Presenting a Holistic Framework for Scalable, Marker-less Motion Capturing: Skeletal Tracking Performance Analysis, Sensor Fusion Algorithms and Usage in Automotive Industry

Michael M. Otto*, Philipp Agethen*,
Florian Geiselhart‡, Michael Rietzler‡, Felix Gaisbauer*, Enrico Rukzio‡

* Daimler AG
Virtual Production Methods
Wilhelm-Runge-Str. 11
89081 Ulm, Germany
phone: +49 176 64116266
email: [firstname.lastname]@daimler.com

‡ Ulm University
Institute of Media Informatics
James-Franck-Ring
89081 Ulm, Germany
email: [firstname.lastname]@uni-ulm.de

Abstract

Even though there is promising technological progress, input is currently still one of virtual reality's biggest issues. Off-the-shelf depth cameras have the potential to resolve these tracking problems. These sensors have become common in several application areas due to their availability and affordability. However, various applications in industry and research still require large-scale tracking systems e.g. for interaction with virtual environments. As single depth-cameras have limited performance in this context, we propose a novel set of methods for multiple depth-camera registration and heuristic-based sensor fusion using skeletal tracking. An in-depth accuracy analysis of Kinect v2 skeletal tracking is presented in which a robot moves a mannequin for accurate, reproducible motion paths. Based on the results of this evaluation, a distributed and service-oriented marker-less tracking system consisting of multiple Kinect v2 sensors is developed for real-time interaction with virtual environments. Evaluation shows that this approach helps in increasing tracking areas, resolving occlusions and improving human posture analysis. Additionally, an advanced error prediction model is proposed to further improve skeletal tracking results. The overall system is evaluated by using it for realistic ergonomic assessments in automotive production verification workshops. It is shown that performance and applicability of the system is suitable for the use in automotive industry and may replace conventional high-end marker-based systems partially in this domain.

First presented at EuroVR 2015, extended and revised for JVRB

Keywords: Scalable, marker-less, skeletal tracking, full-body motion capture

1 Introduction

Interactive virtual and augmented reality (VR, AR) assessments rely on robust, real-time tracking. With the rise of affordable depth cameras, marker-less body tracking has become a feasible option for a number of application areas, not only for gaming but also in research and industry. Being an alternative to more expensive and cumbersome marker-based motion capture systems, depth cameras are used for gestural interaction, natural user interfaces and motion capture for film making. In industry, where e.g. interaction with virtual product models and process simulations have already been common using conventional motion capture systems, depth camera based systems soon also became an appealing alternative for marker-based full-body motion capture.

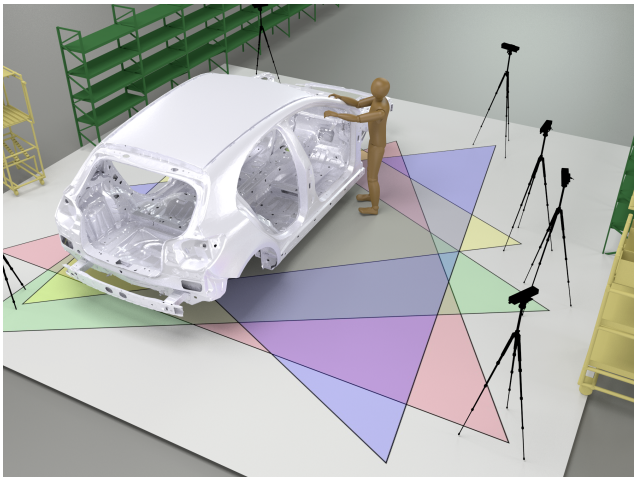

Figure 1: Example setup of full body skeletal tracking in automotive car assembly

However, considering spatially large use cases like car assembly (see Figure 1), the limitations of single depth cameras impede their use. Limited sensing range, a high susceptibility to self and external occlusions and a greatly varying sensing performance depending on the user's posture and position are some of the major drawbacks that need to be faced in order to use such systems in the mentioned scenario.

In order to overcome these limitations, this paper presents a concept of a distributed multi-depth-camera system consisting of multiple Kinect v2 sensors which extend sensing range and improve tracking performance. As the proprietary sensor has to be regarded as a black box, it is crucial to initially gain deeper knowledge about the utilized hardware. Based on these results, additionally, a number of new challenges need to be addressed: First of all, a common coordinate frame for the cameras has to be established by registering them to each other. Afterwards, the data coming from different cameras need to be combined in a meaningful way to actually gain improvements in tracking performance and range. Lastly, fused skeletal data has to be provided to the VR/AR systems via standardized tracking protocols, such as VRPN, trackD, dTrack or ARVIDA[1].

The remainder of the paper is structured as follows: We start with a review of the current state of the art on multi-depth-camera systems. Afterwards, an in-depth evaluation on the properties of the Kinect v2 and its related SDK is presented. Based on the results, we propose a set of registration and fusion techniques and extend those to a complete, ready-to-use tracking system. The evaluation of this system in the last section shows spatial accuracy of registration performance. Subsequently an evaluation of a specific automotive use case is described. The paper concludes with an overall assessment and outlook on further optimizations.

2 Related Work

Various research has already been carried out in the field of multi-depth-camera systems, however mostly focusing either on certain applications or specific aspects of such systems, thus leaving some of those challenges unaddressed to a great amount. These aspects can be clustered into following groups:

- System architecture
- Interference handling
- Registration
- Fusion algorithms

Depending on the use case, it is often also necessary to handle additional application specific issues like user identification or world coordinates registration, which we however do not consider further within this work.

2.1 System architecture

Most of the previous work is based on two or more Kinect cameras (1st gen.) which can be connected to a single computer, thus simplifying the required amount of infrastructure to a moderate level. However, works

[1] http://www.arvida.de/

presented by Schönauer [SK13] or Martínez-Zarzuela et al. [MZPHDP+14] also implement distributed systems, in which skeletal and depth data is gathered on camera nodes and being sent to a central fusion unit. This component handles the creation of a common view of the tracking space. Additionally, solutions have emerged in early states, which allow to stream Kinect data via network, e.g. by Wilson [Wil15]. As distributed systems, those approaches greatly improve scalability, however at the cost of increased complexity. In our system, also a distributed, yet a slightly different approach has been chosen. All information can be requested via service-oriented REST interfaces as e.g. presented by Keppmann et al. [KKS+14] in order to handle additional complexity, while still maintaining scalability.

2.2 Interference Handling

As depth cameras actively illuminate the scene, interferences can occur as soon as tracking frustums overlap, since any camera also receives light emitted from other cameras. There are two main approaches to interference handling which can be found in literature, (1) optical multiplexing (e.g. presented by Butler [BIH+12] or Faion et al. [FRZH12]) and (2) post-processing algorithms e.g. hole-filling as in Maimone and Fuchs [MF11]. Often, it is also possible to simply ignore interferences when using certain camera types and setups, especially in skeletal tracking applications where high frequency noise is not directly affecting tracking performance. As our proposed system uses Time-of-Flight (ToF) depth cameras, which generate only negligible interference noise due to their working principle and slightly different modulation, no countermeasures against cross-camera-interference have been implemented. Direct sunlight exposure turned out to be another source of interference, presumably due to overexposure of the IR imaging system. Again, no sensible countermeasures are possible to deal with this kind of interference besides avoiding direct sunlight exposure in the tracking space if possible.

