TUM

TECHNISCHE UNIVERSITÄT MÜNCHEN
INSTITUT FÜR INFORMATIK

Using Gamification for Stereo Camera Calibration in Terms of Augmented Reality

Andreas Langbein, David A. Plecher, Frieder Pankratz, Chloe Eghtebas, Fabrizio Palmas, Gudrun Klinker

TUM-I2080

Using Gamification for Stereo Camera Calibration in Terms of Augmented Reality

Andreas Langbein* David A. Plecher† Frieder Pankratz‡ Chloe Eghtebas§ Fabrizio Palmas¶
Gudrun Klinker‖

Technical University of Munich

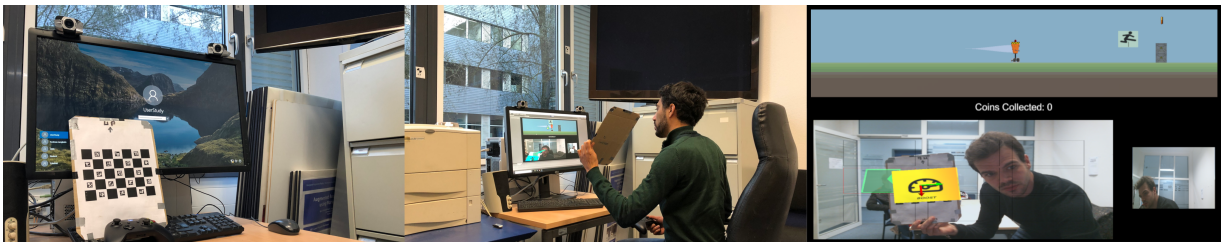

Figure 1: Gamified Calibration Setup and Procedure

Abstract

Augmented Reality (AR) is being adopted in industry for communication and in society as a whole. While development continues, challenges still remain. Calibration tasks especially remain a hassle for developers and end users alike. They are time-consuming, follow strict rule-sets and require great care to ensure a usable end result.

In this report, the extended version of Langbein et al. [16], we propose a solution to ease calibration tasks on users and keep them motivated by adding game concepts to the procedure. We show that, in a gamified application, participants can be incited to perform longer procedures with up to four times the amount of measurements taken without a necessary decrease in the resulting calibration quality.

Index Terms: Gamification [User Interfaces]: Graphical user interfaces; Calibration [Information Interfaces and Presentation]: Gamification Calibration Registration Augmented Reality

1 Introduction

Sensor calibration and registration are very important issues in Augmented Reality, in robotics [6] in automation [14], and even the medical field [17]. In many cases, sensor calibration and registration requires large quantities of measurements to cover a large parameter space with sufficient samples.

Furthermore, with physical setups drifting over time, these tasks need to be repeated frequently. (Re-)calibration of essential equipment is thus a routine activity for personnel working in laboratories and factories with sensory systems.

Even though schemes towards automation, like online self-(re)calibration exist [12], these procedures are still mostly performed manually or semi-automatically. To achieve the required high standards of precision and accuracy, measurements need to be taken with diligence and care. The process is often intensive, time consuming, and tiring. Various efforts have been undertaken to ease these routines on users [21], but problems still remain. Firstly, people grow tired of the same repetitive calibration routines. Second, workers who have to execute those procedures in their regular work may not be as familiar with details and issues related to achieving usable results as a photogrammetry expert would be. Typically, users in industry will receive an introductory tutorial by an expert informing them about the principal steps and general good practices. Then, the workers are left on their own. Depending on the level of feedback provided by the system, workers may not be fully aware of the mistakes they make. Their performance may degrade over time. In consequence, it is even more important that large quantities of measurements be taken - enough to filter out bad measurements.

In this report, we investigate how users can be enticed to spend more than the minimal amount of time on calibration routines. Gamification was already successfully used in Virtual Reality Training [20]. Therefore:

We conjecture that gamification has the potential to motivate people to conduct longer routines, gathering more data (without expecting users to deeply understand the underlying photogrammetric issues), due to the fact that people simply enjoy the experience.

Yet, there is the danger that such gamification will defocus people from carefully adhering to the precise/registration procedure - a potential problem that must be avoided.

We report on first efforts towards designing a gamified calibration system to register a pair of stereo cameras, with the goal to increase users' frequency and length of performing the procedure without significant loss of precision.

2 Related Work

Literature review on related work leads us to two focus questions: What elements are known to make an application intrinsically motivating? And, what has been done in this respective field of gamified calibration procedures before? A pre-study has been performed in our own accord and its results will be built upon [7].

2.1 Motivational game components

In 1981, Malone et al. published their theories about inciting peoples' motivation in video games [18]. They provided a theory describing a multitude of factors that influence user motivation.

*e-mail: langbeia@in.tum.de
†e-mail:plecher@in.tum.de
‡e-mail:pankratz@in.tum.de
§e-mail:eghtebas@in.tum.de
¶e-mail:fabriziopalmas@gmail.com
‖e-mail:klinker@in.tum.de

Highest ranking among them were a clear, understandable game-goal, a persistent scoring system, a factor of randomness in obstacles, and audio effects throughout the experience. Malone narrowed this theory down to three motivational categories: *Challenge, Fantasy and Curiosity*. Serious Games are using these positives effects on motivation to transfer knowledge to the player about various topics like languages [22], nutrition [25] and cultural heritage [23] or to change the user's behaviour [10]

The three motivational categories were reworked by Flatla et al. into: *Theme, Challenge, Reward and Progress* as the main motivating forces [11]. Flatla wrote that *Challenge*s are obstacles or goal elements tied to a reward. A player overcomes a hindrance and is rewarded in-game. The *Theme* provides a fictional context to the application. Enemies, objects, music etc., all provide mental imagery motivating people to interact. *Reward*s are given for positive behavior within the game's set of rules. Handle a challenge, find the right path, anything that motivates players to progress. *Progress* contains different sorts of feedback for the player progression. Points, levels, worlds, achievements, etc. Everything that indicates how well a player is doing in-game. It is important to note, that not all features fall into one game element category. Many ideas like progress bars (progress, theme, challenge) or sound effects (theme, reward) cover multiple categorizations at once. These criteria will be addressed when gamifying our calibration task.

2.2 Player types

In 1996, Bartle et al. conducted research categorizing different kinds of players in gaming scenarios [8]. Bartle's work was focused on a specific genre: M.U.D.'s (Multi-User-Dungeons) - early multiplayer games allowing exploration and enemy encounters with, or against, other players. They provided a *taxonomy* that classified player types with their own sets of interest- and focus-points. Namely *Achievers, Explorers, Killers and Socializers*.

