Wide Area Optical User Tracking in Unconstrained Indoor Environments

Annette Mossel* Hannes Kaufmann†

Interactive Media Systems Group
Vienna University of Technology

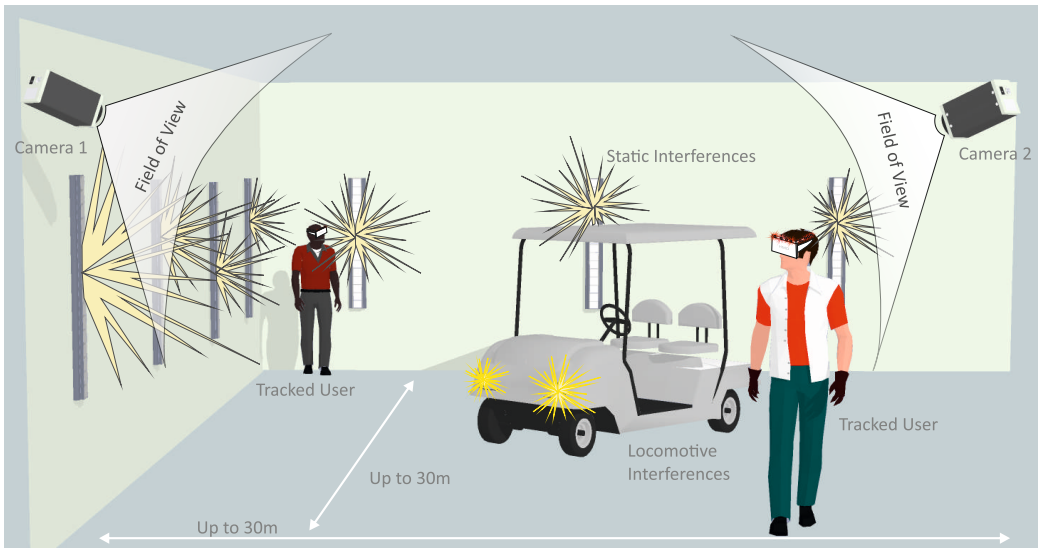


Figure 1: Multiple User Tracking in a large unconstrained indoor environment using two cameras.

ABSTRACT

In this paper, we present a robust infrared optical 3D position tracking system for wide area indoor environments up to 30m. The system consists of two shutter-synchronized cameras that track multiple targets, which are equipped with infrared light emitting diodes. Our system is able to learn targets as well as to perform extrinsic calibration and 3D position tracking in unconstrained environments, which exhibit occlusions and static as well as locomotive interfering infrared lights. Tracking targets can directly be used for calibration which minimizes the amount of necessary hardware. With the presented approach, limitations of state-of-the-art tracking systems in terms of volume coverage, sensitivity during training and calibration, setup complexity and hardware costs can be minimized. Preliminary results indicate interactive tracking with minimal jitter $< 0.0675mm$ and 3D point accuracy of $< 9.22mm$ throughout the entire tracking volume up to 30m.

Index Terms: I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism —Virtual Reality; I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Tracking;Stereo; I.4.9 [Image Processing and Computer Vision]: Applications—;

1 MOTIVATION & CONTRIBUTION

Virtual reality (VR) has applications in numerous domains, such as training, medicine, psychological and physical rehabilitation, education, edutainment and manufacturing. In each VR application domain, estimation of the user's position and orientation (tracking) is a crucial part to enable interaction within a virtual environment (VE). Various techniques and systems exist to estimate position and orientation – yielding in six degrees of freedom, 6DOF – of objects in space (Section 2). Recently emerged low-cost hardware such as the head mounted display Oculus Rift, the Razer Hydra for 6DOF interaction as well as the Microsoft Kinect for full body motion capture massively lowered the initial costs to build a small but fully immersive VE. However, low-cost wide area tracking with high precision remains a challenge. To extend the tracking coverage, existing vision-based approaches usually employ a large amount of cameras in the volume yielding significant high costs as well as complex setup and maintenance routines, making it impractical for general use. Furthermore, state-of-the-art tracking systems are sensitive to environmental interferences such as lights and reflexions, especially during target training and camera calibration. Using such systems in unconstrained indoor environments results in error prone and hence inaccurate tracking data. To our best knowledge, no research has been published about infrared optical-tracking of volumes larger than a few cubic meters (e.g. [23], 4x4x3m) with a cost efficient setup and a minimum of necessary vision hardware while providing accurate tracking data, robust training, camera calibration and tracking in unconstrained indoor environments.

The following three limitations, (1) tracking coverage, (2) system sensitivity and (3) costs impede the further employment of virtual reality scenarios for applications that are located in unconstrained environments such as rooms with wall illumination, entertainment stages, manufacturing workshops or even construction sites. Therefore, there is a need for a low-cost, high precision, reliable and robust wide area tracking system.

*e-mail: mossel@ims.tuwien.ac.at
†e-mail: kaufmann@ims.tuwien.ac.at

1.1 Contribution

In this paper, we present a novel robust infrared optical tracking system that provides high-precision and low-latency 3D position tracking of multiple targets at distances up to 30m. As depicted in Figure 1, the system only requires two cameras so that necessary hardware can be reduced to a minimum to provide cost efficacy as well as ease of use during setup and maintenance. At each stage of the system's workflow – during target training, extrinsic camera calibration and 3D position tracking – no pre-conditioning of the tracking volume is necessary. This enables our system to fully function in unconstrained indoor environments with static and moving light sources. The proposed 2D geometric target design allows for quick (re)-configuration and enhances robustness against accidental break-offs compared to 3D rigid body targets [24]. Targets are equipped with standard infrared light emitting diodes and are re-used for camera calibration.

By overcoming the limitations of existing vision-based tracking systems, we expect our approach to advance future VEs by 1) providing wide area tracking of multiple targets while lowering the costs, 2) making the complete workflow of an optical tracking systems robust against interferences to allow for quick setup and maintenance even by non-experts, and 3) enabling tracking in large unconstrained application scenarios such as entertainment stages, workshops or construction sites that opens novel fields for VR applications.

