
Acquisition of high-quality images for camera calibration in robotics applications via speech prompts

PREPRINT

✉ Timm Linder^{*1}, ✉ Kadir Yilmaz^{†2}, David Adrian¹, and ✉ Bastian Leibe²

¹Bosch Corporate Research & Bosch Center for AI, Renningen, Germany

²Computer Vision Group, RWTH Aachen University, Germany

ABSTRACT

Accurate intrinsic and extrinsic camera calibration can be an important prerequisite for robotic applications that rely on vision as input. While there is ongoing research on enabling camera calibration using natural images, many systems in practice still rely on using designated calibration targets with e. g. checkerboard patterns or April tag grids. Once calibration images from different perspectives have been acquired and feature descriptors detected, those are typically used in an optimization process to minimize the geometric reprojection error. For this optimization to converge, input images need to be of sufficient quality and particularly sharpness; they should neither contain motion blur nor rolling-shutter artifacts that can arise when the calibration board was not static during image capture. In this work, we present a novel calibration image acquisition technique controlled via voice commands recorded with a clip-on microphone, that can be more robust and user-friendly than e. g. triggering capture with a remote control, or filtering out blurry frames from a video sequence in postprocessing. To achieve this, we use a state-of-the-art speech-to-text transcription model with accurate per-word timestamping to capture trigger words with precise temporal alignment. Our experiments show that the proposed method improves user experience by being fast and efficient, allowing us to successfully calibrate complex multi-camera setups.

1 INTRODUCTION

Many applications in 2D and 3D perception for robotic systems rely on accurate camera calibration, for instance methods for object detection [Liang et al., 2021; Guan et al., 2024], simultaneous localization and mapping (SLAM) [Damjanović et al., 2025], or robotic manipulation [An et al., 2024]. One concrete example is the monocular 3D human pose estimation method MeTRAbs [Sárádi et al., 2021], that we utilize on our mobile robotics platform (Figure 1) for human-robot interaction tasks and human-aware navigation [Stefanini et al., 2024] in agile production and intralogistics scenarios. The approach requires knowledge of the pinhole camera intrinsics as input in order to perform absolute pose recovery. Likewise, its recently proposed extension to fisheye optics [Käs et al., 2025] requires calibration with a fisheye camera model.

^{*}first.last@de.bosch.com

[†]lastname@vision.rwth-aachen.de



Figure 1: Our use-case is the calibration of pinhole and fisheye cameras in robotics applications using dedicated calibration targets (e. g. checkerboard patterns, April tag grids).

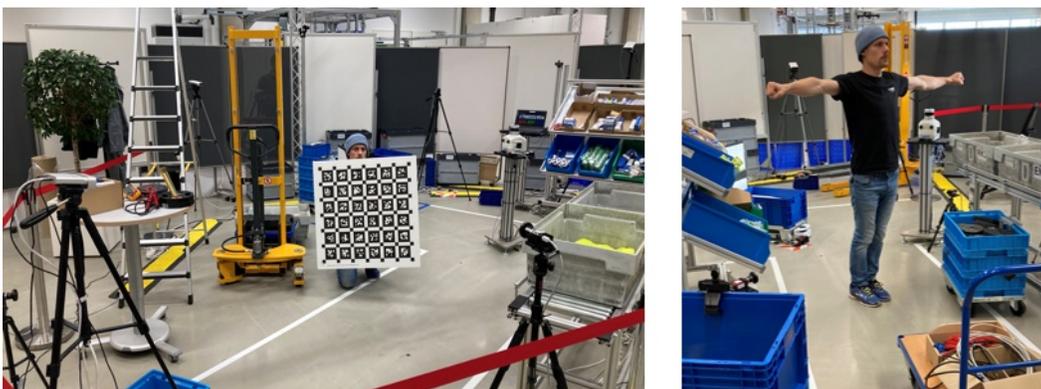


Figure 2: The proposed speech-guided image acquisition method can also be used to calibrate complex multi-view systems, as in the example setup here with >10 cameras on tripods to obtain groundtruth 3D human poses via multi-view triangulation.