Achievers look for rewards in any shape and form. *Explorers* scan their environment for information wherever they go. They search for the most efficient approaches and secrets. *Socializers* seek interaction with other players. To them, the game is an interface to the existing player base. *Killers* search for competitive interaction with others. They find reward in victory and superiority over competitors. Bartle's model doesn't encompass all player types for all games. It was based on observations in a specific target group, the M.U.D.-players. For us, this model indicates that more than one gamification scheme needs to be developed such that different player types can be motivated according to their varying interests.

Dixon et al. wrote a review of Bartle's work in 2011 [9], noting that the it had become overextended over time; often applied outside the genre it was focused on and that it was used on games and players too distant from its original domain. Bartle's player types were mutually exclusive. Users would either be socializers or killers with no overlapping behaviour expected. Yet, players do share interests across multiple type definitions. A user might socialize with others, yet also put all his/her efforts into discovering secrets in-game. Accordingly, we strive towards a game design that is amenable to several player types - a widely acceptable common denominator.

2.3 Gamified Calibration

Flatla et al. have conducted an investigation of combining calibration tasks with game elements and playfulness [11]. They presented a framework that simplifies the design of calbration games and shows the broad applicability of the idea (see section 2.1). As part of this framework, they distinguished a large number of different calibration types and tasks and associated them with suitable game mechanics and game design elements. They presented three exemplary gamification approaches for selected calibration tasks: Calibrating color with respect to just-noticeable-differences (JNDs), calibrating control-to-display (C:D) parameters for targeting, and calibrating a physiological sensor. They concluded that calibration tasks, more precisely the color calibration of monitors (JND), can be made more enjoyable by the addition of game elements. Adding these elements did not degrade the overall calibration quality significantly.

In this paper we build on Flatla's framework and experimental results. Our work is most closely related to the calibration of C:D parameters for targeting, however, we provide novel contributions in several aspects. **1) Gamification methods for 3D tasks:** While Flatla calibrated 2D pointing activity on a screen, our work investigates gamification principles for 3D user interaction in a tracked 3D environment for Augmented Reality. Design and setup of the workspace and the associated gamification concepts for our 3D stereo camera calibration differ significantly from determining the C:D ratio of mouse movement w.r.t. cursor motion on a screen. 3D stereo camera calibration involves many more parameters, regarding the users' understanding of how their interaction is measured by the cameras and how they need to act within the 3D environment. Thus, this paper presents novel contributions towards the use of gamification methods for 3D Augmented Reality. **2) Quantitative analysis:** Flatla presented subjective evidence on users' increased motivation (provided by questionnaires). In contrast, we show on the basis of runtime measurements that test persons interact significantly longer and produce significantly more results in the gamified version of our camera calibration process. **3) Application area:** Flatla showed vague evidence that gamification does not have a negative impact on measurement quality on long term performance. In our experiments, we want to stay as closely as possible to the basic gamification rules as well as build upon Flatla's progress but shift focus from the color space calibration towards AR stereo camera calibration.

3 The Problem: Stereo Camera Calibration

There are many different calibrational and registrational procedures for different setups of sensors. Examples being tip calibration [27], absolute orientation calculation [15], or hand-eye calibration [26]. Many of these require intensive user interaction to function. The kind of interaction depends on the sensor properties and the spatial relationships between the sensors and their environment that need to be determined by the calibration or registration process.

In this paper, we focus on stereo camera calibration - another widely applied calibration task for Augmented Reality. In this sensor setup, two cameras are placed side by side, rigidly attached to one another (see figure 2). Their fields of view overlap with a large enough section, so that users can move around while remaining visible to both cameras. The goal of this procedure is to determine the precise positional and rotational offset (with respect to a connecting baseline) between both cameras.

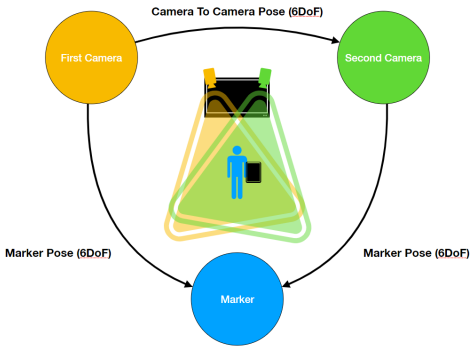

Figure 2: Stereo Camera Setup: Two cameras mounted on a monitor. Users move marker in cameras' common fields of view

3.1 Theoretical requirements

The positional offset between two stereo cameras can be determined by an interactive procedure which requires a user to move a recognizable target (such as a calibration grid) to various locations throughout the area covered by both cameras' fields of view. Both cameras take pictures of the marker at synchronized points in time, triggered in intervals or by direct user input.

For each snapshot, camera calibration routines such as Zhang et al.'s nonlinear refined approach (see ref. [28]) calculate each camera's pose with respect to the marker. With this, we receive a marker pose for each camera on each triggered snapshot. The difference between these poses determines the positional and rotational offset between both cameras have towards (see figure 2).

Even though, in principle, a single such marker position suffices to compute the offset (baseline) between the two cameras, best-practice guidelines of photogrammetry recommend that a large number of measurements be taken from different marker locations in order to account for non-linear lens properties of the cameras and to minimize the effects of noisy marker detection in video images: The marker should be moved to widely spread positions (x,y) in the fields of view of both cameras, covering different areas of their respective (potentially distorted) areas on the sensor chips. They should also be placed at different depths (z) to involve perspective distortion properties at different fields of depth at a given focal length. To provide maximal measurement variation, the marker should not be placed in an orientation that is orthogonal to the viewing axis, since non-orthogonal views provide valuable information about the perspective distortion properties (e.g focal length) of the cameras.

Altogether, users are required to take a large number of measurements of a marker that is placed at widely spread-out positions and orientations within the field of view of the stereo cameras. This opens the floor for two issues:

1. The user has no appreciation for the required 3D spread of measurements - and neither for the required variation in marker orientation. This can be taught on a theoretical basis - but might be forgotten over time. Instead, such variational spread may also be obtained by the system via live visualizations, telling the user where and how to take measurements. We claim that such visualization is not enough and that enticing such user interaction with a computer game improves the overall results significantly.

2. The spatial setup limits the user's operating space. The marker must be within the respective field of view of both stereo cameras. Depending on distance from the camera rig (z), as well as on lateral and horizontal movement (x,y), the marker may leave the visible area of one or both cameras. Depending on the measurement environment, the marker may contrast well or poorly against the background. If the background is very busy (highly textured with objects similar in appearance to the marker pattern), the marker tracker may determine a wrong marker location. This can also occur due to partial marker occlusion by the users' hands or other objects in the environment. Spatially or temporally changing illumination conditions (e.g. under changing daylight illumination) can overexpose or underexpose images of one or both cameras. The marker should be held steadily during the picture taking process. Fast, erratic motion induces motion blur, resulting in imprecise marker detection.