2 Related Work

For object tracking in large volumes, different techniques exist from commercially available products to on-going research prototypes. Extensive research has been performed to develop indoor location systems (ILS) for enabling context aware applications, user tracking and surveillance [12]. The most relevant wide area tracking technologies are radio frequency (RF), ultra-sonic and vision-based systems. Since they all have advantages and disadvantages regarding accuracy, latency, reliability, scalability and cost, no de-facto standard has been established yet.

2.1 Radio Frequency & Ultra Sound

RF systems based on Wi-Fi infrastructure or radio-frequency identification (RFID) [7] require a number of readers within the measurement volume to enable object tracking with low latency in large volumes [16]. However, WiFi signals tend to be extremely noisy and signal strength highly depends on surrounding building structures and materials. Thus, precise position estimation cannot be guaranteed even with multiple readers in the volume. In addition, the extensive pre-conditioning of the tracking volume is cost-intensive due to the amount of necessary hardware. Recently, a number of commercially available ILS applications such as Google Indoor Maps [8], SensionLab [27] as well as Indoo.Rs [14] emerged to localize a smartphone (and thus its user) by fusing mobile cellular data, WiFi and inertial measurements to minimize position jitter from WiFi data. Google Indoor Maps optimizes the position accuracy by pre-measuring and mapping the signal strength of the WiFi spot within the volume. However, this process takes time before the actual tracking can start. Furthermore, all systems require pre-built indoor floor plans for position visualization and only provide – in best case – several meter accuracy.

Ultra-sonic location systems such as [21, 11] rely on time-of-flight measurement of ultra-sonic signals, calculated using the velocity of sound. Such systems are scalable and can track multiple moving objects. However, current systems offer in the very best case meter-level accuracy under optimal conditions for 3D position estimation [13]. Furthermore, precision and range are not reliable since velocity of sound in the air is highly dependent on environmental conditions, especially humidity and temperature. Especially at long ranges, ultra-sonic systems are often extremely noisy and for that reason not a proper solution for our system's objectives. Compared to ultrasound, the RF-based Ultra Wide Band (UWB) technology enables distance measurements without line-of-sight requirements. An example for such a system is Ubisense [31] that employs TDoA[1] and AoA[2] measurements between mobile tags and a minimum of four fixed base stations. It offers fast signal speed and hence high sample rates (approximately 135 Hz) and provides an accuracy of down to 0.2m. The LPM system by Abatec [29] offers a sample rate of 1 kHz with an accuracy down to 0.15m. It measures the distance between fixed base stations and mobile tags based on the frequency modulated continuous wave principle. Although large distances can be covered, the systems are expensive and the resulting accuracy is not sufficient for precise user tracking in immersive VEs.

2.2 Vision-based ILS

Vision-based 3D tracking systems require the target to be within the line-of-sight of multiple cameras to estimate its 3D coordinates from the 2D image-projections. Optical tracking, as a sub-system of vision-based approaches, is robust against magnetic, electric and acoustic interference and works with light-emitting (active) or retro-reflective (passive) targets. The near infrared (NIR) spectrum based systems, such as Vicon [35], A.R.T [3] or iotracker [23, 25] offer (sub)-millimeter accuracy in standard room sized environments ($4x4x3m$) and provide tracking of multiple targets with very low latency. To enlarge the tracking volume, those systems increase the number of employed cameras (up to 50 in A.R.T). However, this causes a growth of costs and setup complexity. The PPT-E system [37] is able to cover areas up to $20x20m$ with a minimum of four cameras but sub-millimeter tracking accuracy is guaranteed only for volumes up to $3x3x3m$. No accuracies are provided for larger volumes. The Prime41 system [20] offers multiple user tracking by detecting passive targets up to 30m, using a perimeter setup with multiple cameras. However, no further details on accuracy nor the number of cameras are given to cover this volume. Furthermore, as the most cost efficient systems of the above mentioned, one Prime41 camera still costs about €5000. A minimal 4-camera perimeter setup results in pure camersa costs of €20.000 (without software), which is a multiple of our complete system costs. Summarizing, existing NIR optical systems for wide area user tracking require a complex system setup and thus are cost intensive. Furthermore, related work in NIR tracking is sensitive to interfering lights, especially during camera calibration, making those systems incapable of being deployed in unconstrained indoor environments.

To overcome these limitations, we describe a robust wide area multi-user tracking system in Section 3 that requires only two cameras to track targets up to distances of 30m.

3 Methodology

Existing infrared optical tracking systems provide highly accurate 3D measurements with low latency (Section 2.2). The presented approach partly builds upon our infrared tracking system *iotracker* [23, 25] and extends it to robust wide area optical tracking.

3.1 Requirements

Optical tracking systems are highly sensitive to the reliability of their inputs. According to [33], lighting conditions and camera calibration are two major sources of errors. When tracking at larger distances, image processing aberration must be taken in stronger consideration compared to short distance tracking. The emitted light from the optical target results in a circular pixel-blob (Blob) in the camera image. These blobs must be robustly segmented to accurately determine their centroid coordinates. Since an image sensor

[1] Time-Difference-of-Arrival

[2] Angle-of-Arrival

consists of discrete pixels, rasterization causes inaccuracies during blob detection and subsequent centroid estimation. Furthermore, thermal deviation influences the amount of noise on the image sensor and causes jitter. Depending on pixel size and density, the sensor temperature and thus jitter can increase. High sensor noise decreases the quality of a blob's centroid determination. Since the blob centroids are used for 3D position reconstruction, they heavily influence 3D position accuracy.

To optimize image processing aberration, feature segmentation algorithms must be highly accurate and the underlying vision hardware should provide a large sensor and high resolution.

3.2 Vision System

The vision component of the proposed tracking system comprises two cameras, lenses and filters. Following [19], we derived an optimal balanced optics setup (sensor size, focal lengths, aperture) for the intended tracking volume that minimizes optical aberration and rasterization effects while providing a sufficient field-of-view (FOV) as well as depth-of-field to cover the intended tracking volume with objects in focus. The coverage depends on focal length f, the distance between the cameras (baseline) as well as the amount of yaw-rotation β of each camera, as depicted in Figure 2.