2.3 Registration

One of the main challenges in multi-depth-camera systems lies in establishing a common coordinate frame by determining rotation and translation of the cameras to each other. Various approaches have been used for this, ranging from methods adopted from the 2D computer vision domain, horn-based methods like presented by Wilson and Benko in [WB10] or checkerboard-based approaches like those presented by Berger et al. [BRB+11] or Zhang et al. [ZSCL12], over iterative closest point (ICP, see [RL01]) approaches [PMC+11] to skeleton based (ICP-like) methods in more recent publications by Faion et al. [FRZH12], Asteriadis et al. [ACZ+13], Baek and Kim [BK15] and Kaenchan et al. [KMWS13]. Most of the methods yield comparable results according to their evaluation, however strongly differing in the ease-of-use and setup time with different approaches. Our proposed approach focuses on reduced setup times and an easy setup procedure while maintaining high precision. Thus, we are using a combination of multiple registration approaches from the work described above.

2.4 Fusion

After establishing a valid registration, skeletal tracking data from different cameras exist in a common coordinate space; nevertheless, body tracking skeletons are still individual and separate. To gain advantages of such a setup, data fusion methods can be employed to gather an improved view on the tracking space. The possible methods range from simple best-skeleton approaches, over joint-counting approaches [CYT+11], substitution approaches ([ACI15], [KDDG14]), weighted averaging methods [FRZH12] or [KMWS13], to dedicated fusion algorithms e.g. by Yeung et al. [YKW13] or Asteriadis et al. [ACZ+13], which respect data quality and the specific tracking situation. This helps in dealing with occlusion and sensing limitations. Combining the advantages of each mentioned previous works, a set of novel fusion heuristics will be presented and analytically evaluated regarding its performance against a ground truth.

2.5 Assessment of related work

While covering many of the relevant aspects each, most of the previous works leave out important factors of a multiple depth-camera system for universal use. In general, registration and fusion approaches also lack end-user optimization as well as comprehensive evaluation of underlying assumptions, e.g. for factors influencing registration and fusion methods and quality. With this work, some currently missing insights and concepts will be provided, which have proven to

be useful for implementing a multiple, scalable depth-camera system.

For this, an in-depth analysis of tracking performance is conducted as a first step, similar to earlier work for different hardware or properties, e.g. by Wang et al. [WKOB15] or Yang et al. [YZD+15].

2.6 FusionKit

Besides the related work mentioned above, a publicly available[2] system named FusionKit was developed based on similar assumptions by some of the authors of this article [RGTR16]. FusionKit is also closely related to the system presented in the following chapters, yet still a separate system. Differences and common attributes are discussed more in-depth in chapter 6, for the sake of completeness, however, this system is also listed as related work here.

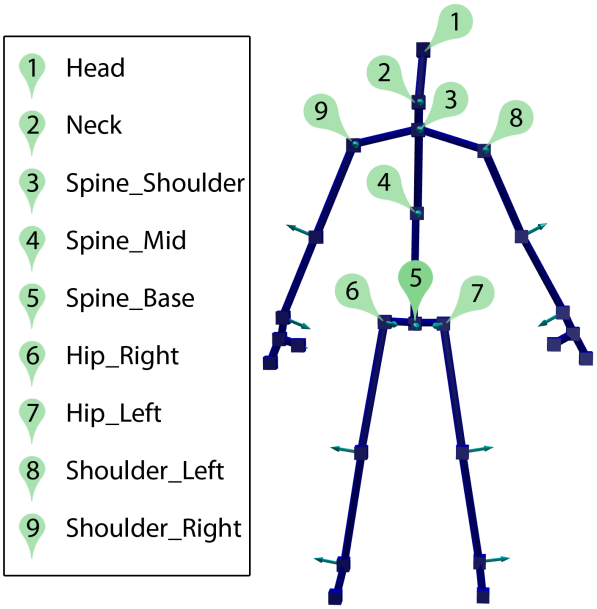

Figure 2: Relevant joints nomenclature of SDK body tracking

3 Performance Measurements of Kinect v2 SDK Skeletal Tracking

The proposed scalable multiple depth-camera system relies on the Kinect v2 sensor and its Software Development Kit (SDK) provided both by Microsoft. Since the combination of hardware and SDK is proprietary and therefore a closed system, several features and properties of the Kinect v2 sensor cannot be determined or influenced directly. For example skeletal tracking algorithms can not be influenced or extended due to proprietary training sets and algorithms [SGF+12]. These body tracking algorithms are estimating a skeleton consisting of 25 3D-positions including 18 orientations of up to six users (see Figure 2) at approximately 30 Hz refresh rate.

However, it is possible to analytically analyze this system in order to derive tailored approaches for the steps mentioned in section 2. In 2012 Obdržálek et al. [OKO+12] already performed similar evaluations for the first Kinect hardware generation which is based on structured light sensing but without the focus on building a multiple depth-camera system. Since the second generation Kinect v2 sensor is based on ToF depth sensing technology, the results have only limited validity for the new sensor which is therefore analyzed in the following. The presented evaluation results have directly influenced design considerations for the proposed skeletal fusion algorithms presented in the following chapters.

3.1 Experimental setup

To gain insights into the runtime behavior and sensing performance of the Kinect body tracking system, a setup for reproducible human postures and trajectories is created to conduct blackbox-like tests on it. In order to achieve reproducible trajectories a high precision robot UR10 of "Universal Robots"[3] is used, being mounted horizontally on a table in $1.0\,\mathrm{m}$ height (see Figure 3). This Robot has six degrees of freedom (DoF) and an operating radius of $1.3\,\mathrm{m}$. Each axis can be rotated by $720°$. All tests are carried out at slow speeds with a maximum velocity of $0.1\,\mathrm{m/s}$ to avoid tracing effects. The repeatability of each trajectory is specified to $\pm\,0.1\,\mathrm{mm}$. The so-called tool center point (TCP) is defined in the center of the mounting plate directly linked to the last rotational axis.

Mounted to the robot arm, a $1.75\,\mathrm{m}$ tall mannequin is used for all experiments. The clothing is chosen to be regular t-shirt and jeans. As dark black clothes are causing problems in the depth image (and in the body tracking results) due to low IR reflection, bright colors are chosen. In all experiments the mannequin has a symmetric posture with open hanging arms.

[2] https://github.com/fg-uulm/fusionkit

[3] http://www.universal-robots.com

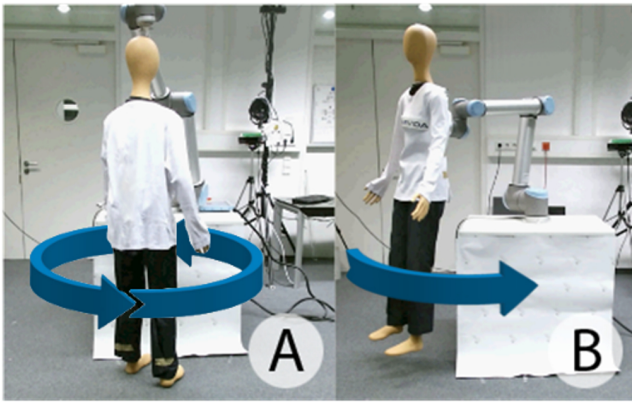

Figure 3: Experimental setups: A) 360° rotational test: Mannequin hanging on robot, B) Mannequin's spine mounted on robot

In order to retrieve precise and low jitter input data of the mannequin's movements, two experiments with reproducible trajectories are carried out: The 360° and 85° experiment. The latter is carried out in order to bring the rotation axis out of the skeletal center-axis, which is not possible in the 360° experimental setup due to axis limitations and occlusions by the robot itself.

