Challenges of Large-Scale Augmented Reality on Smartphones

Clemens Arth* Dieter Schmalstieg†

Graz University of Technology, Austria

ABSTRACT

Smartphones have been identified as most promising future devices for an Augmented Reality (AR) mass market. However, their use puts considerable constraints on the design and composition of AR applications. The key problem is to find a registration mechanism for accurate six degree of freedom (6DOF) self-localization with respect to the environment. Approaches based on Computer Vision (CV) have been shown to be promising, but the feasibility of many CV methods on smartphones is questionable. In this paper we discuss current and future challenges faced in developing AR on smartphones, in particular for large and unconstrained outdoor environments. We focus on the registration task, giving a survey and an assessment of existing approaches from AR and CV. From this survey, we identify a set of important issues still seeking for practical solutions, both in terms of the fundamental registration problem and for making AR on smartphones a unique experience. As will become apparent, despite recent advances, we are still far from arriving at a universal solution to the problem.

Index Terms: I.2.10 [Artificial Intelligence]: Vision and Scene Understanding—3D/stereo scene analysis I.4.8 [Image Processing And Computer Vision]: Scene Analysis—Tracking; I.5.4 [Pattern Recognition]: Applications—Computer Vision C.5.3 [Computer System Implementation]: Microcomputers—Portable devices (e.g., laptops, personal digital assistants)

1 INTRODUCTION

Advanced mobile phones, now commonly called smartphones, are clearly the most promising platforms for an Augmented Reality (AR) mass market. Due to continuous miniaturization, today's mobile phones are surprisingly powerful. However, the relative performance gap to desktop computers is not closing. Consequently it is difficult to deploy modern compute and memory intensive algorithms designed for desktop computers on smartphones.

For AR, solving the registration task is the primary key exercise. Precision requirements are stringent: since AR applications fuse virtual and real information in 6DOF and in real time, even slight inaccuracies in registration can cause intolerable distortions in the combined view. Consequently, there are two hard problems for mobile AR: registration must be both very accurate and efficient in terms of computation and memory usage.

We consider this problem the ultimate challenge put before the deployment of AR for the masses. We claim that despite contrary appearance, most currently known solutions to the registration task are not truly suitable for use on mobile phones. Thus, all AR researchers are prompted to develop solutions specific to the problem of wide-area application of AR on smartphones.

The aim of this paper is to analyze the current state of research in AR on smartphones, in particular concerning Computer Vision (CV) based localization and registration. In this respect we go beyond previous considerations (see *e.g.* [7]). In Section 2, we briefly discuss the current state of the art in smartphone technology. In Section 3, we outline the task of localization and reflect on previously proposed solutions. In Section 4, we summarize challenges currently faced in CV based localization in general. In Section 5, we discuss specific challenges of deploying localization for AR. Section 6 draws some conclusions.

*e-mail: arth@icg.tugraz.at

†e-mail: schmalstieg@icg.tugraz.at

2 SMARTPHONES AS A PLATFORMS FOR AR

Wagner and Schmalstieg [26] were the first to identify the potential of handheld computers for AR in 2003. As the comparison of 2003 and 2011 given in Table 1 indicates, smartphones have improved in every aspect relevant for AR: faster processors, more memory, improved input interfaces, larger and superior displays, more sensors and improved network capabilities.

With the advent of *iPhone* and *Android OS*, the term *smartphone* was coined to indicate the capability of these devices for advanced computer applications. Due to improved touchscreen interfaces in particular and the easy availability of smartphone software ("apps") through online stores, smartphones have had an overwhelming commercial success. The strong commercial interest has also brought along improved software development tools targeting smartphones.

All in all, the smartphone ecosystem provides all ingredients to deploy AR as a software-only solution to a mass audience. However, one should not overlook that despite all technical and logistic improvements, there are still major obstacles for a large scale deployment of AR applications:

Camera quality and handling. Imaging capabilities of camera sensors typically deployed in smartphones are poor under bad lighting conditions. Images are blurry and colors start to suffer from significant aberration. Low-level access to the camera sensor hardware is usually prohibited. APIs only provide a high level of access to the camera sensor, rendering control of exposure, aperture, or focal length impossible. Small CCD sensors are the cause of increased amounts of noise in the camera feed, hurting the performance of subsequent CV algorithms significantly. Quality lost during image acquisition can hardly be compensated by further processing steps.