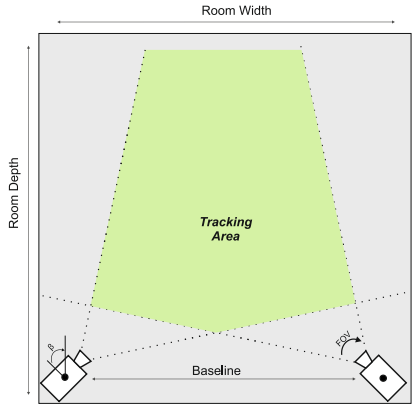


Figure 2: Tracking area coverage with two cameras.

Our system uses high-resolution machine vision cameras in combination with low-distortion lenses that feature large aperture and minimal optical aberrations. The high quality cameras provide low heat evolution and large image sensors yielding little sensor noise, so jitter in the camera image can be minimized. Together with high resolution image sensors, precise segmentation can be provided even at long distances. The cameras offer high global shutter speed to allow for low-latency tracking and to minimize motion blur when the target is moving fast. Both cameras form a *Stereo Camera Rig* and are shutter-synchronized by an external trigger signal to guarantee temporal synchronous image pairs. To enhance robust target identification, a long-wave pass filter is inserted into the optical path to ensure light transmission only in the NIR spectrum. To provide wide area tracking in width and depth, the baselines can span up to $30m$ in the intended tracking environment. Thus, we propose to use the GigE Vision standard [1] to guarantee lossless image transmission while providing long cable lengths. Both cameras are connected to one workstation for image processing and position estimation.

3.3 Target Design

Within the whole intended tracking volume, the target must be reliably visible in the cameras' images to ensure robust feature segmentation. For infrared tracking systems, two types of optical markers exist. While passive markers reflect infrared light back to the camera, active markers directly emit light towards a camera. Passive markers require special retro-reflective surface coating as well as an additional infrared light emitter to illuminate the whole tracking volume, while in case of active markers, multiple infrared light emitting diodes (IR-LEDs) must be individually powered. For small room-size tracking systems, passive markers are usually sufficient. To ensure precise feature segmentation in scenarios with interferences as well as at larger distances of $20m$ and beyond, only active markers can guarantee reliable target visibility [19].

The geometric constellation of our target design constitutes a 2D line approach, as depicted in Figure 3. Our target design offers continuously adjustable positioning of the IR-LEDs by fixing each LED separately with nuts on a rigid bar. This ensures a rapid arrangement of the required IR-LEDs in a permutation invariant geometric constellation to ensure robust target identification (Section 3.5) and occlusion recovery (Section 3.6.1).

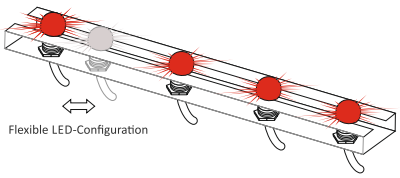


Figure 3: Target design with continuously adjustable positioning.

Furthermore, multiple unique constellations can be easily designed to simultaneously track multiple targets in the same tracking volume. Existing 3D rigid body targets (e.g [24]) also offer permutation invariant geometric constellations to track multiple targets. However, our 2D line approach has three advantages over 3D targets that are crucial for our intended research goals. (1) We can re-purpose the tracking target as calibration target by detecting the two outermost IR-LEDs during extrinsic camera calibration (Section 3.4). Thereby, the amount of necessary hardware for setup and maintenance can be reduced. (2) Even during training and calibration, the target can robustly be tracked despite interfering lights, since the 2D characteristics of the target allows for *Model Fitting* (Section 3.5) already in the image domain instead of in 3D space, as it is common in competing approaches [23, 35, 3]. (3) Fixing the IR-LEDs in a 2D manner increases the physical robustness of the target against accidental breaking off when touching the target during usage; this is especially an issue for tracking at larger distances since the target requires enlarged dimensions as well. Accidental breaking off is a common problem with the sensitive 3D rigid targets that need frequent replacement or repair by experts.

3.3.1 Tracking in a VR-Scenario

Applying the proposed target design to a semi-immersive VR scenario in which the user is tracked in front of a projector wall, a single line target is sufficient to determine the user's (head) 3D position.

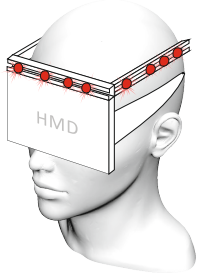


Figure 4: Target design for head tracking.

However, in a fully immersive VR environment the user freely moves in space and wears a head mounted display for visualization. In such a scenario, using a single 2D line target for tracking in combination with two cameras results in occlusions as soon as the user turns around. Since we want to minimize the amount of (costly) vision hardware, the occlusion problem can be compensated by applying a redundant target setup for user head tracking, as depicted in Figure 4.

3.4 Camera Calibration

Calibration [6, 10] for a multi-camera system consists of two steps (intrinsic, extrinsic) and is, as described in Section 3.1, one of the most crucial factors in vision-based tracking systems. A precise intrinsic calibration is required for robust feature segmentation while intrinsic and extrinsic calibrations heavily influence the accuracy of projective triangulation (Section 3.6), especially at large tracking distances. Our approach treats the two calibration routines separately, since the camera's intrinsic parameters have only to be determined if the optical configuration has changed.

3.4.1 Intrinsic Calibration

To enhance the estimation of the intrinsic camera parameters, represented by the *Camera Calibration Matrix* K, all optical components (camera, lens, and filter) of the final tracking setup should be included in the calibration procedure yielding more accurate internal parameters. We included toolboxes [4] and [15] in our calibration software pipeline as they have been proved to be highly accurate and reliable. However, they require a standard chessboard plane as calibration target that is not visible in the NIR spectrum.

(a) IR light (b) Reflective pattern (c) Camera image

Figure 5: Intrinsic camera calibration with a retro-reflective pattern.

Therefore, we extended the intrinsic calibration routine by developing a chessboard plane using a retro-reflective foil that is illuminated with an infrared light source to provide chessboard images in the NIR spectrum. The complete intrinsic setup is illustrated in Figure 5.