Furthermore, to record training data e. g. for 3D human pose estimation approaches, complex multi-view camera setups are typically used (Figure 2), often containing around a dozen cameras that are spread out through the scene. In addition to the camera intrinsics, also extrinsics are required in this case, in order to be able to triangulate accurate groundtruth 3D human poses.

The same applies when calibrating multiple cameras on a robot relative to each other, for example to achieve 360-degree surround view coverage on an omnidirectional drive robot like ours, for downstream tasks such as multi-modal 3D object detection [Piekenbrinck et al., 2024], or multi-modal human detection using cameras and lidar [Linder et al., 2021].

While ongoing research explores on how to perform calibration using in-the-wild images [Jin et al., 2023], in our work we assume that a calibration target (April tag grid or checkerboard pattern) is used for camera calibration, by placing it in different poses relative to the camera. Based upon our extensive experience with calibrating the aforementioned multi-camera systems, using both open-source [Schneider et al., 2014; Furgale et al., 2013] and commercially available [Wilm and Eiriksson, 2018] calibration software, we observe that it is crucial to acquire high-quality calibration images to obtain good results. To achieve this goal, we propose a novel technique for speech-guided calibration image acquisition, which allows the operator to securely hold the calibration target with both hands, without requiring any additional support person during the calibration process.



Figure 3: Alternative approaches that we evaluated, with different shortcomings that make them impractical when calibrating complex multi-camera systems spread out over larger areas.

2 METHOD

Problem definition The main challenge that we aim to address with our method is the elimination, or reduction, of calibration images that exhibit motion blur and rolling shutter artifacts, which can cause feature extraction and subsequently camera calibration to fail. In our experience, this issue occurs frequently when recording continuous calibration video sequences while the operator is in motion along with the calibration pattern, as illustrated in the example of Figure 3c.

We also observe that motion blur can become more severe under difficult lighting situations that are often encountered in industrial factory and warehouse environments (Figure 1, right).

2.1 Alternative approaches

Before introducing the proposed method, we briefly review other related approaches that we have experimented with, and that did not yield satisfactory results.

Adjusting acquisition parameters By reducing exposure time, increasing sensor gain, improving lighting, or switching to a more light-sensitive camera/lens combination, the amount of motion blur can be reduced. However, not all of these parameters are always under the user’s control. For example, in our case, we are also interested in experimenting with very low-cost sensors and lenses, that start to exhibit motion blur (or heavy pixel noise at high gain) already under benign lighting conditions during daylight.

Wired triggering Using a wired keyboard or mouse as input (Figure 3a) requires the operator to put the calibration target to rest, e. g. using a tripod, and then walk to the computer to press a key in order to trigger image acquisition. As this process can be very time-consuming, it is not a practical option in multi-camera setups that extend over a larger space, unless a second person assists with the calibration effort, which increases operation cost.

Remote-controlled triggering During initial experiments with our multi-view capture setup, we therefore tried out several remote-controlled, bluetooth-based triggers (e. g. a wireless presenter, see Figure 3b). However, these proved to be unreliable at distances over 3-4m, or when the line-of-sight to the receiver was obstructed by environmental structures or the calibration target. Also, triggering delays sometimes appeared non-deterministic, reaching up to 2-3 seconds in some cases. In our experiments, this often led to the situation that the operator had already started moving again, while the image had not yet been recorded. Furthermore, the operator may need both hands to securely hold the calibration target, thus hitting keys on a remote while not introducing additional motion jitter can be an intricate endeavour.

Automated detection of low-quality calibration images Various metrics for detecting and quantifying blurriness of images with (motion) blur artifacts (Figure 3c) have been proposed [Andrade, 2021], e. g. based upon variance of the Laplacian, other handcrafted metrics [Liu et al., 2008], or

