Leveraging the User's Face for Absolute Scale Estimation in Handheld Monocular SLAM

Sebastian B. Knorr* Daniel Kurz†

Figure 1: In monocular visual SLAM the structure of a scene as well as the motion of the world-facing camera are only estimated up-to-scale but can be brought to absolute scale by simultaneously capturing and tracking the user's face in the user-facing camera (left). This enables superimposing virtual objects, e.g. the green wire frame model of a parcel, at absolute scale (right).

Abstract

We present an approach to estimate absolute scale in handheld monocular SLAM by simultaneously tracking the user's face with a user-facing camera while a world-facing camera captures the scene for localization and mapping. Given face tracking at absolute scale, two images of a face taken from two different viewpoints enable estimating the translational distance between the two viewpoints in absolute units, such as millimeters. Under the assumption that the face itself stayed stationary in the scene while taking the two images, the motion of the user-facing camera relative to the face can be transferred to the motion of the rigidly connected world-facing camera relative to the scene. This allows determining also the latter motion in absolute units and enables reconstructing and tracking the scene at absolute scale.

As faces of different adult humans differ only moderately in terms of size, it is possible to rely on statistics for guessing the absolute dimensions of a face. For improved accuracy the dimensions of the particular face of the user can be calibrated.

Based on sequences of world-facing and user-facing images captured by a mobile phone, we show for different scenes how our approach enables reconstruction and tracking at absolute scale using a proof-of-concept implementation. Quantitative evaluations against ground truth data confirm that our approach provides absolute scale at an accuracy well suited for different applications. Particularly, we show how our method enables various use cases in handheld Augmented Reality applications that superimpose virtual objects at absolute scale or feature interactive distance measurements.

Keywords: SLAM, monocular, handheld, absolute scale, user-facing, face tracking, distance measurements, true size.

*e-mail: knoseb@gmail.com
†e-mail: d@nielkurz.de

1 Introduction and Motivation

Visual simultaneous localization and mapping (SLAM) describes the process of observing a scene with at least one camera from different viewpoints and simultaneously building a 3D model of the scene as well as estimating the poses of the camera at the observations. There exists a multitude of SLAM methods based on different types and setups of cameras (e.g. passive or active stereo cameras [10]) but in the simplest case only a single intensity camera is used to observe and reconstruct the scene, which is referred to as monocular SLAM, e.g. [4].

If initially both the scene as well as the camera poses corresponding to the observed images are unknown, any 3D reconstruction and camera pose is only determined up-to-scale. This means that it is unknown which *absolute* distance (e.g. in millimeters) in the real world corresponds to a unit of the coordinate system in which the reconstructed scene model and the estimated camera poses are defined. This scale ambiguity results from the fact that images only measure a projection of the scene. Therefore usually an arbitrary scale is assigned.

Many applications, however, require a scene reconstruction or camera poses at absolute scale, e.g. vision-based navigation or Augmented Reality (AR) applications that superimpose virtual objects (e.g furniture) at absolute scale in a previously unknown real environment. Sometimes not *absolute* scale, i.e. the relation to real-world distances, is necessary but still it is beneficial to perform SLAM at *repeatable* scale, e.g. to overcome scale drift or to obtain a consistent scale of separately mapped parts of a scene.

This paper proposes to obtain absolute scale of a scene being reconstructed and tracked by a world-facing camera of a handheld device by simultaneously capturing the user's face with a rigidly attached user-facing camera. If the absolute dimensions of the face (e.g. the interpupillary distance) are known, the motion of the user-facing camera can be determined at absolute scale. We then can infer the absolute scale for the poses of the world-facing camera relative to the scene under the assumption that the face did not move with respect to the scene.

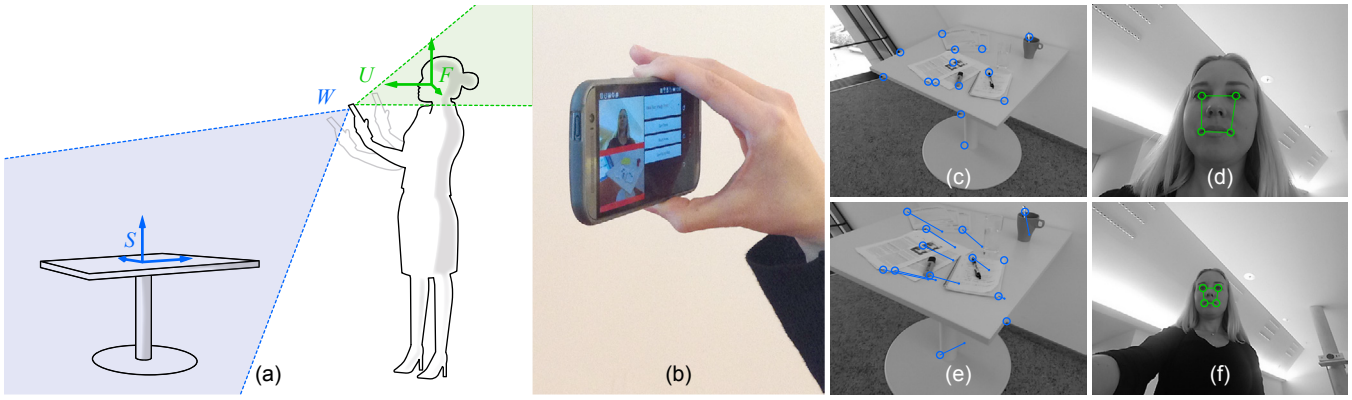

Figure 2: Simultaneous capturing with the world-facing camera W and the user-facing camera U (a) of a mobile phone (b) delivers a sequence of image pairs of the scene S (c,e) and the face F (d,f).