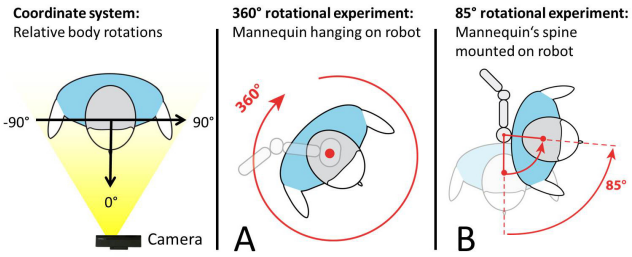

Figure 4: Drawing of coordinate system and two experimental setups: A) 360° rotation, B) 85° rotation

For the 360° experimental setup the mannequin is mounted on the TCP on the top of it's head with a 10 cm long separator in between (see Figure 3A and Figure 4A). The robot is spinning the mannequin around it's vertical axis at a constant angular velocity. With this experimental setup, skeletal orientation and positional stability can be assessed with regard to the user's rotation.

In the 85° experiment the mannequin's spine is mounted on the robot's TCP at 1.3 m height (see Figure 3B and Figure 4B). The mannequin's feet are 1 mm above the floor. The sensor is leveled and stands in a distance of approx. 2 m at a height of 1 m. The robot rotates the mannequin around the TCP from 0° (orthogonal) to 85° (close to side view)- see Figure 4B. The 85° experimental setup is used to quantify the positional precision of the center-axis joints (see Figure 2) and therefore the applicability of joints for multi-camera registration purposes.

3.2 Evaluation 360° experiment

In general, the body tracking is originally designed for gaming use cases in which the user is facing the camera. Therefore, a strongly decreasing tracking performance when slowly rotating to a lateral or even rear view is expected.

Figure 5 shows the course of shoulder joints in a top view over rotation from 0° and 180°. At 0° the mannequin is facing the camera as depicted in Figure 4. In contrast to the two expected continuous semicircles (the lower semicircle representing the right shoulder whereas the upper semicircle is related to left shoulder), the SDK results show a discontinuous behavior. It can be observed that the position of the right shoulder jumps in the mid section in Figure 5 from point $[-0.05\,\mathrm{m}, 2.2\,\mathrm{m}]$ to $[-0.15\,\mathrm{m}, 2.4\,\mathrm{m}]$. This ambiguity is attributable to the SDK assuming that the user's body is always oriented towards the camera. Consequently, an ambiguity appears when the user turns his back towards the camera. The analysis shows, that this effect can be observed for all non-center-axis joints such as arms and legs. Even the non-centered torso joints, e.g. Hip_Left, Hip_Right, Shoulder_Left and Shoulder_Right, suffer from the same problem. However, center-axis joints like Head, Neck, Spine_Base, Spine_Mid and Spine_Shoulder are invariant to the front/rear ambiguity and could be therefore used for registration purposes.

Having excluded non-center-axis joints, the estimated user orientation is subsequently analyzed with regard to it's validity. Figure 6 depicts the expected SDK behavior (turquoise), given by the rotating robot axis. Moreover, this plot shows the estimated vertical orientations for the center joints Neck, Spine_Base, Spine_Mid and Spine_Shoulder. Joints of arms and legs are not included in this figure since they only provide valid data up to 30° due to occlusion.

The SDK reliably estimates the orientation for the center joints within a range of 0° to ± 40°. Beyond this point, the user's vertical orientation is increasingly underestimated until the mentioned ambiguity takes place at approx. 130°. Between 40° and 130° Spine_Shoulder and Neck orientation perform signif-

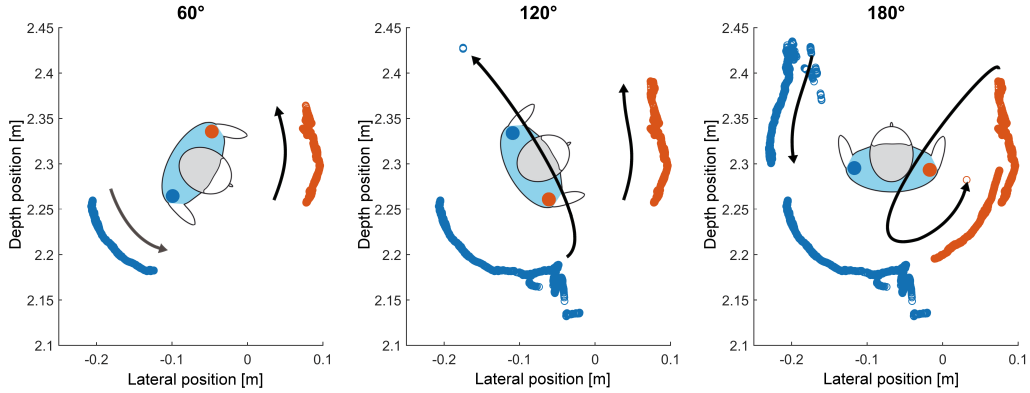

Figure 5: 360° rotational experiment: Front/back ambiguity during rotation visualized for the non-center axis joints Shoulder_Left and Shoulder_Right. Their color change indicate the front/back ambiguity.

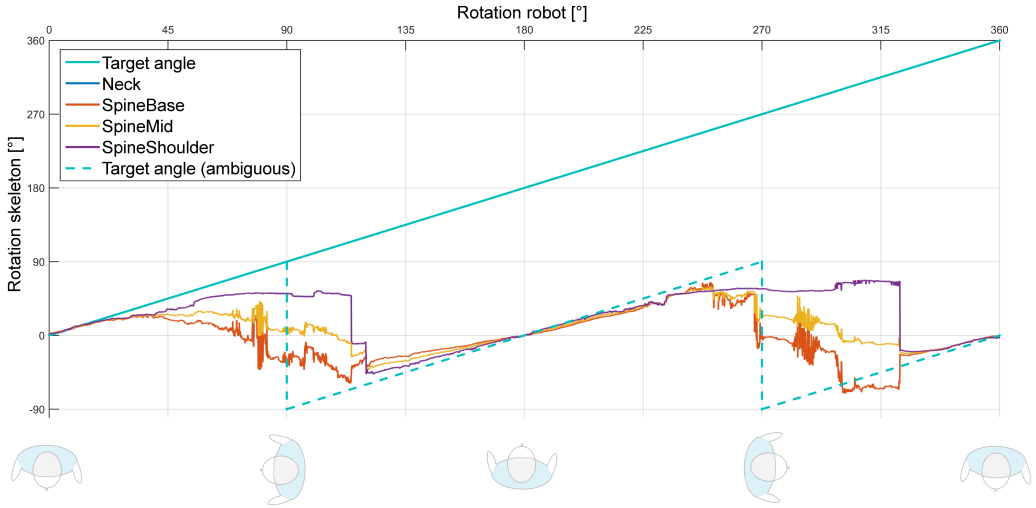

Figure 6: 360° rotational experiment: Expected SDK behavior vs. estimated shoulder- and hip- orientation

icantly better than Spine_Mid and Spine_Base. From 130° to 260° an angular behavior, similar to the range of 0° to ± 45° can be observed, whereas between 260° and 320° the user's orientation switches again. In Figure 6, the dotted line additionally depicts the expected ambiguous behavior with two discontinuities at 90° and 270°. However, comparing this expected behavior to the described results, it can be seen that these distinctive points are located 30° to 40° beyond their predicted location due to the SDK trying to maintain the current orientation of the person. Consequently, the analysis reveals, that the vertical orientation of each joint can be considered as ambiguous and error-prone outside the range of 0° to ± 40°.