Energy consumption. Battery power has not increased significantly in recent years. Camera sensors require a lot of energy when running constantly at high frame rates, which is mostly owed to their intended use for still photography rather than video recording. Moreover, CV algorithms are computationally demanding and tend to drain the battery of smartphones quickly. Likewise, sensors and network interfaces are heavy energy consumers. Running fully featured AR applications causes batteries to rapidly discharge. Consequently, AR applications must be designed to be used for short time durations only, rather than as "always-on" features.

Network dependency. Accessing large amounts of data remotely suffers from several issues. First, network latency can harm the instant behavior of AR applications, causing displeasing lags. Second, accessing remote data is only possible with data plans that may be expensive or unavailable. Last, network coverage may be insufficient in certain areas. This leaves fully autonomous AR applications as the only viable option, implying heavy use of on-device storage capacities.

	2003 [26]	**2011**
CPU	400 MHz Intel xScale	ARM Cortex A8/A9 800 MHz - 1GHz single-core 1.0 - 1.2 GHz dual-core
RAM	64 MB	512 MB - 1 GB
Hardware GFX Support	none	OpenGL ES 1.1/2.0
Camera	single 320x240 color camera attached via CompactFlash jacket	front and back cameras, 2 - 8 MegaPixel color
Display	3.8", 16-bit 240x320	4.3", 32-bit 800x480
Interface	Stylus, Buttons	Touchscreen, Keyboard
Sensors	Fingerprint	GPS, compass, RFID accelerometer, gyroscope, proximity sensor, light sensor
Network	WiFi Bluetooth GPRS	WiFi Bluetooth 3G, GPRS, EDGE
Battery	900 - 1200 mAh	1400 - 1700 mAh
OS	Windows Pocket PC Windows Mobile	Android iOS
Price	$\sim$ 900 €	400 - 700 €

Table 1: Comparison of mobile phone hardware for AR, back in 2003 and today in 2011 (based on typical specifications for smartphones, such as Samsung Galaxy S2, HTC Sensation or Apple iPhone 4).

Visualization and interaction possibilities. The form factor of smartphones plays a major role in purchases. In fact, the maximum acceptable device size puts severe constraints on how large the display can be. Similar considerations restrict applicable interaction techniques. Multi-touch interfaces are probably the most evolved interaction mechanisms but their usability for certain tasks, such as pixel-accurate selection, is poor.

In theory, the relevant aspects for improving future smartphone hardware for AR are clearly known. In practice, however, developers of AR applications are at the mercy of hardware vendors and service providers, who make hardware development decisions based on market predictions that may not include the needs of AR. However, hardware development is generally moving in the right direction, partially driven by new application sectors such as mobile games or mobile navigation system that share many technical requirements with AR. Moreover, researchers are aware of the current limitations concerning camera control, which leads to improved camera APIs, such as the *Frankencamera* work for example [1].

Tablets as alternative mobile platform While tablets are also emerging as popular mobile platforms, we consider these devices to be oversized smartphone platforms basically. The visualization and interaction constraints are slightly relaxed due to the increased form factor, but the size and weight of these devices at the same time limit their applicability in AR as their handling is more exhausting (*i.e*. putting and holding the device up for a longer period of time probably requiring both hands, in turn limiting the interaction possibilities). Apart from that, current tablets share the same issues as smartphones. Case by case, smartphones or tablets are more suitable for a given AR application.

3 State of the art in self-Localization

Self-Localization denotes the process of registering a device in 6DOF with respect to an environment. We must distinguish localization relative to a local reference frame from absolute orientation in global coordinates. In AR, solutions have been mostly proposed for small local scenarios, such as the work of Reitmayr and Drummond [20], *PTAM* by Klein and Murray [13, 14], or the work of Wagner *et al*. [25], amongst others. The localization task is usually split into two parts: an *"initialization"* phase and a subsequent *"incremental tracking"* phase[1]. It is commonly accepted that the initialization phase poses a trickier problem than the incremental tracking phase, and that initialization is also the computationally more demanding problem. For the rest of this section, we therefore only focus on solutions proposed for the *initialization* step.