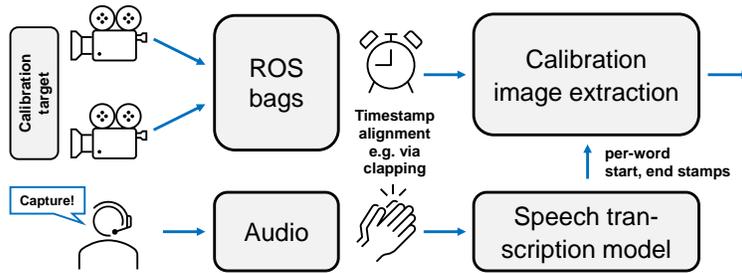


Figure 4: Flow diagram of the proposed approach. In our implementation, we use WhisperX [Bain et al., 2023] as the speech transcription model, which provides highly accurate per-word timestamps.

deep neural networks [Kim et al., 2018]. The more general task is called blind image quality assessment [Saha et al., 2023], if no groundtruth image is given. The challenge here is to find an approach that generalizes well over different camera/lens combinations and scene compositions (foreground, background objects), without having to adjust manually thresholds for every scene or frame.

Motion detection Finally, one could also attempt to curate images from a video where the calibration target is static, by detecting motion across a neighboring set of frames using a sliding-window based technique, e. g. using optical flow [Alfarano et al., 2024]. However, this easily fails if there are objects in the scene that naturally exhibit motion, e. g. ceiling fans, moving people in the background, or the constantly blinking LED strips that we use for intention communication on our robot.

2.2 Our proposed method

Overview We propose to utilize speech commands to trigger image acquisition whenever the operator has positioned the calibration target in a new steady pose (for example by placing it on a stand, chair, or a forklift in our case). We use a configurable trigger word like “Capture!” for this purpose. By using speech as opposed to a remote control, the operator has both hands free to move around or securely hold the calibration target. For full control over the image extraction process, we record aligned, continuous video and audio sequences, and then extract candidate frames based upon the speech prompts offline in post-processing. For multi-camera setups where also extrinsics need to be computed, we ensure to perform image extraction in a synchronized fashion, based on ROS message header timestamps.

Audio-video synchronization To align separately recorded audio from a clip-on microphone with the video recordings from one or multiple cameras, the operator marks the beginning of a new calibration sequence by clapping the hands, which is clearly discernible in both audio and video. While we currently manually determine the claps timestamps, which is required only once per session, this could also be automated using e. g. hand/body pose estimation and audio signal detection.

Detection of trigger words To detect and temporally localize the trigger words, we use a speech transcription model based upon Whisper [Radford et al., 2023], a state-of-the-art general-purpose speech recognition model trained using weak supervision on a large-scale multilingual audio dataset (~680,000 hours). As a multitasking model, it can perform speech recognition, speech translation, and language identification, however, its per-word timestamps for longer audio sequences are not very accurate, and in our experiments it was sometimes prone to hallucinations. Therefore, we utilize the more recent WhisperX [Bain et al., 2023], which extends Whisper with accurate word-level timestamping through forced phoneme alignment via wav2vec2 [Baevski et al., 2020] as well as voice activity detection to reduce hallucinations, both of which proved to be very helpful in our scenario. We use the mean of the start and end timestamps of the detected trigger word as timestamp for image extraction.

