ORIENTATION OF POINT CLOUDS FOR COMPLEX SURFACES IN MEDICAL SURGERY USING TRINOCULAR VISUAL ODOMETRY AND STEREO ORB-SLAM2

O.Kahmen [1]*, N. Haase [1], T.Luhmann [1]

[1] Jade University of Applied Sciences, Institute for Applied Photogrammetry and Geoinformatics (IAPG), Ofener Str. 16/19, 26121 Oldenburg, Germany - (Oliver.Kahmen, Niklas.Haase, Thomas.Luhmann)@jade-hs.de

Commission II - WGII/1

KEY WORDS: Trinocular visual odometry, ORB-SLAM2, point cloud orientation, knee joint replacement

ABSTRACT:

In photogrammetry, computer vision and robotics, visual odometry (VO) and SLAM algorithms are well-known methods to estimate camera poses from image sequences. When dealing with unknown scenes there is often no reference data available and also the scene needs to be reconstructed for further analysis. In this contribution a trinocular visual odometry approach is implemented and compared to stereo VO and ORB-SLAM2 in an experimental setup imitating the scene of a knee replacement surgery. Two datasets are analysed. While a test-field provides excellent conditions for feature detection algorithms with its artificial texture assembled, extracted images show the knee joint itself solely in order to use only the homogenous, but in real application stable, region of the knee joint. The camera trajectories of VO and ORB-SLAM2 are transformed to corresponding coordinate systems and are subsequently evaluated. The tracking algorithms show poor quality when only the inappropriate surface of the knee is used but perform well when the artificial texture of the test-field is used. The third camera does not lead to a significant advantage in this setup using our implementation. Possible reasons, e.g. less overlap, are discussed in this contribution. Nevertheless, the quality of the oriented point clouds, obtained by trinocular dense matching, is less than 1mm for most of the analysed data. The experiment will be used to focus on further developments, e.g. dealing with specular reflections, and for evaluation purposes using different SLAM/ VO algorithms.

1. INTRODUCTION

Visual Odometry (VO) and Simultaneous Localization and Mapping (SLAM) are sequential orientation techniques, well-known in robotics, computer vision and photogrammetry. As an example, a robot tries to estimate its own position and orientation (SLAM and VO) while simultaneously estimating a map of its environment (SLAM).

Especially with the increasing importance of automation, VO and SLAM play crucial roles for navigation tasks to researchers of many disciplines. Robots, cars and other sensor systems use them to navigate themselves in often unknown environments. Especially when other sensors like GNSS or inertial measurement units (IMU) are inappropriate, like in high-accuracy medical surgery applications, visual navigation becomes increasingly important.

This contribution focusses on the spatial orientation of cameras using VO and the state-of-the art SLAM implementation ORB-SLAM2 (Mur-Artal, Tardos, 2017). The exterior orientation of cameras (6DOF) can be compared to a reference, generated by bundle adjustment using circular photogrammetry targets. The goal is to obtain robust and accurate camera orientations without any need for external markers in order to register single point clouds, created by dense matching. The camera system is designed for knee replacement surgeries. In general, the goal of the provided application is to reconstruct the surface of the knee joint in order to navigate medical instrumentations. The absolute location of the reconstructed 3D point cloud with respect to the real time position of the camera is important, since it must be registered with data from computer tomography and it is used to locate medical instruments to cut the bone at the right position. For those reasons, classical post-processing SfM approaches are not suitable for such applications. A fast (not strict real-time) (semi-) dense point cloud calculation as well as real-time pose estimation is needed for the application of a knee replacement. Furthermore, visual localisation techniques have the potential to be used in Augmented Reality (AR) or Virtual Reality (VR) (Khor et al., 2016) for the purpose of marking surgical incisions virtually and locate surgical instruments.

The application behind this research is the surgery of a knee joint. Single camera poses (translation and rotation) are tracked during the scanning process. The optical system consists of three cameras. Some of the image triples are used to conduct a dense matching of the surface of a knee joint. Single dense clouds are oriented via visual odometry to get one oriented 3D surface reconstruction. The scan itself takes only a few seconds and a few image triples. Beyond that, the camera system is tracked along in order to place medical instruments during and past scanning. The target quality of both 3D reconstruction and real-time tracking of medical instruments is specified to ~1mm.

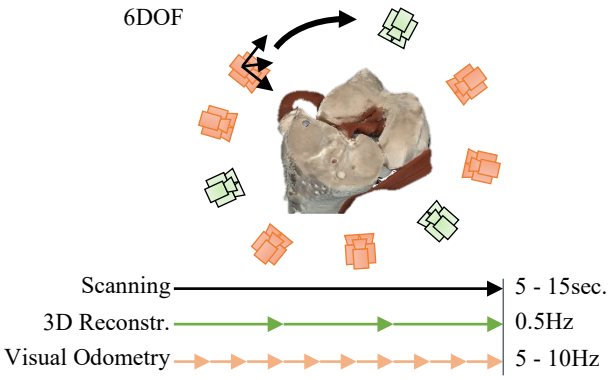

Figure 1. Schematic graphic of the scanning process and chronological arrangement of reconstruction and tracking

* Corresponding author

2. VISUAL ODOMETRY AND ORB-SLAM2

Two methods are used in this work to determine camera trajectories. Firstly, visual odometry is used and implemented as stereo and trinocular solution and secondly, we computed the 6DOF of moving cameras via ORB-SLAM2. These two algorithms differ in their approach of camera pose estimation. While pose recognition is a main feature of ORB-SLAM2 algorithm to calculate loop closures, VO concatenates single estimated poses without conducting a global adjustment of all frames. In general, for both techniques similar requirements do exist in order to establish point correspondences between image frames. The scene should be geometrically stable during the acquisition, since reconstructed 3D points need to be re-identified over time. Furthermore, radiometric and perspective changes of feature points between frames should be small to be able to track these. For SLAM the scene needs to be stable between several key frames in order to perform loop closure, whereas for VO the scene needs to be stable only between single frames. As stated in Song et al. (2018), loop closing is difficult in surgical vision due to deforming scenarios.