Figure 7 is showing the mean Euclidean distances of spatial inter-frame jitter during the 360° experiment. Throughout the whole 360° rotation, one can see jitter differences depending on the respective joints and the user's orientation towards the camera. The head joint has significantly lower jitter compared to the other joints (mean jitter $<0.072\,\mathrm{mm}$). Consequently, it can be assumed that the Kinect SDK filters head joint position in order to reduce spacial jitter enabling viewpoint control applications. The joints Spine_Shoulder and Neck have modest jitter below 35° on both sides (mean jitter $<0.2\,\mathrm{mm}$), whereas in rear facing orientation (180°), the jitter is increasing (mean jitter $<0.5\,\mathrm{mm}$). In contrast to that, Spine_Mid and Spine_Base joints reveal an increasing jitter on the side view at around 90° and 270° user orientation (mean jitter $<4.1\,\mathrm{mm}$). Therefore Spine_Mid and Spine_Base joints cannot be used for flexible multi-sensor skeletal registration purposes.

As a result, Head, Neck and Spine_Shoulder could be used for registration purposes since they are independent from the positional ambiguity and offer the

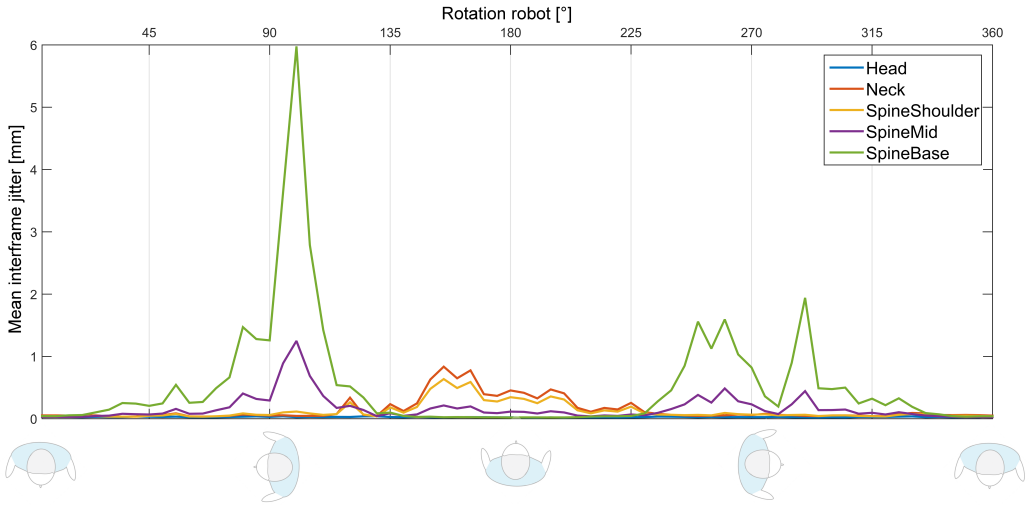

Figure 7: 360° experiment: Mean inter-frame jitter for all center-axis joints

lowers jitter during user rotation. Information about orientation cannot be used since each joints show rotational ambiguity.

3.3 Evaluation 85° experiment

The 85° experimental setup compares the trajectories of the center-axis skeletal joints (Head, Neck, Spine_Shoulder, Spine_Mid and Spine_Base) to the ground truth robot trajectory to gain insights in their positional precision.

During the analysis, the mannequin's back is mounted on the TCP with a distance of approx. 27 mm, thus rigidly coupling both trajectories. Consequently, the mannequin trajectory is known in advance and can be directly compared to the result of the skeletal joints. Figure 8 shows the resulting paths of the baseline and the five joints in a top view (see Figure 4B).

The joint Spine_Base does not follow a circular trajectory. Above 60° the trajectory is getting noisy and unstable. Spine_Mid shows equal behavior. For both of these points the distance is overestimated. Spine_Shoulder has a good circular performance up to 80° whilst the distances at low angles are still overestimated. Head and Neck joints match the baseline for lower angles, whereas at more than 45° the head joint tends to move forward in the mannequin's perspective and away from the sensor (almost 90°). These effects could be reproduced throughout all repetitions.

Overall, these experiments lead to the conclusion that for registration purposes the Neck joint of the Kinect SDK skeletal tracking provides the most stable and suitable positional data even at high rotation angles relative to the sensor. Additionally the head joint offers a low jitter 3D joint position.

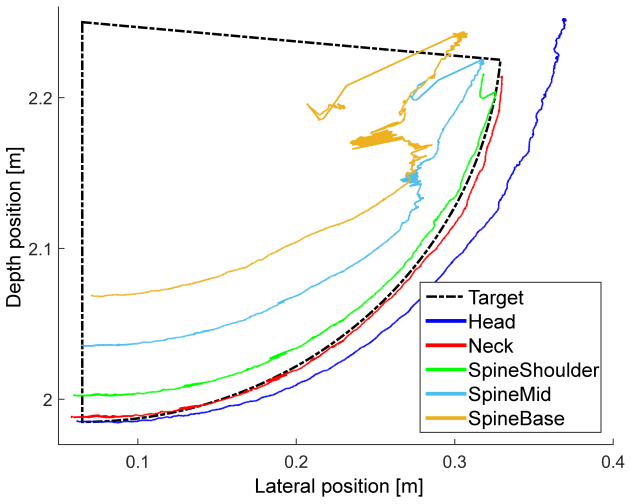

Figure 8: Evaluation of 85° experiment: top view on center-axis joints in relation to ground truth real trajectory

4 Hardware setup

The proposed multiple depth-camera system consists of several Kinect v2 sensors. In comparison to the first generation Kinect, the novel sensor uses the depth sensing technology Time of Flight, which induces improved depth accuracy and better interference resistance. The proposed system consists of several tracking computers accommodating the tracking services

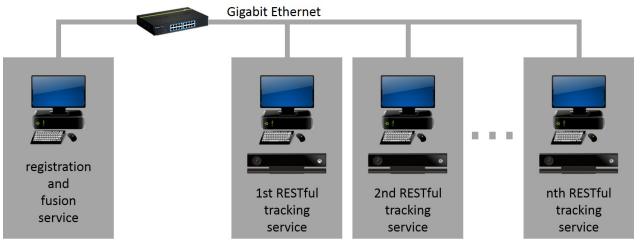

Figure 9: Hardware setup for the tracking system with service-oriented, distributed sensor services

and one central fusion computer (see Figure 9). Each tracking computer is connected to one Kinect v2 sensor via USB 3.0 and additionally to a fast local area network.

5 Software

There are two main software components: The tracking software and the fusion software which is described in the following

5.1 Service-oriented tracking software

Implementing a service oriented RESTful tracking service instead of conventional streaming architecture has several advantages: Third party integrators have the possibility of easily reusing the services for implementing clients. Additionally, using standardized and publicly available tracking vocabulary and Resource Description Framework (RDF) one can achieve interoperability between tracking devices which is also the goal of the ARVIDA project. In this context, the presented tracking services are using a RESTful polling-based approach with linked data which is conforming to the ARVIDA standard. It has been shown by Keppmann et al. [KKS+14] that RESTful Linked Data resources can be applied for virtual reality environments.

In the tracking service, information is gathered by the event-based Kinect SDK. The web service offers all skeletal information, the status of each skeleton, the floor plane estimation and color and depth camera views as RESTful resources. RDF datagrams are serialized using Turtle format. Each datagram contains time stamps for synchronization afterwards.

