ISSN 2464-4617(print)
ISSN 2464-4625(CD)
Computer Science Research Notes
CSRN 2801
Full Papers Proceedings
http://www.WSCG.eu

Markerless Structure-based Multi-sensor Calibration for Free Viewpoint Video Capture

Alexandros Papachristou Nikolaos Zioulis Dimitrios Zarpalas Petros Daras
Information Technologies Institute, Centre for Research and Technology - Hellas
papachra@iti.gr nzioulis@iti.gr zarpalas@iti.gr daras@iti.gr

ABSTRACT

Free-viewpoint capture technologies have recently started demonstrating impressive results. Being able to capture human performances in full 3D is a very promising technology for a variety of applications. However, the setup of the capturing infrastructure is usually expensive and requires trained personnel. In this work we focus on one practical aspect of setting up a free-viewpoint capturing system, the spatial alignment of the sensors. Our work aims at simplifying the external calibration process that typically requires significant human intervention and technical knowledge. Our method uses an easy to assemble structure and unlike similar works, does not rely on markers or features. Instead, we exploit the a-priori knowledge of the structure's geometry to establish correspondences for the little-overlapping viewpoints typically found in free-viewpoint capture setups. These establish an initial sparse alignment that is then densely optimized. At the same time, our pipeline improves the robustness to assembly errors, allowing for non-technical users to calibrate multi-sensor setups. Our results showcase the feasibility of our approach that can make the tedious calibration process easier, and less error-prone.

Keywords

Spatial Alignment, External Multi-Sensor Calibration, Semantic Segmentation, Free-viewpoint Capture, RGBD

1 INTRODUCTION

Capturing the complete appearance of real people and general scenes has matured and attracted much interest lately. Be it either offline for high quality free viewpoint video [Ye13] and streamable 3D content [Col15], or in real-time for tele-presence [Esc16, Bec13] and tele-immersion [Zio16] scenarios, it can open up the potential for new immersive experiences in a variety of applications like gaming [Zio16] or remote interactions [Esc16, Bec13].

The backbone of these new experiences is the acquisition of a full 3D representation of general scenes or performances. While a variety of single sensor methods exist, some focusing only on geometry information [New15, Inn16, Zol14], and others also producing fully textured outputs [Guo17, Cao17], truly immersive experiences can only be facilitated by complete 360^o captures via multi-sensor systems. These systems present with both high-quality but expensive solutions [Dou16, Dou17], as well as lower cost ones [Ale17]. Either option utilizes color and depth (RGB-D) information acquired from multiple viewpoints that are spatially aligned, or otherwise externally calibrated, to a common coordinate system. Therefore, this external calibration step is a necessity for all performance capture methods alike.

However, multi-sensor calibration is typically a complex procedure that requires trained users, a requirement that inhibits the applicability of this technology to the consumer public. The complexity arises from the fact that most methods require capturing a calibration object in numerous poses into the captured area [Bec15, Bec17, Fur13, For17], and in some cases this is also performed in a sensor pairwise manner [Hei97]. Some recent methods utilize a static calibration structure to spatially align all viewpoints. In [Col15], calibrating a large amount of cameras is accomplished by using a very complex octagonal tower. There also exist lower complexity structures for setups with less sensors [Kow15, Ale17]. These structure-based multi-sensor calibration methods are more suitable for non-technical users as they require minimal human intervention apart from assembling the structure.

In this work, we lift the requirement of using markers or patterns when utilizing a known structure for calibrating multiple RGB-D sensors. Our main contribution is a correspondence identification step that requires no feature extraction or marker identification. We exploit only the structure's geometrical semantics and segment input depth maps into labeled regions. Therefore, lifting the requirement of markers or patterns, our pro-

ISBN 978-80-86943-40-4
https://doi.org/10.24132/CSRN.2018.2801.10

posed method does not require color input and operates on a markerless basis.

In the remainder of the paper we review related work in Section 2 and present our method in detail in Section 3. Then, in Section 4 a detailed evaluation follows, with a concluding discussion presented in Section 5.

2 RELATED WORK

Precise spatial alignment of multi-sensor 3D capturing systems is essential for the creation of realistic 3D human models and assets. Preliminary methods relied only on distributed RGB cameras and were based on the simultaneous capture of a planar rigid printed checkerboard with known dimensions from at least 2 cameras [Hei97, Zha00]. This technique, which is typically used for stereo calibration, is the de facto method to estimate the intrinsics parameters along with the relative pose between neighboring cameras. Despite producing high accuracy results, it requires great effort from the user because the printed checkerboard must be slowly moved and (re-)positioned inside the capturing area. In addition, it also requires knowledge of the technical details behind the calibration process to avoid hard-to-detect checkerboard poses and partial views. Furthermore, it requires hardware synchronized sensors or otherwise, a precise synchronization step for all cameras should precede. Furthermore, the solution is anchored on a selected reference camera as it is not possible to transform all viewpoints to a common global coordinate system. In systems composed of more than 2 sensors, a potential erroneous estimation could be accumulated during the aggregation of relative transformations.

To address the aforementioned limitation, state-of-the-art systems based on the same checkerboard pattern, have incorporated additional optical tracking systems [Bec15, Bec17] or IMU sensors [Fur13]. These alternative methods rely heavily on the tracking systems which are mainly responsible to track the checkerboard's location and define the global coordinate system. Nonetheless, tracking systems require special technical knowledge to mount and operate. Moreover, capturing of the moving checkerboard still requires human intervention, which potentially introduces errors. Further, temporal alignment of the sensors is also needed to synchronously capture the input images. Another associated challenge is the motion blur introduced by the moving checkerboard that can deteriorate the calibration's overall performance.