Measurements taken under any such conditions are not suitable for stereo calibration and must be discarded. It is thus very valuable, if workers take a large quantity of measurements.

3.2 Implementation

Using Unity3D [5] as rendering engine and OpenCV [2] for tracking functionalities we have implemented a guiding application for the calibration procedure. The test person is provided with a tracked physical marker and a wireless controller to start/stop the registration procedure and to confirm any measurements taken. The program leads the test person through the process, highlighting different guided positions (see figure 3) with which the test person should to align the marker. The application shows the video stream of one of the cameras and augments it with an auxiliary (screen-based) 3x3 grid. The grid indicates different calibration zones (pyramid stumps of off-centered parts of the camera frustum).

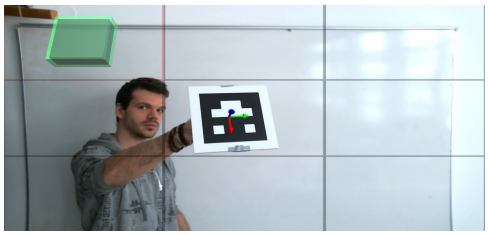

Figure 3: Basic Guiding Application View. Green Model Indicates Desired Marker Positioning

For every measurement, a translucent green marker is rendered onto one of the grid cells. The chosen cell highlights and test persons are tasked to align their marker with the 3D position that optimally matches the green virtual pendant (see figure 3). Once aligned, the test persons press a button to confirm the selection, the indicator disappears and the loop reiterates. The green marker model that appears in one of the 3x3 cells coincides, when aligned correctly, with the size of the black pattern of the physical marker. Its translucency helps test persons to better adjust the depth difference between the physical and the virtual marker. Note that the calibration procedure, in principle, does not require perfect alignment with the virtual green marker - it only entices test persons to cover a larger 3D volume with their measurements. Each time the user is satisfied with the alignment and confirms this via a button press, the system takes a snapshot with both cameras, calculates the marker pose relative to each camera, and estimates the baseline between them. The system then places the green marker in another calibration zone. We take care that each cell is covered once. This loop repeats for a randomized order of selected 3x3 grid areas until the user presses a button to stop the calibration process (indicating that he considers the calibration satisfied). For every user we save the number of measurements taken, the total application running time, all marker poses received by each camera, the final pose estimated and the images taken by each camera. These datasets allow for a detailed post-evaluation.

4 Design Guidelines for Gamifying Stereo Camera Registration

4.1 Challenges and Obstacles

Gamifying our application provides challenges and obstacles:

First, we want to keep our applications similar enough for them to remain comparable in empirical user studies. That means we cannot change the basic user interaction concept of marker alignment and confirmation. Yet, we have to add enough new factors to increase the motivational feeling for test participants and make use of the possibilities gamification offers.

Second, we need to give a gamified meaning to the test person's actions. Every measurement taken has to trigger behavior in the game, giving these actions an actual impact in the gamified world. Sound cues, visual effects on confirmations - anything that helps immerse the test person into the game world.

Third, we need to avoid a steep challenge curve. For our registration purposes we want participants to take as many measurements as possible. A harsh difficulty from the start causes frustration [19] and leads to less measurements taken and shorter application runtime in general. If conditions are harsh enough, this could lead to worse results in our gamified version than our non-gamified application.

Fourth, in a gamified environment the chance of defeat has to be present. A main driving factor has been clarified as the overcoming of obstacles, risk and reward of a challenge. If a participant cannot lose the game, and takes note of this situation, motivation will drop significantly. Yet, adding an "End Screen" under given loss conditions implies a stopping point for the calibration procedure. People will perceive this screen as a good point to end the registration and consider it over even if only a few measurements have been taken so far. We have to balance this situation by highlighting the option to continue, showing users their achieved score again and the overall progress they've made so far, to incite them to continue with a button press. Also, our background routines handling the calibration do not end on a simple "Game Over" screen. Player's restarting the application does not mean restarting the calibration. It continues taking measurements throughout multiple trials until the user ends the calibration application.

We stayed as close as possible to these basic rules when we developed the gamified version of the stereo camera calibration system.

4.2 Game Design for Stereo Camera Registration

Early discussions about the gamified design were dominated by our first basic rule: Keep both applications (gamified and non-gamified) comparable without decreasing the potential benefits each individual version could have. The non-gamified application served as a base onto which gamified elements were added.

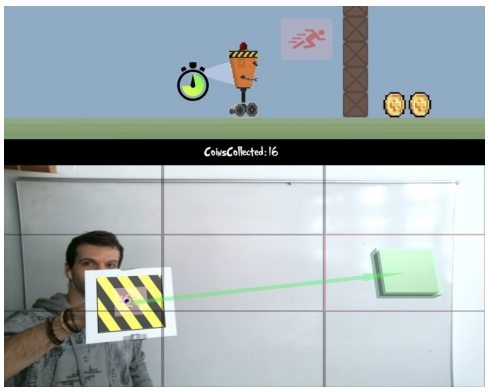

Figure 4: Full Gamified View of Registration Procedure

We scaled down the display area showing the camera view and add a new visual segment containing most game elements (see figure 4). This way, we were able to add a fantasy setting to the user interaction without changing the original registration procedure. To maintain comparability, we applied the same scaling to the non-gamified version, leaving the top third empty.

The authors agreed for the new element to host an *endless runner*-game. An endless runner is a game genre that has become very popular during recent years - especially as mobile apps on smartphones (example *TempleRun* [4]). The basic premise has a protagonist following a side-scrolling environment with randomly generated obstacles. Player interaction comes down to jumping, sliding or switching the active running-lane. Endless runners tend to stick to a very simple input scheme that functions with just a few button presses, something easily integrated into our own application. We added a simple implementation of this concept and scaled it to fit the blank space of our gamified rendering view (see figure 4).

We expect the game to be amenable for most player types since it appears well-known and does not require much effort. Future investigations will need to reevaluate this decision, based on interviews, sociological studies and experiments with real users who do perform sensor calibration tasks on a regular basis.

For a functioning design of the endless runner-game and to properly classify the game elements that we added, we reference back to Flatla et al.'s remodeling of Malone's work [11, 18]: The four basic game elements: Challenge, Theme, Reward and Progress. We categorize our game elements as follows:

World Setting (Theme) We added a pixel-style robot protagonist *RoB* that would slowly move from left to right, a white trail renderer indicating his active speed. RoB has to overcome obstacles spawned in his way and collect coins wherever possible. Music was included in tone with the game world's atmosphere: A happy tune with robotic beeps and 16-bit style effects. Sprites were kept in a pixel-art fashion with all sound effect reminiscent of old Nintendo titles. In general, the game world fits a modernized 16 bit setting with custom made graphics and sounds.

