Cake Cam: Take Your Photo and Be in it Too

CANDICE LUSK, Brigham Young University
MICHAEL D. JONES, Brigham Young University

Tourists often turn to strangers when they need a photographer while traveling; however, they do so at a cost. Strangers are not typically trained photographers, nor are they telepathically intuiting what composition the tourist wants. Existing smartphone camera interfaces do not communicate the desired framing to the stranger, and prior work in mobile photography guidance does not manage the 3D movement required when composing the tourist's ideal photo. We offer a new kind of mobile interaction for communicating the intended photo to a stranger without instructions. In our methodology, the tourist first composes a photo with the desired framing. The app, Cake Cam, then stores the camera position and orientation. Finally, 3D augmented reality markers guide the stranger to retake the photo with the tourist now standing in the frame. Our study resulted in more accurate camera placements and required fewer additional instructions than the traditional tourist photography method (n=40).

CCS Concepts: • **Human-centered computing → Computer supported cooperative work**; *Mobile devices*; User interface design.

Additional Key Words and Phrases: Collaborative photography; Augmented reality

ACM Reference Format:
Candice Lusk and Michael D. Jones. 2020. Cake Cam: Take Your Photo and Be in it Too. 1, 1 (April 2020), 19 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Tourists often hand their cell phones to complete strangers and ask them to take the tourists' picture in front of a landmark. However, the result is often unsatisfying for the tourist. The image may include framing problems, such as cutting off the landmark or the tourist in the image. The issue is that the tourist is far more interested in the quality and composition of the image than the stranger who took the picture. The stranger does not have the time or motivation to recreate the specific photo the tourist had in mind.

For clarity, we will refer to the person who frames and appears in the picture as the *tourist* and the person who takes the final picture as the *local*—whether or not the two people collaborating to take a picture are in fact tourists or locals.

Figure 1 depicts this scenario. The tourist wanted a photo of her parents and herself, with the Golden Gate bridge extending across the photo in the background. The picture on the left is the photo captured by the local. While the picture does have the Golden Gate Bridge extending into the background, it is not as prominent as she would have liked. At first glance, it would be easy to miss the landmark completely. Additionally, the local composed the photo with the white sky covering half of the image causing the camera's automatic exposure settings to compensate for its brightness and underexpose the rest of the image.

Authors' addresses: Candice Lusk, Brigham Young University, Provo, Utah, candice.lusk@byu.edu; Michael D. Jones, Brigham Young University, Provo, Utah, jones@cs.byu.edu.

Fig. 1. The poorly framed, crooked tourist photo on the left was taken by a local in a hurry. The properly exposed, well-framed photo on the right was taken by the tourist's dad after receiving and following thorough verbal instructions. Most locals are not willing to invest the time to receive and follow instructions in this scenario.

This photo was not what the tourist intended, and she wanted another picture taken. The tourist's dad captured the photograph on the right of Figure 1. Her dad was more willing to invest the effort required to understand and capture the specific photo the tourist intended. This scenario is based on the experience of a friend of the author. General observations around tourist photography come from personal understanding rather than an explicit ethnographic study.

As we reflected on our own impromptu tourist photography experiences, we realized that the core problem is guiding the user to take a photo from a specific location; otherwise, the image may be taken from a different angle or position than intended by the tourist resulting in an unsatisfying photo. Prior work explores methods of replicating or guiding participants into taking a photo with a smartphone in various contexts ([Bourke, McCarthy, and SmythBourke et al.2011], [Li and VogelLi and Vogel2017], [McAdam, Pinkerton, and BrewsterMcAdam et al.2010], [Brewster and JohnstonBrewster and Johnston2008], [Xu, Ratcliff, Scovell, Speiginer, and AzumaXu et al.2015], [Carter, Adcock, Doherty, and BranhamCarter et al.2010]).

The most similar work explores overlay-based interfaces to give guidance. This kind of interface shows a semi-transparent copy of the target photo overlaid on the camera preview. Figure 4 shows an example of this kind of interface from our work. Bourke et al. studied this kind of interface as part of compositional guidance in the Social Camera project [Bourke, McCarthy, and SmythBourke et al.2011] and found "some interface challenges." This is consistent with our evaluation of the same interface in which users found them confusing. Much of the other work focuses on

Fig. 2. The photograph on the left was taken by the tourist using Cake Cam. This photo represents the composition the tourist intended. The photograph on the right was taken by a local with no verbal instructions. Cake Cam guided the local into taking the same photograph set by the tourist, but the tourist is now standing in the frame.

guiding participants to take an aesthetically "better" photo based on the current viewing screen of the smartphone. This guidance is typically limited to panning motions along a 2D plane and does not handle the 3D movement (6 degrees of freedom: x, y, z, roll ϕ, pitch θ, yaw ψ) needed to replicate a specific photo. Additionally, many of the interfaces require training to use.

Our main contribution is a novel mobile interface that quickly guides the user into replicating a photo in the context of having a photo taken at a tourist location. The app, titled Cake Cam, lets the tourist take his or her photo and be in it too. In contrast to current mobile guidance methods, Cake Cam uses 3D augmented reality markers to guide participants to capture a specific photo with no verbal instructions. In addition to guiding camera position (x, y, z), 3D markers, with exaggerated depth, give critical feedback for matching camera orientation (ϕ, θ, ψ).

To use the app, a tourist takes a photo of the carefully framed scene. The tourist then asks a local to take their picture, hands the phone to the local, and moves into the frame. 3D Augmented reality alignment markers guide the local to correctly frame the picture with the tourist in it, producing the intended photo.

Figure 2 illustrates this scenario. The photograph on the left was the initial framing photograph taken by the tourist, representing the tourist's intended composition of the photo. The photograph on the right is the photo captured by the local using the app Cake Cam. The two photos are nearly identical in composition with the major difference being that

the tourist is now standing in the frame. By simply aligning the 3D markers, the local was able to replicate and capture the unique angle the tourist wanted, without needing instructions.

After presenting related work, we evaluate the usability of several existing mobile photography guidance methods and describe how this informed our design of Cake Cam. Then we discuss the technical aspects that went into implementing the app. Following that, we present the results from a user study with 40 participants. These participants were recruited on site as they walked towards a university campus. Each participant was asked to take a picture of the research team member in front of a landmark. No other verbal instructions were given. We found that in comparison to the usual process of handing a local a camera, Cake Cam allowed participants to capture the intended photograph faster and more accurately than a verbal description could have. We close with a discussion of possible explanations for key results, limitations of the study and opportunities for future work.

2 RELATED WORK

This work builds on prior work in real-time photography guidance, collaborative photography and computer vision. The following sections outline the techniques and insights that have contributed to this work.

2.1 Real-Time Photography Guidance

Prior work has explored methods of replicating a photo. The Social Camera [Bourke, McCarthy, and SmythBourke et al.2011] is a mobile app that guides users into taking photographs based on their current location and scene context. The Social Camera uses overlaid images to guide users into replicating example photographs. Overlaid images contain too much detail in many cases and are difficult to align. The RePhoto app [1] and the SOVS app [2] also use the target image to guide the photographer to the correct pose. Panorama guidance uses boxes or arrows on the camera preview to help the photographer keep the camera level with panning.