Specifically for RGB-D sensors, some methods exploit the availability of depth measurements as well as the color information to detect a set of 2D features [Low04, Bay08, Rub11, Alc11] within the capturing area, which can then be converted to 3D points using the depth data. When matched between neighboring viewpoints 3D-to-3D correspondences are established [Dou14], that are used to estimate the relative pose between the sensors. Several works have attempted to enrich the capturing area with features or markers placed on a structure to establish robust 3D correspondences [Ale17, Kow15, Kai12]. Using a common structure offers the advantage of avoiding pairwise calibration and instead, spatially aligns all sensors onto the same coordinate system directly.

However, establishing only sparse correspondences based on detected 2D features or markers is frequently prone to errors due to measurement inaccuracies. To overcome this, dense alignment methods are used that exploit the overlap between viewpoints. Albeit, these still require a rough initial alignment that is given by sparse feature correspondences. Dense methods are usually developed using a variant of the Iterative Closest Point (ICP) algorithm [Kow15, Kin05], graph-based optimization [Ihr04] or bundle adjustment [Van17]. A comprehensive review of refinement methods can be found in [Pom13]. In a similar fashion, other approaches densely estimate the viewpoints of spatially distributed sensors by initially detecting lines and planes [Den14, Owe15, Xu17]. This is succeeded by a post-refinement step to find a globally optimal solution. More recently, a color-based object was utilized and tracked to simultaneously align multiple RGB-D sensors both in the spatial and temporal domain [For17]. It still remains though, a complex process that requires a user to move the object within the scene.

While machine learning algorithms are now abundantly used in various computer vision tasks due to their high performance, they have found little use in calibration tasks. They have been mostly used in localization tasks utilizing decision trees on pure color [Sho13] or RGB-D [Bra14, Bra16] information. Similarly, deep learning variants of these methods have emerged [Ken15, Zam16, Mel17, Nak17, Poi16]. Despite having displayed promising initial results, their accuracy and robustness have not been put to the test of multi-sensor alignment in order to demonstrate their applicability to this specific problem.

3 MARKERLESS STRUCTURE-BASED SPATIAL ALIGNMENT

Our goal is to perform a multi-sensor extrinsic calibration aiming to spatially align the generated point clouds into a common, global, coordinate system. We rely on an easy to deploy calibration structure that is assembled by four equally sized boxes, and more specifically, low-cost commercially available packaging boxes. This approach requires minimal human intervention and is inspired by [Ale17]. Unlike the structure assembled in [Ale17] though, we opt for a simpler assembly process where the boxes are positioned on top of each other,

 ISBN 978-80-86943-40-4

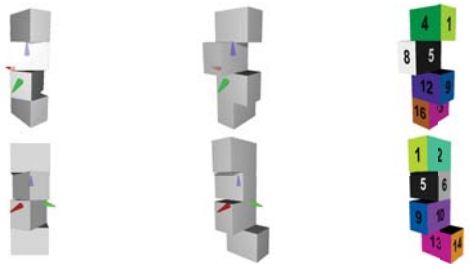

Figure 1: **Left**: The symmetric calibration structure used in [Ale17]. **Middle**: Our asymmetric structure with its corresponding semantic labels per face **(Right)**. Instead of aligning the boxes on top of each other using their diagonals, they are now snapping on their top sides' corners, following a 90° rotational pattern.

while using their side corners instead of their diagonals to snap each box with the one placed on top of it. Fig. 1 showcases the structure of [Ale17], as well as our modified assembly. Another advantage associated to this modification is that the structure is now fully asymmetric, compared to the previous symmetric (i.e. mirrored) assembly. The calibration structure serves as a spatial anchor as all sensor viewpoints' relative pose to its coordinate system (depicted in Fig. 1) will be estimated. This simplifies the calibration process as it removes the necessity of complex pairwise alignments.

Our approach differs from similar approaches that utilize boxes [Kow15] or structures [Ale17], as it does not rely on feature extraction or marker detection. Instead, our correspondence estimation is only reliant on the structure's geometry, as observed by the depth sensor. We exploit the a-priori knowledge related to the structure's shape by training a Fully Convolutional Network (FCN) [Lon15] to identify the structure's boxes' sides. In this way, we perform an initial viewpoint estimation which we then densely refine via a global optimization.

3.1 Semantic Correspondences

Given the now asymmetric geometry, we assign a unique label to each distinct box side and train an FCN for a dense classification task that aims to identify each side in an input depth image. The updated structure's asymmetry allows for easier learning of unique feature descriptors for each viewing direction and is free of any ambiguities that would arise from a symmetric one. The multi-view spatial alignment process can potentially involve a very wide variety of different captured depth data as it involves the full 6 DOF of both the structure and sensor. Training an FCN means we don't have to rely on hand-crafted features or a customized methodology. Training a network for the task of labeling each side, given the numerous possible poses that a sensor can observe the structure, requires a very large dataset. We circumvent the difficult task of manually labeling such a large dataset

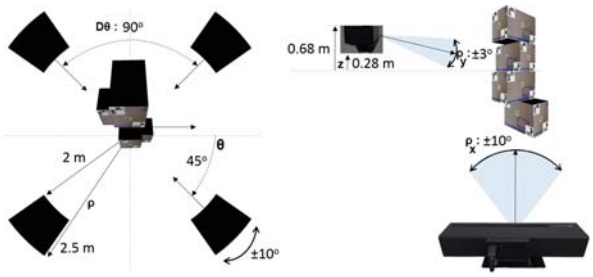

Figure 2: Pose generation process. The range limits of each parameter in equation (1) used to sample/generate the poses are visually presented, showcasing the possible sensor positions around the calibration structure.

by synthesizing it. This is accomplished by building a virtual model replica of the calibration structure using the boxes' known dimensions. Nonetheless, creating such a dataset requires a very large amount of storage. Therefore, we chose to simply generate the depth images and labels on-the-fly, using the graphics pipeline to render our data and simulate realistic depth data capturing conditions.