3.4.2 Extrinsic Calibration

After the stereo rig is physically set up, the geometric relation between the cameras is estimated by the extrinsic calibration process, yielding the definition of the epipolar geometry that is encapsulated in the *Fundamental Matrix* F. Standard techniques, such as [4, 22] estimate F by using a chessboard plane. For our calibration scenario, such a pattern would have to be extremely large to be visible at distances of $10m$ and more as well as highly planar to provide precise corner extraction. Such a target would neither be transportable nor suitable, so we propose to use the target's IR-LEDs to estimate F.

The two-camera calibration approach [5] estimates F by evaluating the screen-space coordinates of two blobs – that corresponding physical markers have a known distance – over a sequence of camera images. The affine transformation to obtain real-world distance units [mm] is not only computed once as it is common in existing approaches (e.g. [23]) and which can result in inaccurate position estimation at larger distances, but is optimized with every processed camera frame pair. However, this powerful approach is sensitive to false input data such as interfering lights, which results in highly inaccurate external calibration parameters. To avoid pre-conditioning before calibration by masking out infrared interferences of the tracking environment, we extended [5] to make it robust against static or locomotive light sources. Therefore, we prepared our proposed tracking target (Section 3.3) to act as an extrinsic calibration target by measuring the physical distance between the two outermost IR-LEDs to sub-millimeter accuracy in a special laboratory setup with a total station (Leica TPS700). This yields the exact distance between the two necessary input blobs. Furthermore, we developed a pipeline (Section 3.5) that filters interfering lights during tracking as well as calibration. It returns a set of ordered target points p for both cameras L and R of a frame at time t, denoted as:

$$S_L^t = \{p_{L,1}^t, p_{L,2}^t, p_{L,3}^t, p_{L,4}^t\},\ S_R^t = \{p_{R,1}^t, p_{R,2}^t, p_{R,3}^t, p_{R,4}^t\} \quad (1)$$
$$p_{L,i}^t, p_{R,i}^t \in \mathbb{R}^2, i = 1...4\,.$$

For each frame at time t, we calculate distances $d_R^{t-1} = \|p_{L,4}^{t-1}, p_{L,1}^{t-1}\|$ and $d_L^{t-1} = \|p_{R,4}^{t-1}, p_{R,1}^{t-1}\|$, where $\|.\|$ denotes the Euclidean norm. If the condition $|(d_L^t - d_R^t) - (d_L^{t-1} - d_R^{t-1})| \leq \lambda$ is fulfilled for a given threshold λ, the blob sets S_L^t, S_R^t are used for calibration, otherwise neglected. This ensures correct rotation-invariant blob input into the calibration algorithm. With this extended calibration approach, we achieve a lightweight calibration target and avoid additional bulky equipment. No pre-conditioning is necessary that enhances the system's ease of use during setup and maintenance.

3.5 Interference Filtering

To provide robust target identification at each stage of a vision-based tracking system workflow (target training, calibration, 3D tracking) static and locomotive interfering lights must be robustly filtered out. Since these light sources frequently emit in the NIR spectrum, they are visible in the camera images and result in bright blobs even if a long-wave pass filter is inserted into the optical path. To overcome this problem, we developed a software-based identification pipeline, as depicted in Figure 6.

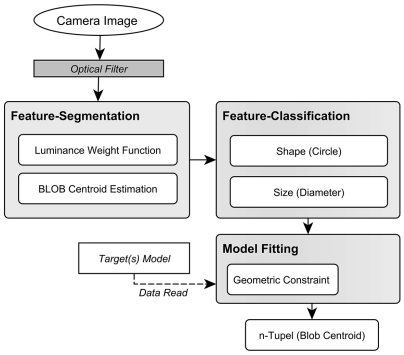


Figure 6: Pipeline for unique target identification.

After a new image (*frame*) is captured from the camera with the attached long-wave pass filter, all blobs are segmented (*Feature Segmentation*) by performing luminance filtering, as proposed in [23]. This adaptive algorithm combines three segmentation techniques (threshold, luminance-weighted centroid computation, and circular Hough transform) to efficiently locate the centers of all circular shapes in a monochrome image with sub-pixel accuracy.

In the next step, each resulting blob is classified by performing shape- and size-based classification (*Feature Classification*). The minimum and maximum values for the size-filter can be manually defined to provide quick configuration for different tracking ranges. The classification results in circular-shaped blobs (*Blob Candidates*) that diameters lie within the specified range. In practice however further filtering must be performed since interfering lights can have a similar size as the target's IR-LED blobs.

By combining the approaches [32, 28, 26, 18], we perform a two-dimensional *Model Fitting* within the set of remaining blob candidates. Thereby, we exploit the permutation and perspective invariant properties of our target. When projecting 3D points onto the 2D camera plane, neither distances nor ratios of distances are preserved [10]. However, the *cross-ratio* as a ratio of distances as well as the collinearity of points sets [34] is preserved. Based on [32], we compute the p^2-Invariants, which represent properties of point sets that are insensitive to projective transformations and to permutations in the labeling of the set. The collinear properties of the target, respectively its corresponding blobs, allows a computationally lightweight and thereby fast way to reject false blobs candidates. By calculating the cross-ratio and comparing it with a certain range to account for noise in the feature segmentation, the recognition of a known geometric target's IR-LED constellation can be performed, resulting in an ordered set of blobs $S^t = \{p_i^t\}, i = 1...N, p \in \mathbb{R}^2$ for each image at time t.

To obtain the unique properties of a target, it is trained once in an off-line process to determine its *Model*. First, the distances between the target's LEDs are precisely measured using a total station. Based on the physical distances, an initial distance heuristic as well as a collinearity model is estimated. Applying this initial estimate, the target's blobs are segmented from a sequence of camera images in various distances to refine the model's parameter profile that includes the target's geometric constellation, a collinearity error model as well as the cross-ratio and its range. Based on this profile, the system is capable of uniquely identifying the target's model during calibration and tracking even in the presence of disturbing light sources.

3.6 3D Position Tracking

The online image-processing pipeline for continuous tracking is depicted in Figure 7.