2 Related Work

Monocular SLAM methods, e.g. MonoSLAM [4] or PTAM [7], are widely used for 3D scene reconstruction from a sequence of images captured by a single moving camera. One reason for their popularity is that they only require an intensity camera and no instrumentation of the scene. A well-known shortcoming of monocular visual SLAM is its ambiguity in scale. To overcome this under-determined problem additional information on scale must be provided.

Davison *et al.* [4] propose to add spatially fix calibration objects of known size into the scene for determining absolute scale of the camera motion and scene structure in monocular SLAM systems. As the scene and the calibration objects are both captured by the same camera, the added objects change the appearance of the scene to be reconstructed. Additionally this approach relies on the user to have specific calibration objects at hand, as well as to *actively* position them in the scene and to capture them with the camera.

Scale information can also be provided by a known baseline between two camera poses. Klein and Murray [7] for example manually provide the absolute distance between the two camera positions where the two images for the initial 3D triangulation are captured.

Besides monocular SLAM, for rigidly connected stereo camera systems (i.e. two cameras with overlapping frusta) the baseline between the two cameras can be calibrated offline as proposed by Lemaire *et al.* [9]. Clipp *et al.* [3] describe how to solve for the absolute scale using a multi-camera setup with *non-overlapping* camera frusta. They leverage the known baseline and a single point correspondence within the images of the second camera by exploiting differences in translations between the motions of the two cameras induced by rotations. The displacement between the two cameras has to be significant, making this approach unsuitable for handheld devices, e.g. mobile phones.

Information about absolute scale can also be provided by special sensors as e.g. presented by Lieberknecht *et al.* [10] who integrate depth information from an RGB-D camera into monocular vision-based SLAM. Such kind of sensors, however, are not commonly available in handheld devices. Additionally active depth cameras projecting and measuring infrared light do not work reliably outdoors during daylight.

Sensor fusion of vision with an Inertial Measurement Unit (IMU) is used by Nützi *et al.* [13] to estimate absolute scale in monocular SLAM for moving vehicles by double integrating acceleration measurements over time yielding a position in meters. Tanskanen *et al.* [14] employ inertial sensors in mobile off-the-shelf handheld devices for estimating metric scale. Those IMUs tend to be somewhat inaccurate resulting in an error of about 10-15% in scale estimates. Also IMU-based approaches require a certain amount of motion over a period of time (15-30 seconds [13, 14]) to estimate the scale.

Our proposed method lies in between methods relying on objects of known size, a determined baseline, and sensor fusion. We employ the user-facing camera of a handheld device as additional sensor. Even though we are using two rigidly connected cameras with non-overlapping camera frusta, we do not directly utilize baseline information between the two cameras like [3] to estimate absolute scale, as this baseline in most handheld devices is negligibly small. Instead, we capture images of the user's face and use the face as kind of a known object providing us with camera poses relative to the face at absolute scale.

By that we substitute the extra object of known size by a body part of the user comparable to Lee and Höllerer [8] who derive scale information for their markerless tracking approach by an initial camera pose estimation from the user's outstretched hand captured by the world-facing camera. While their approach uses a single camera and requires the user to reach out, such that their hand becomes visible in the image of the camera, our approach in contrast relies on the user-facing camera in which the user's face is automatically present most of the time.

Face tracking algorithms work universally over almost all humans because the appearance of facial fiducials can be well approximated by a limited range of variation. Many methods for detecting facial fiducials [2] and for determining the pose of a face [12] exist, which however often deliver a pose at arbitrary scale. For unknown subjects, Flores *et al.* [6] estimate the absolute distance of a face from the camera using the perspective distortion visible in the 2D images in combination with knowledge about how facial fiducials are distributed across people, learned from a small training set of exemplary 3D models of human faces. Similarly, Burgos-Artizzu *et al.* [1] estimate the distance of the camera from an unknown person, based on training in image space on a dataset of frontal portraits of 53 individuals each captured from seven distances. They also investigate which facial landmarks are suitable for the estimation.

Combining a standard 6DoF face tracking method with knowledge about the absolute dimensions of some part of the particular face allows inducing absolute scale for the tracking. For this purpose we employ, in our current implementation, the human interpupillary distance (IPD). While a generic face model with mean IPD can be used, the facial dimensions can additionally also be calibrated for a particular user's face to further enhance the accuracy of the induced scale.

Dogson [5] presents a collection of available statistics on IPD, with a mean adult IPD around 63 mm, and the vast majority of adults having IPDs in the range of 50 mm to 75 mm.

The method we propose in this paper takes advantage of the ability to estimate camera poses relative to human faces at abso-

lute scale. It further makes use of commonly available handheld devices comprising a world-facing camera and additionally a user-facing camera, which captures the user's face. Our method works non-intrusively, not affecting the appearance of the scene to be reconstructed, and works well even in outdoor scenarios during daylight. It neither requires a separate calibration object, such as a marker, to be available and added to the scene, nor does it rely on dedicated sensing hardware, such as depth sensors, stereo cameras with overlapping frusta, or IMUs.