5.2 Fusion and multi sensor tracking service

The fusion and multi-sensor service is running as a central component and handles registration, fusion and

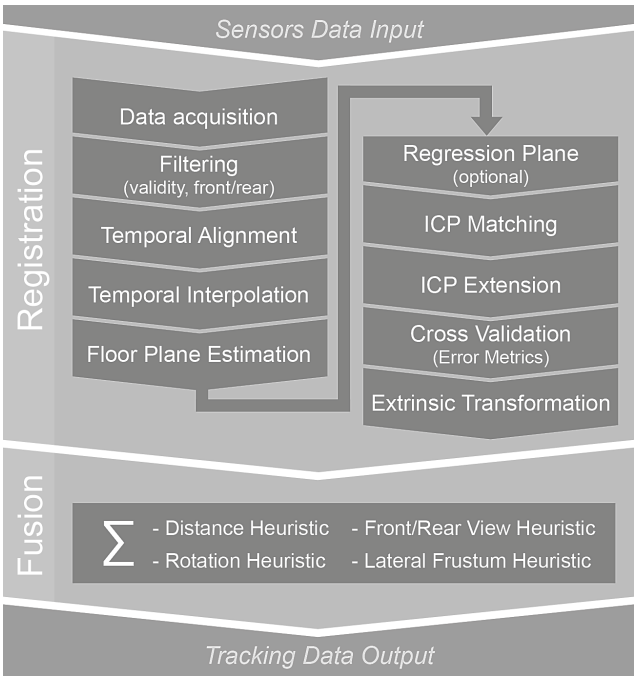

Figure 10: System architecture of fusion service

data input/output in the tracking environment. Figure 10 depicts the architecture of the registration and fusion service.

The fusion service polls the data of the tracking services. This data is used for calculating extrinsic transformations between the cameras for the subsequent fusion process. Several pre-processing steps have to take place in advance (see Figure 10) which are described in detail in the following paragraphs. Whereas the fusion component processes all joints of the skeletal data, the registration process only uses the neck joint information. During registration, neck joint data is being captured over time from each camera and is used as an input point cloud for the ICP algorithm. The algorithm then iteratively minimizes the difference between two point clouds gathered by the sensors. The result of the algorithm is the refined extrinsic transformation between a pair of cameras.

After the registration, the heuristic-based fusion component is able to combine skeletal data from all registered cameras and provides them as an output to possible domain-specific application components.

5.2.1 ICP extension

Gathering only the neck joint information has a drawback which has to be compensated: Since the user's movement takes place on the flat floor plane and the height of the user's neck joint does not vary a lot, the

gathered point cloud data lies almost on a single plane. To compensate this lack of variance, additional information is used. The floor plane estimation compensates the missing information by using an approximation of the distance to the floor and the pitch angle of the sensor. Fusing this information with the ICP data offers an improved transformation for extrinsic registration between one sensor relative to the master sensor. In addition to that, we propose to use a regression plane to further enhance the ICP results, if enough feature points have been gathered during the user's movement.

5.2.2 Front/rear detection

In order to achieve maximum flexibility for the hardware sensor setup, the fusion service has to recognize, whether the user is facing the Kinect or if he is turning his back on the corresponding sensor. The SDK always presents the skeleton data as if the user was facing the camera directly. Even in a rear view, the skeleton is recognized robustly but data being presented laterally reversed. Additionally, a robust indicator if the user is turning his back towards the camera, is to evaluate the angle between the shoulder joints. Evaluating the discrete skeletal states of the collar joints, one can determine the user's orientation to the camera in each frame.

5.2.3 Scalability

To achieve a fully scalable system with a common coordinate frame, extrinsic transformation chains have to be built (see Figure 11). The above described method is used to calculate the transformation matrix for each camera pair with an overlapping tracking frustum. For N sensors sharing an overlapping tracking area there are $(N-1)!$ transformation matrices. Having more than two sensors sharing the same tracking area, the system is over-determined and a cross-validation of transformation chains has to be carried out with regards to the absolute transformation precision. Therefore an error metric is introduced which consists of the summed up and normalized Euclidean distances of the reprojection error. Based on this error value, the best interlinked transformation chain between master and each client can be determined.

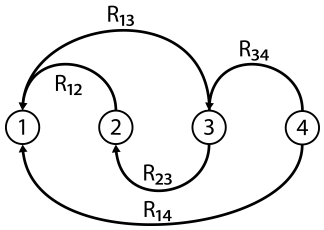

Figure 11: Concatenated transformation chains for a four sensor arrangement

5.2.4 Time Alignment and Interpolation

The time synchronization algorithm is crucial for interpolating the asynchronously captured body tracking information generated by the depth camera sensors. Since the user has to walk slowly during registration process and Kinect v2 only captures skeletal data at $30\,\mathrm{Hz}$, a worst case offset of several centimeters is induced just by event-based, non-synchronized image acquisition. To generate synchronized timestamps within the whole sensor network NTP protocol is utilized. Based on these precise timestamps, skeletal body tracking frames are virtually synchronized within the fusion software through interpolation. The depth camera's skeleton acquisition time is assumed to be constant over all sensors. Since the user's body has a certain inertia and the refresh rate is approximately $30\,\mathrm{Hz}$, the inter-frame trajectory between two skeleton datagrams can be assumed as linear movement.

5.2.5 Fusion Process with quality heuristics

Having registered all cameras via extrinsic transformation chains, the tracked skeletons generated from different views are in the same coordinate frame and need to be fused. For large-scale human tracking and posture analysis we propose a set of quality heuristics for the skeletal fusion process. Each skeleton within each sensor is given a certain weight. The higher the weight the higher the influence of the certain sensor on the user's fused skeleton. A comprehensive set of quality measures will be presented for real time skeletal fusion:

First, we propose the distance between the user and a sensor as the distance quality measure. At a distance of approximately $2.5\,\mathrm{m}$ tracking results are most reliable. This quality measure weights the user's skeleton

over the distance to the neck joint respectively.

$$w(d)=\begin{cases}0 & \text{for } d \leq 1.5\,\mathrm{m} \\ 1-(d-2.5\,\mathrm{m})^2 & \text{for } 1.5\,\mathrm{m}<d\leq 3.5\,\mathrm{m} \\ 0 & \text{for } d > 3.5\,\mathrm{m}\end{cases} \quad (1)$$

Second, we introduce rotation quality heuristics for robust human activity analysis. If there is multiple data on the user's posture we propose to weight the front facing skeletons highly and to set all rear views to weight zero. The user has to stand as orthogonal to the sensor as possible, since 30° has been found to be the maximum vertical user orientation for reliably tracking limbs:

$$w(\phi) = \begin{cases}1 - \frac{|\phi|}{30^\circ} & \text{for } |\phi| \leq 30^\circ \\ 0 & \text{for } |\phi| > 30^\circ\end{cases} \quad (2)$$

Lastly the lateral frustum quality heuristic limits the tracking frustum to a horizontal field of view of 50° so that the limbs are still probable to be within the tracking area of the sensor (70°). We propose zero weight if the user's center axis joints exceed 50° in horizontal axis of the local camera coordinate frame:

$$w(\alpha) = \begin{cases}1 - \frac{|\alpha|}{25^\circ} & \text{for } |\alpha| \leq 25^\circ \\ 0 & \text{for } |\alpha| > 25^\circ\end{cases} \quad (3)$$

The weights determined by the above mentioned heuristics can be integrated and further utilized by subsequent fusion algorithms. For this purpose two different fusion approaches were used. The first approach computes the fused skeleton by the calculation of a joint-wise, weighted mean for the position and the orientation by calculating the corresponding quaternions using the LERP algorithm. Secondly an implementation of a linear Kalman Filter was use [Kal60]. The gathered weights from the heuristics are integrated within the measurement covariance matrix of the filter, increasing or decreasing the variance of the measurement. Since the tracking data of the Kinect sensor are always afflicted with jitter, the Kalman Filter can improve the overall tracking quality by its probabilistic modeling and in case of partially or full loss of tracking data, it can predict the skeleton state.