Pose Generation: The structure's coordinate system, and therefore the global coordinate system that all sensors' data will be transformed to, resides on the virtual structure's origin as shown in Fig. 1. Each sensor i'th pose $[\mathbf{R}|\mathbf{t}]^i \in \mathbb{SE}^3$ with respect to the structure, consists of a rotation $\mathbf{R}^i \in \mathbb{SO}^3$ and translation $\mathbf{t}^i \in \mathbb{R}^3$. To generate a large amount of poses we sample positions $\mathbf{t^i} = (t_x^i, t_y^i, t_z^i)$ in a circular pattern around the structure looking towards its origin. We use a cylindrical coordinates sampling (ρ, θ, z) for the position of each sampled viewpoint, omitting the superscripts i for the remainder of this section. A free viewpoint capture setup requires its sensors to look inwards towards its capturing space's center. In our case, this resides on the structure's center position, i.e. the global coordinate frame's origin. Therefore, the poses' rotations $\mathbf{R}$ are set to look at $(0,0,0)$. In practice, though, one cannot achieve such an accurate positioning of the sensor. To compensate for this, we compose additional rotational perturbations ρ_x and ρ_y to each sampled viewpoint, to further augment the variety of sampled poses and capture realistic positioning conditions. These are rotations around the x and y axis respectively, which essentially represent the sensor's pan(right/left) and tilt(up/down) rotations as shown in Fig. 2. For each of these variables we generate discrete samples from a uniform distribution $U(a,b,c)$ at the interval $[a,b]$ in steps of c units:

$$\begin{aligned} \theta &\overset{\alpha}{\sim} U(\alpha - 10^\circ, \alpha + 10^\circ, 2.5^\circ), \\ z &\sim U(0.28m, 0.68m, 0.02m), \\ \rho &\sim U(2.0m, 2.5m, 0.02m), \\ \rho_x &\sim U(-10^\circ, 10^\circ, 2.5^\circ), \\ \rho_y &\sim U(-3^\circ, 3^\circ, 3^\circ). \end{aligned} \tag{1}$$

ISBN 978-80-86943-40-4

where $\alpha = \{45^\circ, 135^\circ, 225^\circ, 315^\circ\}$. The bounds for range ρ and height z, confine the viewpoint within the limits of fully capturing a human subject given a reasonable vertical field of view. In addition, the sensor originated rotations ρ_x and ρ_y, are set in a range of values that are reasonable to position the captured subject close to the sensor's center. Regarding the distribution of the viewpoints around the structure as offered by the cylindrical angle θ, we restrict it around specific 90° intervals, thus focusing only on the case of 4 viewpoint capture. The 4 sensor case is the most optimal solution in terms of cost against quality when aiming for full 360° coverage with the least amount of sensors. By considering an approximate positioning of the sensors around the structure, offered by the selected range of θ angles, we add a restriction in order to decrease the number of input poses and increase the robustness of our predictions. This restriction is a structure placement guideline: "*to have the sides of all boxes looking in between of two sensors*", as illustrated in Fig. 2. The same figure also presents the aforementioned sampling spaces, as well as the relative to the structure positioning of the sensor poses that are generated for creating the training data. We generate a total of $N = 530712$ poses $[\mathbf{R}|\mathbf{t}]$.

Data Generation: The on-the-fly data generation process takes as input the 3D virtual model which is decomposed into parts, each part being one side of each box comprising the structure. We generate N samples using the poses $[\mathbf{R}|\mathbf{t}]$ to position the virtual camera and render the model, acquiring the generated z-buffer as the input data depth map $\mathbf{D}(u,v)$. Each part is also assigned a unique label for a total of 25 distinct labels, six sides for each box plus the background. Each labeled part is rendered with a unique color. By also acquiring the swapped color buffer we obtain the ground truth per pixel labeled image $\mathbf{L}(u,v)$. To simulate more realistic input, we add noise on the rendered depth map randomly choosing the noise function for each sample. We use a noise model better suited for disparity based depth maps (e.g. structured light) as presented in [Bar13] as well as a random noise simulation scaled with the depth value of each pixel:

$$\mathbf{D}_n(u,v) = \text{sign}(U(-1,1)) * \mathbf{D}(u,v) * \sigma_d * (1 - e^{\frac{-U(0,1)^2)}{2}}) \quad (2)$$

where $\mathbf{D}_n$ and $\mathbf{D}$ are the noisy and rendered depth maps respectively, U denote random uniform distributions, and σ_d is a depth scaling factor. In addition, we composite the rendered model onto random backgrounds. These are selected uniformly from various cases: i) white noise, ii) Gaussian noise, (both scaled appropriately to produce values within the expected depth ranges), iii) 159 backgrounds drawn from the database of [Was16] (selected one per 30 frames) and iv) 326 backgrounds drawn from in-house recordings with actual people being captured. Therefore, we augment our online generated training corpus using a mix of noisy and real backgrounds as well as two distinct depth noise models, with some examples presented in the supplementary material. It should be noted that we also generate a smaller test dataset from sensor positions not included in the train data, as a result of choosing different starting values and step units in the same ranges as those presented in (1).

Architecture: The detailed deep Fully Convolutional Network (FCN) configuration and architecture used is presented in Figure 3 (Left). It comprises of a multilayer convolution and a symmetric deconvolution network. The first part learns to extract various features from the input depth map, while the second learns to produce the semantic segmentation of the input into its distinct labels, i.e. box sides, out of the extracted features. The final densely predicted labels are computed out of a probability feature vector of size equal to the amount of labels (25) for each pixel, which constitutes the output of our network that is estimated via a softmax function. The resulting prediction map matches the resolution of the input depth map, as while the convolution part reduces the size of the activations, the following deconvolution part enlarges them back to their original size.