3 Approach

In order to enable monocular SLAM at absolute scale, our proposed method requires a handheld device comprising a world-facing camera and a user-facing camera as shown in figure 2(a).

Instead of adding a marker of known size to the scene and capturing both – scene and marker – with the world-facing camera, we propose to use the user's face as scale reference, which is usually visible in the user-facing camera. Taking advantage thereof renders any instrumentation of the scene unnecessary.

The absolute dimensions of the user's face can be calibrated once as described in section 5.2.3 and can then be re-used subsequently. If no calibration data is available, a generic average face model can be selected instead as a fallback since facial dimensions, such as the IPD, vary only moderately among different adult humans [5].

With the face model defined at absolute scale, the pose of the user-facing camera relative to the face can be determined at absolute scale in real time based on the image of the user-facing camera by means of a 6DoF face tracking method [12].

Since the world-facing camera and the user-facing camera are rigidly connected to each other and the transformation in between is at least approximately known, also the pose of the world-facing camera can be determined at absolute scale relative to the face, as derived in section 3.1. Under the assumption that the face has not moved relative to the scene over a period of time, the meanwhile determined absolute poses relative to the face are also valid relative to the scene, which enables reconstructing the scene at absolute scale. Analogously it enables transforming an existing reconstruction of the scene from arbitrary scale to absolute scale.

We introduce the following notation. At a time t the user-facing camera captures the image $U(t)$ and the world-facing camera captures the image $W(t)$. The pose of the user-facing camera relative to the coordinate system of the user's face is referred to as $\mathbf{U}_F(t)$, see figure 3(a). The pose of the world-facing camera in the coordinate system of a reconstruction of the scene at arbitrary scale is called $\mathbf{W}_S(t)$ (figure 3(b)) while the pose of this camera in the coordinate system of the user's face is referred to as $\mathbf{W}_F(t)$ (figure 3(a)).

3.1 Absolute Scale From Two Keyframes

At a first keyframe t_1, we store the image $W(t_1)$ of the world-facing camera (figure 2(c)), and the corresponding image $U(t_1)$ of the user-facing camera (figure 2(d)). After moving the camera to a different viewpoint, a second keyframe t_2 with images $W(t_2)$ (figure 2(e)) and $U(t_2)$ (figure 2(f)) is stored. We then use image $U(t_1)$ to determine pose $\mathbf{U}_F(t_1)$ and image $U(t_2)$ to determine pose $\mathbf{U}_F(t_2)$ using a face tracking method at absolute scale (figure 3(a)). Given the extrinsic rigid body transformation $\mathbf{E}$ (see section 4.1) between the user-facing and the world-facing camera we can determine $\mathbf{W}_F(t_1)$ as $\mathbf{E}\mathbf{U}_F(t_1)$ and $\mathbf{W}_F(t_2)$ as $\mathbf{E}\mathbf{U}_F(t_2)$.

Under the assumption that the face stayed stationary in the scene, the poses $\mathbf{W}_F(t_1)$ and $\mathbf{W}_F(t_2)$ are valid relative to the scene, which enables reconstructing the scene at absolute scale using triangulation. It further allows to compute the scale factor a from the arbitrary units of the coordinate system of an up-to-scale model of the scene S to the absolute units of the coordinate system of the user's face F as

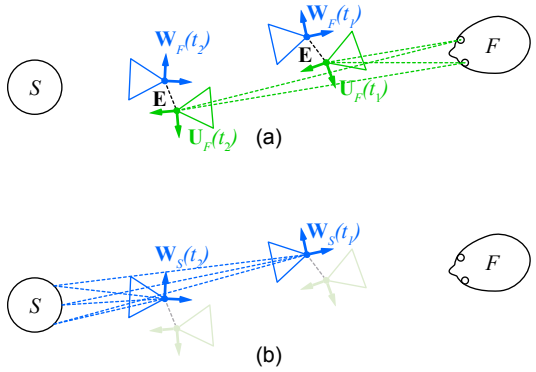

Figure 3: Face tracking (a) allows to determine poses relative to the face at absolute scale for the user-facing and the world-facing camera, which then (b) can be used to transform poses from monocular SLAM relative to the scene from arbitrary to absolute scale.

$$a = \frac{\|\tau(\mathbf{W}_F(t_1)) - \tau(\mathbf{W}_F(t_2))\|}{\|\tau(\mathbf{W}_S(t_1)) - \tau(\mathbf{W}_S(t_2))\|} \quad (1)$$

where the operator τ extracts the translation vector of a pose and $\mathbf{W}_S(t_1)$ and $\mathbf{W}_S(t_2)$ are determined using visual camera localization relative to the model of the scene at arbitrary scale (figure 3(b)).

3.2 Absolute Scale From Multiple Keyframes

While in theory two keyframes t_1 and t_2 suffice, a longer sequence of keyframes of user-facing and world-facing camera images lets us compute one scale factor a_i for each pair of keyframes (t_i, t_j), which gives us a set of scale factors. Some of them are more accurate than others. It is important to keep in mind that the calculation of the factor only works reliably, if the keyframes t_i and t_j differ sufficiently in terms of translation of the cameras. To determine a more reliable and robust overall scale factor based on more than two keyframes we randomly select from the sequence a set of N pairs of keyframes (each keyframe comprising of an image of the user's face and an image of the scene) such that the user-facing camera moved at least a distance of d_{min} between the two keyframes of each pair and compute a scale factor a_i per pair of keyframes using equation (1). Finally we compute a factor $\widetilde{a} = \text{Median}(A)$ of the set of all scale factors $A = \{a_1, a_2, ..., a_N\}$ and use $\widetilde{a}$ to scale the reconstruction.