Many feature detectors exist in field of computer vision having varying properties regarding the needs for geometric and radiometric parameters of image sequences, as evaluated for some by Schmidt et al. (2010).

2.1 Visual odometry

Visual odometry is a well-known technique for estimating visual sensor poses by an acquired image sequence. In literature, usually monocular and stereo VO are separated. In stereo VO, thus also in trinocular VO, forward intersection can be applied, since a relative orientation of the stereo system is available and provides scale. To locate a subsequent epoch, spatially intersected points can be used to perform a 3D transformation. An extensive overview of related work can be found in Yousif et al. (2015).

The system used in this work consists of two monochromatic cameras (1 and 2), and one RGB camera (3) as explained further in chapter 3. The RGB image differs in contrast to the two monochromatic images after converting to an 8-bit greyscale image. Due to different sensors, the images undergo a histogram equalisation before features are detected (see Figure 2). Also, for reasons of comparability, the equalisation is performed when using just two images in stereo VO even though it would not be necessary. The feature detector used is *Good features to track (gftt)* by Shi and Tomasi (1994). The optical flow is performed for each camera between epoch *e* and *e-1* using the algorithm of Bouguet (2001).

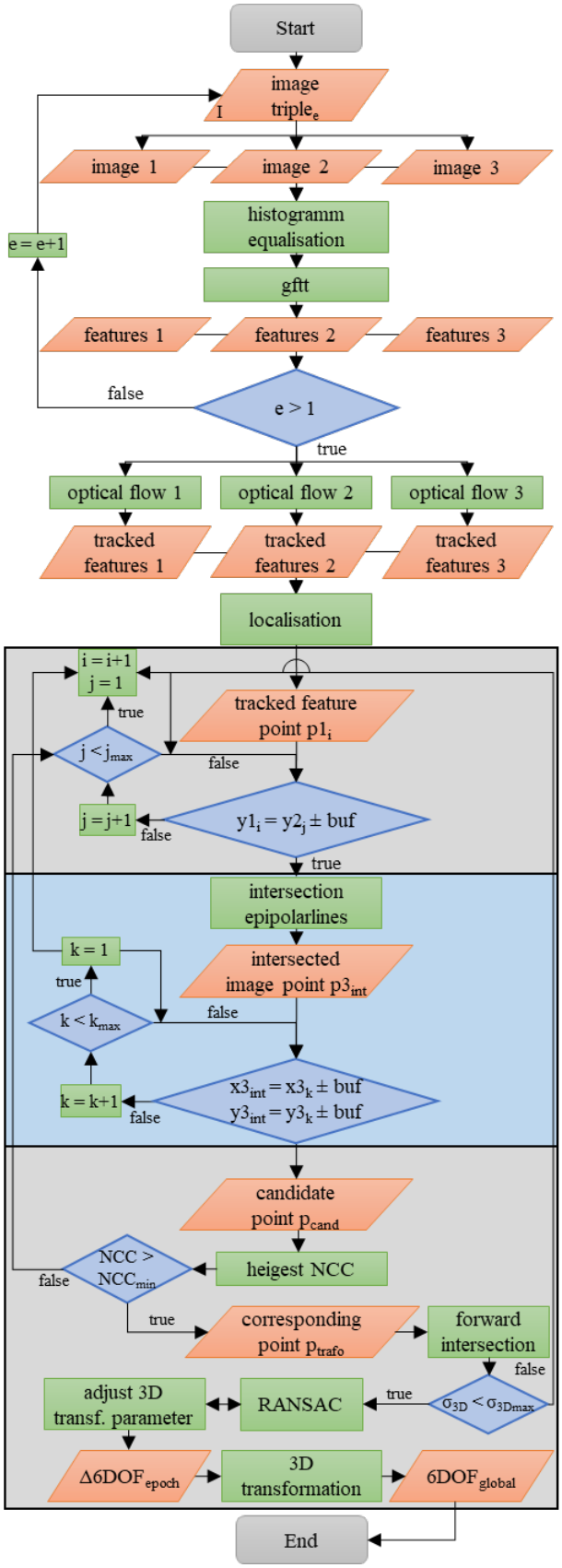

Figure 2. Flowchart of trinocular visual odometry algorithm. The part of the localisation is highlighted in grey. The part with blue background is applied when trinocular VO is conducted. In stereo VO this part is skipped, since only two images are present.

2.1.1 Stereo: Figure 2 shows the flowchart including the main steps to get the 6DOF trajectory of a sequence of image triples. The chart shows the process for image triples (image 1, 2, 3) but is also valid for a stereo system when the third image and the trinocular part of the localisation (marked blue in Figure 2) are excluded. Also in stereo, just one normalised cross correlation (NCC) between image 1 and 2 is checked, while in trinocular VO all three NCC (1→2, 1→3 and 2→3) are tested against the minimum.