Other work focuses on guiding the user to take an ascetically "better" photograph using photography principles, rather than following the personal preferences of the photographer. Li et al. [Li and VogelLi and Vogel2017] developed a system that interactively guides the user into taking a better self-portrait ("selfie"). Three different icons guided the user into taking a better photo based on empirical models of three parameterized composition principles: face size, face position, and lighting direction. McAdam et al. [McAdam, Pinkerton, and BrewsterMcAdam et al.2010] researched guidance for low-level features like exposure, luminance, and motion blur. NudgeCam [Carter, Adcock, Doherty, and BranhamCarter et al.2010] is a smartphone app that provides real-time feedback based on standard photography heuristics to encourage higher-quality video. It places text and colored boxes around the face on the image to indicate problems with the video feed. Brewster et al. [Brewster and JohnstonBrewster and Johnston2008] developed a system that guides subject positioning within a landscape photograph, using a visualized rectangle. However, this system was not implemented on a smartphone or tested for interactive use. Xu et al. [Xu, Ratcliff, Scovell, Speiginer, and AzumaXu et al.2015] developed a photo-taking interface using a three-camera array that provides real-time feedback on how to position the subject of interest according to the rule of thirds. The commercial app Camera51 [3] guided participants with a phone icon and rectangle using scene context and photography principles.

[1] http://projectrephoto.com/
[2] https://itunes.apple.com/us/app/sovs-composition-camera/id1326747827?mt=8
[3] https://www.camera51.com/camera51-app

In contrast, Cake Cam takes no position on aesthetics but enables tourists to capture pictures of themselves based on their preferences. Additionally, the markers used in Cake Cam are uniquely designed to handle 3D movement with no instructions.

2.2 Collaborative Photography

Existing research has explored various aspects of collaborative photography; however, prior work in this area does not address the problem of communicating a target composition to a stranger who then takes the picture. Kim et al. presented LetsPic [Kim, Kang, and LeeKim et al.2017], a group-based photoware, as a way to support group awareness for in-situation collaborative photography over a large physical space. Other research has explored specifically tourist-centered photo collaboration. Recently, Jarusriboonchai et al. [Jarusriboonchai, Olsson, Lyckvi, and VäänänenJarusriboonchai et al.2016] studied asymmetry in interaction capabilities; One person controlled the camera trigger, and the other controlled the viewfinder. ImageSpace [Lucero, Boberg, and UusitaloLucero et al.2009] allowed users to contextualize their photos spatially and share with others. The application introduced the idea of Scenes to help guide users to capture photos that tell a narrative. Brown et al. [Brown, Chalmers, Bell, Hall, MacColl, and RudmanBrown et al.2005] studied photo collaboration between participants who were separated by a distance with a mobile app that allowed tourists to share pictures with their remote friends and family instantly. The instant photo sharing in the system allowed immediate remote feedback that led the friends and family to request particular photos. Another project, a mobile application called Yousies [Wen and ÜnlüerWen and Ünlüer2015], allowed for its user to get his or her picture taken by a stranger, another user, without needing to pass the device around.

2.3 Computer Vision

Cake Cam depends on a computer vision algorithm to continuously estimate the difference between the intended camera pose and the current pose. Within the field of computer vision, Davison et al. [Davison, Reid, Molton, and StasseDavison et al.2007] proposed MonoSLAM, one of the first real-time 3D monocular localization and mapping frameworks. Since then, many improvements have been contributed to various research groups. Specifically in the context of mobile devices, Li et al. [Li and VogelLi and Vogel2017] implemented a monocular visual inertial state estimation for robust camera localization on a smartphone for mobile augmented reality. Shelley [ShelleyShelley2014] implemented visual inertial odometry (VIO) on a smartphone. Recently, further research has been done to improve the accuracy of VIO ([Bloesch, Omari, Hutter, and SiegwartBloesch et al.2015], [Bloesch, Burri, Omari, Hutter, and SiegwartBloesch et al.2017]).

3 CAKE CAM DESIGN AND DEVELOPMENT

The mobile app Cake Cam was designed to mediate the interaction between the tourist and local in impromptu tourist photography. To use Cake Cam, the tourist follows these four steps which are also summarized in figure 3:

(1) The tourist takes and approves a photo of the carefully framed scene.
(2) The tourist then hands the phone to a local and asks the local to take his or her picture.
(3) The tourist moves into the frame.
(4) Augmented reality alignment markers guide the local to correctly frame the picture with the tourist in it, producing the intended photo.

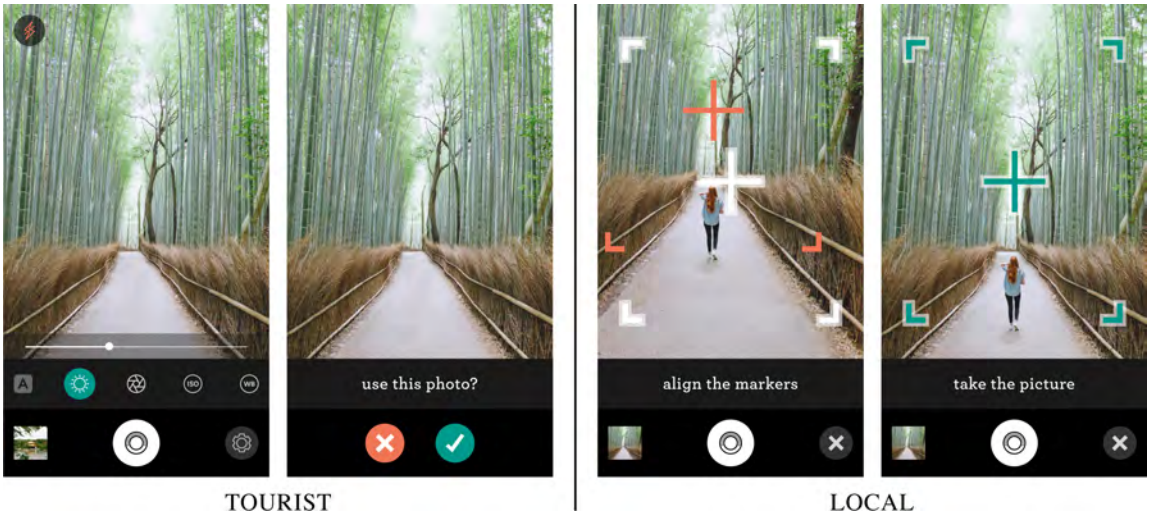

Fig. 3. The tourist captures the intended photo by taking the picture they want to be in (frames 1 and 2). The app communicates the intended framing by guiding the local to align the red and white markers (frame 3). When the markers are aligned, the local snaps a photo of the tourist (frame 4).

Step one of this process will be familiar to anyone who has taken a photo with a mobile camera. Step two changes nothing about how the tourist usually asks a local to take their photo. Step four introduces a new interaction using augmented reality and is the core functionality of the app. The 3D markers are used to communicate the tourist's intended photo composition to the local. The user experience and underlying computer vision algorithms used to align the markers are discussed in the next subsections.

3.1 User Experience Design

Cake Cam introduces a new interaction in mobile photography making the user experience a critical part of the app development. Because the problem is time-sensitive for the local, the interaction cannot involve additional verbal instructions but must allow the local to reproduce the intended photo quickly.

Prior work in overlay-based interface led us to test a design that merely overlaid the first image taken by the tourist and the current camera preview feed as shown in Figure 4. The idea was to align the two images and take the photo. In our evaluation of this process, we asked 10 participants to "align the two photos" and handed them the phone. This process confused participants. They did not understand why there were two images on the screen and struggled to align the photos correctly.

