Urban Pointing: Browsing Situated Media Using Accurate Pointing Interfaces

Tobias Langlotz
University of Otago
Dunedin, 9054
New Zealand
tobias.langlotz@otago.ac.nz

Elias Tappeiner
UMIT - Private University for Health Sciences, Medical Informatics and Technology
Hall in Tirol, 6060
Austria
elias.tappeiner@gmx.at

Stefanie Zollmann
Holger Regenbrecht
University of Otago
Dunedin, 9054
New Zealand
stefanie.zollmann@otago.ac.nz
holer.regenbrecht@otago.ac.nz

Jonathan Ventura
University of Colorado Colorado Springs
Colorado Springs, CO 80918
USA
jventura@uccs.edu

CHI 18 Extended Abstracts, April 21–26, 2018, Montreal, QC, Canada
ACM 978-1-4503-5621-3/18/04.
https://doi.org/10.1145/3170427.3188565

Abstract
Given the advancements in ubiquitous computing we can nowadays link information to places and objects anywhere on the globe. For years, map-based interfaces have been the primary interfaces to browse and retrieve this situated media. While more recently also other interface concepts for situated media have found their way out of the research labs, most notably Augmented Reality, this work looks into a concept that has widely been ignored: Accurate pointing in outdoor environments. This works presents Urban Pointer a phone-based pointing interface that utilizes computer vision to enable accurate pointing in urban environments together with some first insights on the implementation.

Author Keywords
Pointing; Augmented Reality; Situated Media; Location-based systems; Mobile Interfaces; Tricorder.

ACM Classi cation Keywords
H.5.2 [User Interfaces]: (Input devices and strategies); H.5.1 [Multimedia Information Systems]: (Artificial, augmented, and virtual realities)

Introduction
Interfaces like 2D maps allow the user to navigate through a static representation of the physical world and explore the digital content linked to places or objects. This spatially

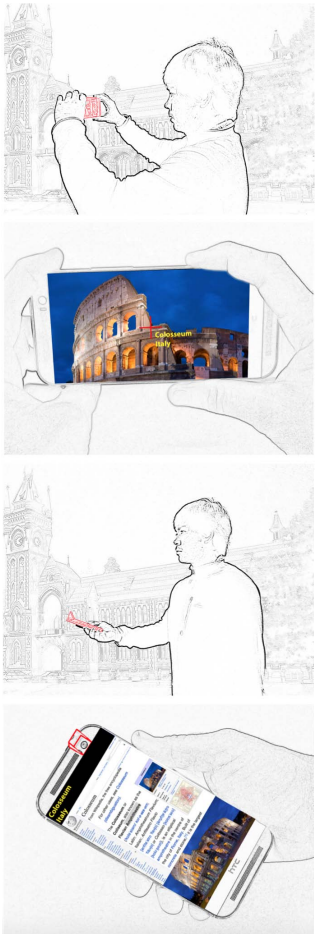

Figure 1: Browsing situated media. Top) AR Browser. The user browses situated media using the video feed of the mobile device. Bottom) Urban Pointing. By pointing towards objects of interest situated media will be displayed on the device.

anchored content is often referred to as situated media. Unfortunately, map-based interfaces have also a number of distinctive shortcomings for example by being usually constrained to a 2D or 2.5D birds-eye view and a minimum distance that prohibits annotation of smaller objects. Several researchers have proposed Augmented Reality (AR) interfaces [5] for browsing situated media and even showed how they could be integrated with map-based interfaces [6, 5]. These AR interfaces are often referred to as AR browsers, in particular if they are used to access situated media in outdoor environments. While there is a large number of works on AR browser, they all share the same idea of blending a graphical representation of the situated media with the physical environment.