After detecting good features to track in each image, the optical flow of each feature is calculated if possible, according to the pyramidal implementation of the Lucas-Kanade tracker (Bouguet, 2001). The tracked features of image 1 and 2 become part of the localisation. A feature of image 1 is checked against all features of image 2. If a feature of image 2 lies on the epipolar line (± buffer), it is saved as a potential corresponding point. Since normalised images are used, one row of pixels in the image indicates the horizontal epipolar line, thus only the y value needs to be checked in correspondence analysis. After this step, each feature point of image 1 has a certain number of candidates (0-n) in image 2. If candidates are found, the NCC is calculated for each candidate using a certain number of grey values in a window surrounding the feature points. The candidate with the highest value, thus representing the best match of all candidates, is checked against a threshold (e.g. 0.7). If the NCC is larger than that, a corresponding point was found and the image coordinates of the features in image 1 and 2 are used in forward intersection. Then, the 3D point is checked against a minimum for its standard deviation. This check correlates highly with the difference in y coordinates between image 1 and 2, which is limited by the buffer. In contrast to the check of the epipolar line correspondence, the standard deviation of a 3D point is dependent on the disparity, thus the depth in object space. Passed points then becomes part of the point set for the 3D similarity transformation. To avoid errors in correspondences and ensure a certain quality level, a RANSAC search is implemented before calculating the final 6DOF transformation from one epoch to another (Δ6DOF in Figure 2). To calculate the global position, each set of transformation parameters of three translations (X, Y, Z) and three rotations (ω, φ, κ) is concatenated to 6DOF of the previous epoch by applying a 3D transformation in each epoch.

$$\boldsymbol{T}_{\boldsymbol{glob}_e} = \boldsymbol{T}_{\boldsymbol{glob}_{e-1}} \cdot \boldsymbol{T}_{\boldsymbol{epoch}_e}^{-1} \quad (1)$$

where

$$\boldsymbol{T} = \begin{pmatrix} r_{11} & r_{12} & r_{13} & t_X \\ r_{21} & r_{22} & r_{23} & t_Y \\ r_{31} & r_{32} & r_{33} & t_Z \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

describes respectively the global transformation matrix ($\boldsymbol{T}_{glob}$) or the one between two consecutive epochs ($\boldsymbol{T}_{epoch}$).

2.1.2 Trinocular: The principle of trinocular VO is similar to stereo VO. Only by the optical flow tracked features in camera 1, 2 and 3 become part of the localisation (according to Figure 2). Besides the third image, the blue part of Figure 2 is added in trinocular VO compared to stereo VO, and the NCC check is conducted for each combination of image pairs (1→2, 1→3 and 2→3).

If a point is tracked in camera 1, it is checked for tracked features of camera 2 if they lie on the epipolar line. If this check is positive, the epipolar lines 3→1 and 3→2 are intersected to get the corresponding position of these feature points in image 3. The NCC is calculated for all three image combinations. Only if all three NCC are above the threshold NCC_{min}, the corresponding feature points become part of the forward intersection. The standard deviation of the intersected 3D point is tested against a threshold. When forward intersection is applied using six observations instead of four as in stereo, the standard deviation is expected to be smaller. Also, this indicator can be seen as more reliable than in stereo case, because poorly matched features might have a good standard deviation as long as they lie on the epipolar line ± buffer. Stereo points can still be incorrect in absolute coordinate values when incorrect matches are used in forward intersection.

Trinocular rectification and intersection of epipolar lines is performed as described by Conen et al. (2016) (Conenetal.,É. The idea of the introduction of a third camera regarding the VO is to obtain more robust feature points and exclude errors in correspondences. In literature, the additional value of a third camera was demonstrated in different articles. For instance, Maas (1997) gives an extensive overview of trinocular and multi-ocular configurations and their characteristics. Conen (2020) implements a trinocular dense matching for a surgical endoscope. Proven benefits of using a third camera are an improved dense matching and the possibility to colourise point clouds independently, which offers possible classification solutions. In this contribution the effect in visual odometry and the possibilities to overcome typical ambiguity of stereo systems are investigated.

2.2 ORB-SLAM2

ORB-SLAM is well-established in robotics and computer vision. It was published by Mur-Artal et al. (2015) and further developed in Mur-Artal and Tardos (2017). ORB-SLAM2 in this work uses the two monochromatic cameras as a stereo system. This method is used to compare it against the implemented VO, since it is supposed to work with invariance to illumination and perspective variations. Based on the ORB feature detector (Rublee et al., 2011), ORB-SLAM2 can be performed for mono, stereo and RGB-D camera systems in real-time. The main three aspects of this algorithm are tracking, mapping and loop closing. During the tracking process, key frames are extracted for the mapping. Tracked features are mapped in 3D using these key frames based on a local bundle adjustment. Key frames are checked constantly, if a loop closure is possible. If a loop is found, a global bundle adjustment determines the camera poses of key frames and the exact 3D coordinates of already mapped feature points.

3. EXPERIMENTAL SETUP WITH TRINOCULAR SYSTEM

The experimental setup is realised in order to test algorithms in a known environment. The reference for exterior orientations is calculated via bundle adjustment using circular photogrammetry targets on a multi-functional test frame (see Figure 6) with the software AICON 3D Studio. With this setup, flexible image configurations can be realised in order to identify optimisation potentials, and find the best setup and parameters for certain tracking and mapping methods. The accuracy can be quantified as absolute values, since corresponding coordinate systems of different localisation methods and an absolute reference dataset are created. A pre-calibrated prototype of a trinocular system is used to acquire image triples.

3.1 Trinocular system

The trinocular system consists of two monochromatic cameras (camera 1, 2) and one RGB camera (camera 3). Figure 3 shows the prototype of the trincoular system mounted on a variable friction arm observing an artificial knee joint.