Localization as an image recognition problem. Solving the localization task is often treated as an image recognition problem. These approaches solve the task using a database of images, natural features extracted from the images, and some kind of matching algorithm. To speed up the entire procedure, approximate search structures, like vocabulary trees, are commonly used.

A work in the area of indoor SLAM was proposed by Se *et al*. [23] for robots, with a final registration in 3DOF. Agrawal and Konolige describe a localization system based on stereo vision and GPS [2][2]. A system for indoor localization using probabilistic models can be found in the work of Li and Kosecka [16], while Zhang and Kosecka proposed a system to match the current image versus a database of GPS-tagged outdoor images [28]. A similar system was proposed by Schindler *et al*. dealing with an urban-scale database and a large visual vocabulary [22]. Knopp *et al*. use Google Streetview images for approximate place recognition [15], similar to Zamir and Shah [27]. A very special application was proposed by Hays and Efros matching the query image based on image statistics to recover likely positions of where images where acquired [8]. Later, Kalogerakis *et al*. enhanced the system with additional priors of likely traveler visits [11]. Zheng *et al*. build a database from community photo sites by finding clusters of images from commonly visited landmarks [29]. Baatz *et al*. use rectified images resembling facades of buildings [4]. Upon this system, Chen *et al*. build a landmark identification system, estimating an approximate GPS position [5]. Note that while some of these works use images acquired with phones, none of them deal with actually running CV software on smartphones.

The first localization system for mobile phones was proposed by Robertson and Cipolla [21]. Takacs *et al*. presented a system for approximate localization using natural features partitioned into smaller feature bags [24]. Hile *et al*. clustered images from similar locations and built local 3D reconstructions to register individual views. The actual camera position is estimated directly from the top image matches via triangulation [9]. Note that all these approaches rely on a client-server setup with the mobile phone as the client and the major processing load located on the server side.

It is also important to mention that *none* of the approaches mentioned so far actually calculate an *accurate absolute 6DOF pose*. The approaches of Hile *et al*. [9] and Baatz *et al*. [4] mention the theoretical capability of delivering approximate camera poses, however, no quantitative evaluation results are given.

Localization from models Registration is one part of the SLAM problem, thus, a large number of approaches directly stem from this area. Examples are the work of Davison [6], the work of Klein and Murray [13], or the work of Karlekar *et al*. [12]. Other important works can be found in the area of large-scale real-time reconstruction. These works include Mouragnon *et al*. [19], Irschara *et al*. [10] and Lothe *et al*. [18]. Reitmayr and Drummond proposed a system fusing GPS and model-based tracking in outdoor localization [20]. In the work of Zhu *et al*., localization is computed using a landmark database and multi-stereo visual odometry [30]. Li *et al*. solve the task as part of urban reconstruction, giving errors in the range of several meters [17].

[1]Note that as opposed to *"incremental tracking"*, *"tracking by detection"* is achieved by performing an *"initialization"* step for each new input sample.

[2]In this work natural features are used for depth perception rather than image retrieval.

To the best of our knowledge, there are only two approaches facilitating instant 6DOF registration on smartphones without heavy server-side processing. PTAM was shown to run on smartphones targeting small workspaces [14]. Arth *et al*. proposed a system for instant localization in large-scale urban environments based on 3D reconstructions and database partitioning [3].

The vast amounts of literature about solving the *localization* task suggest that there are plenty of solutions available for AR. However, as can be seen, most approaches only solve the problem at a lower level of complexity than necessary for AR. They, for instance, solely consider coarse position with fewer than 6DOF, only estimate longitude/latitude, or just deliver the inaccurate GPS tags of images retrieved from a database. They are dependent on very large (Gigabytes) databases and have high computational requirements. Moreover, few of the techniques operate in real-time, not even on desktop computers.

These limitations inherently render such approaches useless for the purpose of smartphone AR. It must also be observed that the CV and AR communities appear to work from a different set of basic assumptions concerning goals and outcomes, and that there is little cross-fertilization between communities. Therefore, we have to conclude that despite an apparently rich field of related works, the task of localization for AR is still far from being solved satisfactorily.