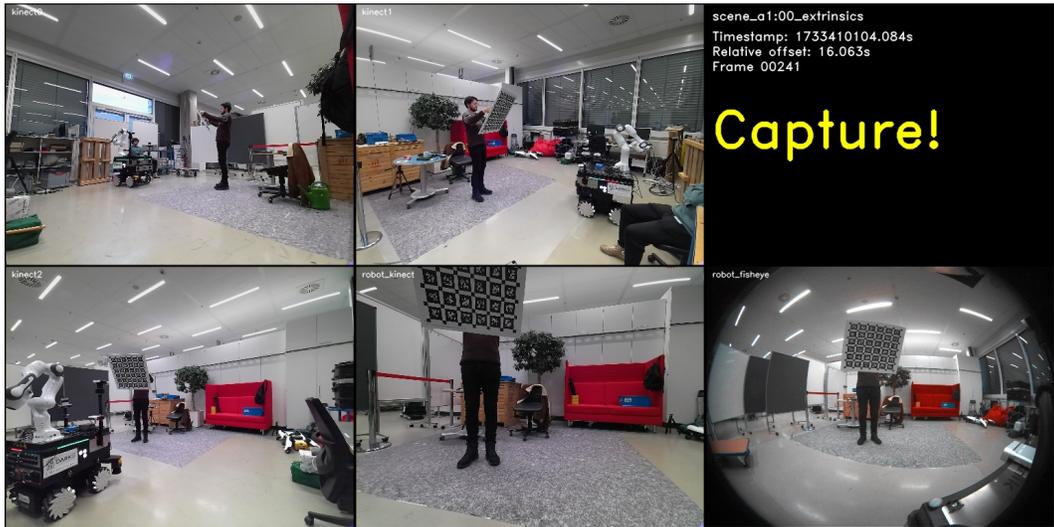


Figure 5: Our user interface shows an operator recording calibration images in a synchronized multi-camera setup involving our robot platform (lower left). He prompts image acquisition during steady poses via speech commands, while having his hands free to securely hold and move the calibration target. The recognized trigger word is shown on the right in yellow.

Further processing As shown in Fig. 4, once the timestamps of all “Capture!” commands have been determined using the proposed method, we extract the corresponding images and feed them into a commercially available camera calibration suite [Wilm and Eiriksson, 2018], to obtain intrinsics and extrinsics for our pinhole and fisheye cameras.

User interface To visualize the synchronized image frames from our multi-view camera setups, along with the extracted trigger words, we have further implemented a graphical user interface using OpenCV, shown in Figure 5, helping us verify the precise alignment of extracted trigger timestamps.

3 EXPERIMENTS

We now want to demonstrate that the proposed calibration image acquisition technique can be successfully used to calibrate cameras on a robotic system.

Experimental setup For our experiments, we recorded calibration video sequences of around 5 minutes length using 4 Azure Kinect pinhole RGB cameras and one 5.1MP machine vision fisheye camera as ROS bagfiles. One Kinect and the fisheye camera were mounted on our robot and connected to its internal PC, while three of the Kinects have been placed on tripods and connected to an external computer that records also the audio from a low-cost clip-on microphone via a USB radio receiver. Both computers’ clocks have been synchronized via network time protocol (NTP), the clock of the machine vision camera is synchronized via precision time protocol (PTP). Around 50 speech prompts for image acquisition have been recorded per session and were stored as Ogg Vorbis audio files. We use an April tag grid with 6 rows and 6 columns, printed on a 1x1 meter rigid surface to which we attached handholds using aluminium profiles, as calibration target.

Camera models for calibration For calibration of pinhole images, we use the standard OpenCV pinhole camera model with 3 radial and 2 tangential distortion coefficients. For calibration of fisheye images, we use the Double-Sphere model by Usenko et al. [2018], which yielded robust calibration results on 5 different models of fisheye lenses that we experimented with.

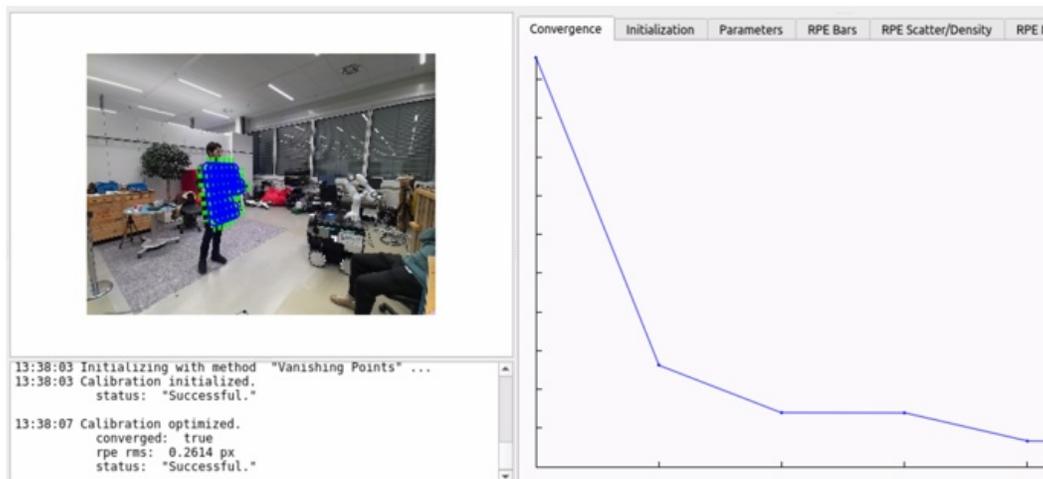


Figure 6: Successfully converging calibration using an existing calibration software [Wilm and Eiriksson, 2018] and motion blur-free calibration images extracted using our proposed approach.