In their study of the same kind of overlay based interface for compositional guidance in the Social Camera, Bourke et al. [Bourke, McCarthy, and SmythBourke et al.2011] also found, "some interface challenges." On average, participants in this study rated the features ease of use at about 3 out of 5. This is consistent with our evaluation and speaks to the difficulty of aligning an image with the camera preview. The photo on the left of Figure 4 shows a more complicated image; this type of scenery would be challenging for a local to align as there are no distinct features to match. However, even when the picture is simpler, the image on the right, it still is not quickly obvious what direction the local needs to move the camera to align the two images.

Fig. 4. We tested overlaying the initial image and the current camera feed and asked the local to align the images. Consistent with [Bourke, McCarthy, and SmythBourke et al.2011], participants in our usability study found this interface difficult to use.

These results led us to create paper prototypes with transparent "screens" showing different arrangements of arrows, circles, and crosshairs placed on the app screen as shown in Figure 5, to guide the stranger. These markers are similar in design to markers used in [Li and VogelLi and Vogel2017], [McAdam, Pinkerton, and BrewsterMcAdam et al.2010], [Brewster and JohnstonBrewster and Johnston2008], [Xu, Ratcliff, Scovell, Speiginer, and AzumaXu et al.2015], [Carter, Adcock, Doherty, and BranhamCarter et al.2010], the commercial app Camera51, and panorama mode guidance. However, there are some key differences.

We found that a crosshair in the center with corner markers (second from the right in Figure 5) was the most intuitive setup. Users instinctively aligned the two sets of crosshair markers during testing. The first set of markers are 3D augmented reality markers that are fixed in the world while the other set of markers are placed on the image plane of the camera. Unlike other guidance methods, no instruction or explanation of the markers was given to the users before testing the app. Although the users in our testing instinctively aligned the markers, we later added the instruction "align the markers" at the bottom of the screen for further clarification.

Panorama guidance and aesthetic guidance use 2D markers on the image plane of the camera to guide users along the x and y axis. 2D markers on the image plane are unable to convey any direction along the z axis. This makes it difficult to guide the user to move the camera toward or away from the target camera pose.

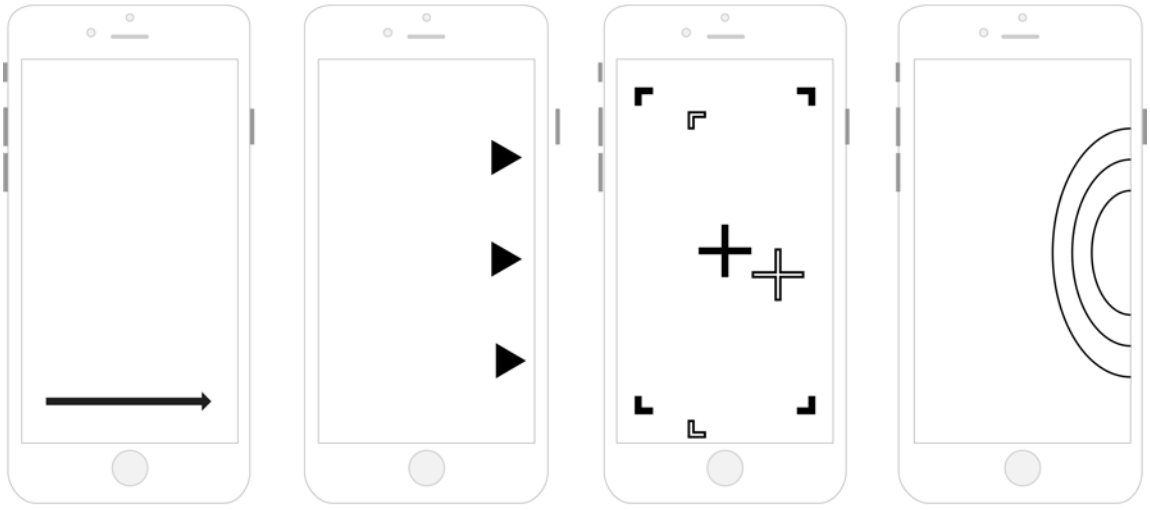

Fig. 5. We tested various markers to understand which ones best guided the local into taking the picture in the correct position.

The guidance used in work by Li et al. [Li and VogelLi and Vogel2017] used a combination of 2D markers on the image plane of the camera to guide the users into taking a better "selfie." Participants were given a tutorial before using the app to understand the meaning of the markers. These markers were limited to guiding participants to orbit the phone around themselves. This meant that one marker, an arrow, was used to guide both the orientation of the phone and the position along the x axis.

In contrast, we designed our markers to support motion in 3D dimensions by using both 3D AR markers and 2D markers on the image plane of the camera. Additionally, we exaggerated the depth of the 3D markers. This small change gave critical feedback for matching the camera orientation. Because of the exaggerated depth, users were able to quickly see when the phone was not correctly oriented and adjust accordingly. The effect is similar to looking down a long straight rod. If the rod is not aligned with the eye's gaze, the side of the rod is visible. However, when the rod is perfectly aligned, only the end is visible. The final 3D marker design is shown in Figure 6. We have not found another interface that accounts for changes in both orientation and camera position using 3D markers.

3.2 Visual Inertial Odometry

The key technical challenge in Cake Cam is detecting the camera pose. A camera pose is the combination of its position and orientation in world coordinates. The camera pose was obtained in real time using visual inertial odometry (VIO) [ShelleyShelley2014]. This algorithm uses a monocular camera coupled with linear acceleration and angular velocity from an inertial measurement unit (IMU) to create a robust estimate of the camera's pose.

When the user finishes framing his or her picture, the VIO algorithm is initialized, and the current camera position and orientation are marked as the initial pose. When the local is handed the smartphone, the error between the current pose of the phone and the initial pose increases. Visible alignment markers guide the local to minimize pose error and correctly frame the picture. Because this interaction must be efficient enough to be responsive, we use optimized VIO routines provided by Apple's iOS AR Kit[4].

[4]https://developer.apple.com/arkit

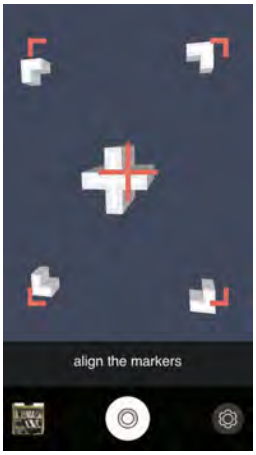

Fig. 6. Final 3D marker design. The white markers move in 3D and appear to be anchored to the real world in the camera preview. The local photographer aligns the white markers with the red markers.

While VIO can handle people moving into and out of the frame as is common at tourist locations, the algorithm performs poorly in low light, against moving backgrounds, or near blank walls, because the algorithm cannot find and match feature points. Furthermore, the algorithm fails to account for sudden large movements.

3.3 Distance Calculator

When the local photographer moves the phone to the correct position, the alignment markers turn green indicating the photo should be taken. Cake Cam uses a combination of the difference in Euler angles and the difference in the x, y, z axes to calculate the error between the initial camera pose and the current camera pose. The margin of acceptable error varies in the x, y, and z axes and the roll ϕ, pitch θ, and yaw ψ as each movement has a different effect on the framing of the photo.

