(3). https://doi.org/10.1145/2487228.2487237.

Kutulakos, K., Seitz, S., 1999. A theory of shape by space carving. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, 1, 307–314 vol.1.

Lindenberger, P., Sarlin, P.-E., Larsson, V., Pollefeys, M., 2021. Pixel-Perfect Structure-from-Motion with Featuremetric Refinement. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 5967–5977. https://api.semanticscholar.org/CorpusID:237194807.

Romanoni, A., Matteucci, M., 2019. TAPA-MVS: Textureless-Aware PAtchMatch Multi-View Stereo. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 10412–10421. https://api.semanticscholar.org/CorpusID:85517913.

Sayed, M., Gibson, J., Watson, J., Prisacariu, V., Firman, M., Godard, C., 2022. Simplerecon: 3d reconstruction without 3d convolutions. *European Conference on Computer Vision*, Springer, 1–19.

Schönberger, J. L., Zheng, E., Frahm, J.-M., Pollefeys, M., 2016. Pixelwise view selection for unstructured multi-view stereo. *European conference on computer vision*, Springer, 501–518.

Schönberger, J. L., Frahm, J.-M., 2016. Structure-from-motion revisited. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 4104–4113.

Seitz, S., Curless, B., Diebel, J., Scharstein, D., Szeliski, R., 2006. A comparison and evaluation of multi-view stereo reconstruction algorithms. *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, 1, 519–528.

Seitz, S., Dyer, C., 1997. Photorealistic scene reconstruction by voxel coloring. *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1067–1073.

Sorkine, O., Cohen-Or, D., Lipman, Y., Alexa, M., Rössl, C., Seidel, H.-P., 2004. Laplacian surface editing. *Proceedings of the 2004 Eurographics/ACM SIGGRAPH symposium on Geometry processing*, 175–184.

Szeliski, R., 1999. A multi-view approach to motion and stereo. *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, 1, 157–163.

Taylor, 2003. Surface reconstruction from feature based stereo. *Proceedings Ninth IEEE International Conference on Computer Vision*, 1, 184–190.

Treuille, A., Hertzmann, A., Seitz, S. M., 2004. Example-based stereo with general brdfs. T. Pajdla, J. Matas (eds), *Computer Vision - ECCV 2004*, Springer Berlin Heidelberg, Berlin, Heidelberg, 457–469.

Tsai, R., 1987. A versatile camera calibration technique for high-accuracy 3D machine vision metrology using off-the-shelf TV cameras and lenses. *IEEE Journal on Robotics and Automation*, 3(4), 323–344.

Wang, Y., Zeng, Z., Guan, T., Yang, W., Chen, Z., Liu, W., Xu, L., Luo, Y., 2023. Adaptive patch deformation for textureless-resilient multi-view stereo. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1621–1630.

Xu, Q., Tao, W., 2019. Multi-Scale Geometric Consistency Guided Multi-View Stereo. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5478–5487. https://api.semanticscholar.org/CorpusID:118679778.

Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L., 2018. Mvsnet: Depth inference for unstructured multi-view stereo. V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (eds), *Computer Vision – ECCV 2018*, Springer International Publishing, Cham, 785–801.

Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., Quan, L., 2019. Recurrent MVSNet for High-Resolution Multi-View Stereo Depth Inference. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5520–5529. https://api.semanticscholar.org/CorpusID:67855970.

Zhang, Z., 2000. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11), 1330–1334.

Zheng, E., Dunn, E., Jojic, V., Frahm, J.-M., 2014. Patchmatch based joint view selection and depthmap estimation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1510–1517.