Qualitative results To understand whether our method helps in obtaining high-quality calibration images, we visually inspect the extracted image frames and manually assess them for their quality (sharpness, lack of motion blur). Thanks to our proposed approach, no noticeable motion blur has been observed in any image, as the operator consistently issued speech prompts while he was standing still, such that the calibration target was resting in a stationary pose.

Quantitative evaluation We then use the extracted calibration images to perform actual calibrations using existing commercial calibration software [Wilm and Eiriksson, 2018]. As visualized exemplarily in Figure 6, the calibration using the images extracted via speech prompt quickly converges. We obtain root mean square reprojection errors of on average <0.5 px, which can be considered a successful calibration result.

Video material Our supplementary video³ provides further examples and insights into our proposed calibration image acquisition process. At the end, the video also includes a working demonstration of our user interface from Figure 5, showing trigger words and associated timestamps as they are being extracted by the proposed approach.

Study on user satisfaction In a small user study, three different subjects that we interviewed noted that the proposed method for calibration image acquisition is more convenient to use than 1) a bluetooth-based trigger, 2) running back and forth between calibration board and PC to trigger frame capture, 3) performing calibration on the entire video sequence. It is noteworthy that the latter can take around 6 hours due to the lengthy feature detection process with such a large number of frames and multiple cameras. Therefore, prior extraction of relevant key frames is essential, which is greatly simplified by our proposed technique.

4 CONCLUSIONS

In this paper, we proposed a novel technique for acquisition of high-quality calibration images via speech prompts. Our method allows the operator to securely hold the calibration target still with both hands, without requiring any additional support person during the calibration process, making the process robust, fast and efficient.

³<https://www.youtube.com/watch?v=HFjTqGHMxIw>

Using the proposed technique, which leverages a state-of-the-art AI-based method for accurately timestamped speech recognition, we have been easily able to calibrate multiple multi-camera setups. The calibrated camera setups have been used to record novel datasets for our ongoing research. We believe that our method and lessons learned can help robotics researchers to streamline their camera calibration workflows.

In future work, we want to explore how we can utilize the proposed setup also for speech-guided 3D scene labelling tasks, where precise per-word timestamps are of equally high importance as semantically accurate speech-to-text transcription quality, and powerful multi-task language models such as WhisperX can demonstrate their full potential.

Acknowledgments

The preparation of this manuscript and its experiments have been supported by the EU Horizon 2020 research and innovation program under grant agreement 101017274 (DARKO). Support in data acquisition and ideation was provided by the project “Context Understanding for Autonomous Systems” by Robert Bosch GmbH.