4 METHODS

We designed a study to evaluate the accuracy, efficiency, and usability of Cake Cam compared to a standard phone camera from the local's perspective in the context of tourist photography. Ethical and privacy issues related to the study design were reviewed and approved in advance by our institutional review board (IRB). After completing the study, participants were given a candy bar as compensation for their time.

In this study, members of the research team stood by a bell tower on a university campus and asked 40 passersby to take the researcher's photo in front of the tower. We specifically chose the campus bell tower because students pass by it in a hurry. Of the participants, 22 were female, and 18 were male. The average age was 21 years old. The sampling method was a convenience sample where the research team randomly stopped people who were walking near the tower. We invited anyone who passed by to participate. Only one person was invited to participate at a time to avoid biasing future participants.

20 participants used the Cake Cam app, and 20 participants used a camera app that was designed to be similar to the standard iOS camera app, titled the "normal camera" app. The normal camera app uses VIO to track camera pose error exactly like Cake Cam, but the 3D alignment markers are not visible. We used a between-subjects design to prevent bias

caused by learning effects. The participant group size was determined using a power analysis with a critical difference of 70% between mean scores [Sauro and LewisSauro and Lewis2012].

The participants took on the role of the local, and a member of the research team acted as the tourist. We chose to pose as the tourist so that we could replicate the same experience for each participant while understanding the interaction from the "local's" perspective. Testing in a controlled environment, like a laboratory, would not involve people passing by a landmark on the way to somewhere else. This sample includes people who may be likely to stop and take someone's photo in a tourist setting as well.

To prevent the participants from viewing the correct framing, we first asked the participant to turn around while the researcher was taking the initial framing photograph for both Cake Cam and the normal camera app. In a typical tourist/local interaction the tourist does not first take a framing photograph and we did not want the participant to be biased by watching this interaction. The participant was then asked to take the researcher's photo in front of the bell tower.

After the "local" participant handed the phone back to the researcher, the researcher checked the pose error. If the error was too large relative to predefined thresholds, a translational error less than 26 cm and a rotational error less than 0.02, then the participant was given one instruction from a list of instructions (such as such as "place the bell tower on the right of the photo") and asked to take another picture. This process repeated until the translational and rotational error was within the specified range. We chose these thresholds as it was the maximum error that did not cut off specific elements—the bell tower, walkway, and trees—in the photo. We chose this to replicate a common issue in tourist photography where the local cuts off part of the landmark.

After taking the first photo, the participant was asked question 1 from the following list. After taking the last photo, the participant was asked questions 2–3, and the study concluded. Participants required less than 1 minute to take each photo. Including recruitment and follow up questions, the study lasted less than 15 minutes.

- On a scale of 1 to 10, how confident are you that this is the photo I wanted?
- On a scale of 1 to 10, how difficult was it to get the photo I wanted?
- Please describe your experience.

5 ANALYSIS METHODS

The study generated data in the form of participant responses and pose errors. We next describe the analysis of each kind of data.

5.1 Analysis of Participant Responses

We chose to use an unlabeled 10-point scale in questions 1 and 2 rather than 5-point Likert items because, in preliminary trials, we found that people compressed their rating to the top end of a 1 to 5 scale. having more values between 5 and 10 would help us better differentiate the participants' experiences. This is consistent with the distribution of scores given by users to software systems on the Software Usability Scale (SUS) test, which also skews high [SauroSauro2011].

We analyzed responses to question 3 using thematic [Lazar, Feng, and HochheiserLazar et al.2010, p. 208][Warren and KarnerWarren and Karner2010, p. 218] and critical incident analysis [Preece, Rogers, and SharpPreece et al.2015, p. 298]. Thematic analysis followed a three-step process based on Warren and Karner [Warren and KarnerWarren and Karner2010, chapter 9]. The first step is to repeatedly read, think about, and discuss participant responses to

identify emergent themes. The second step is to identify a small set of central themes. The third step is to code, or mark, occurrences of the central themes in the data.

Critical incident analysis identifies important or interesting responses from a single participant [Preece, Rogers, and SharpPreece et al.2015]. This can provide insights about where to "dig deeper" [Lazar, Feng, and HochheiserLazar et al.2010, p. 211] in the study of participant responses.

We used these methods to analyze the responses to question 3 by having two members of the study team read all participant responses three times and then meet several times to discuss themes that emerged from the comments. We later identified a small set of central themes, listed below, and marked occurrences of the themes in the data. Several critical incidents were identified during and after data collection and are discussed in section 6.2.

5.2 Analysis of Pose Error

The app stored the 3D pose of the camera for each photograph. This data was then used to calculate the relative pose error between the target photo and the second captured photo.

5.2.1 Translational Error. The initial pose $\mathbf{x_i}$ is captured when the reference photo is taken. Upon taking the final picture, the final pose $\mathbf{x_f}$ was used to calculate the Euclidean translational error.

5.2.2 Rotational Error. Rotational space is not well defined in Euclidean space. Instead, we compute rotation matrices from the captured Euler angles $\mathbf{\Theta_i}$ and $\mathbf{\Theta_f}$. We use the 3-2-1 Euler angle sequence [Bloesch, Omari, Hutter, and SiegwartBloesch et al.2015], [ShelleyShelley2014]. Orientation errors are naturally expressed as a 3×3 matrix which describes the differences between the actual and intended camera pose rotations. However, to simplify the comparison of rotational errors, we reduce the rotational error to a single scalar value using the method in [Lee, Leok, and McclamrochLee et al.2010].

$$R_e = \frac{1}{2} tr\left(I - R_i R_f^T\right), \tag{1}$$

Where the matrix trace $tr(\cdot)$ is defined as the sum of the main matrix diagonal. The values that this error metric can have varies between 0 and 2. A rotational error $R_e = 0$ represents no rotational error. An error $R_e = 2$ represents a 180° error about the principal rotation axis—the axis about which all rotation occurs.

6 RESULTS

We found that with Cake Cam, the intended photo was matched in fewer tries and the final camera pose was closer to the intended pose than with the normal camera app. We also found that participants were more confident that they had taken the intended photo and found the experience less difficult.

6.1 Matching Camera Pose

All participants that used Cake Cam took 1 photo to capture the intended picture, an average of 3.1 fewer photos than normal camera participants ($p < 0.0001$). See Figure 7. [5]. It is unsurprising that, if left unprompted, participants will not match a target camera pose. What is surprising is that Cake Cam users matched the target pose on the first try, every time with no additional verbal instructions. Comparatively normal camera users took and average of 4.1 photos. Figure 7 shows the exact number of photos taken per normal camera participant.