Figure 4 plots the distribution of scale factor estimates a_i and the median $\widetilde{a}$ for an example sequence. In section 5 we quantitatively evaluate how accurate our proposed method estimates the absolute scale of a scene based on the median scale estimate over a sequence.

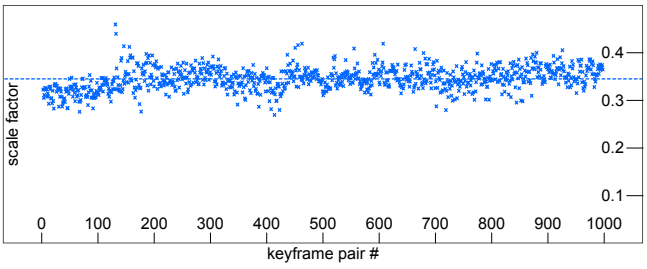

Figure 4: Distribution and median of estimated scale factors for a set of 1000 pairs of keyframes capturing a scene and a face.

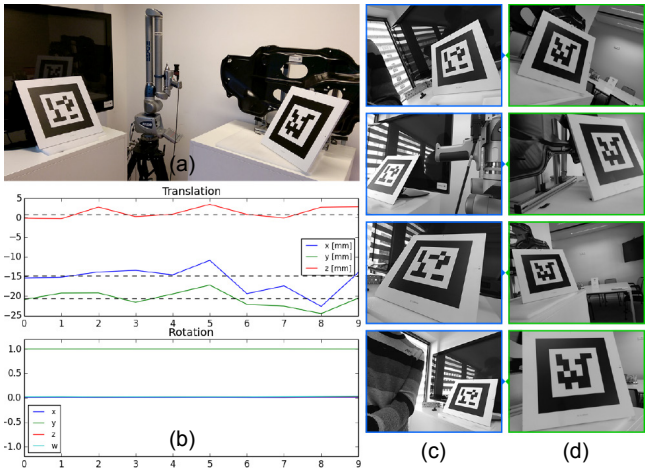

Figure 5: The marker setup (a), calibrated by a mechanical measurement arm, which we used to calibrate the extrinsic parameters (b) between the two cameras using a set of image pairs (c,d).

4 Implementation

To proof our proposed method, we implemented it based on an HTC One M8 mobile phone (figure 1), which allows simultaneous capture from the world-facing camera and the user-facing camera. We use a resolution of (640×480) pixels for both cameras and determine intrinsic parameters of each camera using images of a checkerboard [16].

4.1 Extrinsic Inter-Camera Calibration

The user-facing camera on the front and the world-facing camera on the back of a handheld device are not located at the exact same spot. For evaluation purposes we calibrated the extrinsic parameters $\mathbf{E}$, i.e. translation and rotation, between the two cameras of the employed phone.

For that we first accurately determined the positions and rotations of two markers (figure 5(a)) by touching their respective corners with the tip of a mechanical measurement arm. We then moved the mobile phone between the two markers such that the user-facing camera captures the first marker (figure 5(c)) while the world-facing camera sees the second marker (figure 5(d)). For each image pair, the camera poses were determined in a common coordinate system based on marker tracking, and the resulting 6DoF transformation between the two cameras was computed (figure 5(b)). Finally we computed the median of the coordinates of the translation vector and the rotation expressed as quaternion to determine the extrinsic parameters $\mathbf{E}$ transforming from the coordinate system of the user-facing camera to world-facing camera coordinates.

The results show that the two cameras are facing nearly perfectly in opposite direction with a translational offset of length 26 mm along the image plane.

4.2 Offline Evaluation

Our experiments – both for evaluation and real-time applications – are based on a proprietary monocular SLAM system from the Metaio SDK [11]. The SLAM system is capable of running in real-time on the mobile phone mapping a real scene and tracking the pose $\mathbf{W}_S(t)$ of the world-facing camera relative to it at arbitrary scale. Besides poses, the SLAM system provides the 3D coordinates of reconstructed points.

We use two approaches in our evaluations for determining the pose $\mathbf{U}_F(t)$ of the user-facing camera relative to the user's face. To simulate perfect 6DoF face tracking at absolute scale we place, in section 5.1, a square marker where the user's face would usually be and track it using the marker tracking framework of the Metaio SDK. In section 5.2 we further use a proprietary face tracking method which provides the 6DoF pose of a camera relative to a face given an image of it. This method is based on a generic face model which can be adjusted by one parameter, the IPD, to account for the faces of different users.

For the quantitative evaluations in section 5 we merely use the phone as a capturing device. A custom app allows to store synchronized video sequences of the world-facing camera and the user-facing camera to files at a framerate of $\sim$30 Hz. All the further processing then is performed offline on a PC. This enables repeatable evaluations and systematically testing the impact of different parameters on the estimated absolute scale.

4.3 Real-Time Applications

In addition to the quantitative offline evaluations we present in section 6 different applications that are enabled by our proposed method. These applications run in real-time on the mobile phone without any additional PC. The deployed SLAM system is the same as in the offline evaluation, while we use a mobile-specific proprietary method for the face tracking. In the real-time application the scale is estimated using equation (1) for the first and last keyframe of the sequence instead of the median over the whole sequence.