Correspondence Establishment: After segmenting the depth map into regions that correspond to each distinct box's sides, we can use this semantic information to establish correspondences between the acquired depth map and the virtual structure model. Initially, we discard the regions labeled as background or the box sides that are facing upwards/downwards. Then, for each remaining segmented region L, we back-project all depth map pixels to 3D (in the sensor's local coordinate system), and extract their median 3D position $\mathbf{m}_L$. Small area labeled regions are heuristically discarded when containing less than n elements. The 3D point $\mathbf{m}_L$ corresponds to the labeled box's rectangular side center point and therefore, we can establish a correspondence in 3D with the known point's coordinates in the asymmetric structure's virtual model. This 3D correspondence establishment is illustrated in Fig. 3 (Right), where the matching of the corresponding median points between two real views and the virtual structure model is presented.

3.2 Global Spatial Alignment

Given the 3D-to-3D correspondences as an input, we determine an initial alignment of all viewpoints by using the generalized Procrustes analysis [Ken05]. For each viewpoint s we obtain its pose $\mathbf{P}_s$ with respect to the global coordinate system that is originated in the structure's virtual model. However, this initial viewpoint estimation may often present slight errors as a result of the sparseness or inaccuracy of correspondences

 ISBN 978-80-86943-40-4

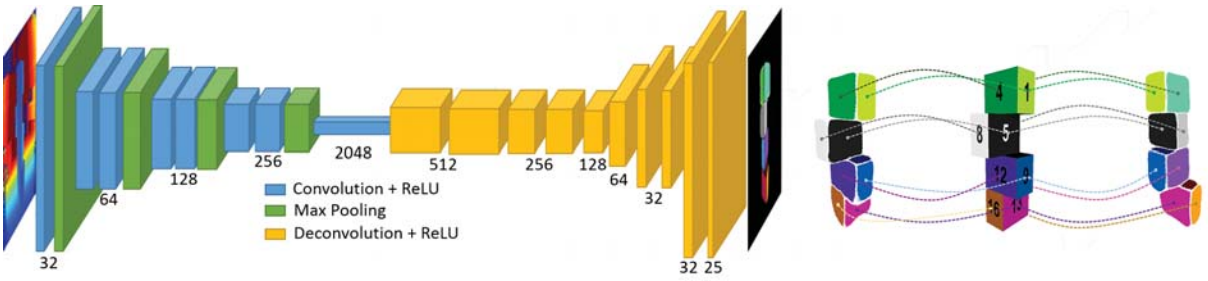

Figure 3: **Left:** The architecture of our semantic segmentation FCN. Having a single depth image as an input, it segments and densely classifies it in order to identify the per-pixel labels of the observed calibration structure. **Right:** Correspondences are established by extracting the 3D medians of the detected labeled regions. These 3D points are then matched against the midpoints of their corresponding virtual structure's box sides to establish 3D-to-3D correspondences. (best viewed in color)

used. This will lead to a visible drop of quality for the produced / captured content. Another reason for inaccurate correspondences is the possibility of an imperfect assembling of the structure, which is more typical when using markers (misplacement) or features (imprecise localization). Thereby, a second step is needed to refine the initial viewpoint estimations by densely aligning the point clouds of adjacent sensors. Instead of a simple pairwise optimization, we solve for an optimal global solution using all viewpoints simultaneously. We use a graph-based optimization where the spatial relationships between the sensor set S are represented by a graph $G = (P,E)$.

The nodes of the graph are the estimated poses $\mathbf{P}_s \in \mathbb{SE}^3$ of each sensor s in the global (virtual structure) coordinate system. The edges E_{ij} represent constraints in the poses between the nodes i and j in the form of observations of j from node i. These observations are established as correspondences $\mathbf{P_i}\mathbf{v}_i \leftrightarrow \mathbf{P}_j\mathbf{v}_j$ with $\mathbf{v} \in \mathbb{R}^3$ being a point in the sensor's local coordinate system. These correspondences are acquired by nearest neighbor searches between viewpoints i and j after transformed to the common coordinate system. Each correspondence / edge is encoded as point-to-plane distance:

$$E_{ij} = \|(\mathbf{P}_i^{-1}\mathbf{P}_j\mathbf{v}_j - \mathbf{v}_i)^T\mathbf{n}_i\|_2 \qquad (3)$$

where $\mathbf{n}_i \in \mathbb{R}^3$ is the normal vector of $\mathbf{v}_i$. As depth maps can be noisy around edges, in order to reduce the effect of outliers, we only establish 3D point correspondences within a radius r_{cutoff} between adjacent sensors, with their adjacency estimated by their initial sparse spatial alignment. The graph-based optimization uses the Levenberg-Marquardt method [Mar63] to solve the underlying system, with an iterative scheme. We perform a fixed number of iterations while also dropping r_{cutoff} after a set number of iterations. Solving for all poses simultaneously instead of in a pairwise fashion, we get a globally optimum solution. The refined poses of the viewpoints effectively maximize the overlap between neighboring point clouds. Overall, this dense refinement step rectifies any human-related or systematic errors and improves the quality of the spatial alignment.

4 RESULTS AND DISCUSSION

We evaluate our multi-sensor external calibration method under a variety of 4-sensor setups all focused on free viewpoint capture of human performances. Consequently, the sensors are all looking inwards, towards the center of the capturing area.