4 Challenges in Computer Vision

One advantage of smartphones is that localization does not have to rely on a camera sensor alone but can use any of the other available sensors, such as GPS, compass, accelerometers and gyroscopes. While the use of additional sensors is often considered as "cheating" in core CV communities, additional sensors provide essential contribution to the development of fast and robust localization that works outside of laboratory conditions. Even with the help of fusion from multiple sensors, CV based localization remains a very hard task for a number of reasons, as detailed in the following:

Textureness. Most approaches rely on natural features in the form of interest points, which require sufficiently well-textured areas in the environment. A major issue with interest points is that the presence of texture is crucial. Especially in indoor scenarios, where blank walls commonly occur, using localization approaches based on natural features, is difficult.

Lighting and weather conditions. While natural feature descriptors are usually designed to be lighting invariantly, this assumption can only hold for observations describing actual physical features. Unfortunately, it turns out that in outdoor environments a large number of features present in natural images does not relate to real physical features. Shadows cast by objects in the scene cause blobs, corners and lines to occur and to dynamically move as the lighting or weather conditions change. As a result, an overwhelming number of outliers and mismatches affect localization quality, independent of the choice of matching algorithm.

Large and volatile databases. For outdoor environments, vast amounts of data have to be acquired and processed to form an initial model prior to localization. Real-time approaches using expensive equipment can handle this issue: however, inaccessible areas can still cause holes (*i.e*., unmapped regions) in the final model. Furthermore, the acquired model can only represent a static snapshot at a given point in time. Any change in the environment, *e.g*., a shop window being redecorated, open or closed umbrellas in a café, or parked cars cause the model to become outdated immediately after data collection. Another important aspect is the distribution of the final models over (possibly mobile) communication channels. As these models are usually of considerable size, their distribution as a whole, or in part, imposes technical difficulties.

Inaccurate and missing sensor information. In outdoor localization, GPS and compass information provides valuable absolute information about the rough position and orientation of the device. Unfortunately, sensors are brittle: depending on the actual location, the accuracy of the sensor information can vary significantly. Especially in narrow urban canyons, GPS information can be off up to 100 meters or even be unavailable. Similarly, electronic compass readings are strongly affected by magnetic disturbances that are unavoidable in man-made environments.

Accurate localization is the primary and most important task to be solved for AR. Yet, as outlined above, there are significant challenges remaining, for which truly practical solutions still have to be found. Recent deployments of SLAM in tablet AR demonstrate that localization in small-scale environments works sufficiently well if certain conditions mentioned above (*i.e*. sufficient textureness) are met[3]. However, localization in large-scale environments only exists as proof-of-concept work. The associated problems appear to be of very hard nature, so only slow progress can be assumed.

5 Challenges in Augmented Reality

Beyond academic goals such as achieving precision and scalability of the researched algorithms, there is a set of practical concerns that strongly affect the usability of AR experiences. These considerations are only relevant for real-world applications of AR, and are therefore not widely discussed in scientific literature. This may lead to the incorrect observation that these problems are not difficult or not relevant for the success of AR. The following issues are relevant for smartphones, but also for general purpose AR:

Real hardware development versus the "AR wish list": As mentioned in Section 2, the quality of cameras and other sensors in current smartphone hardware is insufficient for high quality AR. Developers of AR applications would greatly benefit from hardware advancements, such as stereo cameras, unified CPU/GPU memory with random access, or WiFi triangulation. Unfortunately, it is naïve to assume that mobile phones will be optimized for AR without a large established market. Any change in hardware configuration costs millions of dollars in development, even more if market expectations cannot be met afterwards. Today, customers buy mobile phones mainly for voice communication, gaming and web browsing. These markets will drive the near to middle term evolution of smartphone capabilities. To make a claim for AR, we will have to convince manufacturers that AR is the upcoming market for mobile phone applications. Fortunately, there is sufficient excitement about AR, nowadays, and therefore this could be happening in the near future.