6 Advanced Error prediction in FusionKit

Additionally to the presented registration and fusion algorithms, Rietzler et al. [RGTR16] show the impact of an accurate error prediction model used for weighting the various information given by different sensors. In their experiments, different error prediction models were compared. In the first experiment only the three Kinect states *tracked, inferred* and *untracked* were used in a linear regression model to predict the errors. This simplified weighting strategy proved to be the less accurate one, leading to a mean prediction error of $8.5\,\mathrm{cm}$ which was resulting in an error of $6.5\,\mathrm{cm}$ when used for fusing the joints.

For the enhancement of the error prediction quality, they use a set of heuristics similar to the ones, presented in the last chapter (e.g. angle between user and camera, distance to the camera, periphery of the sensor). To predict the influence of the different heuristics, they first used a linear regression model, resulting in a prediction error of still around $6.4\,\mathrm{cm}$. Using a linear regression model with assumption of a gamma distribution, this prediction error was reduced to around $4.4\,\mathrm{cm}$ ($1.9\,\mathrm{cm}$ for centered joints, but around $7.5\,\mathrm{cm}$ for wrist joints).

Regression does not model influences between heuristics and assumes a linear error distribution. They therefore conducted a last experiment using a non-linear boosting model. The result of the boosting is a large decision tree, with edges being the representation of a certain heuristic value and the leaves being a fix error under the given circumstances. This model allows to describe dependencies, since e.g. the impact on the predicted error of the angle between user and sensor may be different and depending on e.g. the distance of the user towards the sensor. Using this model, the error could be predicted with an offset around $3.2\,\mathrm{cm}$ ($1.2\,\mathrm{cm}$ for centered joints and $5.5\,\mathrm{cm}$ for wrists and feet), which proved to be the most accurate tested prediction model.

This experiment also showed the influence of the heuristics on the error prediction per joint. The influence of the angle between user and sensor proved to be most important and increasing with the distance of the respective joint towards the body center. The tracking quality of the spine base joint was not even influenced by this heuristic. Interestingly, the Kinect state had little influence, being even close to irrelevant for the centered joints.

Another advantage of this model is the possibility of training various scenarios. Different rating profiles can be used to optimize the fused result under special circumstances. The rating strategy could therefore vary in a laboratory setting without any objects occluding

the user, or a more complex setting as depicted in the use case driven evaluation in the following chapters. For example in automotive process verification workshops, physical mock-ups standing in the middle of the workshop environment may weaken the tracking quality of some sensors.

7 Evaluation of registration accuracy and validity

To determine the accuracy of extrinsic transformations and therefore the spatial registration error, a series of experiments has been carried out.

7.1 Experimental setup

Since an absolute accuracy evaluation is needed, a high precision marker-based tracking system was chosen as a ground truth. The system consists of 16 'OptiTrack Flex 13' cameras which reported a residual mean error of $0.624\,\mathrm{mm}$ for the whole tracking volume. Rigid bodies, consisting of multiple markers, were applied on the Kinect sensors. The pivot point translation of the rigid body markers was defined to be in the Kinect's depth camera focal point to match the origins of Kinect body tracking and the OptiTrack rigid body markers.

7.2 Design of Experiments

All registration scenarios were conducted using two Kinect. The registration process has been recorded 100 times; for each of the five scenarios 20 measurements were performed. During each experiment point cloud movement data was gathered for 10 seconds within the overlapping tracking area. The scenarios differed by the angles around the vertical axis: $0°$, $45°$, $90°$, $135°$ and $180°$. No outliers were removed for the following evaluation.

7.3 Results

Figure 12 highlights the registration performance of the fusion service. Circles depict the calculated ideal Optitrack positions. For these scenarios the Euclidean distance in the floor plane is always less than $15\,\mathrm{mm}$ to the ground truth position. The vertical axis reveals maximum deviations of $1.5°$ for the sensor's pitch axis. The body tracking estimator within the SDK reveals uncertainties especially in the vertical axis. The uncertainty of the joints can vary more than $20\,\mathrm{mm}$, depending on the angle between the user and the sensor.

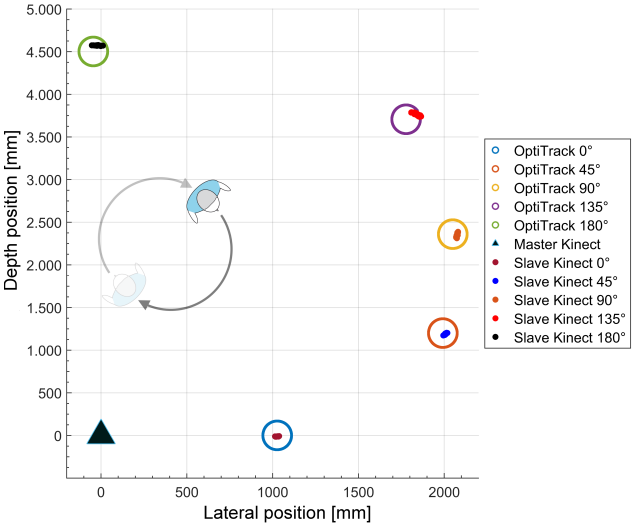

Figure 12: Top view on the registration results: Master sensor at [0,0], 5 scenarios with 20 registrations each, circles indicate the ground truth of the OptiTrack measurements

Scenario	Mean Error	Standard Deviation
0°	9.6 mm	4.5 mm
45°	16.0 mm	7.1 mm
90°	17.8 mm	13.6 mm
135°	26.6 mm	10.0 mm
180°	12.3 mm	10.0 mm

Table 1: Reproducibility and deviation of registration results for the 5 scenarios (N=100)

The reproducibility of this approach is summarized in Table 1. The mean deviation from the center point is ranging from $9.6\,\mathrm{mm}$ to $26.6\,\mathrm{mm}$ with a standard deviation ranging from $4.5\,\mathrm{mm}$ to $13.6\,\mathrm{mm}$. The maximum deviation was 42.2 mm in the 135 scenario.

8 Evaluation of applicability in ergonomic assessments

Full body motion capture data is frequently used in automotive industry for various use cases such as process verification, visibility checks or tool assessments. Moreover, ergonomics experts are using motion capture technology to virtually audit end-assembly workplaces and worker postures. Three specific workplaces have been chosen for an virtual ergonomic evaluation.

While being tracked, a worker is performing the pre-planned assembly routines in the virtual environment whereas the ergonomics expert is evaluating the movements, weights and resulting forces. The following three experiments have been performed during real production planning workshops:

- Reachability check for mounting an antenna on the roof
- Posture definition for screwing work tasks
- Stress screening for battery assembly

8.1 Experimental hardware & software setup

During these assessment workshops six Kinect sensors are utilized which are all facing the center of the workplace and are evenly distributed on the edges of the tracking area. This area covers 6 m x 6 m since movements within real workplaces in automotive end-assembly lines match these dimensions.

Having registered all cameras to a common world coordinate frame, the presented system architecture in combination with the fusion heuristics enable the worker to be constantly tracked regardless of his position and his orientation within the concatenated tracking frustum.

To carry out the mentioned ergonomic assessments in real time the virtual manufacturing software Delmia V5-6 R23 has been used in combination with Haption RTID plugin. The following pipeline was used in this case: The fusion service exposes all fused tracking joints via the A.R.Tracking protocol as 6DoF tracking data. The Haption suit and configuration maps this tracking data onto the fully flexible virtual human. 20 tracking joints are used to modify the DHM interactively.

As depicted in Figure 13 the virtual scene in Delmia V5 included a car body in the assembly status for the respective station. Dynamic parts have been simulated and attached to the right hand joint. The anthropometry of the virtual human was adjusted to the real worker's size and weight.