Implementation details: Our experiments are based on the Microsoft Kinect 2.0, a Time-of-Flight RGB-D sensor. Our semantic labeling FCN was trained on a NVIDIA Titan X using the Caffe framework [Jia14]. We rendered the generated data using the average Kinect intrinsics parameters (512×424 resolution, a 366.66 focal length baseline and placed the principal point at the depth maps center) to create the projection matrix and trained the FCN on the full resolution images. We train our network for $100k$ iterations with an initial learning and batch size of 0.001 and 5 respectively. We increase the batch size to 15 after $50k$ iterations and linearly decay the learning rate with a gamma of 0.9 every $10k$ and $15k$ iterations when the batch size is 5 and 15 respectively. We use the ADAM optimizer [Kin14] with its standard momentum and epsilon parameters. The threshold for discarding labeled regions n_{pixel} was heuristically selected to be 2000 pixels to discard potential erroneous estimations predicted by the FCN. After training is over, our model achieves a mean Intersection over Union (mIoU) of 86.23% on the generated test dataset. For the refinement step we use the *g2o* framework [Kum11] for 10 iterations and initially set r_{cutoff} to $0.05m$ and drop it to $0.01m$ after 5 iterations.

Data acquisition: As the data acquisition requires capturing a static structure object, the process is free of temporal synchronization or motion blur issues. We exploit this to aggregate frame information within a time window of N frames. Thus, we obtain a median depth map out of 30 frames capturing the static structure. The

ISBN 978-80-86943-40-4

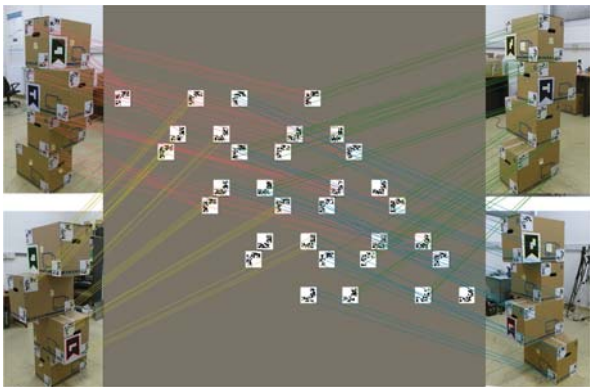

Figure 4: Example calibration views as captured by the color sensor. The SIFT correspondences are also shown, as well as the marker placement for [Ale17] and [Kow15]. SIFT features are matched against the texture applied to the virtual calibration structure model. LiveScan3D marker detections are highlighted on the color images.

median depth map cancels out noise and has less holes, while also being robust to any interference between the sensors. Further, the acquisition process requires minimal human intervention which is limited to assembling the structure near the center of the capturing area, so as to be visible from all sensors simultaneously.

Metric: We measure the accuracy of the registration using the Rooted Mean Squared Euclidean (RMSE) distance between the closest points of overlapping areas of adjacent point-clouds (back-projected from the corresponding depth maps). Given that we seek to measure how well the overlapping surfaces fit and since the viewpoints' overlap is limited as a result of their 90^o intervals placement around the capturing space, we only use those correspondences with distances less than $0.02m$ for each viewpoint pair. We calculate the RMSE error across all adjacent viewpoint pairs for each sensor and average the overall error.

4.1 Evaluation

We compare our method against the structure-based method [Ale17] and "LiveScan3D" [Kow15] that is similarly reliant on attaching markers on rigid surfaces (i.e. boxes). For [Ale17] we utilize the publicly offered markers offered that we attached on the structure following the available instructions. This method only performs spatial alignment based on sparse correspondences. For extracting the SIFT [Low04] correspondences we opt for a brute force matching strategy, instead of approximate versions as we are not bounded by timing constraints. The marker placement and feature matching process is shown in Fig. 4.

For [Kow15], we use the offered set of markers which are used to obtain initial pose estimates. These are then refined by a dense optimization step using pairwise ICP. The "LiveScan3D" markers were also attached on the same calibration structure to allow for simultaneous comparison between all methods as shown in Fig. 4. Markers' positions in the structure's (i.e. global) coordinate system were calculated as they are required as input by [Kow15] to drive the initial registration. For box assemblies that are not perpendicular (e.g. the structure of [Ale17]), this would require some effort by the users to calculate the markers' positions using trigonometry.

Both [Kow15] and our method utilize a post dense refinement step to improve the spatial alignment results, while [Ale17] does not. As a result, we also offer results for [Ale17] by adding a graph-based dense refinement step after the initial alignment obtained by the sparse feature correspondences. We refer to the sparse version as "Sparse-Only" and to the extended post-refinement version as "Sparse+Graph".

We conduct experiments for a variety of setups in order to evaluate all methods in terms of accuracy and robustness. We even purposefully include defective assemblies of the calibration structure to assess each method's efficacy with respect to mis-assemblies (namely f, g and h in Table 1). Given that our markerless correspondence estimation focuses on a particular capturing setup and was trained on these poses only, we use an approximate 4 sensor placement at 90^o intervals around a circle. It should be noted that markers for [Ale17] and [Kow15] were placed on the structure without overlapping as seen in Fig. 4.

Table 1 presents quantitative results of our experiments while also offering each setup's approximate sensor placements. Fig. 6 displays the qualitative results for the same setups. Even though experiments (a) and (b) included sensor poses that were out of the training range, there was no meaningful accuracy degradation compared to other setups, demonstrating how our model has generalized efficiently, well-behaving even in unseen poses. More importantly, while trained on synthetic data with an assortment of augmentations (noise and backgrounds) it has managed to produce high quality segmentation results in realistic data acquired from various sensors as seen in Fig. 5. Segmentation results for all experiments are available in the supplementary material.