Dynamic scenery versus AR realism. Current AR applications assume everything in a scene to be static. However, the reality is the exact opposite. Especially in outdoor scenarios, almost everything is subject to change: people passing by, lighting and weather conditions, even buildings may be painted in different colors every few years. For localization, this causes a severe problem. In dynamic scenes, basic assumptions that most algorithms make are violated right from the start. Assuming you are augmenting a building façade while people pass by and partially occlude your view. Due to missing occlusion reasoning, noticeable errors will become apparent, no matter how good the visualization of augmented content actually is and how powerful the hardware platform gets in the future. The lack of interaction between dynamic objects and virtual content unconditionally harms realism in AR applications. Thus, the inclusion of object dynamic detection and tracking techniques currently researched in CV are the key for high-quality AR in the future.

[3] *e.g*. Ballinvasion game: `http://13thlab.com/ballinvasion`

Content creation versus registration: A large portion of the excitement about AR comes from the potential involvement of end users in content creation. Personal content creation is a key to actively integrate users rather than leaving them as passive observers. However, basic mechanisms to facilitate this concept are still missing. While interaction methods on mobile phones have improved greatly, the question how to conveniently and accurately register even simple content in 6DOF using a 2D interface and no accurate global model of the environment is still open. Assume the task of augmenting a window on a building façade. Current methods might not even suffice for the task of simple tagging. There is no mechanism to input an arbitrary 3D position in open space, let alone specifying orientation. Current approaches typically use the (inaccurate) GPS position of the user, rather than of the object of interest to determine the tag. For realistic and satisfactory content creation through end users, accurate registration of arbitrary locations in the user's vicinity must be made simple and robust – yet, another challenging research topic beyond basic CV.

A solution to the registration problem only provides a small component in the complex AR ecosystem. For many other impeding issues, there is room for improvement using existing technologies, and even more room for further research. Hardware related issues could disappear quickly if industry picks up AR as a developing market. Adoption of new CV algorithms in industry works slower - we see that algorithms proposed in the CV research community usually start being used in practice roughly 5-10 years after publication. Bringing the latest CV techniques to bear on AR will take this amount of time. For successful deployment, it will further be necessary to combine algorithmic improvement with research results from the mobile human-computer interaction community, so that the technological capabilities can be transformed into tools that benefit an actual user.

6 Conclusion

In this work we have discussed challenges in adopting smartphones as the most important AR platform of the future. This vision faces several tough challenges, which are discussed at some length in the paper. The reader may have noticed that not all issues mentioned here are of technical nature. Some of the problems have a more subtle and obscure source, such as market strategies of hardware vendors.

A major driving force for the development of AR on smartphones are improvements in hardware, which lead to more powerful and more energy-aware devices. However, even with much improved hardware, significant challenges are remaining. We therefore conclude that we are still far from being able to use AR as a generally available feature on smartphones. Especially the localization task is currently a major issues demanding more research. Several difficulties of localization performed outdoors and in the real world are barely addressed by existing CV research. We argue that the diversity of issues proves the complexity of the application of AR on smartphones, making it a worthy field for future scientific research.

Acknowledgements

This work was partially sponsored by the Christian Doppler Laboratory for Handheld Augmented Reality.

References

[1] A. Adams and *et al*. The Frankencamera: an Experimental Platform for Computational Photography. *ACM Trans. Graph.*, 29:29:1–29:12, July 2010.

[2] M. Agrawal and K. Konolige. Real-Time Localization in Outdoor Environments using Stereo Vision and Inexpensive GPS. In *ICPR*, volume 3, pages 1063 –1068, 2006.

[3] C. Arth, D. Wagner, M. Klopschitz, A. Irschara, and D. Schmalstieg. Wide Area Localization on Mobile Phones. In *ISMAR*, pages 73–82, 2009.

[4] G. Baatz, K. Köser, D. Chen, R. Grzeszczuk, and M. Pollefeys. Handling Urban Location Recognition as a 2D Homothetic Problem. In *ECCV*, pages 266–279, 2010.

[5] D. Chen, G. Baatz, Köser, S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk. City-scale Landmark Identification on Mobile Devices. In *CVPR*, 2011.

[6] A. J. Davison, W. W. Mayol, and D. W. Murray. Real-Time Localisation and Mapping with Wearable Active Vision. In *ISMAR*, pages 18–, 2003.

[7] J. B. Gotow, K. Zienkiewicz, J. White, and D. C. Schmidt. Addressing Challenges with Augmented Reality Applications on Smartphones. In *Mobile Wireless Middleware, Operating Systems, and Applications*, volume 48, pages 129–143. 2010.