8.2 Results

All mentioned manufacturing tasks could be carried out without having any prior physical mock-ups. Limitations of the pre-planned process and unfavorable ergonomic situations could be identified for all three experiments with this virtual methodology. Additionally

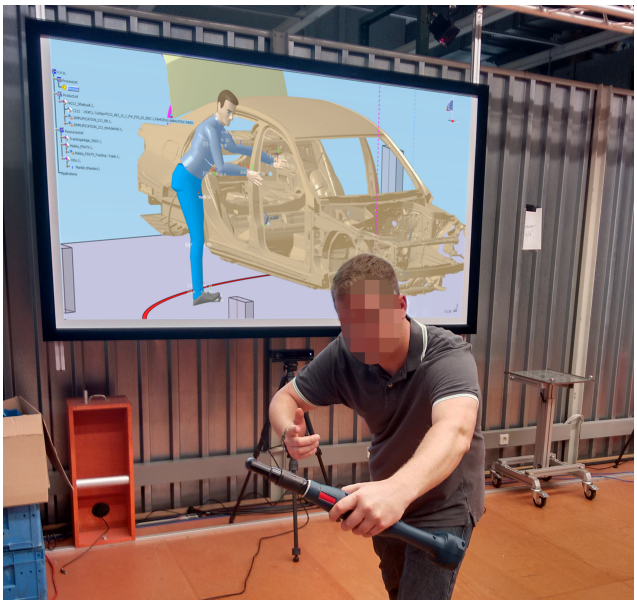

Figure 13: Delmia V5 DHM directly manipulated by marker-less motion capture approach. All assessments can be carried out in real time.

the results gathered could be verified by subsequent traditional hardware workshops.

Comparing common marker-based tracking systems to this novel approach, ergonomic experts pointed out several effects: First of all, the users do not have to put on a special suit with retro-reflective markers. This is time-consuming and cumbersome for the tracked persons. User's movements may be influenced by the marker suits and seem not as natural as in regular working clothes. Secondly, users can swap immediately without any preparation time, so that multiple users can test the process without any prior work. Lastly, the marker-less system induces more latency and jitter to the tracking data than the marker-based tracking system. Ergonomic experts pointed out that the motion capture data quality is still sufficient to identify and solve the issues related to ergonomic assessments. Latency of several frames is considered to be irrelevant, since there is no immersive feedback to the user causing motion sickness. It became apparent that the registration and fusion precision are sufficient for human posture analysis, for profound ergonomic simulations and for large-scale view point control applications in virtual environments.

Additionally for an automatic recognition of digital human postures, the ErgoToolkit was utilized in this pilot case that was presented by Alexopoulos et al. in 2013 [AMC13]. With this additional plugin a

rough stress screening could be carried out automatically and critical postures could be detected reliably. Furthermore, experts appreciated the side benefits of this tracking approach like visibility checks through interactive viewpoint control and validation of assembly and disassembly routines for dynamic virtual objects via hand joint tracking. Follow-up processing times like documentation can be reduced significantly, by pre-filling assessment sheets automatically. All of these use cases will directly profit of advances in multi depth-camera tracking technologies.

9 Conclusion

A holistic framework for large-scale, marker-less motion capture is presented using multiple depth-cameras. The given flexibility of sensor arrangements enables a multitude of use cases.

Using the Microsoft Kinect v2 sensor, an in-depth analysis of the skeletal tracking performance is carried out. For evaluation purposes, an UR10 Universal robot in combination with a mannequin is used to generate reproducible trajectories and constant velocity. Using the skeletal tracking of the Kinect SDK for extrinsic registration, the best results can be achieved processing the Neck joint information. Interestingly, the head shows significantly less inter-frame jitter compared to the remaining joints of the tracked skeleton.

Several novel registration-relevant techniques are presented and evaluated like time-synchronous interpolation, front/rear detection and error measures. Additionally, a comprehensive set of quality heuristics has been derived for the skeletal fusion process, which showed to improve skeletal tracking. An additional approach for further error prediction is described.

Technically the number of possible tracking nodes is limited by computing power and network throughput. Sensor setups up to ten tracking services have been successfully tested. The fusion service itself can be addressed transparently and acts externally as if it was a single sensor tracking service. Standardized tracking protocols (VRPN, dTrack, trackD, ARVIDA) are implemented to achieve interoperability with virtual assessment environments and game engines.

Three pilot test cases within the automotive industry are carried out to evaluate the system's performance with real ergonomic use cases. The requirements in terms of occlusion robustness (e.g. when working with car bodies in the tracking area), tracking range and tracking precision could be fulfilled in each of the pilot test cases. Since the novel system has proved its applicability, reduced costs and the ease-of-use, it can complement the variety of existing industrial tracking systems.

In future work, both the presented system and FusionKit will be integrated for an industrialization of the system. Merging all advantages of such similar systems will lead to further improvements in terms of tracking quality, usability and stability.

10 Acknowledgments

We would like to thank Brian Moore, Microsoft Germany, for supporting this work. This work was additionally supported by the German Federal Ministry of Education and Research (ARVIDA project, grant no. 01IM13001N).

References

[ACI15] Nur Aziza Azis, Ho-Jin Choi, and Youssef Iraqi, *Substitutive skeleton fusion for human action recognition*, 2015 International Conference on Big Data and Smart Computing (BigComp), 2015, DOI 10.1109/35021BIGCOMP.2015.7072828, pp. 170–177, ISBN 978-1-4799-7303-3.

[ACZ+13] Stylianos Asteriadis, Anargyros Chatzitofis, Dimitrios Zarpalas, Dimitrios S. Alexiadis, and Petros Daras, *Estimating Human Motion from Multiple Kinect Sensors*, MIRAGE '13 Proceedings of the 6th International Conference on Computer

Citation

Michael M. Otto, Philipp Agethen, Florian Geiselhart, Michael Rietzler, Felix Gaisbauer, Enrico Rukzio, *Presenting a Holistic Framework for Scalable, Marker-less Motion Capturing: Skeletal Tracking Performance Analysis, Sensor Fusion Algorithms and Usage in Automotive Industry*, Journal of Virtual Reality and Broadcasting, 13(2016), no. 3, February 2017, urn: nbn:de:0009-6-44811, DOI 10.20385/1860-2037/13.2016.3, ISSN 1860-2037.

Vision / Computer Graphics Collaboration Techniques and Applications (New York, NY, USA), ACM, 2013, DOI 10.1145/2466715.2466727, p. Article No. 3, ISBN 978-1-4503-2023-8.

[AMC13] Kosmas Alexopoulos, Dimitris Mavrikios, and George Chryssolouris, *ErgoToolkit: An Ergonomic Analysis Tool in a Virtual Manufacturing Environment*, International Journal of Computer Integrated Manufacturing **26** (2013), no. 5, 440–452, ISSN 0951-192X, DOI 10.1080/0951192X.2012.731610.

[BIH+12] D. Alex Butler, Shahram Izadi, Otmar Hilliges, David Molyneaux, Steve Hodges, and David Kim, *Shake'n'sense: reducing interference for overlapping structured light depth cameras*, CHI '12 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (New York, NY, USA), ACM, 2012, DOI 10.1145/2207676.2208335, pp. 1933–1936, ISBN 978-1-4503-1015-4.

[BK15] Seongmin Baek and Myunggyu Kim, *Dance Experience System Using Multiple Kinects*, International Journal of Future Computer and Communication **4** (2015), no. 1, 45–49, ISSN 20103751, DOI 10.7763/IJFCC.2015.V4.353.