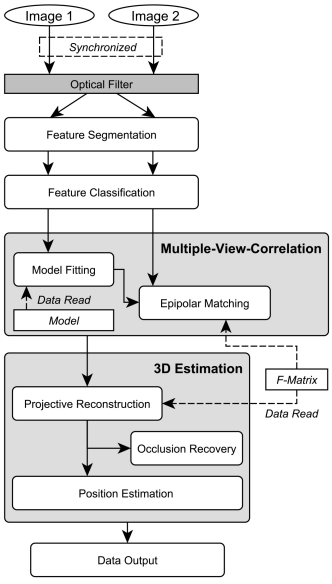


Figure 7: Pipeline for 3D position tracking.

Given an intrinsically and extrinsically calibrated, shutter-synchronized stereo camera rig, the tracking is performed as follows. After a new frame is received from each camera, blob candidates are segmented and classified in both frames, as described in Section 3.5. To minimize computational load, unique target identification is only performed in *Image 1* by applying model fitting within the set of all blob candidates. After the target blobs have been determined in *Image 1*, their correspondences have to be identified in *Image 2* amongst all blob candidates that result from the feature classification. To apply a time- and cost-efficient search routine, we exploit the properties of the epipolar geometry that is encapsulated in F. For each target blob in *Image 1*, a search for its corresponding blob can be performed along its epipolar line (*Epipolar Matching*) in *Image 2* [6, 10]. Thereby, corresponding features over multiple camera views can be robustly identified (*Multiple-View-Correlation*). Using epipolar matching, computational complexity can be heavily optimized because the original 2D search problem is reduced to a 1D problem. Since our target model features a unique geometric permutation invariant property in the 2D domain, we can already apply model fitting within the 2D projections of the target's IR-LED. Thereby, we obtain a drastically reduced set of correspondence candidates and can considerably decrease the combinatorial complexity of the multiple-view correlation problem.

By performing a projective triangulation between each correlated 2D blob-tuple (*Projective Reconstruction*), the 3D-coordinate of each target IR-LED can be reconstructed. As described in [9, 10], multiple solutions exist. Following [23], we apply the standard linear singular value decomposition (SVD) method to obtain the initial 3D estimate for each blob-tuple, followed by bundle adjustment [30] with a Levenberg-Marquardt non-linear least squares algorithm. This results in a 3D point cloud of all target's IR-LEDs $T = \{P_1, P_2, P_3, P_4\}, P \in \mathbb{R}^3$. Based on T and λ as the actual distance between the outermost IR-LED and the epicenter of the target, the target's epicenter $C \in \mathbb{R}^3$ can be calculated (*Position Estimation*) as follows:

$$C = P_4 - (\lambda * \hat{m}). \tag{2}$$

Therefore, we normalize the vectors $\vec{a} = \overline{P_2P_1}$, $\vec{b} = \overline{P_3P_2}$, $\vec{c} = \overline{P_4P_3}$, resulting in $\hat{a}, \hat{b}, \hat{c}$. Calculating the arithmetic mean of $\hat{a}, \hat{b}, \hat{c}$, we determine the mean direction $\hat{m}$ which is applied according to Equation 2.

3.6.1 Occlusion Recovery

If a target's IR-LED and an interfering light source lie on the same line of sight of the camera, their corresponding blobs can overlap in the images. Furthermore, parts of the target can be occluded, i.e. when the target gets partly hidden behind an object in the scene. Our model fitting approach requires four optical markers. Currently, the proposed target identification pipeline can compensate one occluded marker while retaining the capability of robustly detecting the target within the set of blob candidates. After projective reconstruction, the 3D positions of occluded markers can be reconstructed based on the target's geometric model and the resulting 3D point cloud. The recovery of occluded IR-LEDs optimizes the accuracy of the 3D position estimate of the target's epicenter. With this recovery functionality, loss of tracking can be reduced in cases of occlusions or over-blooming by (stronger) interfering light sources.

4 Implementation

Based on the methodological approach, we developed a hardware- as well as software prototype to test our tracking system in large unconstrained indoor environments.

4.1 Hardware Prototype

Our hardware protoytpe comprises vision system, target and a notebook as main processing unit. The vision system consists of two

GigE Vision [1] Dalsa Genie HM1400/XDR cameras[3]. Both cameras are equipped with a high-resolution and high-speed lens (F1.4-F16) and a focal length of $f = 25mm$, resulting in a diagonal FOV of 35.49° to adequately cover the intended tracking range. A long-wave pass filter is inserted into the optical path to block wavelengths smaller than $780nm$.

The target prototype has a total length of $687mm$ and is equipped with four IR-LEDs in a permutation invariant constellation. Each IR-LED emits a at peak wavelength of $850nm$ with a radiant intensity of 20 mW/sr[4] and features a viewing half angle of $\pm 23°$. Thereby, robust feature segmentation up to a distance of $50m$ can be performed. A minimum distance of $130mm$ between two neighboring LEDs is advisable with a shutter speed of $1ms$ to avoid blob overlaps in the camera image during rotations and at larger distances. With this prototype, tracking in the intended volume can be provided while minimizing the total length of the target. To protect the LEDs and to prevent optical aberrations (flare artifacts on the blob edges), each IR-LED is covered with a translucent diffuse plastic sphere.

The processing core unit is a portable workstation with two Gigabyte Ethernet host adapters to interface via Category 6 cable with the cameras. Both cameras are shutter-synchronized from a square-wave current loop signal that is generated by the trigger unit with a built-in programmable oscillator. The trigger unit comprises two BNC connectors[5] and the trigger signal, generated by an Arduino Uno board [2].

4.2 Software Framework

Our software framework follows a three-tier-architecture comprising hardware abstraction, a processing layer as well as data visualization on a graphical user interface. The processing core consists of loosely-coupled modules for calibration, tracking and unique target identification. The modules and their functionalities are centrally accessed by the controller component that delivers data from the processing layer to the GUI. Our software framework prototype is implemented in C/C++ and MATLAB. For the intrinsic camera calibration, the third-party software MATRAX [15] as well as the open-source MATLAB Camera Calibration Toolbox [4] were integrated. With the open-source Arduino IDE [2], we developed the embedded component for camera synchronization.