Overall, the results presented in Table 1 show that our method outperforms others, except for the SIFT-based one enhanced with the graph-based refinement step. However, our method removes the need for markers, which is a cumbersome and error-prone procedure during the assembly of the structure. Moreover, in the mis-assembly experiments the semantic based method outperforms the marker-based one and, as seen in Fig. 6 it was able to converge in all cases despite the errors, compared to the other methods that did not converge in all cases. This is due to a robuster initial alignment (and

 ISBN 978-80-86943-40-4

	a	b	c	d	e	f	g	h
LiveScan3D	6.2	6.42	7.07	6.44	7.35	9.57	**6.30**	10.40
Sparse-Only	10.85	11.94	7.50	8.42	8.66	11.44	11.01	12.36
Sparse+Graph	**5.8**	6.52	**5.84**	**6.18**	**6.29**	9.50	7.43	12.25
Ours	6.39	**6.30**	6.31	6.61	7.56	**6.67**	6.68	**7.15**

Table 1: RMSE results (in mm) of our method and the compared ones. Approximate sensors' placements were: $a \sim \{\rho : 1.7m, z : 0.5m\}$, $b \sim \{\rho : 1.7m, z : 0.28m\}$, $c \sim \{\rho : 2.0m, z : 0.28m\}$, $d \sim \{\rho : 2.0m, z : 0.5m\}$, $e \sim \{\rho : 2.0m, z : 0.5m$ globally rotated compared to d $\}$, $f \sim \{\rho : 2.0m, z : 0.5m$ with translational error$\}$, $g \sim \{\rho : 2.0m, z : 0.5m$ with rotational error$\}$ and $h \sim \{\rho : 2.0m, z : 0.5m$ with both rotation and translational errors$\}$

by extension correspondence estimation) that helps the dense post-refinement step rectify any potential errors.

5 CONCLUSIONS

In this work, we have presented a markerless structure-based external calibration method for multi-sensor setups oriented towards 3D performance capture. Instead of relying on markers to establish correspondences, we exploit the known structure's geometry and train a CNN to semantically label perspective depth maps acquired when viewing the calibration structure. It is an innovative alternative to sparse feature-based spatial alignment that only works with depth input instead of relying on color information. We have demonstrated that this is indeed an effective approach that minimizes human error when assembling the structure and simplifies the overall process. In addition, we showcase how machine learning can be used in the task of multi-sensor spatial alignment. Overall, our method offers an easier and more practical multi-sensor calibration process that is more appropriate for a wider offering of free-viewpoint capture technologies.

Regarding the limitations of our method, it cannot be used to spatially align viewpoints that are looking outwards like showcased in [Kow15]. Additionally, its effectiveness in greater distances is questionable, but that is also a concern in general for depth sensors, whose accuracy degrades proportionally to the measured depth. Further, generalization to any position around the structure is something that should be explored in the future to allow for arbitrary positioning of the sensors (e.g. setups focusing on frontal captures only). Moreover, sparse alignment is reliant on a good segmentation result as erroneous estimates would cause the median calculation to drift. Finally, given that training is coupled to the selected sensor, applicability to a variety of sensor types might require re-training using new intrinsic parameters, however re-training is an one-time requirement.

6 ACKNOWLEDGEMENTS

This work was supported and received funding from the European Union Horizon 2020 Framework Programme funded project *Hyper360*, under Grant Agreement no. 761934. We are also grateful and acknowledge the support of NVIDIA for a hardware donation.

REFERENCES

[Alc11] P. F. Alcantarilla and T. Solutions. Fast explicit diffusion for accelerated features in nonlinear scale spaces. *IEEE Trans. Patt. Anal. Mach. Intell*, 34(7):1281–1298, 2011.

[Ale17] D. S. Alexiadis, A. Chatzitofis, N. Zioulis, O. Zoidi, G. Louizis, D. Zarpalas, and P. Daras. An integrated platform for live 3d human reconstruction and motion capturing. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(4):798–813, 2017.

[Bar13] J. T. Barron and J. Malik. Intrinsic scene properties from a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 17–24, 2013.

[Bay08] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Speeded-up robust features (surf). *Comput. Vis. Image Underst.*, 110(3):346–359, 2008.

[Bec13] S. Beck, A. Kunert, A. Kulik, and B. Froehlich. Immersive group-to-group telepresence. *IEEE Transactions on Visualization and Computer Graphics*, 19(4):616–625, 2013.

[Bec15] S. Beck and B. Froehlich. Volumetric calibration and registration of multiple rgbd-sensors into a joint coordinate system. In *2015 IEEE Symposium on 3D User Interfaces (3DUI)*, pp. 89–96, 2015.

[Bec17] S. Beck and B. Froehlich. Sweeping-based volumetric calibration and registration of multiple rgbd-sensors for 3d capturing systems. In *2017 IEEE Virtual Reality (VR)*, pp. 167–176, 2017.

[Bra14] E. Brachmann, A. Krull, F. Michel, S. Gumhold, J. Shotton, and C. Rother. Learning 6d object pose estimation using 3d object coordinates. In *European conference on computer vision*, pp. 536–551. Springer, 2014.

[Bra16] E. Brachmann, F. Michel, A. Krull, M. Y. Yang, S. Gumhold, and C. Rother. Uncertainty-driven 6d pose estimation of objects and scenes from a single rgb image.

 ISBN 978-80-86943-40-4

Figure 5: Qualitative results of our model on realistically acquired depth maps. (best viewed in color)