Obstacles and Lifes (Challenge) To provide a meaningful challenge we added randomly generated *Jump*- and *Boost*-obstacles along with corresponding input-options (see figure 5).

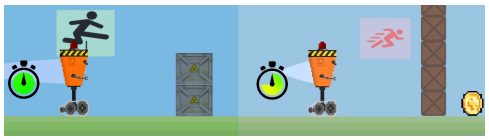

Figure 5: In-Game Jump- and Boost-Obstacle

Once the protagonist closes in on an obstacle, the camera zooms in and RoB comes to a halt. Each obstacle can be cleared with its respective action: Jump-obstacles have to be leaped over, Boost-obstacles have to be charged through. If the player doesn't clear an obstacle right away, a stopwatch will appear and count down the time left to clear it.

This play element accounts for the fact that we do not want to rush test persons. They should take their time to initiate the jump or boost action (by placing the marker in the pose proposed by the green virtual marker and pressing the accept button). Thus, we deisigned RoB to slow down and stop when reaching an obstacle. If the obstacle is not cleared during the duration of the stopwatch, the player will lose one of three lives, indicated by heart symbols.

This game element is intended to address the test person's interest in a measurable (yet not too demanding) challenge. When the test person presses the button or when the stopwatch time allowance is exceeded, the obstacle will disappear and RoB will move back on track with the camera zooming out again to standard distance.

Test persons take measurements in our applications by aligning their physical marker with the marker indicator on-screen and confirming their placement by a button press. Since the marker pose is not evaluated online, participants thus have a lot of freedom on how far they tilt the marker upwards. This measured tilt is used for one of our additional input options: We estimate the precise angle the marker has upon confirmation and use it to distinguish different interaction modes (Jump = Tilt Upwards, Boost = Hold Flat (see figure 6)). Thus, if a jump-obstacle appears, the player aligns the marker and tilts it upwards until it turns to blue. He/she then confirms by pressing a button and RoB will leap over the obstacle.

Challenge only consists of a feeling of risk and reward. The worth of a challenge depends on the risk avoided. We thus needed to allow players to take damage and eventually lose. They lose a game when then fail three obstacles and results in an end game screen blending in. The player can continue with a simple button press, but has to start over on collected coins and achievements.

Particle- and Sound-Effects (Theme, Reward) Every user action has to receive some form of feedback. This adds a feeling of significance to every command put in. To achieve this, we've added sound effects to every confirmation, every jump and every boost action. Additionally a particle cloud is instantiated and rendered on

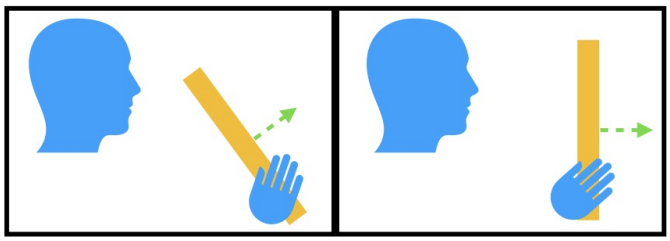

Figure 6: Different Marker Orientation Modes. (Left) Backwards Tilt - Jump Mode, (Right) Flat - Boost Mode

top of the marker's screen position on each alignment.

Coins / Points (Reward, Progress) With new obstacles spawned along the runner path, we've also added collectible coins. Rotating coin objects are placed in direct proximity of each obstacle and form an immediate reward for challenge cleared. To incite people's motivation the active coin-count is kept visible in the middle of our screen (see figure 4). A steadily increasing score doubles as a feeling of reward and a sense of progression through growing numbers.

High-Score (Reward, Progress, Challenge) To incite a sense of competition each player is asked to enter a fictional individual name for our high-score list. While playing, the game indicates the closest superior participant on our list. If the player scores higher than his/her competitor the mechanism moves on to the next best participant and shows the new score to beat. Once participants quit the application, their name and score are saved on our list.

Achievements (Reward, Progress) A longer lasting trend in gaming has been the awarding of virtual trophies and prices for in-game challenges often referred to as *achievements*. These achievements are visualized using simple graphic notifications and are highly customizable for developers. This allows them to link trophies to very specific in-game events like winning a certain level or even causing a very rare and specific scenario. We wanted to add this functionality for its great sense of reward and progression. Every time a specific event is triggered in our gamified application a small notification appears with a sound cue. Our total of 20 achievements range from *Collect 3 Coins* up to *Outmatch 10 High-Score Competitors*.

With these elements added, our gamified application has a large amount of motivational factors beyond the standard version.

5 User Study and Evaluation

Based on the gamified and non-gamified calibration implementations, we have conducted a user study.

5.1 Hypotheses

Our respective hypotheses for this paper are the following:

- **H1:** Users will use a gamified version for a longer period of time than its non-gamified counterpart.
- **H2:** Users will take more measurements in a gamified version.
- **H3:** The gamified application will not reduce the calibration quality.

We expect that effects will occur not only during one-time use but also for repetitional tasks, such as daily re-calibrations spread across several days.

5.2 Participant Groups

We gathered 18 people from several institutes of the university. They were mixed students of different fields of computer science and mechanical engineering at junior and senior levels (see table 1).

The participants were evenly divided into two separate groups (between-subject test design): Not-Gamified and Gamified.

Table 1: User Study - Participant Distribution

	Gamified	Not-Gamified
# Participants	9	9
Gender (f / m)	2 / 7	1 / 8
Calibration experience (y / n)	2 / 7	4 / 5

5.3 Experimental Setup

The test apparatus consisted of two identical cameras, mounted rigidly on top of a television monitor. The camera positions were carefully calibrated to provide a ground truth reference against which the calibration results of the test persons could be compared.

At the start of each calibration process, test persons were asked to sit approximately one meter in front of the monitor (see the right image in Figure 1). The participants received a marker (17.7 cm x 12.5 cm) in one hand and held an Xbox-One controller for button inputs in the other. They were instructed to move the marker in front of the setup according to the calibration and/or gaming tasks.

The video monitor showed the augmented video feed from one of the two cameras, gamified or not (see Figures 3 and 4). Furthermore, a miniature video feed of the second camera was shown in a corner of the monitor – to help the test persons verify whether the marker was visible for both cameras.

5.4 Execution of the test procedure

Participants were asked to select unique IDs, representing them anonymously in our data sets. They received identical descriptions of the stereo calibration problem and their respective application (Gamified or Non-Gamified). They were informed that a minimum of three measurements was required for the calibration to work – while more snapshots should result in a better overall quality.

Altogether, test persons each went through five runs. They were free to undertake these runs on their own scheduling over the course of two weeks.

5.5 Measurements

We collected both objective and subjective measurements.

