Towards ubiquitous tracking: Presenting a scalable, markerless tracking approach using multiple depth cameras

MICHAEL M. OTTO[1], PHILIPP AGETHEN[1], FLORIAN GEISELHART[2], ENRICO RUKZIO[2]

[1]Daimler AG, Wilhelm-Runge-Str. 11, D-89081 Ulm, [firstname.lastname]@daimler.com
[2]Ulm University, James-Franck-Ring, D-89081 Ulm, [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 heuristics-based sensor fusion using skeletal tracking. Based on a distributed, service-oriented and scalable system architecture, a markerless tracking system consisting of multiple Kinect v2 sensors has been developed for real-time interaction with virtual environments. Evaluation showed that a system based on the proposed techniques help in increasing tracking areas, resolving occlusions and improving human posture analysis. This system is used for ergonomic assessments in production planning workshops and it was 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.

Categories and Subject Descriptors (according to ACM CCS): H.5.2 [User Interfaces]: Input devices and strategies—scalable, markerless tracking and full body motion capture

1. Introduction

Interactive virtual and augmented reality 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. However, considering spatially large use cases like car assembly, 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. One way to overcome these limitations is the use of multiple depth cameras which extend sensing range and improve tracking performance. But with this approach, there are also a number of new challenges which need to be addressed in order to successfully implement such a system. First of all, it is necessary to establish a common coordinate frame for the cameras 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.

In this paper, we propose a novel system consisting of multiple Kinect v2 cameras for the use in ergonomic assessments. We present a concept of a distributed multi-depth-camera system, whose improvements were also quantified in a systematic evaluation. 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. Then we propose

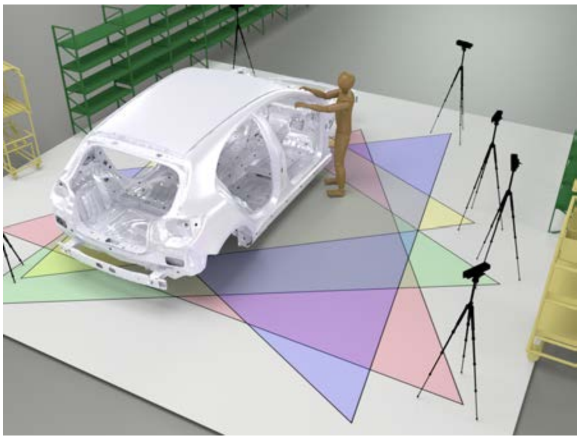

Figure 1: *Example setup of full body skeletal tracking in automotive car assembly*

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 concrete 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 aspects of such systems, thus leaving others unspecified to a great amount. As already lined out, different challenges have to be faced in order to successfully implement such a system: Architecture, interference, synchronization, registration and fusion. Depending on the use case, it is often also necessary to handle additional application specific issues like user identification or world coordinates registration, not being further considered within this work.

2.1. 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 presented by Schönauer [SK13] or Martínez-Zarzuela et al. [MZPHDP*14] 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]. In our system, a different approach has been chosen. All information can be requested via service-oriented and scalable RESTful tracking services as presented by Keppmann et al. [KKS*14].

2.2. Interference

As time-of flight (ToF) and structured light 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. As the proposed system uses ToF depth cameras, which generate negligible interference noise due to their modulation, no countermeasures against interference have been implemented.

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] and Baek und Kim [BK15]. Most of the methods yield comparable results, however strongly differing in the ease-of-use and setup time with different approaches. The proposed approach focuses on reduced setup times and an easy setup. Thus, we decided to implement a combination of skeleton-, regression plane- and ICP-based registration.

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 (see Caon et al. [CYT*11]), weighted averaging methods [FRZH12], 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.

2.5. Assessment of related work

While covering many of the relevant aspects, most of the previous works leave out important factors of a multiple depth-camera system for universal use. Generally, registration and fusion approaches 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.