Even though AR Browser are already a widely used method to access the rising amount of digital media placed in the physical world surrounding us, many challenges remain. For example, we know from the literature that AR browsers have problems in usability and social acceptance [3]. Part of it stems from the ergonomics that require the user of AR browser to "look through" the phone (see Figure 1, top). Similarly, Colley et al. observed advantages when the camera for AR browsers is angled and stated that "ergonomic considerations are dominant when users judge the overall experience of an AR browser" [1]. In their research they also had an AR browser which is more similar to a 'torchlight' as the camera was angled by 90° showing completely different ergonomics. However, their research was within a lab environment with perfect optical tracking completely ignoring real-world constraints. There are also other interfaces concepts than AR Browsers that can be used to access situated media in 3D. One of those is pointing. Though the concept of pointing is well known and studied (e.g. for virtual environments [2]), it is relatively under-researched for outdoor environments. The idea of urban pointing is in order to query digital information in the physical world, the mobile device aims at the target of interest. For example, to query a situated Wikipedia article of an interesting historical building, users would point their mobile devices towards the building (Figure 1, bottom). In that example, we would use it like a remote controller or the fictional "tricorder" (Star Trek).

While the general idea of pointing in outdoor environments was already conceptually introduced [7][8], these systems exclusively rely on device sensors, such as GPS or magnetic compass, to identify pointing actions. These sensors come with low accuracy when estimating positioning and orientation of the devices and are often affected by the environmental influences, such as shadowing of tall building structures. However, in particular for urban pointing a good estimate of positioning and orientation is crucial to make them usable and previous works have never explored and showed practical limitations. The lack of accurate pointing interfaces so far made it difficult to investigate urban pointing in more detail, even though the concept of urban pointing promises more natural ergonomics and an unobtrusive interface to browse situated media. To our best knowledge, no accurate urban pointing interfaces have been investigated so far.

Urban Pointer

In this work, we introduce the concept and first implementation of an accurate pointing interface for outdoor, urban environments. Similar to existing AR Browsers, we utilize the mobile phone and the integrated camera for browsing and interacting with the situated media. However, instead of browsing and interacting with situated media by "looking through" the phone we use the phone to point at objects or places in the environment (Figure 1, bottom). The anchored information is shown on the screen and pointing towards other objects and places will interactively show

Figure 2: 3D model used for localization. Top) Input images used for creating 3D model. Bottom) 3D point cloud used for localization.

Figure 3: Camera modification for urban pointing using a 90 degree prism on top of the front facing camera.

other information linked to these currently selected objects and places. One of the main reasons that prevented further work on outdoor pointing might be the challenges for accurate estimation of device position and orientation. In order to be able to point and access situated media, we need be able to precisely estimate the location and orientation of our pointing device. This problem is challenging and also often faced when implementing AR browsers. Most commercial AR browser rely entirely on their integrated hardware sensors, such as GPS, magnetometer, accelerometers, and gyroscopes.

However, these sensors have been proven to be relatively error prone. Standard GPS sensors for example, show an accuracy of less than eight meters within a confidence interval of 95% [11]. This is not accurate enough for the purpose of situated media browsing. The uncertainty radius of eight meters only allows to annotate digital content to objects with 16 meters in diameter (see Figure 5). Especially existing pointing prototypes [7, 8] suffer from poor location and orientation awareness, affecting the interface performance. However, researchers have shown that the combination of vision-based methods and sensors can yield accurate global localization and pose tracking [9, 5, 10].

Localization

To compensate the error-prone GPS localization when implementing our pointing interface, we adapted the vision-based approach earlier presented by Ventura et al. [9, 10]. Our localization framework works in four main steps: (1) the reconstruction, (2) the global registration, (3) the online localization and (4) the mobile tracking system. The framework is implemented as a client-server architecture. After an offline reconstruction and global registration step that builds a sparse geo-referenced 3D model from images and is stored on the server (Figure 2), the mobile client requests its position from the server by sending a camera image from the current environment. The server then estimates the global position of the image by matching it against the geo-referenced 3D reconstruction. We realize this by extracting and matching SIFT feature descriptors from the current camera image against the triangulated 3D points of the 3D reconstruction. With a parallel brute force search for each 2D SIFT descriptor, the closest reconstructed 3D point according to the Euclidean distance is found. By using the 2D-3D point correspondences the Location Determination Problem (LDP) given us a pose for the mobile client [9]. Although the global location estimation can be performed on the mobile client itself, it runs faster and scales better for large reconstructions on a remote server. After processing the received images the server responds with the global position estimate as latitude, longitude, altitude. For stability reasons, the position is filtered on the client by a Kalman filter with a dynamic, linear movement model. While similar approaches have been used for estimating positioning and orientation for AR interfaces [12], we face a major challenge when using a phone as a pointing interface. The issue is that the back facing camera is pointing downwards to the ground, and contrary the front camera towards the face. In order to solve this problem and to capture the environment as input for the vision-based localization, we mount a 90° prism on top of the front facing camera. This allows us to capture the environment the user is pointing at (Figure 3). This is of course only a prototypical implementation but there are already phones available which offer angled cameras with a view-direction as simulated in our prototype (e.g. Huawei ShotX).