[BRB+11] Kai Berger, Kai Ruhl, Christian Brümmer, Yannic Schröder, Alexander Scholz, and Marcus Magnor, *Markerless Motion Capture using multiple Color-Depth Sensors*, Vision, Modeling & Visualization. Berlin, Germany, Oct. 04 - 06, 2011 (Goslar) (Peter Eisert, Konrad Polthier, and Joachim Hornegger, eds.), Eurographics Association, 2011, pp. 317–324, ISBN 978-3-905673-85-2.

[CYT+11] Maurizio Caon, Yong Yue, Julien Tscherrig, Elena Mugellini, and Omar Abou Khaled, *Context-aware 3d gesture interaction based on multiple kinects*, AMBIENT 2011: The First International Conference on Ambient Computing, Applications, Services and Technologies, 2011, pp. 7–12, ISBN 978-1-61208-170-0.

[FRZH12] Florian Faion, Patrick Ruoff, Antonio Zea, and Uwe D. Hanebeck, *Recursive Bayesian calibration of depth sensors with non-overlapping views*, 2012 15th International Conference on Information Fusion (FUSION), 2012, pp. 757–762, ISBN 978-1-4673-0417-7.

[Kal60] Rudolph Emil Kalman, *A new approach to linear filtering and prediction problems*, Journal of Basic Engineering **82** (1960), no. 1, 35–45, ISSN 0021-9223, DOI 10.1115/1.3662552.

[KDDG14] Alexandros Kitsikidis, Kosmas Dimitropoulos, Stella Douka, and Nikos Grammalidis, *Dance analysis using multiple kinect sensors*, 2014 International Conference on Computer Vision Theory and Applications (VISAPP), IEEE, 2014, ISBN 978-9-8975-8133-5.

[KKS+14] Felix Leif Keppmann, Tobias Käfer, Steffen Stadtmüller, René Schubotz, and Andreas Harth, *High Performance Linked Data Processing for Virtual Reality Environments*, ISWC-PD'14 Proceedings of the 2014 International Conference on Posters & Demonstrations Track (Aachen, Germany), vol. 1272, CEUR-WS.org, 2014, pp. 193–196.

[KMWS13] Suttipong Kaenchan, Pornchai Mongkolnam, Bunthit Watanapa, and Sasipa Sathienpong, *Automatic multiple Kinect cameras setting for simple walking posture analysis*, 2013 International Computer Science and Engineering Conference

(ICSEC), 2013, DOI 10.1109/ICSEC.2013.6694787, pp. 245–249, ISBN 978-1-4673-5322-9.

[MF11] Andrew Maimone and Henry Fuchs, *Encumbrance-free telepresence system with real-time 3D capture and display using commodity depth cameras*, 2011 10th IEEE International Symposium on Mixed and Augmented Reality (ISMAR) (Piscataway, NJ), IEEE, 2011, DOI 10.1109/ISMAR.2011.6092379, pp. 137–146, ISBN 978-1-4577-2183-0.

[MZPHDP+14] M. Martínez-Zarzuela, M. Pedraza-Hueso, F. J. Díaz-Pernas, D. González-Ortega, and M. Antón-Rodríguez, *Indoor 3D Video Monitoring Using Multiple Kinect Depth-Cameras*, The International Journal of Multimedia & Its Applications **6** (2014), no. 1, 61–76, ISSN 0975-5934.

[OKO+12] Štěpán Obdržálek, Gregorij Kurillo, Ferda Ofli, Ruzena Bajcsy, Edmund Seto, Holly Jimison, and Michael Pavel, *Accuracy and robustness of Kinect pose estimation in the context of coaching of elderly population*, Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (Piscataway, NJ), IEEE, 2012, DOI 10.1109/EMBC.2012.6346149, pp. 1188–1193, ISBN 978-1-4244-4119-8.

[PMC+11] François Pomerleau, Stéphane Magnenat, Francis Colas, Ming Liu, and Roland Siegwart, *Tracking a depth camera: Parameter exploration for fast ICP*, 2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (Piscataway, NJ) (Nancy M. Amato, ed.), IEEE, 2011, DOI 10.1109/IROS.2011.6094861, pp. 3824–3829, ISBN 978-1-61284-454-1.

[RGTR16] Michael Rietzler, Florian Geiselhart, Janek Thomas, and Enrico Rukzio, *FusionKit: A Generic Toolkit for Skeleton, Marker and Rigid-body Tracking*, Proceedings of the 8th ACM SIGCHI Symposium on Engineering Interactive Computing Systems (New York, NY, USA), ACM, 2016, DOI 10.1145/2933242.2933263, pp. 73–84, ISBN 978-1-4503-4322-0.

[RL01] Szymon Rusinkiewicz and Marc Levoy, *Efficient variants of the ICP algorithm*, Third International Conference on 3-D Digital Imaging and Modeling, 2001. Proceedings, 2001, DOI 10.1109/IM.2001.924423, pp. 145–152, ISBN 0-7695-0984-3.

[SGF+12] Jamie Shotton, Ross Girshick, Andrew Fitzgibbon, Toby Sharp, Mat Cook, Mark Finocchio, Richard Moore, Pushmeet Kohli, Antonio Criminisi, Alex Kipman, and Andrew Blake, *Efficient Human Pose Estimation from Single Depth Images*, IEEE Transactions on Pattern Analysis and Machine Intelligence **35** (2012), no. 12, 2821–2840, ISSN 0162-8828, DOI 10.1109/TPAMI.2012.241.

[SK13] Christian Schönauer and Hannes Kaufmann, *Wide Area Motion Tracking Using Consumer Hardware*, The International Journal of Virtual Reality **12** (2013), no. 1, 57–65, ISSN 1081-1451.

[WB10] Andrew Wilson and Hrvoje Benko, *Combining multiple depth cameras and projectors for interactions on, above and between surfaces*, UIST '10 Proceedings of the 23nd annual ACM symposium on User interface software and technology (New York, NY, USA), ACM, 2010, DOI 10.1145/1866029.1866073, pp. 273–282, ISBN 978-1-4503-0271-5.

[Wil15] Andy Wilson, *Roomalive toolkit*, GitHub `https://github.com/Kinect/RoomAliveToolkit`, 2015.

[WKOB15] Qifei Wang, Gregorij Kurillo, Ferda Ofli, and Ruzena Bajcsy, *Evaluation of Pose Tracking Accuracy in the First and Second Generations of Microsoft Kinect*, 2015 IEEE International Conference on Healthcare Informatics. ICHI 2015. Proceedings (Los Alamitos), IEEE, 2015, DOI 10.1109/ICHI.2015.54, pp. 380–389, ISBN 978-1-4673-9548-9.

[YKW13] Kwok-Yun Yeung, Tsz-Ho Kwok, and Charlie C. L. Wang, *Improved Skeleton Tracking by Duplex Kinects: A Practical Approach for Real-Time Applications*, Journal of Computing and Information Science in Engineering **13** (2013), no. 4, 041007, ISSN 1530-9827, Article no. JCISE-13-1078, DOI 10.1115/1.4025404.

[YZD+15] Lin Yang, Longyu Zhang, Haiwei Dong, Abdulhameed Alelaiwi, and Abdulmotaleb El Saddik, *Evaluating and improving the depth accuracy of Kinect for Windows v2*, IEEE Sensors Journal **15** (2015), no. 8, 4275–4285, ISSN 1530-437X, DOI 10.1109/JSEN.2015.2416651.

[ZSCL12] Licong Zhang, Jürgen Sturm, Daniel Cremers, and Dongheui Lee, *Real-time human motion tracking using multiple depth cameras*, 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (Piscataway, NJ), IEEE, 2012, DOI 10.1109/IROS.2012.6385968, pp. 2389–2395, ISBN 978-1-4673-1737-5.