Figures 7, 8, 9 and 10 show box plots of all measurements. They represent the median as a line, the 25% and 75 % percentiles as a box, the 1.5 IQR limits as whiskers, and outliers as small dots. The plots show that our data is not normally distributed. We thus use the Wilcox-RankSum test to determine significant differences.

The statistical data is summarized in table 2. The table shows the means and standard deviations of all measurements that were taken during the five test runs. A sixth column presents the statistics for measurements of all five runs combined. The table further indicates the results of the Wilcox-RankSum significance tests (w and p values). Significant differences at level 0.05 (5%) are annotated with one asterisk. Two asterisks represent significance at level 0.01 (1%).

5.5.1 Application run time

Runtime measurements were collected with a simple timer running alongside the calibration procedure. The time measured covers the period from starting the application to ending it in the menu.

5.5.2 Number of measurements/images taken by the users

A measurement is taken each time a user confirms a new marker alignment by button press. Both cameras then take a snapshot of the environment and save the pictures for later evaluation.

5.5.3 Measurements per minute

Using run time and measurement count, we can derive the *measurements per minute (mpm)*, describing the number of measurements taken for every full minute of runtime.

5.5.4 Calibration quality

Outliers occur because we are tracking participants in motion. Not all camera snapshots are clear and usable. Some might be blurred to the extent that no marker is detected in the image at all. Much worse are blurred images still detecting markers but with less precision: The resulting marker pose becomes falsified.

5.5.5 Outliers

To prevent strong outliers from polluting the evaluation, all measurements taken in our gamified and non-gamified applications are run through a filter. We iterate over each individual run's results and determine mean and standard deviation. Results that are not within two standard deviations of the mean are considered to be outliers. They are not considered in a subsequent recalculation of the mean. This process is run twice. Afterwards, the final mean and standard deviation are set. All outliers after this stage are dropped from further consideration.

For the filtered data sets, we compare the estimated marker poses to the respective ground truth data. The deviation from the ground truth data is calculated using euclidean distance.

5.5.6 Subjective measurements

As an additional utility in understanding our users' behaviour test persons were required to fill out a System Usability Scale form [3] after every trial they went through, as well as a NASA-TLX form [1].

6 Discussion

6.1 Results: Application run time (H1)

All in all, the gamified runtime runs higher throughout all trials, occasionally doubling the non-gamified runtime in comparison (run 1 and 3). The RankSum test, applied to the combined run time measurements of all 5 runs indicates a highly significant difference with W = 1287, p-value = 0.000. We conclude that hypothesis H1 is correct: Gamification is capable of producing significantly longer run times for calibration procedures.

Yet, there is a fast convergence towards shorter run times over the course of our trials and we have to expect that a further continuation with more trial runs would diminish the differences between both versions even more. The tendency indicates that the current gamified application can only sustain user motivation for this long and will eventually fade out in its effect. As discussed in section 2.2, different people react differently to individual game elements [8, 9]. As a result, some test persons quickly loose interest in the game. In a future investigation, we need to investigate in more detail which game elements fit most into the context of calibration tasks – specifically geared towards specific player types.

6.2 Results: Number of images (H2)

In all runs except run 4, the gamified application generated more than twice the number of measurements than the non-gamified version. For several runs, as well as over all runs combined, the RankSum test detected highly significant differences (W = 1424, p = 0.000 in the combined case). We thus conclude that hypothesis H2 is correct: Gamification can incite users to take significantly more measurements.

Yet, we see a decrease in numbers over the gamification runs as opposed to a consistent number of images taken in the non-gamified application. This reinforces the observation made above: our chosen game features seem to be wearing off over time.

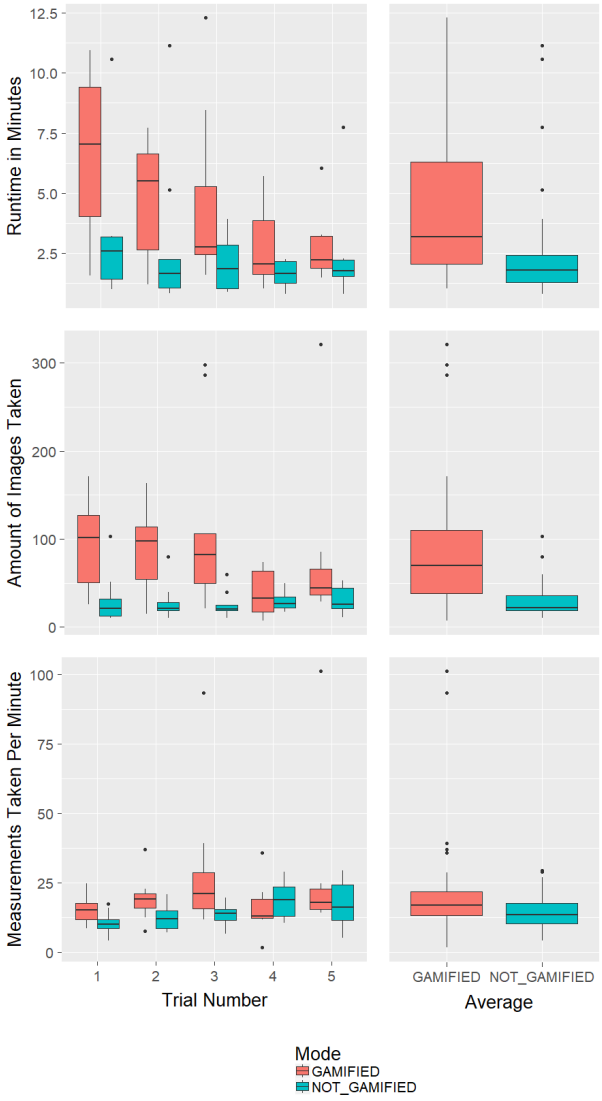

Figure 7: Runtime, number of images and measurements per minute

6.3 Results: Measurements per minute

Interestingly, even though the run time and the measurements taken in the gamified calibration have decreased from run 1 to run 5, the average number of measurements taken per minute remains fairly stable – or even increases. There are some possible reasons for this. Over the course of five runs, participants learn to interact more freely with the gamified application. The feeling of time pressure incites them to move fast and to be able to align the marker instantaneously when necessary. They learn to respond quickly and can take more measurements in a shorter period of time. In contrast, the non gamified application does not induce any sense of haste. Users align their markers and confirm when they are placed correctly without any indication of time being an issue. If this difference in the gamified application persists, then the calibration task can be made to increasingly challenging by adjusting the game-flow to a higher pace to counter act the players skill. This means adding additional challenges and speeding up the application behavior whenever possible. This idea will be further inspected in future works.