[8] J. Hays and A. A. Efros. Im2gps: Estimating geographic information from a single image. In *CVPR*, 2008.

[9] H. Hile, R. Grzeszczuk, A. Liu, R. Vedantham, J. Košecka, and G. Borriello. Landmark-based Pedestrian Navigation with Enhanced Spatial Reasoning. In *Pervasive Computing*, pages 59–76, 2009.

[10] A. Irschara, C. Zach, J. Frahm, and H. Bischof. From Structure-from-Motion Point Clouds to Fast Location Recognition. In *CVPR*, pages 2599–2606, 2009.

[11] E. Kalogerakis, O. Vesselova, J. Hays, A. A. Efros, and A. Hertzmann. Image Sequence Geolocation with Human Travel Priors. In *ICCV*, 2009.

[12] J. Karlekar, S. Zhou, W. Lu, Z. C. Loh, Y. Nakayama, and D. Hii. Positioning, Tracking and Mapping for Outdoor Augmentation. In *ISMAR*, pages 175 –184, oct. 2010.

[13] G. Klein and D. Murray. Parallel Tracking and Mapping for Small AR Workspaces. In *ISMAR*, pages 1–10, 2007.

[14] G. Klein and D. Murray. Parallel Tracking and Mapping on a Camera Phone. In *ISMAR*, pages 83–86, 2009.

[15] J. Knopp, J. Sivic, and T. Pajdla. Avoiding Confusing Features in Place Recognition. In *ECCV*, pages 748–761, 2010.

[16] F. Li and J. Kosecka. Probabilistic Location Recognition using Reduced Feature Set. In *ICRA*, 2006.

[17] Y. Li, N. Snavely, and D. P. Huttenlocher. Location Recognition using Prioritized Feature Matching. In *ECCV*, pages 791–804, 2010.

[18] P. Lothe, S. Bourgeois, E. Royer, M. Dhome, and S. N. Collette. Real-Time Vehicle Global Localisation with a Single Camera in Dense Urban Areas: Exploitation of Coarse 3D City Models. In *CVPR*, 2010.

[19] E. Mouragnon, M. Lhuillier, M. Dhome, F. Dekeyser, and P. Sayd. Real-Time Localization and 3D Reconstruction. In *CVPR*, pages 363–370, 2006.

[20] G. Reitmayr and T. W. Drummond. Going Out: Robust Model-Based Tracking for Outdoor AR. In *ISMAR*, pages 109–118, 2006.

[21] D. Robertson and R. Cipolla. An Image-Based System for Urban Navigation. In *BMVC*, pages 819–828, 2004.

[22] G. Schindler, M. Brown, and R. Szeliski. City-Scale Location Recognition. In *CVPR*, 2007.

[23] S. Se, D. Lowe, and J. Little. Global Localization using Distinctive Visual Features. In *IROS*, volume 1, pages 226–231, 2002.

[24] G. Takacs and *et al*. Outdoors Augmented Reality on Mobile Phone using Loxel-based Visual Feature Organization. In *MIR*, pages 427–434, 2008.

[25] D. Wagner, G. Reitmayr, A. Mulloni, T. Drummond, and D. Schmalstieg. Pose Tracking from Natural Features on Mobile Phones. In *ISMAR*, pages 125–134, 2008.

[26] D. Wagner and D. Schmalstieg. First steps towards handheld augmented reality. In *ISWC*, pages 127–, 2003.

[27] A. R. Zamir and M. Shah. Accurate Image Localization Based on Google Maps Street View. In *ECCV*, pages 255–268, 2010.

[28] W. Zhang and J. Kosecka. Image Based Localization in Urban Environments. In *3DPVT*, pages 33–40, 2006.

[29] Y.-T. Zheng and *et al*. Tour the World: Building a Web-scale Landmark Recognition Rngine. In *CVPR*, pages 1085–1092, June 2009.

[30] Z. Zhu, T. Oskiper, S. Samarasekera, R. Kumar, and H. Sawhney. Real-time Global Localization with a Pre-built Visual Landmark Database. In *CVPR*, pages 1 –8, june 2008.