4.3 System Costs

As stated in Section 1.1, cost efficiency is one of the objectives of the presented work. Therefore, we minimized the amount of necessary hardware and focused on off-the-shelf components as well as open source hardware and software. The current hardware prototype costs in total $\sim$ €7300. This includes both cameras (each €2000 with filter), lenses (each €600), notebook (€2000) and technical parts (€100 for Arduino, battery, wires, IR-LEDs and target material).

5 Evaluation

Based on the developed hard- and software prototype, we evaluate the robustness of target identification and the accuracy of 3D position estimation.

5.1 Test Platform & Environment

We tested our system on a Lenovo W520 notebook, featuring an Intel Quadcore i7 2820QM at 2,3GHz, 8 GB memory and Windows7 (64bit). Since we were lacking access to an indoor environment that features the intended tracking ranges, we deployed the prototype in an outdoor environment during twilight and night.

[3]Sensor: 1" mono with 1400x1024px, @ 60fps

[4]mW/sr: milli watts per steradian

[5]BNC: Bayonet Neill Concelman connector

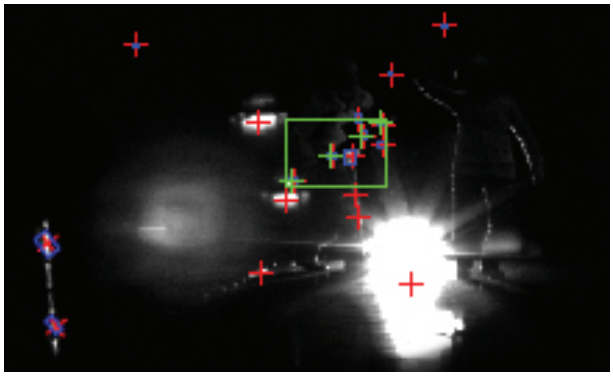


Figure 8: Light situation during extrinsic calibration and tracking.

We added light sources (neon lights, halogen spots up to 1500W) to simulate wall illuminations, reflections and locomotive interfering lights, as depicted in Figure 8. Thereby, we established a controllable realistic simulation of the intended tracking scenario.

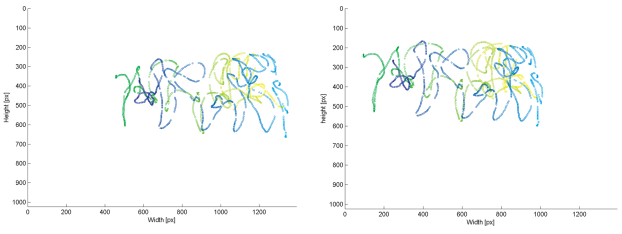
5.2 Test Cases

We assessed our wide area tracking system according to the following parameters: (1) Accuracy and robustness of the extrinsic calibration. (2) Accuracy of 3D position estimation, as well as (3) Tracking performance by examining continuous volume coverage and the system's update rates. We performed all three tests in an environment with static as well as locomotive interfering lights, with a baseline $d_{base} = 10.0m$ and distances between the vision system and target d_{track} of 5.0 to $30.0m$

5.3 Calibration

Both cameras were set up with a baseline $d_{base} = 10.0m$ and a yaw-rotation $\beta_{cam1} = 30°$, $\beta_{cam1} = -30°$ to cover a tracking volume up to $30.0m$. Using the tracking target from Section 4.1, we performed the calibration at a distance of $15.0m$ from the cameras.

To evaluate the extrinsic calibration accuracy, we used a high quality total station (Leica TPS700) to accurately measure the real baseline d_{base}. Although d_{base} is the only available measurement to evaluate the calibration accuracy, it can only be used as approximation because the camera coordinate system's origin coincides with the center of projection of one camera. Since this is a virtual point in the physical camera, the geodesic prisms cannot be positioned with absolute accuracy at both projection centers.



(a) Left camera (b) Right camera

Figure 9: Corresponding blob traces used for extrinsic calibration.

We ran three different calibration tests with ~1100 frames each to evaluate the robustness of the calibration procedure. As depicted in Figure 9, our system robustly identifies the target despite static and locomotive interfering lights, resulting in continuous blob traces of the two outermost IR-LEDs. As illustrated, the blob trace was

interrupted at some points due to complete occlusion of the target because of obstacles in the environment.

An example of a successfully detected target in the camera frame is shown in Figure 8, marked with a green box. Despite this unconstrained calibration environment, our system robustly estimated the fundamental matrix F at each run, yielding consistent 3D point estimates for all tracking distances, as illustrated in Table 1. This demonstrates the robustness of our calibration procedure. F was estimated with an average duration of ~110s. By deriving the translation between both cameras from F, we could compare it with d_{base}. Our calibration approach achieves centimeter accuracy for each run with a mean deviation $\varepsilon_{base} = 0.1763m$.

5.4 Accuracy & Stability

To evaluate the accuracy of 3D position estimation, we performed measurements at six different distances between camera and target, denoted as d_{track} for each calibration procedure. At each accuracy run, the 3D coordinate of each target's IR-LED $L_{1..4} \in \mathbb{R}^3$ as well as of the target's epicenter $C = C_{x,y,z} \in \mathbb{R}^3$ was estimated based on 300 consecutive frames. Thereby, we were able to evaluate the following two parameters, accuracy and stability, for the entire tracking volume.

5.4.1 3D Position Accuracy

To determine the accuracy of the 3D position estimate, we first measured the geometric distance between the two outermost IR-LEDs to millimeter precision using the Leica TPS700 to obtain ground truth d_{bar}. During run-time, we estimated $\hat{d} = \|L_4, L_1\|$, $\|$ denoting the Euclidean norm, to determine the root mean square deviation $x_{RMS}(\hat{d}) = \hat{d} - d_{bar}$. The deviation of a point $x_{RMS}(P)$ is determined by $x_{RMS}(P) = \frac{x_{RMS}(\hat{d})}{2}$. This allows us to evaluate the 3D position accuracy of a single target point throughout the tracking volume. In Figure 10, the arithmetic mean of $x_{RMS}(P)$ over all three calibration runs with respect to the tracking distance is depicted.