[5]All p-values generated by unpaired, two-tailed t-tests on data generated by 40 participants with 20 using Cake Cam and 20 using the normal camera

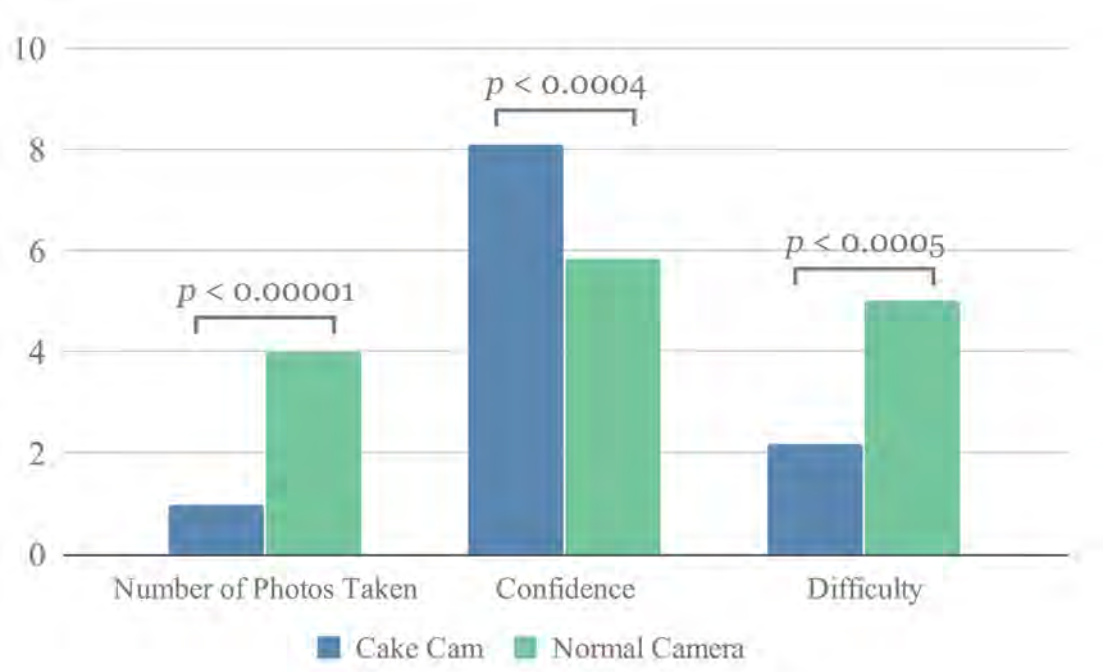

Fig. 7. Compared to normal camera users, participants who used Cake Cam required fewer photos to match the target pose. After taking one photo, Cake Cam users were more confident that they had captured the intended imaged and reported that the experience was less difficult.

After taking the first photo, participants using Cake Cam rated their confidence in having taken the correct photo 2.2 points higher ($p < 0.0004$) than participants who used the normal camera app as shown in Figure 7. Additionally, participants who used the normal camera rated the difficulty 2.8 points higher ($p < 0.00005$) than participants who used the normal camera as shown in Figure 8.

After the first photo was taken, the difference between the pose error of Cake Cam and the normal camera was quite large. Cake Cam had a translational error of $9.2cm$ compared to $115cm$ for the normal camera as shown on the left side of Figure 9. The rotational error for Cake Cam was 0.005 compared to 0.020 as shown on the left side of Figure 10. The difference between the pose error of Cake Cam and the pose error of the normal camera is statistically significant. The right sides of Figures 9 and Figure 10 show the translational and rotational error for the final image (which was also the first, and only image for Cake Cam users).

Figure 11 shows the relative pose error between the reference image and the first photograph taken by each participant. The orange plane represents a pose with no relative error all other planes represent one participant. The top graph shows the relative poses for Cake Cam. The bottom graph shows the relative poses for the normal camera app. These graphs highlight the extreme difference between the pose error of the two groups.

After receiving verbal guidance and taking on average another 3.1 photographs, the final pose error of the normal camera decreased to a translational error of $17cm$ and a rotational error of 0.0039. Even with verbal guidance and multiple attempts, normal camera users produced larger translational errors ($p < 0.0008$) than Cake Cam users. However,

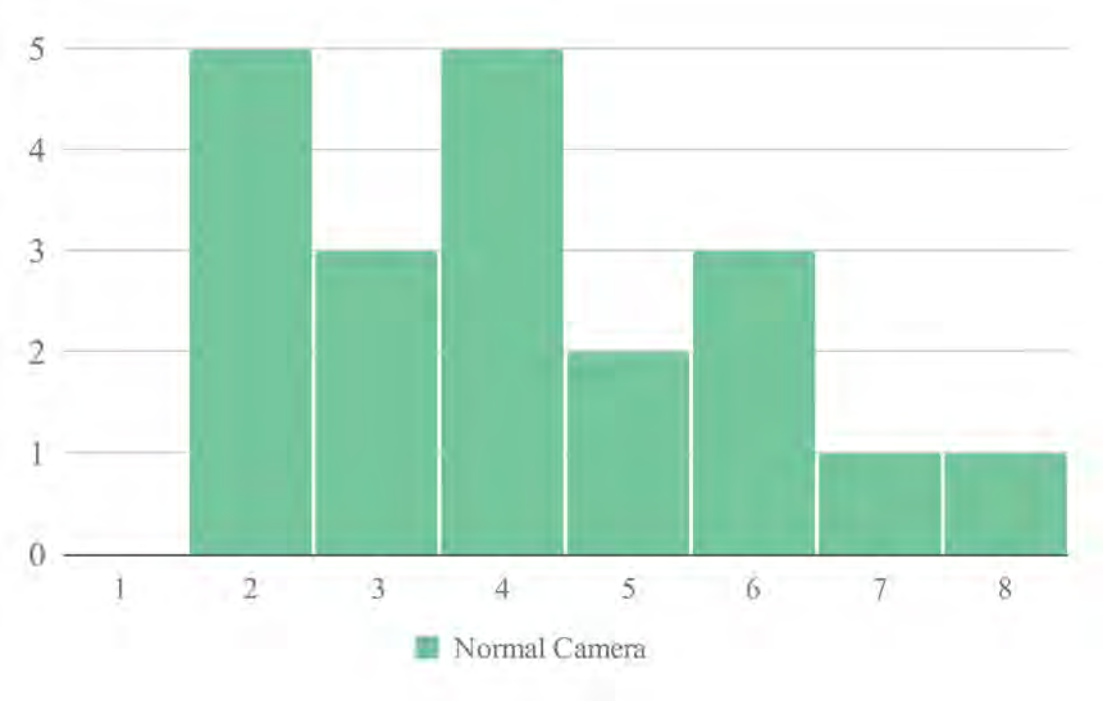

Fig. 8. Number of photos taken by each normal camera users.

for the last photograph, there was no statistically significant difference between the rotational pose error between Cake Cam and the normal camera app. This is likely because the camera pose in the study uses a nearly horizontal camera viewing angle, which is common in hand-held photography. Additionally we found that participants first aligned the position. Only Once they had aligned the camera to the correct position, did they adjusted the orientation.

6.2 Themes in Participant Responses

We identified four primary themes in participant responses to an open-ended question about their experience: worry, communication, resignation, and ease of use. We discuss each of them in the following subsections.

6.2.1 Worry. Many participants expressed a sense of worry, or lack of worry, about taking the photo the tourist wanted. Participants who used Cake Cam gave feedback such as, "It was awesome! I wasn't stressed about getting what you wanted in the frame" (P11). In contrast, participants who used the normal camera app expressed a more negative experience: "It did get a little frustrating, as I knew you had something in mind but I wasn't getting it" (P38). Many of the participants commented that they usually never knew how to get the photo the other person wanted and that this app took away all such worries. This theme aligns with the higher confidence rating observed from participants who used Cake Cam.