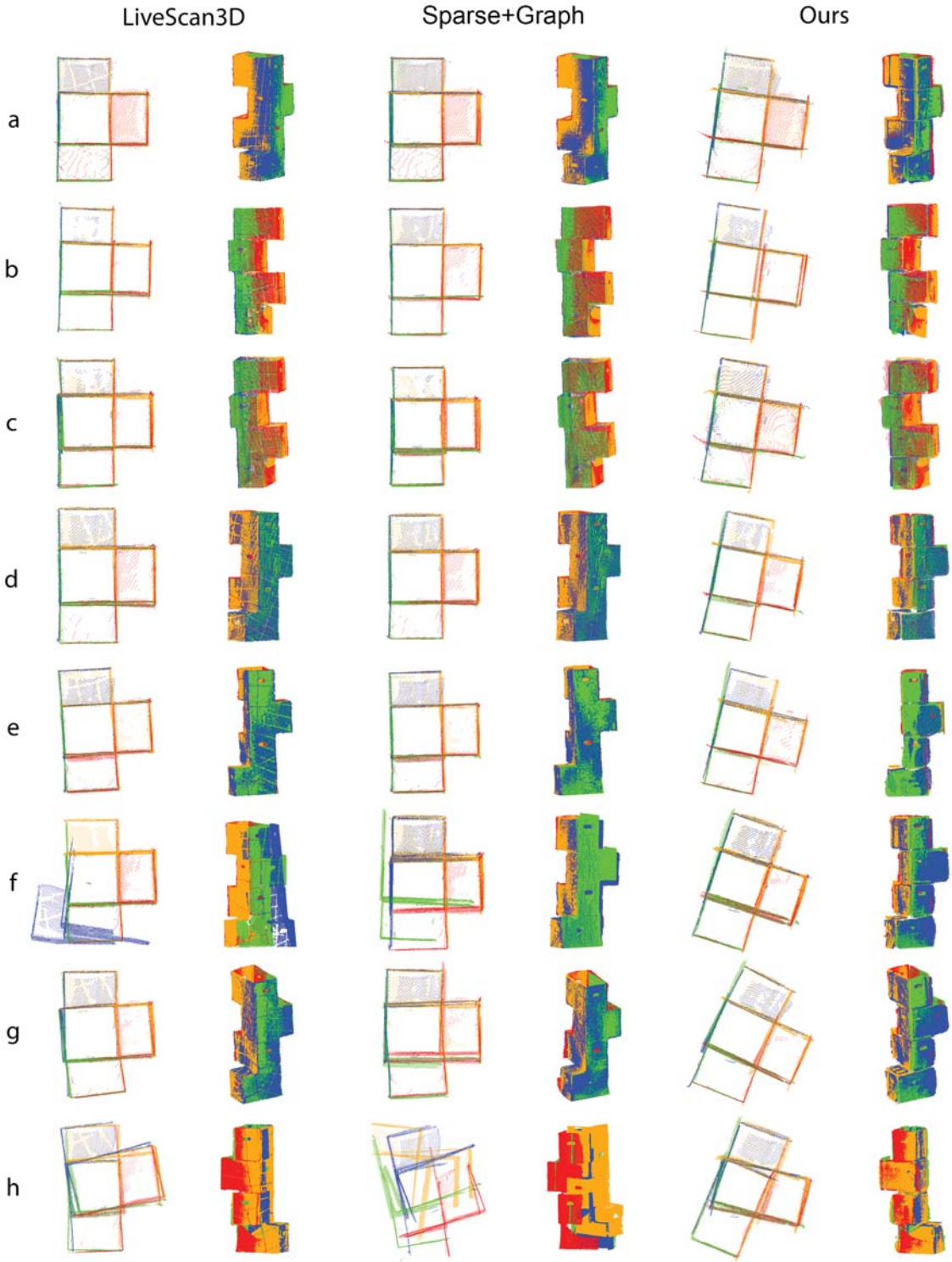

Figure 6: Qualitative results for all experiments. Each row depicts experiment (a-h), whilst each column shows the calibrated point clouds (color-per-sensor) for the evaluated methods in top and side views. (best viewed in color)

ISBN 978-80-86943-40-4

In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[Cao17] H. Zhu, Y. Liu, J. Fan, Q. Dai, and X. Cao. Video-based outdoor human reconstruction. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(4):760–770, 2017.

[Col15] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan. High-quality streamable free-viewpoint video. *ACM Transactions on Graphics (TOG)*, 34(4):69, 2015.

[Den14] D. Teng, J. C. Bazin, T. Martin, C. Kuster, J. Cai, T. Popa, and M. Gross. Registration of multiple rgbd cameras via local rigid transformations. *IEEE International Conference on Multimedia & Expo*, 2014.

[Dou14] M. Dou and H. Fuchs. Temporally enhanced 3d capture of room-sized dynamic scenes with commodity depth cameras. In *2014 IEEE Virtual Reality (VR)*, pp. 39–44, 2014.

[Dou16] M. Dou, S. Khamis, Y. Degtyarev, P. Davidson, S. R. Fanello, A. Kowdle, S. Orts-Escolano, C. Rhemann, D. Kim, and J. Taylor. Fusion4d: Real-time performance capture of challenging scenes. *ACM Transactions on Graphics (TOG)*, 35(4):114, 2016.

[Dou17] M. Dou, P. Davidson, S. R. Fanello, S. Khamis, A. Kowdle, C. Rhemann, V. Tankovich, and S. Izadi. Motion2fusion: real-time volumetric performance capture. *ACM Transactions on Graphics (TOG)*, 36(6):246, 2017.

[Esc16] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou, V. Tankovich, C. Loop, Q. Cai, P. A. Chou, S. Mennicken, J. Valentin, V. Pradeep, S. Wang, S. B. Kang, P. Kohli, Y. Lutchyn, C. Keskin, and S. Izadi. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, UIST '16, pp. 741–754. ACM, 2016.

[For17] A. Fornaser, P. Tomasin, M. D. Cecco, M. Tavernini, and M. Zanetti. Automatic graph based spatiotemporal extrinsic calibration of multiple kinect v2 tof cameras. *Robotics and Autonomous Systems*, 98(Supplement C):105 – 125, 2017.

[Fur13] P. Furgale, J. Rehder, and R. Siegwart. Unified temporal and spatial calibration for multi-sensor systems. In *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 1280–1286, 2013.

[Guo17] K. Guo, F. Xu, T. Yu, X. Liu, Q. Dai, and Y. Liu. Real-time geometry, albedo, and motion reconstruction using a single rgb-d camera. *ACM Transactions on Graphics (TOG)*, 36(3):32, 2017.