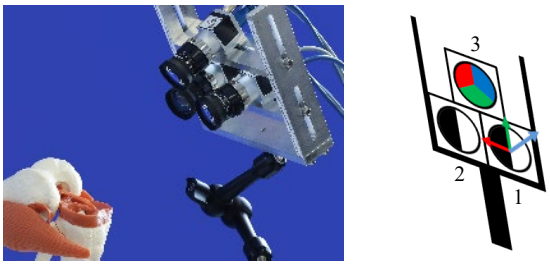

Figure 3. Trinocular camera system observing artificial knee joint (left). Schematic arrangement of two mono- and one RGB camera/s (right).

For the application of the knee replacement surgery, the three cameras are arranged in an equilateral triangle having a baseline of ~54mm. The angle of convergence between camera 1 and 2 is ~14° around Y and 0° around X and Z axis. Between camera 1 and 3 the angle is ~5° around Y, ~13° around X and 0° around Z (see Figure 3). All cameras use a lens with ~10mm focal length. At a distance of 185mm this setup results its highest trinocular relative overlap of 93%. The overlap drops with increasing or decreasing acquisition distance, thus the potential area in images for corresponding features shrinks. Figure 4 shows the trinocular overlap exemplarily for three acquisition distances at 100, 200 and 300mm. The cyan coloured area shows that the overlapping image content is the largest at 200mm.

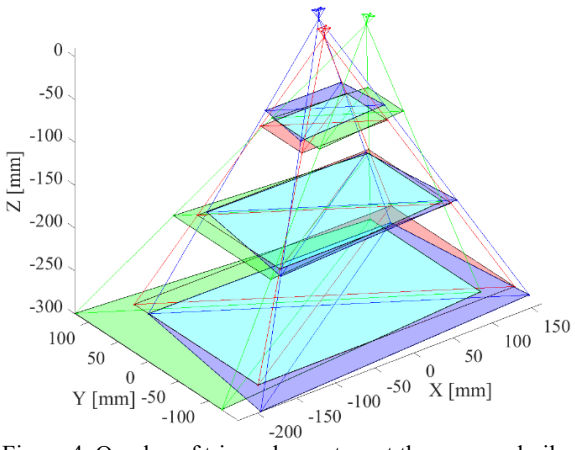

Figure 4. Overlap of trinocular system at three exemplarily acquisition distances. The cyan area indicates the overlapping area of three single images (red, green, blue).

Figure 5 shows the percentage of image overlap for stereo and trinocular configurations for the calibrated setup used in this work. It can be seen that the trinocular overlap is less at all distances. A reasonable overlap can only be reached in a small area around the peak at 185mm. Especially when images are acquired at smaller distances, the trinocular overlap decreases rapidly. Depending on the angular orientation, the length of baselines and the interior orientations, the curve of overlap results to a larger or smaller extent of the acquisition distance.

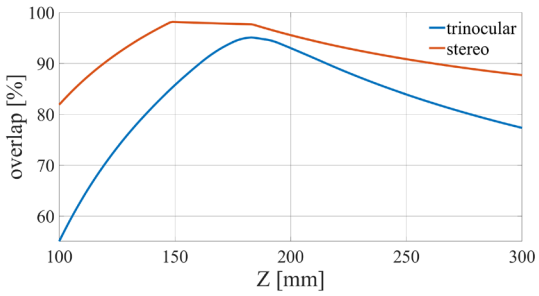

Figure 5. Relative overlap of stereo and trinocular setup as a function of acquisition distance (Z). Orange shows the overlap of camera 1 and 2, blue the overlap of camera 1, 2 and 3.

3.2 Experimental setup

An artificial knee joint is used to get realistic looking images in a controlled environment. Furthermore, a multi-functional test frame provides reference data as well as artificial texture for VO respectively SLAM. The trinocular system is placed in five different positions pointing towards the knee joint.

For each of the five positions of the trinocular system (Figure 6, left, red dots), 360 images at angular step width of 1° of the rotary table are acquired, leading to 5 × 360 = 1800 images for each of the three cameras. As can be observed in Figure 6 (top right), the acquisition positions result in a half-sphere like shaped arrangement consisting of five rounds from round 1 (top view) to round 5 at ~45° observation angle.

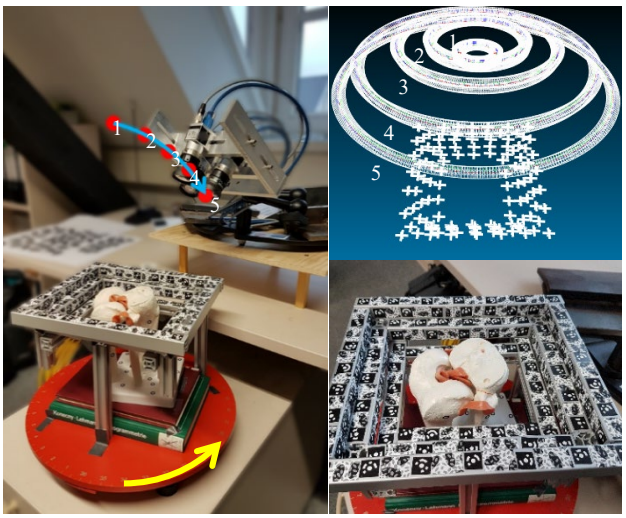

Figure 6. Left: Experimental setup. The trinocular system observes the test-field and the artificial knee joint from five positions (red dots, blue arrow). The test-field rotates in 1° steps around its vertical axis (yellow arrow).

The frame (Figure 6, bottom right) is designed for this experiment. It contains 120 photogrammetric circular targets, which can be detected automatically and measured within the images with high accuracy. In addition, between these targets an artificial texture is installed to provide good corners for feature detection in different image pyramids. Also, the corners of the square-shaped photogrammetric targets can be used by feature detector algorithms.