Orientation Tracking

Beside an accurate localization, the orientation of the mobile device needs to be known to browse the digital information accurately placed in our physical environment. The

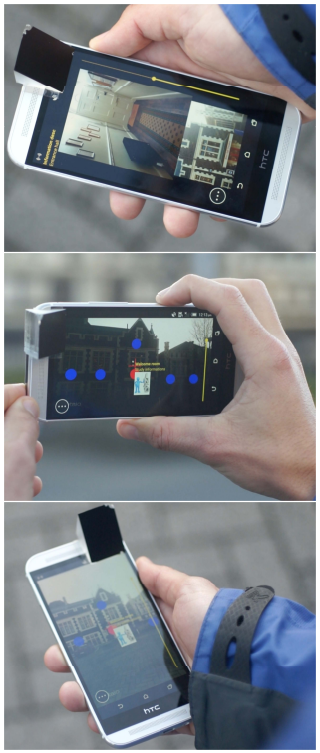

Figure 4: Prototypes. Top) Media Pointing. Middle) AR Browser. Bottom) AR Browser with off-axis camera.

main issues of the integrated sensors are electromagnetic interference (magnetometer) and drift (gyroscope). Contrary to the localization which can tolerate a small delay, the orientation estimation needs to be in real-time to build an interactive interface and thus runs on the locally on the mobile pointing device. To compensate for the inaccuracies of the integrated sensors, we implemented and extended the sensor-fusion concept introduced by Langlotz et al. [4]. The idea is to combine internal sensors with vision-based orientation tracker. As the situated media is placed in the real world, a north-centered orientation estimation of the pointing device is required. We stabilize the absolute but noisy sensor reading from the internal sensors which gice us north-centered orientation by using the accurate but relative vision-based orientation tracking. Contrary to the work of Langlotz et al., we use a full 6 Degrees of Freedom (6DoF) visual pose tracker instead of their panorama tracker. Our tracker automatically detects planes within the camera images using homographies and tracks these planes between images. So for tracking the relative orientation we use the second camera (facing downwards when used as pointing device) and track the ground plane which we can assume to be planar in urban environments with 6DoF. Not all phone support using two cameras simultaneously and we use a HTC M8 which does. Finally, we fuse the vision-based tracking with the sensor one via a Kalman Filter as described by Langlotz et al. [4] reducing the orientational errors.

Graphical Interface

The idea of urban pointing is to point with the tracked mobile device, similar as a remote controller, towards the physical object of interest (Figure 1, bottom). The device which is tracked with the approach outlined earlier is held one-handed at waist level. Similarly to street photography which is often shoot through waist-level viewfinders and more unobtrusive as the target does not feel it is directly aimed. Previous works in AR browser have already shown that users felt often "stupid" because of the ergonomics of AR browser forcing them to directly aim at their target because of the "see-through" ergonomics enforced by the AR interface [3]. We argue that our pointing interface has the potential to offer a more natural gesture of using a mobile device for accessing situated media compared to AR browser. Due to the concept of pointing instead of camera aiming, the camera image is not rendered onto the screen of the device. The unused screen space can now be fully occupied with the situated media content, allowing to display more content than on an AR browser. For instance, blending the whole content of a news article over the camera stream of the AR browser makes the concept of targeting a camera view to select the situated media impractical. By occupying too much of the screen space, there is not enough of the camera stream visible to accurately browse media. Therefore, most AR browsers require to click the label in the AR view switching to a non-AR view showing the actual information. In contrast, the different approach to object selection of our proposed the media pointer interface allows us to give more more screen space to the actual information of interest while the search is done by physically aiming at different objects (see figure 4, top, for a picture showing the final prototype). For comparison we also implemented two other interfaces. Firstly, a traditional AR interface which serves as reference to our outdoor pointing interface. Here the device is hold at head height, used as a video see through to explore the environment and the situated media is shown as small annotations augmented on top of the physical environment (see Figure 1). The final prototype can be seen in Figure 4, middle. Secondly, based on the research by Colley et al. [1] we implemented an AR browser with a camera to screen angle of 90°(see Figure Figure 4, bottom). Colley et al. have shown that changing