[Hei97] J. Heikkila and O. Silven. A four-step camera calibration procedure with implicit image correction. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1106–1112, 1997.

[Ihr04] I. Ihrke, L. Ahrenberg, and M. Magnor. External camera calibration for synchronized multi-video systems. In *WSCG 2004 : the 12th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision 2004 ; short communication papers proceedings*, Journal of WSCG, pp. 537–544, Plzen, Czech Republic, 2004. UNION Agency.

[Inn16] M. Innmann, M. Zollhöfer, M. Nießner, C. Theobalt, and M. Stamminger. Volumedeform: Real-time volumetric non-rigid reconstruction. In *European Conference on Computer Vision*, pp. 362–379. Springer, 2016.

[Jia14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[Kai12] B. Kainz, S. Hauswiesner, G. Reitmayr, M. Steinberger, R. Grasset, L. Gruber, E. Veas, D. Kalkofen, H. Seichter, and D. Schmalstieg. Omnikinect: Real-time dense volumetric data acquisition and applications. In *Proceedings of the 18th ACM Symposium on Virtual Reality Software and Technology*, VRST '12, pp. 25–32, New York, NY, USA, 2012. ACM.

[Ken05] D. G. Kendall. A survey of the statistical theory of shape. *Statist. Sci.*, 4(2):87–99, 05 1989.

[Ken15] A. Kendall, M. Grimes, and R. Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 2938–2946, 2015.

[Kin05] B. J. King, T. Malisiewicz, C. V. Stewart, and R. J. Radke. Registration of multiple range scans as a location recognition problem: hypothesis generation, refinement and verification. In *Fifth International Conference on 3-D Digital Imaging and Modeling (3DIM'05)*, pp. 180–187, 2005.

[Kin14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

 ISBN 978-80-86943-40-4

[Kow15] M. Kowalski, J. Naruniec, and M. Daniluk. Livescan3d: A fast and inexpensive 3d data acquisition system for multiple kinect v2 sensors. In *2015 International Conference on 3D Vision*, pp. 318–325, 2015.

[Kum11] R. Kummerle, G. Grisetti, H. Strasdat, K. Konolige, and W. Burgard. G2o: A general framework for graph optimization. In *2011 IEEE International Conference on Robotics and Automation*, pp. 3607–3613, 2011.

[Lon15] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

[Low04] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[Mar63] D. W. Marquardt. An algorithm for least-squares estimation of nonlinear parameters. *Journal of the society for Industrial and Applied Mathematics*, 11(2):431–441, 1963.

[Mel17] I. Melekhov, J. Ylioinas, J. Kannala, and E. Rahtu. *Relative Camera Pose Estimation Using Convolutional Neural Networks*, pp. 675–687. Springer International Publishing, 2017.

[Nak17] Y. Nakajima and H. Saito. Robust camera pose estimation by viewpoint classification using deep learning. *Computational Visual Media*, 3(2):189–198, 2017.

[New15] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 343–352, 2015.

[Owe15] J. L. Owens, P. R. Osteen, and K. Daniilidis. Msg-cal: Multi-sensor graph-based calibration. In *2015 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3660–3667, 2015.

[Poi16] P. Poirson, P. Ammirato, C. Y. Fu, W. Liu, J. Kosecka, and A. C. Berg. Fast single shot detection and pose estimation. In *2016 Fourth International Conference on 3D Vision (3DV)*, pp. 676–684, 2016.

[Pom13] F. Pomerleau, F. Colas, R. Siegwart, and S. Magnenat. Comparing icp variants on real-world data sets. *Auton. Robots*, 34(3):133–148, 2013.

[Rub11] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pp. 2564–2571, Washington, DC, USA, 2011. IEEE Computer Society.

[Sho13] J. Shotton, B. Glocker, C. Zach, S. Izadi, A. Criminisi, and A. Fitzgibbon. Scene coordinate regression forests for camera relocalization in rgb-d images. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2930–2937, 2013.

[Van17] F. Vasconcelos, J. P. Barreto, and E. Boyer. Automatic camera calibration using multiple sets of pairwise correspondences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99):1–1, 2017.

[Was16] O. Wasenmüller, M. Meyer, and D. Stricker. Corbs: Comprehensive rgb-d benchmark for slam using kinect v2. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pp. 1–7. IEEE, 2016.

[Xu17] C. Xu, L. Zhang, L. Cheng, and R. Koch. Pose estimation from line correspondences: A complete analysis and a series of solutions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1209–1222, 2017.

[Ye13] G. Ye, Y. Liu, Y. Deng, N. Hasler, X. Ji, Q. Dai, and C. Theobalt. Free-viewpoint video of human actors using multiple handheld kinects. *IEEE Transactions on Cybernetics*, 43(5):1370–1382, 2013.

[Zam16] A. R. Zamir, T. Wekel, P. Agrawal, C. Wei, J. Malik, and S. Savarese. *Generic 3D Representation via Pose Estimation and Matching*, pp. 535–553. Springer International Publishing, Cham, 2016.

[Zha00] Z. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1330–1334, 2000.

[Zio16] N. Zioulis, D. Alexiadis, A. Doumanoglou, G. Louizis, K. Apostolakis, D. Zarpalas, and P. Daras. 3d tele-immersion platform for interactive immersive experiences between remote users. In *2016 IEEE International Conference on Image Processing (ICIP)*, pp. 365–369, 2016.

[Zol14] M. Zollhöfer, M. Nießner, S. Izadi, C. Rehmann, C. Zach, M. Fisher, C. Wu, A. Fitzgibbon, C. Loop, and C. Theobalt. Real-time non-rigid reconstruction using an rgb-d camera. *ACM Transactions on Graphics (TOG)*, 33(4):156, 2014.

 ISBN 978-80-86943-40-4