4.4 Stationary Face Assumption

Our current implementation assumes that the user's face remains static with respect to the environment during the scale estimation.

The user at the moment manually triggers the scale estimation procedure by holding down a button at the lower left of the user interface, see figure 8(a). By performing a motion with the smartphone towards or away from the face, the face stays intuitively stationary. Before and after the scale estimation procedure, which roughly takes a second, the user can again freely move around.

If the user's face does not remain stationary in the scene during the procedure, the estimate will be inaccurate. The relative error in measuring the distance traveled by the user-facing camera relative to the world falsifies the estimated scale factor a proportionally.

5 Evaluation

We quantitatively evaluate the accuracy and precision in estimated scale achieved by our method in order to assess which use cases it enables. We compare the dimensions of reconstructed maps of a scene, which we brought to absolute scale using our method, against ground truth information on the dimensions of the scene. The scenes we use in our evaluation are spherical because this allows for an easy and reliable evaluation against ground truth. Our proposed method itself supports scenes of any shape.

We use two styrofoam spheres (figure 6(a)) at two different sizes pasted up with newspaper. The large sphere has a radius of 102 mm, while the small sphere has a radius of 61 mm. By moving the world-facing camera around the respective sphere, we obtain a 3D reconstruction using monocular SLAM at arbitrary scale. We then track the sphere based on this reconstruction and in parallel track the user's face in the user-facing camera. This enables us to determine the scale factor $\widetilde{a}$ (see section 3.2 with $N = 1000$ and $d_{min} = 120$ mm) between the arbitrary scale of the reconstruction and the real dimensions of the scene in millimeters.

The device in these sequences is moved mainly along the optical axes of the cameras as illustrated in figure 3. This movement makes it more convenient to not move the head relative to the scene as opposed to sideways motions. The predominant translational motion also reduces the influence of the transformation $\mathbf{E}$ between user-facing and world-facing camera on the covered distances of the respective cameras.

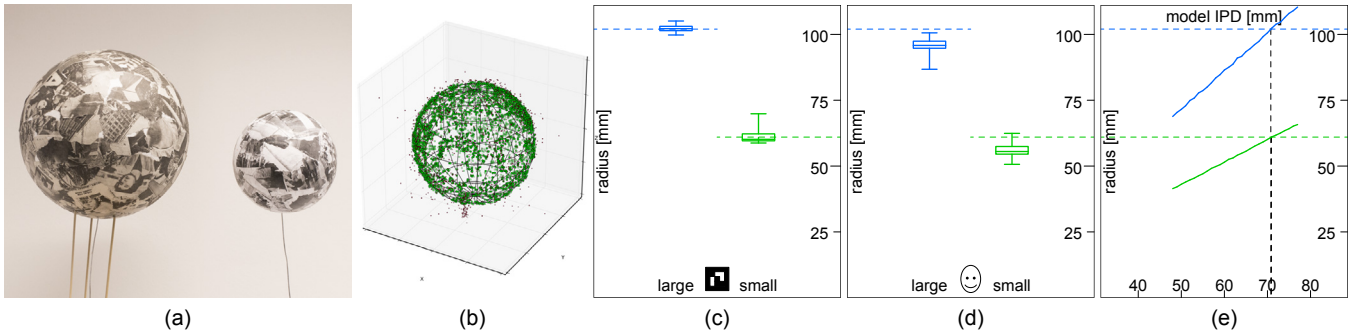

Figure 6: We reconstruct multiple times two spherical scenes (a) with different known radiuses and fit a virtual sphere to each reconstruction to determine its radius at arbitrary scale (b). The distribution of estimated radiuses at absolute scale (mm) by our method is shown for an idealized case using marker tracking (c) and for the real-world case using face tracking (d) in comparison with ground truth. The influence of inaccurate interpupillary distance calibration is plotted in subfigure (e) for both scenes.

To measure the accuracy of the scale estimation, we fit a virtual sphere to each set of reconstructed 3D points (figure 6(b)) using RANSAC. We scale the fitted radius by the estimated scale factor and compare it against the ground truth radius. To evaluate on as much data as possible, we created for each of the two real spheres six reconstructions using SLAM. The radiuses of these arbitrarily scaled reconstructions vary between 95.9 and 718.2 for the large styro sphere and between 68.6 and 454.5 for the small styro sphere. Additionally we captured ten sequences of a few seconds each with the respective sphere being tracked with the world-facing camera while the user-facing camera captures a face or marker. All combinations of reconstructions and sequences result in 60 radius estimates per sphere.

5.1 Under Perfect Conditions – Marker Tracking

To get an idea of the accuracy and precision achievable under perfect conditions, i.e. without any motion between face and scene and with very accurate 6DoF face tracking, we first replace the user's face with a square marker on a tripod at the position where the user's face would usually be and use 6DoF marker tracking instead of face tracking.