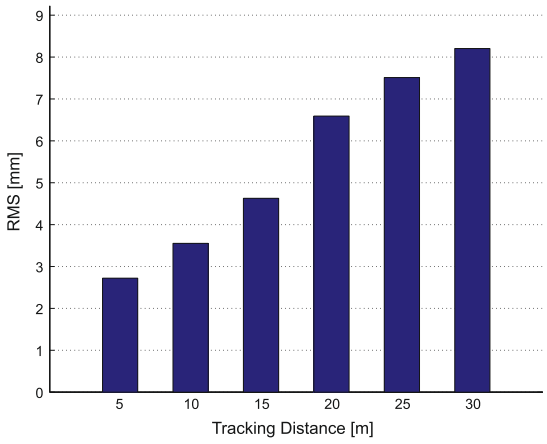


Figure 10: Accuracy (RMS) of a point throughout the tracking volume.

The obtained $x_{RMS}(P)$ values for each calibration run and each tracking distance d_{track} are listed in detail in Table 1.

5.4.2 3D Position Stability

To evaluate static jitter of the system and thus the stability of the 3D point estimation, we determined the standard deviation σ of C_x, C_y, C_z as well as C over the sequence of 300 consecutive frames. Throughout the entire tracking volume and over all three calibration runs, we measured sub-millimeter precision for 3D position estimation with $C_x : \sigma = 0.0545mm$, $C_y : \sigma = 0.0304mm$, $C_z : \sigma = 0.1175mm$ and $C : \sigma = 0,0675mm$.

	Calibration 1	Calibration 2	Calibration 3
d_{track}	$x_{RMS}(P)$	$x_{RMS}(P)$	$x_{RMS}(P)$
05m	3.39 [mm]	2.99 [mm]	1.78 [mm]
10m	4.12 [mm]	3.91 [mm]	2.63 [mm]
15m	4.76 [mm]	4.54 [mm]	4.58 [mm]
20m	6.08 [mm]	6.23 [mm]	7.47 [mm]
25m	6.64 [mm]	6.97 [mm]	8.92 [mm]
30m	7.44 [mm]	7.96 [mm]	9.22 [mm]

Table 1: Accuracy of a single point $x_{RMS}(P)$ throughout the volume with three different extrinsic calibrations.

5.5 Tracking Performance

To determine the system's capability to continuously track a target throughout the entire tracking space, we moved it through the whole volume. The resulting 3D position reconstruction of each target's IR-LED is illustrated in Figure 11.

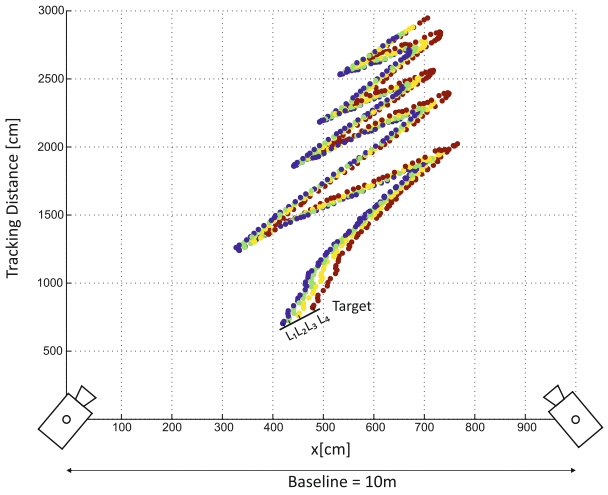


Figure 11: 3D position tracking from $5 - 30m$.

Depending on the number of interfering lights, our system identifies and tracks a target with a latency of ~69ms within the unconstrained test environment, allowing for wide area position tracking with interactive frame rates.

5.6 Discussion

Our results demonstrate 3D point accuracy $x_{RMS}(P) < 9.22mm$ with sub-millimeter static position jitter $\sigma = 0.0675mm$ throughout the entire tracking volume, ranging from $5 - 30m$. We tested our system with several different target constellations, which can be detected within both camera views with rotations yaw and pitch from 0 to 45° as well as roll from 0 to 360°. To our best knowledge, no competing approach and system (Section 2.2) provides comparable accuracy for this range, especially not with the minimal amount of only two cameras.

Our system offers tracking at interactive frame rates, however further run-time optimization will be done to reduce latency. The interference pipeline reliably detects the target during extrinsic calibration and tracking, making the whole system robust to be deployed in an unconstrained indoor environment.

Currently, our system tracks only 3D position because the implicit 2D characteristic of the target (Section 3.3) does not provide

orientation. This can be compensated by combining multiple 2D line targets, resulting in 5DOF, or employing an additional inertial measurement unit for full 6DOF-pose tracking.

6 Conclusion & Future Work

In this paper, we present a novel robust optical tracking system that provides high-precision 3D position tracking with interactive frame rates at distances up to $30m$. Our proposed 2D geometric target design allows for quick (re)-configuration and enhances robustness against accidental break-offs compared to 3D rigid body targets [24]. Targets are equipped with standard infrared light emitting diodes and are re-used for camera calibration. As depicted in Figure 1, the system only requires two cameras so necessary hardware can be reduced to a minimum to provide cost efficacy and ease-of-use during setup and maintenance. At each stage of the system's workflow, no pre-conditioning of the tracking volume is necessary. This enables our system to robustly and accurately track users in large unconstrained indoor environments with static and locomotive light sources.

For future work, we will first evaluate the accuracy with different hardware setups using higher resolution cameras and lenses with smaller focal length for extended horizontal tracking area coverage. Additionally, infrared LEDs with less radiant intensity will be examined to reduce the overall target size. Second, we will evaluate state-of-the-art prediction approaches such as [36, 17] to enhance tracking robustness and latency for real-time multi-target tracking. Third, we plan to test the system in harsh indoor environments with interfering lights and poor visibility (dust, fog, smoke) such as entertainment stages, workshops as well as an underground construction site to further evaluate the system's robustness in real-world application scenarios.

Acknowledgements

We especially thank our colleagues Georg Gerstweiler and Emanuel Vonach for their extensive contribution on the calibration- and model fitting modules as well as Matthias Zeppelzauer for assistance during system evaluation.

References

[1] AIA. GigE Vision. [Online] http://www.visiononline.org, 2013.