Table 2: Evaluation Results

Run	1	2	3	4	5	Combined
			Run Time [minutes]			
Gamified	6.59 (3.60)	4.73 (2.43)	4.53 (3.60)	2.82 (1.71)	2.85 (1.56)	4.39 (2.99)
Non-Gamified	3.10 (2.93)	2.93 (3.35)	1.96 (1.06)	1.62 (0.56)	2.44 (2.19)	2.31 (2.23)
W(P)	56 (0.059)	64 (0.040) *	63 (0.0503)	39 (0.231)	37 (0.336)	1287 (0.000) **
			Number of Measurements			
Gamified	94.25 (51.08)	87.22 (47.15)	118.22 (102.23)	39.43 (28.03)	85.57 (105.40)	86.95 (74.43)
Non-Gamified	31.66 (29.56)	28.33 (21.12)	25.67 (15.20)	29.00 (11.03)	31.50 (15.60)	28.41 (18.64)
W(P)	65 (0.006) **	69.5 (0.012) **	75.5 (0.002) *	31.5 (0.728)	43 (0.093)	1424 (0.000) **
			Measurements per Minute			
Gamified	15.45 (5.12)	19.32 (8.17)	29.59 (25.35)	16.14 (10.57)	29.88 (31.70)	22.59 (18.92)
Non-Gamified	10.78 (4.03)	12.60 (4.56)	13.75 (4.00)	18.99 (6.83)	16.95 (9.09)	13.82 (5.83)
W(P)	56 (0.0592)	65 (0.031) *	68 (0.014) *	21 (0.463)	34 (0.536)	1165 (0.005) **
			Outliers			
Gamified	2.34 (1.51)	2.56 (2.19)	3.56 (4.98)	1.00 (1.53)	1.14 (1.78)	2.23 (2.87)
Non-Gamified	1.22 (1.79)	1.33 (1.73)	0.67 (0.71)	1.63 (1.19)	1.63 (2.07)	1.28 (1.53)
W(P)	52 (0.124)	56.5 (0.160)	60.5 (0.074)	17.5 (0.229)	23.5 (0.626)	1037 (0.097)
			Calibration Error [mm]			
Gamified	0.55 (0.22)	0.54 (0.26)	0.49 (0.29)	0.46 (0.10)	0.72 (0.48)	0.55 (0.29)
Non-Gamified	0.51 (0.23)	0.75 (0.42)	0.55 (0.30)	0.93 (0.91)	0.74 (0.99)	0.73 (0.63)
W(P)	38 (0.888)	29 (0.340)	35 (0.667)	18 (0.281)	33 (0.613)	811 (0.660)
			System Usability Scale (SUS)			
Gamified	64.44 (19.79)	63.89 (26.43)	60.00 (26.10)	55.71 (28.89)	55.00 (30.54)	56.16 (28.70)
Non-Gamified	69.17 (11.25)	70.00 (9.60)	71.11 (11.73)	76.88 (8.53)	73.13 (11.32)	71.85 (10.14)
W(P)	38.5 (0.894)	42 (0.929)	31 (0.426)	14 (0.117)	16 (0.332)	704 (0.019) *
			TLX Mental Demand			
Gamified	9.33 (6.34)	6.11 (4.96)	3.89 (3.95)	8.00 (6.83)	7.50 (8.14)	6.40 (6.05)
Non-Gamified	2.44 (2.70)	1.89 (1.96)	2.78 (3.19)	1.75 (2.25)	1.25 (1.75)	2.22 (2.41)
W(P)	65.5 (0.028) *	62 (0.061)	44.5 (0.749)	46 (0.039) *	35 (0.146)	1370 (0.001) **
			TLX Physical Demand			
Gamified	10.78 (6.46)	10.44 (5.59)	8.56 (6.89)	11.43 (5.59)	11.00 (7.90)	11.02 (6.48)
Non-Gamified	8.33 (6.062)	7.67 (5.48)	7.33 (5.41)	5.25 (4.86)	6.50 (5.24)	6.78 (5.24)
W(P)	54 (0.248)	53 (0.284)	46 (0.658)	46.5 (0.036) *	33.5 (0.242)	1386 (0.001) **
			TLX Temporal Demand			
Gamified	8.78 (5.47)	9.78 (5.93)	9.22 (7.14)	9.00 (6.06)	10.17 (6.49)	8.72 (6.17)
Non-Gamified	5.56 (3.13)	4.11 (4.20)	5.67 (2.96)	2.88 (3.18)	5.63 (5.42)	4.74 (3.71)
W(P)	54 (0.248)	63 (0.049) *	50.5 (0.400)	49 (0.016) *	33.5 (0.242)	1357 (0.002) **
			TLX Performance			
Gamified	10.78 (2.99)	7.33 (5.57)	10.33 (7.05)	15.14 (3.18)	10.17 (8.54)	11.23 (6.24)
Non-Gamified	6.78 (4.71)	9.78 (4.44)	10.00 (4.36)	7.00 (5.38)	11.13 (6.312)	9.59 (5.55)
W(P)	61 (0.076)	29.5 (0.353)	42 (0.929)	51.5 (0.007) **	23 (0.948)	1164 (0.151)
			TLX Effort			
Gamified	11.00 (5.24)	11.11 (5.11)	8.67 (6.30)	10.86 (6.12)	12.00 (7.07)	11.28 (6.01)
Non-Gamified	7.78 (3.99)	7.67 (5.66)	5.33 (4.44)	4.75 (6.73)	3.88 (2.47)	5.89 (4.92)
W(P)	57.5 (0.144)	54 (0.248)	53 (0.286)	44 (0.070)	41.5 (0.027) *	1488 (0.000) **
			TLX Frustration			
Gamified	7.00 (6.305)	8.00 (6.56)	9.67 (6.98)	12.71 (6.42)	13.17 (7.81)	10.47 (7.11)
Non-Gamified	7.22 (5.29)	7.33 (5.15)	8.67 (5.48)	4.63 (5.18)	8.75 (5.18)	7.35 (5.17)
W(P)	36.5 (0.755)	43.5 (0.825)	46 (0.658)	49 (0.017) *	32.5 (0.300)	1249 (0.033) *

6.4 Results: Outliers

Both in the gamified and in the non-gamified setting, only few calibration results had to be filtered out due to poor quality. Combined across all five runs, there is no significant difference (W = 1037, p = 0.097). This is interesting since it indicates that the more complex and faster user interaction of the gamified version did not result in significantly worse measurement quality.

6.5 Results: Calibration quality (H3)

The distance to ground truth averages to 0.55 centimeters in the gamified version and 0.73 centimeters in the non-gamified application. Considering all individual trials differences become more appearant: Both versions remain within a range of roughly 3 mm around one another. The gamified version shows slightly better results on trials 2,3 and 4 whereas the non-gamified application scores slightly better in trial 1 and 5.