6.2.2 Communication. Participants who used Cake Cam made comments about how nice it was that the app communicated the exact photo desired. As one participant said, "It was cool how you [the app] were able to tell me exactly how you wanted the photo. You're never sure what angle they are looking for" (P25). Comparatively, participants who used

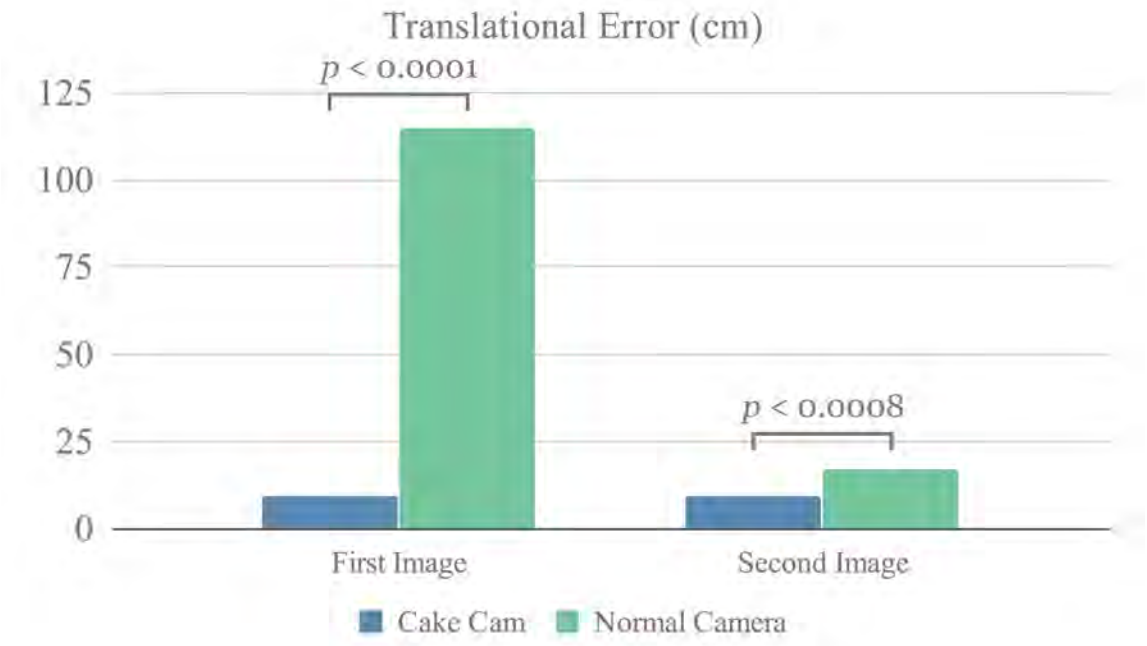

Fig. 9. Translational Error: Cake Cam users had smaller translation errors than normal camera uses on both the first and the last picture taken.

the normal camera app often indicated that more direction would have been helpful. Participants made such comments as, "If you want it a specific way I need more direction since I can't read your mind" (P36). One participant, however, noted that a general lack of communication is typical when someone has his or her photo taken by a local when he said, "This is something that happens all the time. It's hard to get exactly what they want if they don't communicate it" (P29).

Overall we believe that participants wanted more direction about the intended photo; however, we found that many participants were unable to understand the direction that was given quickly. After being given an instruction to correct the photo, the participant would correct this aspect of the photo but simultaneously introduce a new mistake. This struggle can be seen in figure 12.

6.2.3 Resignation. Another theme we identified in our results was a sense of resignation about getting a "tourist"-quality photo. As one participant commented, "I think when you ask someone to take your photo, you have to have 'tourist photo' expectations" (P41). Such a statement implies that, as another participant said, "when you ask someone [to take your photo] you have to assume it's not going to be what you want" (P3).

We found one possible explanation for this resignation to be caused by the social anxiety of asking a local to take multiple photos, as was done in this study. This breaks a social norm and is not something people are comfortable doing. The introduction of social awkwardness is where we started to sense a sort of paradox when it comes to having a photo taken at a tourist location. While the participants acting in the role of the local were actively worried about

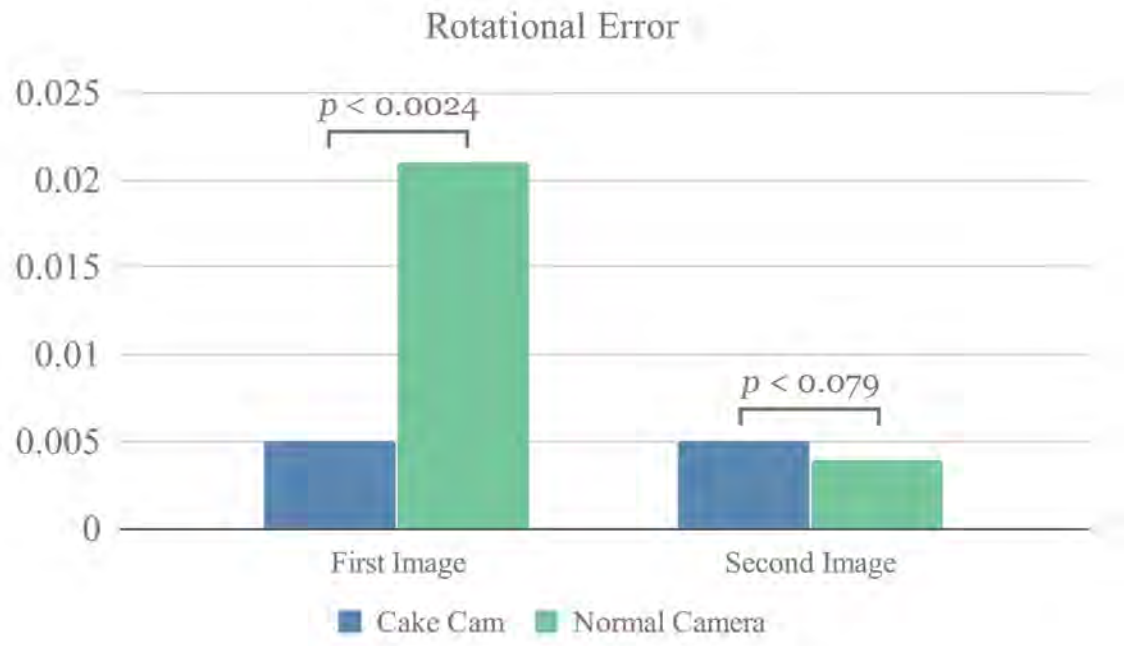

Fig. 10. Rotational Error: Cake Cam users had a smaller translational error than normal camera users on the first picture taken. However, there was no statistically significant difference for the last picture taken.

failing to meet the tourist's expectations and indicated a desire for more communication, they did not want to spend the time needed to receive this communication and potentially take multiple photos.

6.2.4 Ease. Many participants who used Cake Cam commented on how easy it was to get the intended photo. Cake Cam participants made comments such as, "It made it a lot easier: all I had to do was line it up. Knowing what they wanted was way easier!" (P 17). Other participants made very similar general comments about the usability of the app, "It was super easy! As soon as I looked at the screen it was very self-explanatory" (P 9). While some of these comments may be due to a novelty effect, the experiment supports our claim that the Cake Cam interface is more effective, efficient, and natural than the current verbal method for a standard camera app.

7 DISCUSSION

Cake Cam allowed impromptu local photographers to capture the intended photograph with less effort, greater accuracy, and more confidence than the usual process based on verbal descriptions. We hypothesize that there are two factors which contribute to this.