[2] Arduino. Arduino IDE. [Online] http://www.arduino.cc/, 2013.

[3] ART. Advanced Real Time Tracking. [Online] http://www.ar-tracking.de, 2013.

[4] J.-Y. Bouguet. Camera Calibration Toolbox for Matlab. [Online]: http://www.vision.caltech.edu/bouguetj/calib_doc, 2013.

[5] K. Dorfmüller-Ulhaas. *Optical Tracking: From User Motion To 3D Interaction*. Phd thesis, Vienna University of Technology, 2002.

[6] O. Faugeras, Q.-T. Luong, and T. Papadopoulou. *The Geometry of Multiple Images: The Laws That Govern The Formation of Images of A Scene and Some of Their Applications.* MIT Press, USA, 2001.

[7] K. Finkenzeller. *RFID handbook: Radio-frequency identification fundamentals and applications*. John Wiley, New York, USA, 1999.

[8] Google. Indoor Maps. [Online] https://www.google.com/intl/en/maps/about/explore/mobile.

[9] R. Hartley and P. Sturm. Triangulation. *Computer Vision and Image Understanding*, 68(2):146–157, Nov. 1997.

[10] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

[11] M. Hazas and A. Ward. A novel broadband ultrasonic location system. *Ubiquitous Computing*, 2498(September):264–280, 2002.

[12] J. Hightower and G. Borriello. Location Systems for Ubiquitous Computing. *IEEE Computer*, 34(8)(August):57 – 66, 2001.

[13] F. Ijaz, H. K. Yang, A. W. Ahmad, and C. Lee. Indoor Positioning: A Review of Indoor Ultrasonic Positioning systems. In *Proceedings of 15th International Conference on Advanced Communication Technology (ICACT)*, pages 1146 – 1150, 2013.

[14] IndooRs. Location Tracking. [Online] http://indoo.rs/, 2013.

[15] Inoptech GmbH. MATRAX. [Online] http://www.inoptech.de, 2013.

[16] B. Jiang, K. P. Fishkin, S. Roy, and M. Philipose. Unobtrusive long-range detection of passive RFID tag motion. *IIEEE Transactions on Instrumentation and Measurement*, 55(1):187–196, 2006.

[17] J. LaViola. An Experiment Comparing Double Exponential Smoothing and Kalman Filter-based Predictive Tracking Algorithms. In *Proceedings of IEEE Virtual Reality*, number 8, pages 283 – 284, Los Angelas, CA, USA, 2003.

[18] M. Loaiza, A. Raposo, and M. Gattass. A novel optical tracking algorithm for point-based projective invariant marker patterns. *Advances in Visual Computing*, 4841:160–169, 2007.

[19] A. Mossel, T. Pintaric, and H. Kaufmann. Analyse der Machbarkeit und des Innovationspotentials der Anwendung der Technologie des Optical Real-Time Trackings für Aufgaben der Tunnelvortriebsvermessung. Technical report, Institute of Software Technology and Interactive Systems, Vienna University of Technology, Austria, 2008.

[20] NaturalPoint Inc. OptiTrack. [Online] http://www.naturalpoint.com/optitrack/, 2013.

[21] B. P. Nissanka. *The cricket indoor location system*. Phd thesis, Massachusetts Institute of Technology, USA, 2005.

[22] OpenCV. Open Source Computer Vision Library. [Online] http://opencv.org/, 2013.

[23] T. Pintaric and H. Kaufmann. Affordable Infrared-Optical Pose-Tracking for Virtual and Augmented Reality. In *Proceedings of Trends and Issues in Tracking for Virtual Environments Workshop, IEEE VR 2007*, pages 44–51, 2007.

[24] T. Pintaric and H. Kaufmann. A Rigid-Body Target Design Methodology for Optical Pose-Tracking Systems. In *Proceedings of the 2008 ACM Symposium on Virtual Reality Software and Technology*, pages 73–76, New York, NY, USA, 2008. ACM Press.

[25] T. Pintaric and H. Kaufmann. iotracker. [Online] http://www.iotracker.com, 2013.

[26] P. Santos and A. Stork. Ptrack: introducing a novel iterative geometric pose estimation for a marker-based single camera tracking system. In *Proceedings of IEEE Virtual Reality*, pages 149–156, USA, 2006.

[27] SensionLab. Indoor Positioning and Navigation. [Online] http://www.senionlab.com, 2013.

[28] F. A. Smit, A. van Rhijn, and R. van Liere. GraphTracker: A Topology Projection Invariant Optical Tracker. In *Proceedings of the 12th Eurographics Conference on Virtual Environments*, pages 63–70, 2006.

[29] A. Stelzer, K. Pourvoyeur, and A. Fischer. Concept and application of LPMA novel 3-D local position measurement system. *IEEE Transactions on Microwave Theory and Techniques*, 42:2664–2669, 2004.

[30] B. Triggs, P. F. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment - A modern synthesis. In W. Triggs, A. Zisserman, and R. Szeliski, editors, *Vision Algorithms: Theory and Practise*, volume 34099, pages 298–372. Springer, 2000.

[31] Ubisense. Real-Time Localization Systems. [Online] http://www.ubisense.net, 2013.

[32] R. van Liere and J. D. Mulder. Optical tracking using projective invariant marker pattern properties. In *Proceedings of IEEE Virtual Reality*, pages 191–198. IEEE Comput. Soc, 2003.

[33] R. van Liere and A. van Rhijn. An experimental comparison of three optical trackers for model based pose determination in virtual reality. In *Proceedings of 10th Eurographics Conference on Virtual Environments (EGVE'04)*, pages 25–34, Aire-la-Ville, Switzerland, 2004.

[34] A. van Rhijn. *Configurable Input Devices for 3D Interaction using Optical Tracking*. Phd thesis, Technische Universiteit Eindhoven, Netherlands, 2007.

[35] Vicon. Motion Capture. [Online] http://www.vicon.com/, 2013.

[36] G. Welch and G. Bishop. An Introduction to the Kalman Filter. Technical report, University of North Carolina, Chapel Hill, USA, 1995.

[37] WorldViz. PPT E Motion Tracking. [Online] http://www.worldviz.com/products/ppt/, 2013.