H3 proposes that the resulting calibration quality will not decrease through gamification features. With our given data set RankSum significance (W = 811, p=0.660 for combined calibration quality across all five runs), we cannot prove or disprove this assumption. We experienced gamified results closer to the actual ground truth on average, but saw fluctuating results from trial to trial.

In retrospect, we observe that our stereo camera calibration scenario is very basic. It does not require highly dynamic user interactions. Using our non-gamified guiding UI, test persons were able to generate very good calibrations with very few samples. Thus,

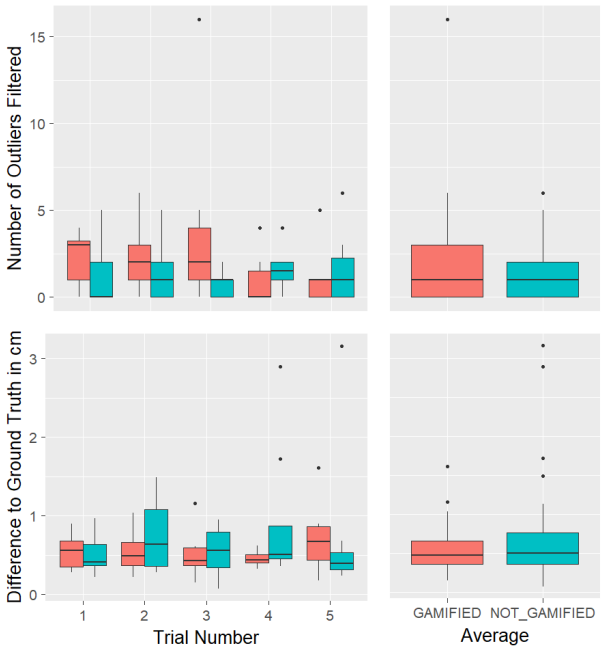

Figure 8: Amount of Outliers filtered and Calibration Error Summary

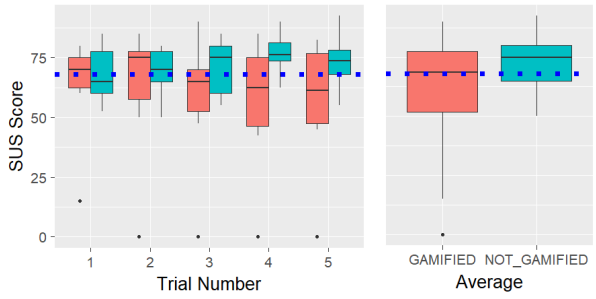

Figure 9: Total SUS Score. Blue line indicates above average 68 points mark

further samples of the gamified routine were mostly redundant. In a future investigation, we need to select a more complex calibration routine requiring time-critical user interaction.

6.6 Results: Subjective measurements

We saw before that gamified participants performed more measurements and acquired longer total runtimes in our application than non-gamified users. All throughout our test groups, participants scored the gamified application as less user friendly than its counterpart. Running these tests over all five trials the scores of the gamification users dropped while the non-gamified applications remained relatively consistent in high scoring areas.

Moving on to the TLX testing, along all axes the gamified application is perceived as putting more pressure on its participants than the not-gamified version. It is likely, that this stress causes the negative usability results back in our SUS testing.

We hypothesize in two directions: For one, games as a whole might just be more demanding on participants than non-gamified applications. This raises the question if demand, both physically and mentally, is just a part of a desirable game experience. A part that might be tolerated or even desired to a certain extent. The other theory, is that our gamification in this procedure is not yet at its

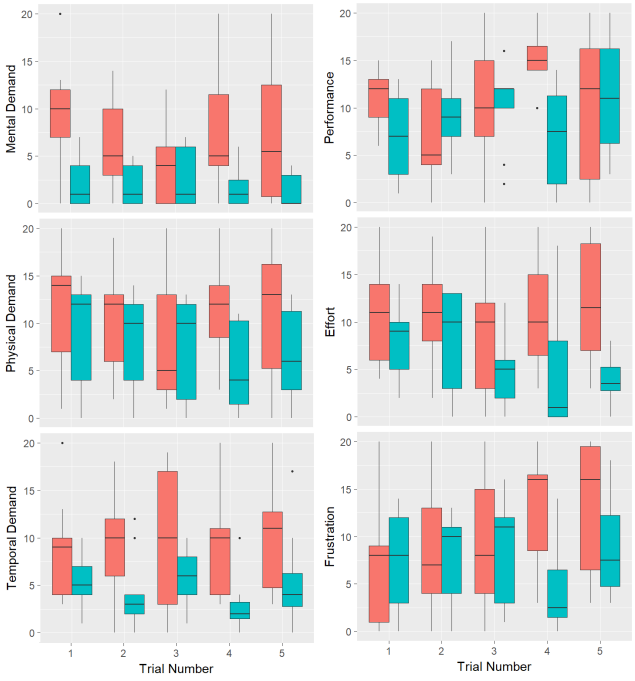

Figure 10: TLX Results

actual full potential. That the effect our current feature-set has, is strong enough to entice users despite its shortcomings, but not strong enough to hide demand and usability problems. We will have to open this up for further discussion and design a new but comparable tool-set for a new calibration game in the future.

7 Conclusions and Future Work

For this paper we wanted to analyze the effect of gamification on calibration tasks. Even more, we wanted to build on the work of Flatla et al. [11] by extending towards a long term evaluation and different procedures. Flatla categorized game elements and added them to a provided calibration task. He inferred, that these elements can lead to motivational benefits for the users without degrading the overall resulting calibration quality. From this starting point we have added two major changes.

First, we moved away from Flatla's calibration procedures (color space calibration, etc.) and demonstrated these concepts on a stereo camera calibration. Second, we evaluated the resulting applications in a long term study, focusing on the resulting user behavior in repeating the same tasks five times over. In this section we conclude by interpreting our results and addressing the three hypothesis this research aims to address. Additionally, we discuss other measures presented in the results such as mpm, SUS and TLX for a deeper evaluation of the experimental data.

The results we recorded showed comparable calibration quality to the non-gamified version, lead to more measurements taken and a longer running time on all runs. While weakening over time, these effects remained throughout the course of multiple repetitions. This fall-off became apparent in both versions (gamified- and non-gamified) with the gamified application maintaining higher values in most cases. Regarding the calibration result, both versions are close to the positional ground truth. With occasional noise in input images resulting from user movement or other hardware/software issues a certain positional offset is unavoidable for both applications. Sensor calibration in general profits highly from filtering algorithms (e.g. RANSAC) that are capable of removing eventual outliers from the study results. Yet, for better quality outcomes extensive filtering

requires larger data sets with more available measurements to choose from. We thus believe that calibration quality profits highly from gamified applications and their larger measurement count overall.