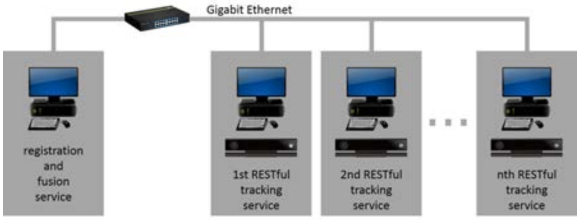
Gigabit Ethernet
registration and fusion service
1st RESTful tracking service
2nd RESTful tracking service
...
nth RESTful tracking service

Figure 2: *Hardware setup for the tracking system with service-oriented, distributed sensor services*

3. Hardware setup

The multiple depth-camera system consists of several Kinect v2 sensors due to affordable ToF hardware costs and improved depth accuracy compared to the first generation Kinect. The distributed system consists of several computers accommodating the tracking services using Kinect hardware and one high-performance computer connected by a 1 Gbit/s Ethernet network (see Figure 2). The sensor computers have only low requirements such as USB 3.0 support for the sensor connection. In this case we used an Intel Core i5-4200U 1.6 GHz with 4 GB DDR3 RAM. The central high-performance computer is based on an i7-4712HQ CPU with 16 GB DDR3L RAM. This computer is calculating the extrinsics and performs the fusion of the sensors.

4. Software

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

4.1. 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.

4.2. Fusion and multi sensor tracking service

The fusion and multi-sensor service is running as a central component on a high performance computer and handles registration, fusion and data input/output in the tracking environment. Figure 3 depicts the architecture of the registration and fusion service.

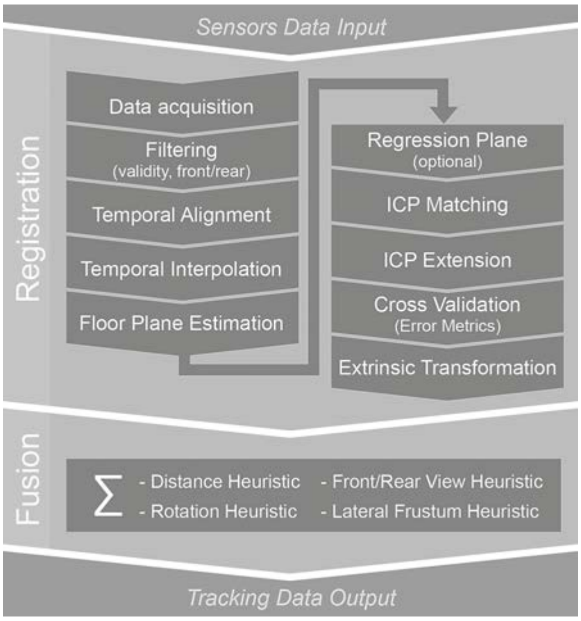

Figure 3: *System architecture of fusion service*

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 Figrure 3) 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. This joint was chosen, due to its advantages compared to the remaining joints: front/rear invariance, orientation-independent position and low overall jitter. 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 heuristics-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.

4.2.1. ICP extension

Gathering only the neck joint information has an 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 lie almost on a single plane. To compensate this lack of variance, additional information is used. The floor plane estimation compensates the missing information. The floor plane is a rough approximation of the distance to the floor and the 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 precise the ICP results, if enough feature points have been gathered during the user's movement.

4.2.2. 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 were 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.

4.2.3. Scalability

To achieve a fully scalable system with a common coordinate frame, extrinsic transformation chains have to be built (see Figure 4). 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.

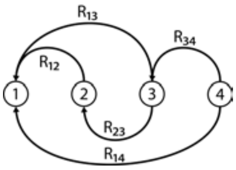

Figure 4: *Registration result for three sensor system in the arrangement for a tracking area range extension*

4.2.4. 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 move during registration process 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 Hz, the inter-frame trajectory between two skeleton datagrams can be assumed as linear movement.

4.2.5. 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 m tracking results are most reliable. This quality measure weights the user's skeleton over the distance to the neck joint respectively.

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

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:

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

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:

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

5. 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.

5.1. 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 mm for the whole tracking volume. On the Kinect sensors, rigid body markers were applied on the top of the sensor. 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.

5.2. 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.