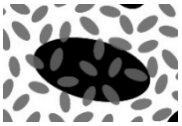

Figure 7. Artificial texture of randomly oriented ellipses in different scale

Due to the grandstand-like form of the test-field, the knee in the centre as well as targets and patterns are visible at all observation angles without too much shadowing. Also, different heights, similar to the physical expansion of a bended knee, are present to stabilise the bundle adjustment.

3.3 Datasets

VO and ORB-SLAM2 use the artificial texture to estimate orientation parameters of the cameras. Furthermore, only the knee itself can be extracted in the images and used for these algorithms to calculate orientation parameters.

Therefore, two oriented image triples from opposite directions are used to reconstruct the surface of the knee joint via dense matching. The obtained 3D model of the knee joint is used to project the object points into every single image of the 3 × 1800 images of the experiment. The point cloud is subsampled to speed up the process. Since the projected points represent only a small number of the relevant points, the mask is filtered. A morphological filter closes the gaps in the mask. Gaussian blur is added at the borders of the mask to make sure that feature detectors will not use these sharp edges to identify feature points. The original image is weighed using the mask as a weighting function. Only the knee and a small blurred edge are visible in the extracted image. Hence, these images have the same orientation as the original images referenced via bundle adjustment. Besides original images, these filtered images represent the second set of images used for evaluation.

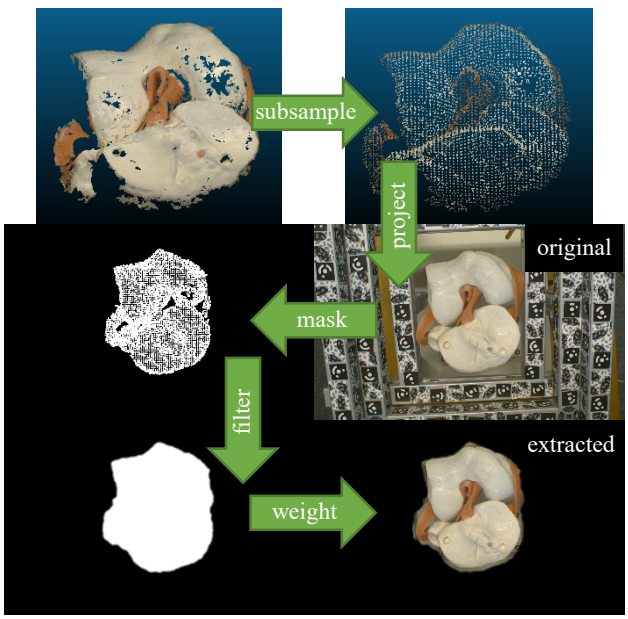

Figure 8. Knee joint extraction from 3D point cloud. A rough cloud is computed by two image triples, subsampled and projected into one original image. The projected points are coloured white as a mask. The mask is closed and blurred at the edges before used as a weighting function to extract only the knee joint of the original image.

The setup provides excellent conditions for feature-based algorithms when the artificial texture is used and also realistic conditions when only the knee surface is used for feature detection. Thus, seven main datasets can be separated, whereas number 1 poses the reference dataset while the remaining six datasets are evaluated in different subsets.

Set	Texture for orientation	Orientation technique
1	Circular targets	Bundle adjustment
2	Frame	VO_stereo
3	Frame	VO_trinocular
4	Frame	ORB-SLAM2
5	Knee joint	VO_stereo
6	Knee joint	VO_trinocular
7	Knee joint	ORB-SLAM2

Table 1. Datasets evaluated in this work. Dataset 1 poses the reference dataset.

4. ORIENTATION OF THE CAMERA SYSTEM

4.1 Transformation

To be able to compare absolute coordinates of the camera positions, the coordinate systems of different orientation techniques need to be transformed so their origin and orientation coincide. The coordinate system of VO is located in the first camera triple (camera 1) at the principal point. It is desired to locate the coordinate systems of all datasets in the first image of camera 1 of the particular dataset.

The coordinate system of the camera positions of dataset 1 according to Table 1 can be transformed to the first camera by a simple 3D transformation.

$$\boldsymbol{T}_{trans} = \boldsymbol{T}_{origin}{}^{-1} \cdot \boldsymbol{T}_{p} \tag{2}$$

$\boldsymbol{T}$ describes respectively the transformation matrix of the origin where the system is supposed to be transformed to, or the individual matrix of a camera pose p of the dataset 1. The transformed rotation matrix, thus the Euler angles, and the translation vector in the transformed coordinate system can then be extracted of $\boldsymbol{T}_{trans}$. Appling this transformation, the dataset of the bundle adjustment of the software AICON 3D Studio now also have its origin in the first image of camera 1.

Although the origin of ORB-SLAM2 is located in the first image of camera 1, we need to consider different orientations of the axes of ORB-SLAM2 definition. In contrast to the VO and bundle adjustment solution, here the coordinate system is defined in the rectified images instead. To transform the trajectory of ORB-SLAM2 into the same coordinate system as VO the following transformation needs to be applied.

$$\boldsymbol{T}_{trans} = \boldsymbol{T}_{otrans}{}^{-1} \cdot \boldsymbol{T}_{p} \cdot \boldsymbol{T}_{otrans} \tag{3}$$

where

$$\boldsymbol{T}_{otrans} = \boldsymbol{T}_{rotX} \cdot \boldsymbol{T}_{rect} \tag{4}$$

where

$$\boldsymbol{T_{rotX}} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

and

$$\boldsymbol{T_{rect}} = \begin{pmatrix} r_{11} & r_{12} & r_{13} & 0 \\ r_{21} & r_{22} & r_{23} & 0 \\ r_{31} & r_{32} & r_{33} & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

where r_{ij} represent the complete rotation of the spatial coordinate transformation, defined by three individual rotations. See e.g. (Luhmann et al., 2020) for further details.

$$\boldsymbol{R} = \boldsymbol{R_{\omega}} \cdot \boldsymbol{R_{\varphi}} \cdot \boldsymbol{R_{\kappa}} \quad (5)$$

The angles ω, φ, κ representing the individual rotations describe the Euler angles of the relative orientation of the rectified camera 2.