5.1.1 Results

The resulting estimated radiuses for the two spherical scenes based on all combinations of six reconstructed maps and ten camera sequences are plotted in figure 6(c). For each scene, a candlestick chart shows minimum, first quartile, median, third quartile, and maximum value of the estimated radiuses. A dashed horizontal line shows the ground truth radius for the reader's reference. We see that in this configuration our method achieves to estimate the radiuses of the two spheres with both high accuracy and precision. The median of 102.17 mm over all estimates for the large styro sphere corresponds to a relative error of 0.16 % (equivalent to 0.17 mm) with a standard deviation of 1.20 mm over all estimates. For the small sphere the median of 60.17 mm corresponds to a relative error of 1.36 %, (equivalent to 0.83 mm) with a standard deviation of 2.59 mm over all estimates.

5.1.2 Influence of the Extrinsic Inter-Camera Calibration

For all the estimations evaluated above as well as plotted in figure 6(c) we considered the extrinsic rigid body transformation $\mathbf{E}$ between the user-facing and the world-facing camera, determined in section 4.1. $\mathbf{E}$ can be assumed to vary between different handheld devices, and potentially there is not always a calibration available. We therefore separately run the estimations on the same sequences *ignoring* the extrinsic calibration, i.e. using a generic extrinsic calibration $\mathbf{E}$ assuming that the two cameras are located exactly at the same position. Note that the rotation between the two cameras is irrelevant for the distances used for the scale calculation.

The simplification of ignoring the extrinsic rigid body transformation $\mathbf{E}$ only slightly affects the results with a median of 102.87 mm for the large and 60.49 mm for the small styro sphere with comparable standard deviations. This negligible influence is partly due to the carried out nearly purely translational motion towards and away from the face, as only rotational motions induce translation, as well as due to the small baseline between the cameras of less than 3 cm. It shows, that our method works even without extrinsic calibration for a particular device.

5.2 Under Realistic Conditions – Face Tracking

We then evaluate our method using face tracking instead of marker tracking. To enable face tracking at absolute scale we provide the IPD of the particular person to the face tracking method.

5.2.1 With Calibrated Interpupillary Distance

For this part of the evaluation the IPD of the user has been calibrated manually using a ruler and a mirror. During capturing the sequences used for the scale estimation the user tried to avoid moving their head but we can assume that small motions occurred.

The distribution of radiuses of reconstructed spheres in 60 runs per scene are plotted in figure 6(d). We observe that in this case estimations are less accurate and the radiuses, and hence the scale of the scene, are mostly underestimated.

The median of 95.81 mm over all estimates for the large styro sphere corresponds to a relative error of 6.07 % (equivalent to 6.19 mm), the median of 55.65 mm over all estimates for the small styro sphere corresponds to a relative error of 8.78 % (equivalent to 5.35 mm). With a standard deviation of 2.35 mm for the large styro sphere and 2.60 mm for the small one, the estimates however are only slightly less precise than those obtained with marker tracking.

5.2.2 Influence of the Interpupillary Distance

We use the IPD to enable face tracking at absolute scale. If for a user the exact IPD is not available the mean IPD of an adult person, which is about 63 mm [5], could be assumed. Hence we evaluate the impact of an inaccurately calibrated IPD on the absolute scale estimate. Therefore we estimate the radiuses of the two spheres based on sequences of a user with an IPD of 68 mm while configuring the face tracking method to use an IPD between 48 mm and 77 mm in steps of 1 mm which covers the vast majority of adults [5].

Figure 6(e) plots the radius of the reconstruction at absolute scale as a function of the assumed IPD. We observe a linear dependency between the two parameters. The introduced percental error in scale estimation is linearly coupled to the error between real and

Figure 7: Arbitrarily scaled reconstructions of a real scene lead to superimposed virtual objects at arbitrary scale (a). Estimating absolute scale enables correctly scaled superimpositions (b).

assumed IPD. Statistically the potential lack of accuracy from relying on statistics for this distance instead of calibrating it for a user follows the same distribution as the IPD. According to the Ansur database mentioned in [5], the IPD (age 17 to 51) follows a normal distribution with mean 63.4 mm and standard deviation of 3.8 mm, corresponding to $< 6\,\%$.

Interestingly the most accurate reconstructions were achieved with IPD 71 mm for both the large and the small styro sphere, while the manually measured IPD for the subject is 68 mm. This suggests some bias in our applied face tracking method.

5.2.3 Per User Calibration of the Interpupillary Distance

The IPD may be calibrated manually using e.g. a ruler. People with glasses may also already have their IPD measured by an optician.

Beside that, a semiautomatic calibration of the IPD or other facial features can be performed using the dual camera setup presented in here by inverting the scale transfer. Camera motion can be estimated at absolute scale by means of e.g. marker tracking using the world-facing camera. By simultaneously tracking the facial features to be calibrated on the user-facing camera, the absolute scale can be transferred to the facial features as long as the user's face stays static with respect to the marker. We implemented a prototype of this semi automatic approach and performed 8 calibration runs for a person with an IPD of 68 mm. The resulting estimations for the IPD had a mean value of 68.9 mm with a standard deviation of 2.3 % (corresponding to 1.6 mm). The deviation towards an overestimated IPD is to a certain extent consistent with our observation of the underestimated sizes for the small and large styro spheres based on face tracking in section 5.2.1. Enhancements with regard to the face tracking method could in future eliminate this bias and additionally lower the standard deviation for improved accuracy and precision in IPD calibration as well as scale estimation.

The calibration for a particular user only needs to be done once. If multiple users share a device, visual face recognition could be employed to select the stored calibrations corresponding to the particular user from the set of available calibrations.