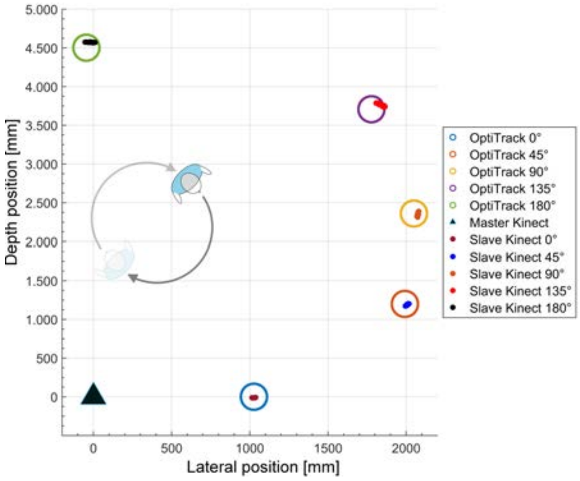

Figure 5: *Top view on the registration results: Master sensor at [0,0], 5 scenarios with 20 registrations each, circles indicate the ground truth of the OptiTrack measurements*

5.3. Results

Figure 5 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 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 mm, depending on the angle of the user.

6. Ergonomic assessments

One specific use case within the automotive industry where full body motion capture data is used are ergonomic assessments of workplaces. Ergonomics experts are using motion capture technology to virtually audit end-assembly workplaces. 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

6.1. Experimental hardware & software setup

During these assessment workshop 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 approximately 6 m x 6 m since movements within real workplaces in automotive end-assembly lines have equal dimensions.

Having registered all cameras to a common world coordinate frame, the presented system architecture in combination

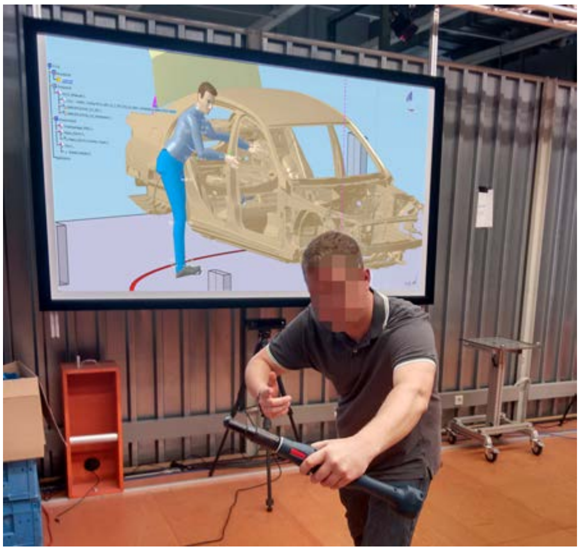

Figure 6: *Delmia V5 DHM directly manipulated by markerless motion capture approach. All assessments can be carried out in real time.*

with the fusion heuristics enable the worker to be constantly tracked regardless of his position and his orientation within the concatenated tracking frustum.

Standardized tracking protocols have been implemented to connect to commercial VR software used: A.R.Tracking protocol, VRPN and ARVIDA linked data protocol. 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 6 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.

6.2. 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 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 markerless 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.

7. Conclusion

In our effort to improve multiple depth-camera systems, we developed a novel large-scale multi depth camera system, which supports scalable setups and different use cases through its service-oriented architecture. The number of possible tracking nodes is limited by computing power and network throughput. Setups up to ten tracking services have been successfully tested but more sensors should be possible as long as network throughput is sufficient. The fusion service itself can be addressed transparently and acts externally as if it was a single sensor tracking service. Standardized tracking protocols have been implemented in order to achieve interoperability. Furthermore, several novel registration-relevant techniques have been 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.

Three pilot test cases within the automotive industry have been 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 proved its applicability, reduced costs and the ease-of-use, it will complement the variety of existing industrial tracking systems.

In future work, we plan to refine the fusion process by extending heuristics with additional criteria and more fine-grained weighting, e.g. on a per-joint or per-bone level instead of the current per-body approach. Additional measurements and analysis may also broaden the insight on the behavior of the proprietary Kinect technology and thus lead to further improvements in this approach.