After transforming the results of all tracking methods, the absolute positions and orientations of the trajectories can be evaluated and absolute deviations can be quantified.

4.2 Quality evaluation

The reference is given by the bundle adjustment using the circular targets. The positions of the targets are known with a higher quality of 0.01mm. The reference exterior orientations of the 5 × 3 × 360 images are given with an accuracy of 0.02mm in translation and 0.002° in rotation, thus much better than VO and SLAM results can be expected.

The data is analysed for each round (see Figure 6) individually, since the perspective difference from one round to another would introduce a large perspective "jump", which would be against the idea of VO respectively SLAM. Even though, SLAM algorithm are able to process such data thanks to place recognition mechanism. For reasons of clarity in this work only round 1 and 5 will be presented as the extrema regarding the acquisition angles from top view (round 1) to ~45° side view (round 5). For each of the 360 positions in one acquisition set, the 3D error of the position of the cameras is calculated. The root mean square (RMS), the maximum error and the used number of transformation points between two epochs are given in Table 2. The number of transformation points represents the average number of points finally used for the rigid 3D transformation between two epochs. Hence, RANSAC and the filtering for standard deviation in forward intersection of matched feature points thin out the tracked points intensively before the final 3D transformation is applied. In case of ORB-SLAM2 data, the sum of matches and VO matches according to the frame drawer is given.

The absolute deviations to the reference camera positions are much smaller when the artificial texture on the frame is used. The gradients in the original images are much higher allowing to detect more and stronger features. Besides the number and quality, the distribution in the images and in object space is much better when the frame is used. Thus, the datasets with frame texture cannot be compared to the knee datasets strictly. Nevertheless, the knee datasets give an idea of how accurate such a setup can be in real applications assuming an unstable, thus not usable, surrounding area. Furthermore, some effects can be observed in these data as discussed in the following.

No.	Set	Round	Technique	Texture	RMS 3D [mm]	Max 3D [mm]	Trans. Pt.
1	2	1	VO_stereo	Frame	0.16	0.25	963
2	3	1	VO_tri	Frame	0.48	0.73	595
3	4	1	OS2	Frame	0.47	1.04	708
4	2	5	VO_stereo	Frame	0.45	0.76	741
5	3	5	VO_tri	Frame	1.18	1.82	467
6	4	5	OS2	Frame	0.30	0.58	712
7	5	1	VO_stereo	Knee	0.86	1.65	232
8	6	1	VO_tri	Knee	0.77	1.59	124
9	7	1	OS2	Knee	6.57	17.35	700
10	5	5	VO_stereo	Knee	2.29	3.70	53
11	6	5	VO_tri	Knee	2.72	4.26	57
12	7	5	OS2	Knee	10.88	25.86	663

Table 2. Deviations of actual trajectories to nominal trajectories by photogrammetric bundle adjustment and used number of transformation points. First six rows represent original image sets (2, 3, 4), last six the extracted image sets (5, 6, 7) according to Table 1.

The trinocular configuration leads to less accurate results when the frame is used. The main reason for that is suspected to be the missing overlap. Since high quality features can be detected easily and stated as geometrically stable, the benefit of a third camera regarding the robustness in correspondence analysis becomes small. When comparing VO in the knee datasets the differences of stereo and trinocular in positioning are more similar. Since the knee is located in the image centre, the overlap of stereo and trinocular is 100% in both cases. However, no improvement can be observed when using three instead of two cameras. Since RANSAC and other filtering methods are implemented, there is no improvement due to more robust features in trinocular configuration. When using three images, less features need to be filtered by RANSAC or other filtering methods, since the chosen features are more reliable. Thus, the trinocular setup gives more flexibility regarding the parameters of RANSAC, epipolar line buffer and NCC threshold. That might be an advantage over stereo when dealing with a more inhomogeneous image acquisition as when a real handheld system is used. However, it is not evaluated further in this work. When tuning VO parameters, the results can be improved slightly for some setups and more features could be found. However, it could be observed that the results are very sensitive for parameter tuning.

The ORB-SLAM2 performs well for the frame datasets. In case of round 5 it results as best localisation method, while in round 1 it performs similar to the trinocular VO. When only the knee joint is used, ORB-SLAM2 struggles as well as VO. The deviations are too high to represent reasonable trajectories. Even when tuning the parameters for the FAST operator to take the lower contrast into account, no reasonable results can be obtained using ORB-SLAM2. It should be noted that this case of moving around a stable scene, thus observing the same scene in every frame, is not a typical SLAM application. Still, ORB-SLAM2 is able to processes a section of these data without additional coding and only fails when the features are weak.