6 Applications

Our proposed method enables a variety of handheld AR applications, which require camera pose estimation or mapping of a real environment at absolute scale, e.g. superimposing virtual objects at absolute scale, or interactive distance measurements in the scene.

6.1 Superimposition at Absolute Scale

When virtual objects in AR act as a substitute of a real object, it is beneficial to superimpose them at absolute scale. Common examples include virtually placing a piece of furniture, e.g. an armchair (figure 7), in the living room to test if it would fit in the room and how it matches with the remaining (real) furniture. Here it is crucial that the armchair is superimposed at correct size (b), compared to a superimposition at arbitrary scale (a), which provides a wrong visual feedback to the user about the potential real appearance.

Virtually placing canvas prints on the wall to pick the right size out of the available selection (figure 8(a,b)) is another example when superimposition of virtual 3D objects requires absolute scale.

Absolute scale is also needed when not visual appearance within the surroundings but the physical dimensions themselves matter, see e.g. figure 8(c) where the size of a parcel is visualized so that the user can visually decide which parcel size is needed for a particular shipment.

6.2 Measurements at Absolute Scale

Additionally to augmentations of virtual objects at correct size, the reconstruction of a scene at absolute scale enables measuring distances within the scene.

Conventionally, different tools are needed for measuring depending on the use case and scale, from rulers to tape measures. It is even harder to measure when either the direct connection between measurement points is not possible e.g. the length of a wall occupied in between by furniture or when measurement points are visible but out of reach.

Our proposed method enables using the same measurement tool – the smartphone – to perform all these measurements – from a few millimeters to many meters – in a convenient non-contact manner. By simply clicking on the touch screen which shows the scene captured by the world-facing camera, the corresponding 3D locations in the scene are selected and the distance between successively selected scene locations is provided to the user. This allows a convenient way to perform measurements for a large variety of settings and object. It can be used to measure a path (figure 8(e)) corresponding to the needed length of a cable. It further allows to measure the linear distance between 3D points (figure 8(d)) that cannot be directly measured with a ruler or a measuring tape including the total length over multiple path segments (figure 8(f)).

Depending on the particular use case, scene locations can be selected on a plane aligned to either the ground or a wall or on an arbitrary object in 3D space. For selecting locations on a plane we project a touch position in the camera image onto the respective plane and thus obtain the 3D coordinates of the corresponding scene point. For general 3D scenes, we project all 3D features of the reconstructed SLAM map into the camera coordinate system and select the feature which projects closest to the touch position.

We evaluate the accuracy of distance measurements performed with the tool explained above. Note that this evaluation includes the accuracy of the SLAM and the user's ability to select points on the screen. Results are listed in table 1. Except for the outcome for the envelope, the achieved measurements have a relative error below 7 % over the whole range of small scale measurements of a stamp up to large scale measurements of a whole room. While the achieved accuracy is sufficient for many use cases we plan to further improve it in the future.

Table 1: Measurement Results in Comparison with GT Distances

	Distances (cm)		
Object	GT	Measurement	Relative Error (%)
Stamp (diag.)	3.6	3.8	5.6
Envelope (diag.)	24.4	27.3	11.9
Book (diag.)	30.0	31.7	5.7
Barbell	34.7	35.8	3.2
Newspaper (diag.)	69.5	72.3	4.0
Table	122.6	131.0	6.8
Room	898.0	902.6	0.5

7 Conclusions and Future Work

This paper presented the first approach to take advantage of the world-facing and user-facing camera in current handheld devices to estimate absolute scale in handheld monocular SLAM. In combination with leveraging the face of the user as trackable object of

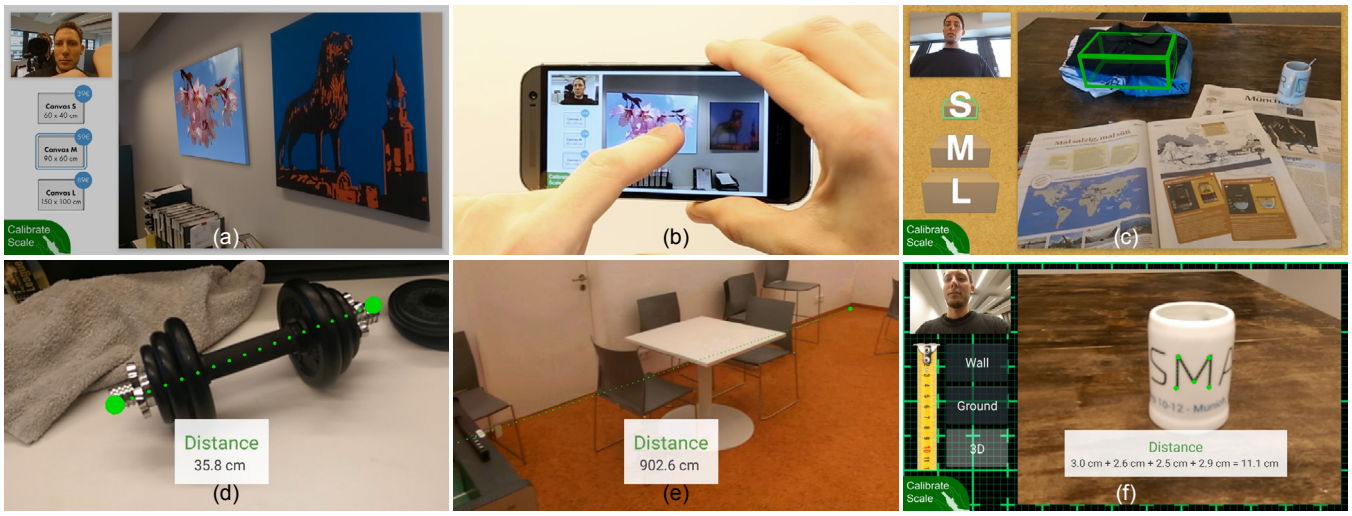

Figure 8: Examples of the various applications which our proposed method enables. Performing SLAM at absolute scale enables superimposing virtual objects at absolute scale (a-c) as well as measuring distances at absolute scale, e.g. in centimeters (d-f).