References

- Andrea Alfano, Luca Maiano, Lorenzo Papa, and Irene Amerini. Estimating optical flow: A comprehensive review of the state of the art. *Computer Vision and Image Understanding*, 249: 104160, 2024. ISSN 1077-3142. doi: 10.1016/j.cviu.2024.104160.
- Boshi An, Yiran Geng, Kai Chen, Xiaoqi Li, Qi Dou, and Hao Dong. RGBManip: Monocular image-based robotic manipulation through active object pose estimation. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 7748–7755, 2024. doi: 10.1109/ICRA57147.2024.10610690.
- J. Andrade. *A Survey of Blur Detection and Sharpness Assessment Methods*. Synthesis Lectures on Algorithms and Software in Engineering. Morgan & Claypool Publishers, 2021. doi: 10.1007/978-3-031-01529-8.
- Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: a framework for self-supervised learning of speech representations. In *Proc. International Conference on Neural Information Processing Systems (NIPS)*, 2020. doi: 10.5555/3495724.3496768.
- Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. WhisperX: Time-accurate speech transcription of long-form audio. *INTERSPEECH*, 2023. doi: 10.21437/Interspeech.2023-78.
- D. Damjanović, P. Biočić, S. Prkljačić, et al. A comprehensive survey on SLAM and machine learning approaches for indoor autonomous navigation of mobile robots. *Machine Vision and Applications*, 36(55), 2025. doi: 10.1007/s00138-025-01673-0.
- Paul Furgale, Joern Rehder, and Roland Siegwart. Unified temporal and spatial calibration for multi-sensor systems. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 1280–1286, 2013. doi: 10.1109/IROS.2013.6696514.
- Jian Guan, Yingming Hao, Qingxiao Wu, Sicong Li, and Yingjian Fang. A survey of 6DoF object pose estimation methods for different application scenarios. *Sensors*, 24(4), 2024. doi: 10.3390/s24041076.
- Linyi Jin, Jianming Zhang, Yannick Hold-Geoffroy, Oliver Wang, Kevin Blackburn-Matzen, Matthew Sticha, and David F. Fouhey. Perspective fields for single image camera calibration. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. doi: 10.1109/CVPR52729.2023.01660.
- Stephanie Käs, Sven Peter, Henrik Thillmann, Anton Burenko, David Adrian, Dennis Mack, Timm Linder, and Bastian Leibe. Systematic comparison of projection methods for monocular 3D human pose estimation on fisheye images. In *IEEE Intl. Conference on Robotics and Automation (ICRA)*, 2025. To appear.

- Beomseok Kim, Hyeongseok Son, Seong-Jin Park, Sunghyun Cho, and Seungyong Lee. Defocus and motion blur detection with deep contextual features. *Computer Graphics Forum*, 37(7), 2018. doi: 10.1111/cgf.13567.
- W. Liang, P. Xu, L. Guo, et al. A survey of 3D object detection. *Multimed Tools Appl*, 80: 29617–29641, 2021. doi: 10.1007/s11042-021-11137-y.
- Timm Linder, Narunas Vaskevicius, Robert Schirmer, and Kai O. Arras. Cross-modal analysis of human detection for robotics: An industrial case study. In *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2021. doi: 10.1109/IROS51168.2021.9636158.
- Renting Liu, Zhaorong Li, and Jiaya Jia. Image partial blur detection and classification. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008. doi: 10.1109/CVPR.2008.4587465.
- Jens Piekenbrinck, Alexander Hermans, Narunas Vaskevicius, Timm Linder, and Bastian Leibe. RGB-D Cube R-CNN: 3D object detection with selective modality dropout. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2024. doi: 10.1109/CVPRW63382.2024.00205.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Int. Conference on Machine Learning (ICML)*, 2023. doi: 10.5555/3618408.3619590.
- Avinab Saha, Sandeep Mishra, and Alan C. Bovik. Re-IQA: Unsupervised learning for image quality assessment in the wild. In *IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. doi: 10.1109/CVPR52729.2023.00566.
- Thomas Schneider, Paul Furgale, Jérôme Maye, and Jörn Rehder. The Kalibr visual-inertial calibration toolbox, 2014. URL <https://github.com/ethz-asl/kalibr>.
- István Sárádi, Timm Linder, Kai Oliver Arras, and Bastian Leibe. MeTRAbs: Metric-scale truncation-robust heatmaps for absolute 3D human pose estimation. *IEEE Trans. Biometrics, Behavior, and Identity Science*, 2021. doi: 10.1109/tbiom.2020.3037257.
- Elisa Stefanini, Luigi Palmieri, Andrey Rudenko, Till Hielscher, Timm Linder, and Lucia Pallottino. Efficient context-aware model predictive control for human-aware navigation. *IEEE Robotics and Automation Letters*, 9(11):9494–9501, 2024. doi: 10.1109/LRA.2024.3461552.
- Vladyslav Usenko, Nikolaus Demmel, and Daniel Cremers. The Double Sphere camera model. In *International Conference on 3D Vision*, 2018. doi: 10.1109/3DV.2018.00069.
- J. Wilm and E.R. Eiriksson. calib.io camera calibrator software, 2018. URL <https://calib.io/>.