Neither VO nor ORB-SLAM2 are able to reach the goal of 1mm quality for the trajectories when only the knee joint is used. For both VO and SLAM the specular reflections and homogenous

surface are barley to handle. While the VO does not care globally for moving reflections, since it transforms frame to frame, ORB-SLAM2 matches features with 3D points from the map. Due to a quasi-dynamic surface created by moving reflections, this might lead to large errors. As stated in Ibragimov and Afanasyev (2017), ORB-SLAM is not recommended to use in flat, homogenous environment. For real applications the texture on the knee could be improved using a stationary artificial light as successfully done in Qiu and Ren (2018).

Since the positioning only represents three of the six DOF, the orientation of calculated point clouds will be evaluated in the following to quantify the final deviation in object space with respect to the determined Euler angles.

5. POINT CLOUD ORIENTATION

The idea of the proposed approach is to have multiple possibilities to generate a dense point cloud whenever it is necessary in terms of accuracy and completeness. At each image triple of a sequence this possibility is given. The point clouds used in this work are computed via trinocular dense matching as described in Conen et al. (2016) and Conen (2020).

To evaluate the 6DOF tracking quality, thus the point cloud orientation quality, in this work three single point clouds are generated at different angles of the circular image sequence. As shown in Figure 9 three single point clouds, oriented via VO respectively SLAM, lead to one overlapping point cloud. Note that the single point clouds contain holes. These holes mainly arise due to specular reflections on the lateral and medial condyle of the femur. In such an application it is advisable to ignore saturated image parts rather than trying to overcome homogenous areas by classical semi-global optimisation. Since the optical system is moved around the object, the saturations will also move in the images and there will be some position at where the part of the object is more reliable to reconstruct.

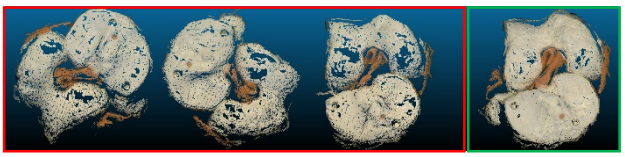

Figure 9. Single point clouds from individual image triplets at positions 120°, 240°, 360° of the circular image sequence (framed red). Oriented point clouds, transformed into first camera according to VO/ SLAM transformation matrix for individual 6DOF of the image triples (framed green).

To quantify the orientation quality of the 3D reconstruction the three single point clouds at 120°, 240° and 360° (last epoch) are transformed using the transformation matrices of determined trajectories (see Table 2) and compared to the transformed point clouds of the reference dataset. This comparison includes all 6DOF obtained by tracking method (absolute error) and shows how the deviations in tracking effect the point cloud orientations.

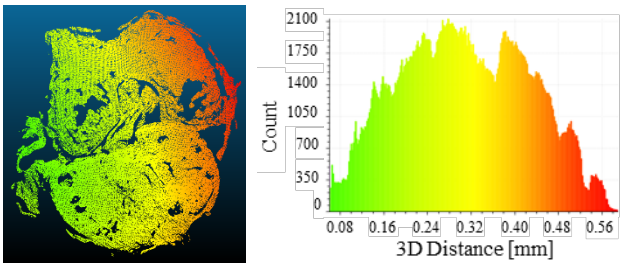

Figure 10. Euclidian distances between corresponding points of two point clouds. Exemplarily, the oriented cloud of the reference dataset at 120° is compared to the oriented cloud of VO_tri (No. 7 in Table 2). The right image shows the histogram of the 3D distances.

In Figure 10, the Euclidian distances between oriented reference and oriented point cloud using VO are visualised. It can be seen that the exemplarily shown point cloud of VO is tilted so that deviations to the nominal point cloud increase diagonal from left to right. The mean 3D distance of 120°, 240° and 360° oriented point clouds for each dataset of Table 2 is given as well as the maximum of all distances of these three oriented clouds in Table 3. Since the ORB-SLAM2 datasets of the knee-only datasets result in poor quality, these datasets are not analysed any further.

No.	Set	Round	Technique	Texture	RMS 3D [mm]	Max 3D [mm]
1	2	1	VO_stereo	Frame	0.13	0.24
2	3	1	VO_tri	Frame	0.12	0.22
3	4	1	OS2	Frame	0.14	0.21
4	2	5	VO_stereo	Frame	0.24	0.35
5	3	5	VO_tri	Frame	0.56	0.77
6	4	5	OS2	Frame	0.22	0.35
7	5	1	VO_stereo	Knee	0.33	0.79
8	6	1	VO_tri	Knee	0.39	0.82
9	7	1	OS2	Knee	/	/
10	5	5	VO_stereo	Knee	0.53	1.41
11	6	5	VO_tri	Knee	0.54	1.20
12	7	5	OS2	Knee	/	/

Table 3. Mean and maximum Euclidean distances of corresponding points of oriented point clouds. RMS and Max are calculated using three transformed point clouds at angular steps 120°, 240° and 360°.

The Euclidean distances in Table 3 are mostly lower than the positioning error of the trajectory (see Table 2). The numbers show that the transformed point clouds fit comparatively well to the reference dataset. For all datasets only the knee, thus mainly the image centre, is part of the point cloud as in Figure 10. The effect of poor orientation is comparatively small at the image centre. Hence, differences in the Euler angles of respective tracking result do have a minor effect on the point cloud. In general, the point clouds orientation quality is within 1mm except for the datasets 10 and 11.

The error of the dense matching is not taken into account in previous analysis. The result of a final point cloud of course does underlie the accuracy discrepancies of single point clouds too. The accuracy of a single point cloud can be expected to be 0.1 - 0.3mm in average as empirical studies based on the data of

this contribution shows. Nevertheless, multiple point clouds can be combined to a redundant and robust final point cloud complementing each other at incomplete or fragile areas on the knee joint.