Seq. 1	Mean	STD	GPS
No filter	0.49	0.29	15.75
Filtered	0.27	0.14	
Seq. 2			
No filter	1.3	1.0	7.5
Filtered	0.8	0.46	
Seq. 3			
No filter	0.99	0.74	4.63
Filtered	0.54	0.33	

Table 1: Evaluation results of the localization using our vision-based approach and purely GPS-based resulting from three test sequences recorded in urban environments. The presented error is the mean distance of the vision-based approach to the ground truth, of each test sequence (surveying marks). The sensor values are the measured GPS offsets to the ground truth. All measurements are in meter.

the camera to screen angle changes how users interact with the device as they start to use it more as a 'torchlight' making it similar to our pointing interface. Overall, these three interfaces are all implemented using the same tracking relying on integrated and vision-based tracking and will be part of a future user study investigating these three interfaces in urban outdoor environments.

Preliminary Results and Discussion

While we are still refining the proposed interfaces we already conducted a technical evaluation which provides some valuable insights on the accuracy that can be expected for pointing interfaces in realistic environments. This is important as it will inform researchers on outdoor pointing interfaces what to expect when using current vision-based technologies for tracking in urban environments and consequently also on how small objects or places can be so that the linked information can be retrieved. In order to test the accuracy of the prototype, we chose three locations on our campus for which we also received multiple differential GPS positions as ground truth. The evaluation results of the different refinement methods (non filtered and filtered) for all three test sequences are summarized in Table 1. Overall, the measured mean error of the unfiltered localization in test sequence 3 is 0.99m (STD 0.74m). However, filtering the location with a sliding window approach over 10 data points results in 0.54m mean distance (STD 0.33m) which is drastically improving the results compared to GPS alone (4.63m error). This improvement is even more visible for the other locations as their GPS error is even higher while having similar sub-meter positioning accuracy when using our vision-based approach (see Table 1). For evaluating the orientation we measured the rotation around the Z-axis (north bearing of the mobile device) at several locations. In order to obtain a reliable ground truth, we mounted the pointing device onto a tripod capable of measuring angles while also having a precisely measured survey marks with known bearings for error with respect to absolute north. The evaluation shows a mean error of 4.16° (STD 3.22°) for the raw sensors north bearing and 4.14° (STD 2.98°) for the hybrid sensor. Even though the evaluation produces similar results for the raw sensor bearing and the hybrid tracker, a small correction of the bearing is recognizable. Our evaluation shows that using our sensor-fusion based approach for improving internal sensors we can achieved, in worst case, sub-meter position accuracy (0.68m) and an expected orientation error of 4.14°. These results have practical relevance as they also allows us to estimate the required size of objects serving as an anchor for situated media. Figure 5 demonstrates the relation of the individual technical limits to the distance and size of real world pointing targets. The green arrow is the ground truth orientation, while the red (naive fusion of integrated sensors) and blue (our sensors fusion combining integrated sensors with vision tracking) show the orientation error. The circle is presenting the localization uncertainty. By combining both average error estimates, localization and orientation error, we can calculate the size of objects we can statistically expect to be able to point and retrieve attached information for different distances. For example using our approach we are able to point and retrieve information for objects 20m away if they are 4.3m large. In contrast when using integrated sensors exclusively, target objects would be required to be 12.1m in size. Over large distances the positive effect of the improved localization wears off and the orientation error that improved less becomes more relevant. However, even in this case objects 200m away need to be approx. 30m in size (our approach) compared to approx. 38m (naive). Overall, this work presents a first implementation of an accurate pointing interface for urban outdoor environments. Different to previous works that only implemented pointing in indoor environments or by using error