8. 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

[ACZ*13] Asteriadis S., Chatzitofis A., Zarpalas D., Alexiadis D. S., Daras P.: Estimating human motion from multiple kinect sensors. In *MIRAGE '13 Proceedings of the 6th International Conference on Computer Vision / Computer Graphics Collaboration Techniques and Applications* (New York, NY, USA, 2013), MIRAGE '13, ACM, pp. 3:1–3:6. 2

[AMC13] Alexopoulos K., Mavrikios D., Chryssolouris G.: Ergotoolkit: An ergonomic analysis tool in a virtual manufacturing environment. *Int. J. Comput. Integr. Manuf. 26*, 5 (May 2013), 440–452. 6

[BIH*12] Butler D. A., Izadi S., Hilliges O., Molyneaux D., Hodges S., Kim D.: Shake'n'sense: reducing interference for overlapping structured light depth cameras. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2012), CHI '12, ACM, pp. 1933–1936. 2

[BK15] Baek S., Kim M.: Dance experience system using multiple kinects. *International Journal of Future Computer and Communication 4*, 1 (2015), 45–49. 2

[BRB*11] Berger K., Ruhl K., Brümmer C., Schröder Y., Scholz A., Magnor M.: Markerless motion capture using multiple color-depth sensors. In *Proc. Vision, Modeling and Visualization (VMV) 2011* (2011), pp. 317–324. 2

[CYT*11] Caon M., Yue Y., Tscherrig J., Mugellini E., Abou Khaled O.: Context-aware 3d gesture interaction based on multiple kinects. In *AMBIENT 2011 : The First International Conference on Ambient Computing, Applications, Services and Technologies* (2011), pp. 7–12. 2

[FRZH12] Faion F., Ruoff P., Zea A., Hanebeck U. D.: Recursive bayesian calibration of depth sensors with non-overlapping views. In *2012 15th International Conference on Information Fusion (FUSION)* (2012), pp. 757–762. 2

[KKS*14] Keppmann F. L., Käfer T., Stadtmüller S., Schubotz R., Harth A.: High performance linked data processing for virtual reality environments. In *Proceedings of the ISWC 2014 Posters & Demonstrations Track a track within the 13th International Semantic Web Conference, ISWC 2014, Riva del Garda, Italy, October 21, 2014* (2014), pp. 193–196. 2, 3

[MF11] Maimone A., Fuchs H.: Encumbrance-free telepresence system with real-time 3d capture and display using commodity depth cameras. In *2011 10th IEEE International Symposium on Mixed and Augmented Reality (ISMAR)* (2011), pp. 137–146. 2

[MZPHDP*14] Martínez-Zarzuela M., Pedraza-Hueso M., Díaz-Pernas F. J., González-Ortega D., Antón-Rodríguez M.: Indoor 3d video monitoring using multiple kinect depth-cameras. *arXiv:1403.2895 [cs]* (2014). 2

[PMC*11] Pomerleau F., Magnenat S., Colas F., Liu M., Siegwart R.: Tracking a depth camera: Parameter exploration for fast icp. In *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2011), pp. 3824–3829. 2

[RL01] Rusinkiewicz S., Levoy M.: Efficient variants of the icp algorithm. In *3-D Digital Imaging and Modeling, 2001. Proceedings. Third International Conference on* (2001), IEEE, pp. 145–152. 2

[SK13] Schönauer C., Kaufmann H.: Wide area motion tracking using consumer hardware. *The International Journal of Virtual Reality 12*, 1 (2013), 1–9. 2

[WB10] Wilson A., Benko H.: Combining multiple depth cameras and projectors for interactions on, above and between surfaces. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology* (New York, NY, USA, 2010), UIST '10, ACM, pp. 273–282. 2

[Wil15] Wilson A.: Roomalive toolkit. https://github.com/Kinect/RoomAliveToolkit, 2015. 2

[YKW13] Yeung K.-Y., Kwok T.-H., Wang, Charlie C. L.: Improved skeleton tracking by duplex kinects: A practical approach for real-time applications. *Journal of Computing and Information Science in Engineering 13*, 4 (2013), 041007. 2

[ZSCL12] Zhang L., Sturm J., Cremers D., Lee D.: Real-time human motion tracking using multiple depth cameras. In *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* (2012), pp. 2389–2395. 2