First, AR markers in 3D space provide continuous feedback to the local who is attempting to replicate the photo. This constant feedback contrasts with the limited verbal descriptions given before and after each picture is taken, which method leaves greater room for error. Furthermore, we learned from our thematic analysis that participants find it to be socially awkward to ask or be asked to take multiple photos or to involve too much instruction; however, they

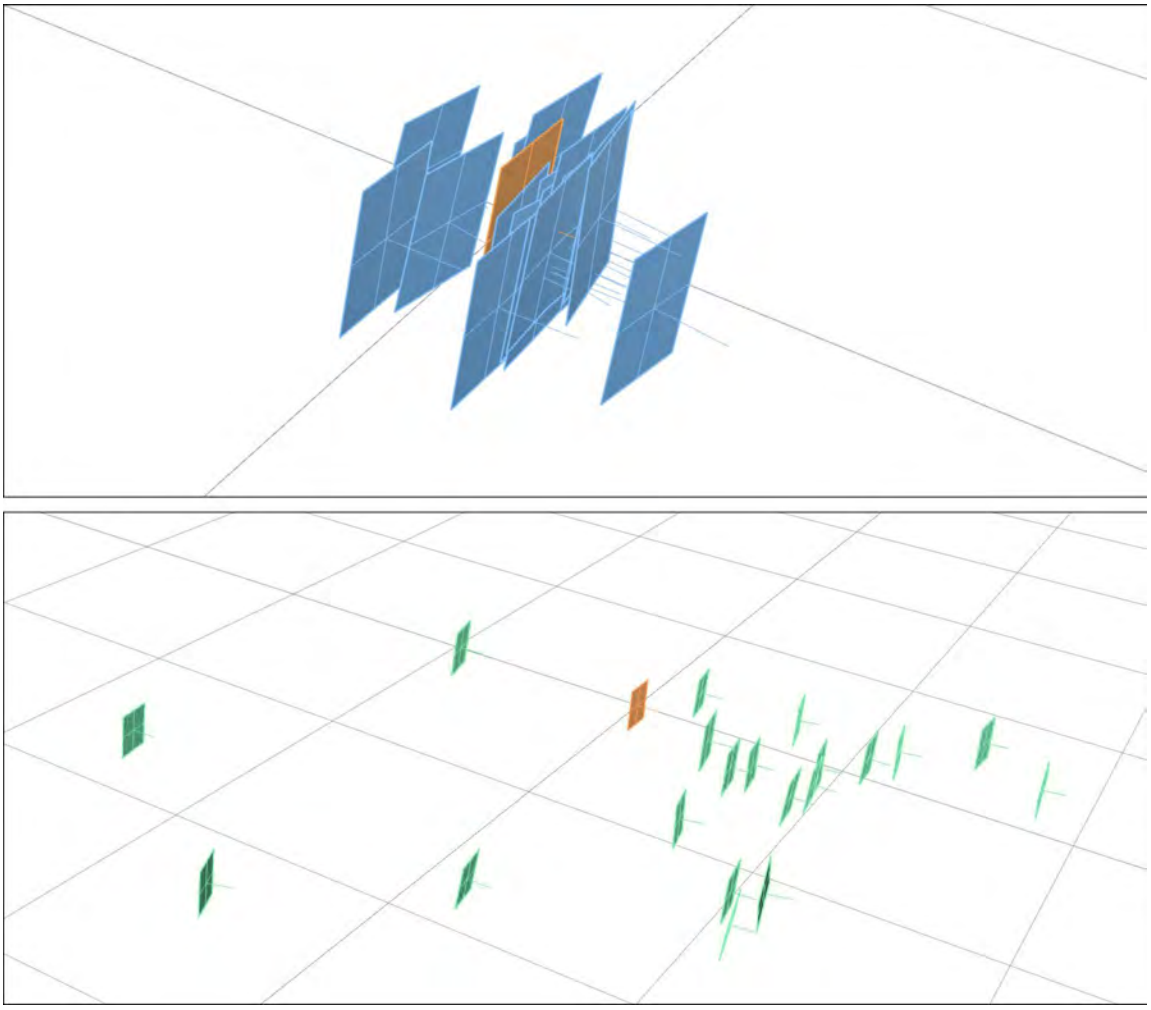

Fig. 11. Each coordinate frame represents the relative pose error between the reference image and the final image for one participant. The orange plane represents a pose with no relative error. The top graph shows participant who used Cake Cam. The bottom graph shows participant who used the normal camera app.

conversely indicated a desire for more guidance on how to take the photograph. These two contradicting needs can be brought into alignment with continuous feedback. The local receive the information they want—what photograph to take—without having to take the time to verify the photograph is correct and possibly retake it.

The second factor is that a low-order representation of the goal is better than a highly detailed representation in this kind of collaborative problem. AR markers provide an abstract idea of the intended photo, as opposed to showing the desired picture and asking the user to replicate it without further instruction. As detailed before, a superimposed copy of the desired image over the screen was confusing to the user and took longer to achieve the desired photo compared with using 3D AR markers. This lag is likely caused by the requirement that the user must process the highly detailed visual instruction, requiring them to reverse engineer the camera's pose from the image before taking one themselves.

Fig. 12. This participant took four photos before capturing the intended photo. After being given told to the photo with the bell tower on the right side of the photo, the participant corrected this aspect of the photo but simultaneously introduce a new mistake, moving in closer.

By communicating abstract instructions and salient features, the user does not need to process so much information. The AR markers, therefore, communicate what needs to done to achieve the intended result—where to position the phone—instead of communicating the exact intended results, the desired photograph. We hypothesize that abstract representation of intent is more effective in this setting for communicating intent than the more detailed alternative.

Future work might consider tools for supporting similar kinds of asymmetric collocated collaboration tasks. We only considered the specific problem of asking a stranger to take one's picture in front of a tourist landmark. A unique asymmetry of this problem is that the tourist cares more about the outcome of the collaboration than the local. Similar scenarios involving asking a stranger for directions or negotiating with an airline gate agent may involve a similar asymmetry in concern for the outcome of the collaboration.

7.1 Limitations

A limitation of our study is that a member of the research team acted as the tourist, which may have changed the dynamics of the interaction between "tourist" and local. We chose to pose as a tourist, rather than recruiting actual tourists, to make the study more feasible and to eliminate variation in how the tourist framed the picture and spoke to the local. We designed the study with more local authenticity by recruiting actual strangers to take photos. We wanted

to understand the local's perspective on photo collaborations. To do that, we needed to mimic an impromptu photo request in a repeatable manner.

Another weakness of our study is that we learned little about the difficulty of generating the target pose. While this is an important question, we assume that tourists will invest in learning if a better photo is guaranteed. This implementation assumes the tourist is visualizing where they want to stand in the frame as they are composing the initial framing shot.

8 CONCLUSION

We compared the novel Cake Cam interface to an interface with no additional guidance. We gave no guidance on composition to the normal camera users in order measure a baseline pose error to which to compare pose error generated by Cake Cam users. We chose "no guidance" for the normal camera users because this is a common practice for impromptu tourist photography. In retrospect, it may seem obvious that guidance given in an AR interface is better than no guidance at all. However, we also point out that some interfaces for guided photography—such as the image overlay in Figure 4 and [Bourke, McCarthy, and SmythBourke et al.2011]—may actually be more difficult to use than no guidance at all. This was not the case for 3D AR markers in Cake Cam which were easier to use than the normal camera interface, as reported by study participants.