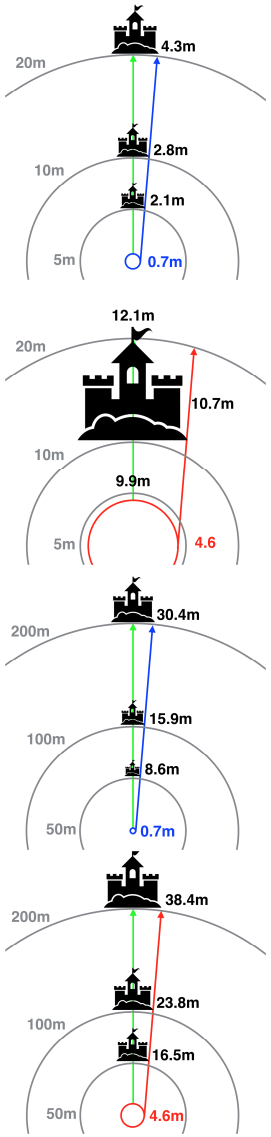

Figure 5: Required object sizes for different distances for a naive approach (red line) and our approach (blue line).

prone sensor-based tracking, we investigated the usage of sensor fusion to track position and orientation of the pointing device. A preliminary technical evaluation highlights the consequences for real world applications showing the relation between object size and distance when using pointing interfaces to retrieve situated media. Future works include a more extensive comparative user study investigating preference, usability, and workload for the implemented interfaces: Urban Pointing, AR browsers, and AR browser with changed camera to screen angle. Our initial hypothesis is that the improved ergonomics and social acceptance of the pointing interface might give it an advantage in user preference.

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1464420.

REFERENCES

1. Ashley Colley, Wouter Van Vlaenderen, Johannes Schöning, and Jonna Häkkilä. 2016. Changing the Camera-to-screen Angle to Improve AR Browser Usage. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services*. ACM, 442–452.
2. N. Dang. 2007. A Survey and Classification of 3D Pointing Techniques. In *International Conference on Research, Innovation and Vision*. 71–80.
3. Jens Grubert, Tobias Langlotz, and R Grasset. 2011. Augmented reality browser survey. *Technical Report, Graz University of Technology* (2011), 1–30.
4. Tobias Langlotz, Claus Degendorfer, Alessandro Mulloni, Gerhard Schall, Gerhard Reitmayr, and Dieter Schmalstieg. 2011. Robust detection and tracking of annotations for outdoor augmented reality browsing. *Computers and Graphics* 35, 4 (2011), 831 – 840.
5. T. Langlotz, T. Nguyen, D. Schmalstieg, and R. Grasset. 2014. Next-Generation Augmented Reality Browsers: Rich, Seamless, and Adaptive. *Proc. IEEE* 102, 2 (Feb 2014), 155–169.
6. Tobias Langlotz, Daniel Wagner, Alessandro Mulloni, and Dieter Schmalstieg. 2012. Online creation of panoramic augmented reality annotations on mobile phones. *Pervasive Computing* 11 (2012), 56 –63.
7. S. Robinson, P. Eslambolchilar, and M. Jones. 2009. Sweep-Shake: finding digital resources in physical environments. In *Proceedings of the International Conference on Human-Computer Interaction with Mobile Devices and Services*. 1–12.
8. R. Simon and P. Fröhlich. 2008. GeoPointing: evaluating the performance of an orientation aware location based service under real-world conditions. *Journal of Location Based Services* (2008), 24–40.
9. J. Ventura and T. Höllerer. 2012. Wide-area scene mapping for mobile visual tracking. *International Symposium on Mixed and Augmented Reality* (2012), 3–12.
10. Jonathan Ventura and Tobias Höllerer. 2015. 8 Urban Visual Modeling. *Fundamentals of Wearable Computers and Augmented Reality* (2015), 173.
11. P. A. Zandbergen and S. J. Barbeau. 2011. Positional Accuracy of Assisted GPS Data from High-Sensitivity GPS-enabled Mobile Phones. *Journal of Navigation* (2011), 381–399.
12. S. Zollmann, C. Hoppe, T. Langlotz, and G. Reitmayr. 2014. FlyAR: Augmented reality supported micro aerial vehicle navigation. *IEEE Transactions on Visualization and Computer Graphics* 20, 4 (2014).