6. CONCLUSIONS AND OUTLOOK

A state of the art SLAM algorithm and an own implementation of a specific trinocular visual odometry are used to track camera positions in a complex environment with the goal not to use any artificially assembled targets. The presented experiment is feasible to evaluate localisation quality and represents a realistic scene of a knee joint surgery. The developed test-field gives optimised conditions for feature tracking, thus for visual localisation methods. Based on two datasets, first evaluation results based on the experiment are presented in this work focussing on stereo and trinocular visual odometry. ORB-SLAM2 as a state of the art SLAM algorithm performs well when the texture frame is used, but it fails when only the knee is used in extracted images. The knee joint itself is a challenging object for image matching methods, since it is homogenous, almost flat and creates specular reflections. Furthermore, it might be partly dynamic in real scenes, hence a classical loop closure seems to be unsuitable. When acquisition conditions (frame rate, moving speed, constant acquisition distance, overlap) and parameter settings of tracking method (RANSAC, NCC, buffer, etc.) are respected, the used setup leads to reasonable tracking results in this laboratory setup. However, the system needs to be evaluated in real application handheld scenarios.

In this work, no major advantages of trinocular compared to stereo configuration could be found regarding the localisation methods for named reasons. Nevertheless, besides a more robust dense matching, the third camera gives other advantages in medical surgery and visual applications. In future work, other methods like LSM should be used to improve the VO using iterative feature matching focussing on the specular reflections inevitable in medical surgery. Beyond that, algorithms dealing with dynamic surfaces (e.g. Song et al., 2018; Xiao et al., 2019) need to be evaluated in more real environments and compared with implemented VO solution.

ACKNOWLEDGEMENT

This work is based on a project funded by the European Regional Development Fund (EFRE, ZW 6-85007826).

7. REFERENCES

Bouguet, J.-Y., 2001. Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the algorithm. Intel Corporation, Microprocessor Research Labs.

Conen, N., Jepping, C., Luhmann, T., Maas, H.-G., 2016. Rectification and robust matching using oriented image triplets for minimally invasive surgery. ISPRS Annals of Photogrammetry, Remote Sensing and Spatial Information Sciences III-3, 27–34.

Conen, N.P., 2020. Konzeption und Entwicklung eines trinokularen Endoskops zur robusten Oberflächenerfassung in der minimalinvasiven Chirurgie, https://nbn-resolving.org/urn:nbn:de:bsz:14-qucosa2-380439. (Accessed 31 January, 2020).

Ibragimov, I.Z., Afanasyev, I.M., 2017. Comparison of ROS-based visual SLAM methods in homogeneous indoor environment. 2017 14th Workshop on Positioning, Navigation and Communications (WPNC). IEEE, 1–6.

Khor, W.S., Baker, B., Amin, K., Chan, A., Patel, K., Wong, J., 2016. Augmented and virtual reality in surgery-the digital surgical environment. Applications, limitations and legal pitfalls. Annals of translational medicine 4 (23), 454.

Luhmann, T., Robson, S., Kyle, S., Boehm, J., 2020. Close-range photogrammetry and 3D imaging, 3rd edition. De Gruyter, Berlin, Boston.

Maas, H.-G., 1997. Mehrbildtechniken in der digitalen Photogrammetrie. Zugl.: Zürich, Habilitationsschrift, ETH Zürich, 1997. Institut für Geodäsie und Photogrammetrie an der Eidg. Technischen Hochschule Zürich, Zürich.

Mur-Artal, R., Montiel, J.M.M., Tardos, J.D., 2015. ORB-SLAM. A Versatile and Accurate Monocular SLAM System. IEEE Transactions on Robotics 31 (5), 1147–1163.

Mur-Artal, R., Tardos, J.D., 2017. ORB-SLAM2. An Open-Source SLAM System for Monocular, Stereo and RGB-D Cameras. IEEE Transactions on Robotics 33 (5), 1255–1262.

Qiu, L., Ren, H., 2018. Endoscope Navigation and 3D Reconstruction of Oral Cavity by Visual SLAM With Mitigated Data Scarcity. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2197–2204.

Rublee, E., Rabaud, V., Konolige, K., Bradski, G., 2011. ORB. An efficient alternative to SIFT or SURF. 2011 International Conference on Computer Vision. IEEE, 2564–2571.

Schmidt, A., Kraft, M., Kasi´nski, A., 2010. An Evaluation of Image Feature Detectors and Descriptors for Robot Navigation. In: Bolc, L., Tadeusiewicz, R., Chmielewski, L.J., Wojciechowski, K. (Eds.), Computer vision and graphics. International conference, ICCVG 2010, Warsaw, Poland, September 20 - 22, 2010 ; proceedings, part II. Springer, Berlin, 251–259.

Shi, J., Tomasi, 1994. Good features to track. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition CVPR-94. IEEE Comput. Soc. Press, 593–600.

Song, J., Wang, J., Zhao, L., Huang, S., Dissanayake, G., 2018. MIS-SLAM. Real-time Large Scale Dense Deformable SLAM System in Minimal Invasive Surgery Based on Heterogeneous Computing, IEEE Robotics and Automation Letters, Vol.3, issue 4, 4068-4075.

Xiao, L., Wang, J., Qiu, X., Rong, Z., Zou, X., 2019. Dynamic-SLAM. Semantic monocular visual localization and mapping based on deep learning in dynamic environment. Robotics and Autonomous Systems 117, 1–16.

Yousif, K., Bab-Hadiashar, A., Hoseinnezhad, R., 2015. An Overview to Visual Odometry and Visual SLAM. Applications to Mobile Robotics. Intelligent Industrial Systems 1 (4), 289–311.