REFERENCES

[Bloesch, Burri, Omari, Hutter, and SiegwartBloesch et al.2017] Michael Bloesch, Michael Burri, Sammy Omari, Marco Hutter, and Roland Siegwart. 2017. Iterated extended Kalman filter based visual-inertial odometry using direct photometric feedback. *The International Journal of Robotics Research* 36, 10 (2017), 1053–1072. https://doi.org/10.1177/0278364917728574 arXiv:https://doi.org/10.1177/0278364917728574

[Bloesch, Omari, Hutter, and SiegwartBloesch et al.2015] Michael Bloesch, Sammy Omari, Marco Hutter, and Roland Siegwart. 2015. Robust visual inertial odometry using a direct EKF-based approach. *IEEE International Conference on Intelligent Robots and Systems* 2015-Decem (2015), 298–304. https://doi.org/10.1109/IROS.2015.7353389

[Bourke, McCarthy, and SmythBourke et al.2011] Steven Bourke, Kevin McCarthy, and Barry Smyth. 2011. The Social Camera: A Case-study in Contextual Image Recommendation. In *Proceedings of the 16th International Conference on Intelligent User Interfaces (IUI '11)*. ACM, New York, NY, USA, 13–22. https://doi.org/10.1145/1943403.1943408

[Brewster and JohnstonBrewster and Johnston2008] Stephen A. Brewster and Jody Johnston. 2008. Multimodal Interfaces for Camera Phones. In *Proceedings of the 10th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '08)*. ACM, New York, NY, USA, 387–390. https://doi.org/10.1145/1409240.1409295

[Brown, Chalmers, Bell, Hall, MacColl, and RudmanBrown et al.2005] Barry Brown, Matthew Chalmers, Marek Bell, Malcolm Hall, Ian MacColl, and Paul Rudman. 2005. Sharing the Square: Collaborative Leisure in the City Streets. In *Proceedings of the Ninth Conference on European Conference on Computer Supported Cooperative Work (ECSCW'05)*. Springer-Verlag New York, Inc., New York, NY, USA, 427–447. http://dl.acm.org/citation.cfm?id=1242029.1242051

[Carter, Adcock, Doherty, and BranhamCarter et al.2010] Scott Carter, John Adcock, John Doherty, and Stacy Branham. 2010. NudgeCam: Toward Targeted, Higher Quality Media Capture. In *Proceedings of the 18th ACM International Conference on Multimedia (MM '10)*. ACM, New York, NY, USA, 615–618. https://doi.org/10.1145/1873951.1874034

[Davison, Reid, Molton, and StasseDavison et al.2007] Andrew Davison, Ian Reid, Nicholas Molton, and Olivier Stasse. 2007. MonoSLAM: real-time single camera SLAM. *Pattern Analysis and Machine Intelligence (PAMI), IEEE Transactions on* 29, 6 (2007), 1052–67. https://doi.org/10.1109/TPAMI.2007.1049

[Jarusriboonchai, Olsson, Lyckvi, and VäänänenJarusriboonchai et al.2016] Pradthana Jarusriboonchai, Thomas Olsson, Sus Lundgren Lyckvi, and Kaisa Väänänen. 2016. Let's Take Photos Together: Exploring Asymmetrical Interaction Abilities on Mobile Camera Phones. In *Proceedings of the 18th International Conference on Human-Computer Interaction with Mobile Devices and Services (MobileHCI '16)*. ACM, New York, NY, USA, 529–540. https://doi.org/10.1145/2935334.2935385

[Kim, Kang, and LeeKim et al.2017] Auk Kim, Sungjoon Kang, and Uichin Lee. 2017. LetsPic: Supporting In-situ Collaborative Photography over a Large Physical Space. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 4561–4573. https://doi.org/10.1145/3025453.3025693

[Lazar, Feng, and HochheiserLazar et al.2010] Jonathan Lazar, Jinjuan Heidi Feng, and Harry Hochheiser. 2010. *Research Methods in Human-Computer Interaction*. Wiley.

[Lee, Leok, and McclamrochLee et al.2010] Taeyoung Lee, Melvin Leok, and N. Harris Mcclamroch. 2010. Geometric Tracking Control of a Quadrotor UAV on SE(3). In *Conference on Decision and Control*. 5420–5425. https://doi.org/10.1002/asjc.567 arXiv:arXiv:1003.2005v1

[Li and VogelLi and Vogel2017] Qifan Li and Daniel Vogel. 2017. Guided Selfies Using Models of Portrait Aesthetics. In *Proceedings of the 2017 Conference on Designing Interactive Systems (DIS '17)*. ACM, New York, NY, USA, 179–190. https://doi.org/10.1145/3064663.3064700

[Lucero, Boberg, and UusitaloLucero et al.2009] Andrés Lucero, Marion Boberg, and Severi Uusitalo. 2009. Image Space: Capturing, Sharing and Contextualizing Personal Pictures in a Simple and Playful Way. In *Proceedings of the International Conference on Advances in Computer Enterntainment Technology (ACE '09)*. ACM, New York, NY, USA, 215–222. https://doi.org/10.1145/1690388.1690424

[McAdam, Pinkerton, and BrewsterMcAdam et al.2010] Christopher McAdam, Craig Pinkerton, and Stephen A. Brewster. 2010. Novel Interfaces for Digital Cameras and Camera Phones. In *Proceedings of the 12th International Conference on Human Computer Interaction with Mobile Devices and Services (MobileHCI '10)*. ACM, New York, NY, USA, 143–152. https://doi.org/10.1145/1851600.1851625

[Preece, Rogers, and SharpPreece et al.2015] J. Preece, Y. Rogers, and H. Sharp. 2015. *Interaction Design: Beyond Human-Computer Interaction*. Wiley.

[SauroSauro2011] Jeff Sauro. 2011. *A Practical Guide to the System Usability Scale: Background, Benchmarks and Best Practices*. CreateSpace Independent Publishing Platform.

[Sauro and LewisSauro and Lewis2012] Jeff Sauro and James R. Lewis. 2012. *Quantifying the User Experience: Practical Statistics for User Research* (1st ed.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[ShelleyShelley2014] Michael Andrew Shelley. 2014. Monocular Visual Inertial Odometry on a Mobile Device. *Thesis* (2014), 92.

[Warren and KarnerWarren and Karner2010] Carol A. B. Warren and Tracy Xavia Karner. 2010. *Discovering Qualitative Methods: Field Research, Interviews and Analysis*. Oxford University Press.

[Wen and ÜnlüerWen and Ünlüer2015] James Wen and Ayça Ünlüer. 2015. Redefining the Fundamentals of Photography with Cooperative Photography. In *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia (MUM '15)*. ACM, New York, NY, USA, 37–47. https://doi.org/10.1145/2836041.2836045

[Xu, Ratcliff, Scovell, Speiginer, and AzumaXu et al.2015] Yan Xu, Joshua Ratcliff, James Scovell, Gheric Speiginer, and Ronald Azuma. 2015. Real-time Guidance Camera Interface to Enhance Photo Aesthetic Quality. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1183–1186. https://doi.org/10.1145/2702123.2702418