It should be noted that camera calibration procedures are still subject to change in many ways. Progress is being made towards requiring less and less measurements for a usable end result (see Rojtberg et. al. [24]), which of course diminishes the need for a large data-set to begin with. This leads us to extend our research into new and different calibration procedures in the future (example procedures see Fuhrmann et. al. [13]). User recommendations, along with our study results, lead us to many new ideas to continue this work. We expect that the competitive aspect was one of the major motivational aspects for users in the gamified version: The sense of accomplishment upon beating a high score, the idea to get "just one more coin" before stopping to further strengthen your own position within the high-score list. We aim to increase this competitive aspect in future tests and explore just how strongly they influence participants in comparison to other factors.

Also, as discussed in our study evaluation, we want to investigate further what it takes to maintain a high mpm throughout all our trials. For now, we've established that gamification can lead to significant changes in user behavior for AR purposes. Higher amounts of measurements taken, longer application usage and a similar if not potentially better calibration quality. We will also investigate further regarding the empirical usability results. Trying to determine how much mental and physical strain is not only acceptable for a gamified application but even desired by participants.

ACKNOWLEDGMENTS

We would like to thank the test persons who participated in the user study.

REFERENCES

[1] Humansystems nasa tlx reference page. `https://humansystems.arc.nasa.gov/groups/TLX/`. Accessed: 2018-11-14.

[2] Opencv homepage. `www.opencv.org`. Accessed: 2018-11-14.

[3] System usability scale information page. `https://www.usability.gov/how-to-and-tools/methods/system-usability-scale.html`. Accessed: 2018-11-14.

[4] Templerun itunes appstore page. `https://itunes.apple.com/de/app/temple-run/id420009108?mt=8`. Accessed: 2018-11-14.

[5] Unity3d engine homepage. `www.unity3d.com`. Accessed: 2018-11-14.

[6] A. Alempijevic, S. Kodagoda, and G. Dissanayake. Sensor registration for robotic applications. In *Springer Tracts in Advanced Robotics*, volume 42, pages 233–242, 2008.

[7] A. Anonymous. Gamifying stereo camera registration for augmented reality. In *Adjunct Proceedings of the IEEE International Symposium for Mixed and Augmented Reality 2018 (To appear)*, 2018.

[8] R. Bartle. Richard A. Bartle: Players Who Suit MUDs. *MUSE web site, April*, page 1, 1996.

[9] D. Dixon. Player Types and Gamification. *CHI 2011 Workshop Gamification Using Game Design Elements in NonGame Contexts*, pages 12–15, 2011.

[10] C. Eichhorn, D. A. Plecher, G. Klinker, M. Lurz, N. Leipold, M. Böhm, H. Krcmar, A. Ott, D. Volkert, and A. Hiyama. Innovative game concepts for alzheimer patients. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2018.

[11] D. R. Flatla. calibration games: Making calibration tasks enjoyable by adding motivating games elements, 2011.

[12] C. S. Fraser. Automatic Camera Calibration in Close Range Photogrammetry. *Photogrammetric Engineering & Remote Sensing*, 79(4):381–388, 2013.

[13] A. Fuhrmann, R. Splechtna, and J. Prikryl. Comprehensive Calibration and Registration Procedures for Augmented Reality. In B. Froehlich, J. Deisinger, and H.-J. Bullinger, editors, *Eurographics Workshop on Virtual Environments*. The Eurographics Association, 2001.

[14] P. Hirmer, M. Wieland, U. Breitenbücher, and B. Mitschang. Automated sensor registration, binding and sensor data provisioning. In *CEUR Workshop Proceedings*, volume 1612, pages 81–88, 2016.

[15] B. K. P. Horn. Closed-form Solution of Absolute Orientation using Unit Quaternions. *Journal of the Optical Society of America*, 4(4):629–642, 1987.

[16] A. Langbein, D. A. Plecher, F. Pankratz, C. Eghtebas, F. Palmas, and G. Klinker. Gamifying stereo camera calibration for augmented reality. In *Adjunct Proceedings of the IEEE International Symposium for Mixed and Augmented Reality (ISMAR)*, pages 125 – 126, 2018.

[17] B. Lange, C.-Y. Chang, E. Suma, B. Newman, A. S. Rizzo, and M. Bolas. Development and evaluation of low cost game-based balance rehabilitation tool using the microsoft kinect sensor. In *Engineering in medicine and biology society, EMBC, 2011 annual international conference of the IEEE*, pages 1831–1834. IEEE, 2011.

[18] T. W. Malone. Toward a theory of intrinsically motivating instruction. *Cognitive Science*, 5(4):333–369, 1981.

[19] Mihaly Csikszentmihalyi. *Flow - The Psychology of Optimal Experience*. 1990.

[20] F. Palmas, D. Labode, D. A. Plecher, and G. Klinker. Comparison of a gamified and non-gamified virtual reality training assembly task. In *2019 11th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*, pages 1–8, Sep. 2019.

[21] K. Pfeuffer, M. Vidal, and J. Turner. Pursuit calibration: making gaze calibration less tedious and more flexible. *Uist*, pages 261–269, 2013.

[22] D. A. Plecher, C. Eichhorn, J. Kindl, S. Kreisig, M. Wintergerst, and G. Klinker. Dragon tale-a serious game for learning japanese kanji. In *Proceedings of the 2018 Annual Symposium on Computer-Human Interaction in Play Companion Extended Abstracts*, pages 577–583. ACM, 2018.

[23] D. A. Plecher, C. Eichhorn, A. Köhler, and G. Klinker. Oppidum - a serious-ar-game about celtic life and history. In A. Liapis, G. N. Yannakakis, M. Gentile, and M. Ninaus, editors, *Games and Learning Alliance*, pages 550–559, Cham, 2019. Springer International Publishing.

[24] P. Rojtberg and A. Kuijper. Efficient pose selection for interactive camera calibration. In *Proceedings of the IEEE International Symposium for Mixed and Augmented Reality 2018 (To appear)*, 2018.

[25] H. Schäfer, D. A. Plecher, S. L. Holzmann, G. Groh, G. Klinker, C. Holzapfel, and H. Hauner. Nudge-nutritional, digital games in enable. 2017.

[26] R. Y. Tsai and R. K. Lenz. A New Technique for Fully Autonomous and Efficient 3D Robotics Hand/Eye Calibration. *IEEE Transactions on Robotics and Automation*, 5(3):345–358, June 1989.

[27] M. Tuceryan, D. S. Greer, and R. T. Whitaker. Calibration Requirements and Procedures for a Monitor-Based Augmented Reality System. *IEEE Transactions on Visualization and Computer Graphics*, 1(3):255–273, 1995.

[28] Z. Zhang. A flexible new technique for camera calibration. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(11):1330–1334, 2000.