known size this brings multiple benefits over common approaches superseding an additional marker or object of known size and being non intrusive to the scene to be reconstructed. Our method enables a variety of AR applications from displaying virtual objects superimposed onto a scene at the correct size (figure 8(a-c)) to distance measurements (figure 8(d-f)).

Our experiments showed for different scenes, that scale could be estimated with a relative median error of less than 9 % which outperforms the IMU based approach by Tanskanen *et al.* [14] who report an error of 10-15 %. However direct comparison to alternative approaches is hard to achieve. The IMU approach requires stronger movements and around 30 seconds to estimate scale, while our implementation at the moment would fail if the head is not kept static and delivers a higher error when face dimensions are not calibrated.

Our method is largely independent of the particular employed systems for monocular SLAM and face tracking as it uses both as black boxes that provide poses. It however depends on the quality of the poses and hence will immediately benefit from any improvements in both the SLAM system or the face tracker in terms of precision, accuracy and robustness. Potential for high accuracy has been demonstrated in section 5.1 where we substituted face tracking with marker tracking and achieved a median relative error <1.4 %.

We showed that using a generic IPD still results in reasonable estimates which are slightly inaccurate but still precise, i.e. repeatable. This allows to map parts of a larger scene separately at a consistent scale. For several applications, e.g. playing (augmented) games, approximate information on the absolute scale of a scene may suffice.

Our method requires the user's face to be stationary in the scene during the scale estimation, which in practice takes about a second. In our evaluations, the user was instructed to not move their face. In future work, we will look into methods to automatically determine when the face did not move relative to the scene for a set of keyframes and then automatically perform (re-)estimation of the absolute scale as a background process of a SLAM system. Continuous scale estimates can not only be combined into a more robust scale factor but also prevent scale drift – an important problem to address in monocular SLAM.

Evaluating if the face remained stationary could be done based on a similarity transformation [15] between the two trajectories from SLAM and face tracking (considering the extrinsics $\mathbf{E}$), which would deliver the wanted scale factor, and remaining inconsistencies would indicate a motion of the face. Motions of the face could also be identified by transforming epipolar constraints from one camera to the other and evaluating if the constraints hold for the moving features of the respective tracked target. For both methods however there remain certain motions of the face that cannot be identified like this, which we plan to address in future work.

References

[1] X. P. Burgos-Artizzu, M. R. Ronchi, and P. Perona. Distance estimation of an unknown person from a portrait. In *Proc. ECCV*. 2014.

[2] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. *IJCV*, 107(2):177–190, 2014.

[3] B. Clipp, J.-H. Kim, J.-M. Frahm, M. Pollefeys, and R. Hartley. Robust 6dof motion estimation for non-overlapping, multi-camera systems. In *Proc. WACV*, 2008.

[4] A. J. Davison, I. D. Reid, N. D. Molton, and O. Stasse. MonoSLAM: Real-time single camera SLAM. *TPAMI*, 29(6):1052–1067, 2007.

[5] N. A. Dodgson. Variation and extrema of human interpupillary distance. In *Proc. SPIE 5291*, 2004.

[6] A. Flores, E. Christiansen, D. Kriegman, and S. Belongie. Camera distance from face images. In *Proc. ISVC*. 2013.

[7] G. Klein and D. Murray. Parallel tracking and mapping for small AR workspaces. In *Proc. ISMAR*, 2007.

[8] T. Lee and T. Höllerer. Multithreaded hybrid feature tracking for markerless augmented reality. *TVCG*, 15(3):355–368, 2009.

[9] T. Lemaire, C. Berger, I.-K. Jung, and S. Lacroix. Vision-based SLAM: Stereo and monocular approaches. *IJCV*, 74(3):343–364, 2007.

[10] S. Lieberknecht, A. Huber, S. Ilic, and S. Benhimane. RGB-D camera-based parallel tracking and meshing. In *Proc. ISMAR*, 2011.

[11] Metaio GmbH. Metaio SDK, http://www.metaio.com/sdk, Mar. 2015.

[12] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *TPAMI*, 31(4):607–626, 2009.

[13] G. Nützi, S. Weiss, D. Scaramuzza, and R. Siegwart. Fusion of IMU and vision for absolute scale estimation in monocular SLAM. *IROS*, 61(1-4):287–299, 2011.

[14] P. Tanskanen, K. Kolev, L. Meier, F. Camposeco, O. Saurer, and M. Pollefeys. Live metric 3d reconstruction on mobile phones. In *Proc. ICCV*, 2013.

[15] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *TPAMI*, 13(4):376–380, 1991.

[16] Z. Zhang. A flexible new technique for camera calibration. *TPAMI*, 22(11):1330–1334, 2000.