Exploring the Evolution of Mobile Augmented Reality for Future Entertainment Systems

KLEN ČOPIČ PUCIHAR and PAUL COULTON, Lancaster University, School of Computing and Communications, UK

Despite considerable progress in mobile augmented reality (AR) over recent years, there are few commercial entertainment systems utilizing this exciting technology. To help understand why, we shall review the state of the art in mobile AR solutions, in particular sensor-based, marker-based, and markerless solutions through a design lens of existing and future entertainment services. The majority of mobile AR that users are currently likely to encounter principally utilize sensor-based or marker-based solutions. In sensor-based systems, the poor accuracy of the sensor measurements results in relatively crude augmentation, whereas in marker-based systems, the requirement to physically augment our environment with fiducial markers limits the opportunity for wide-scale deployment. While the alternative online markerless systems overcome these limitations, they are sensitive to environmental conditions (i.e. light conditions), are computationally more expensive, and present greater complexity of implementation, particularly in terms of their system-initialization requirements. To simplify the operation of online markerless systems, a novel, fully autonomous map initialization method based on accelerometer data is also presented; when compared with alternative move-matching techniques, it is simpler to implement, more robust, faster, and less computationally expensive. Finally, we highlight that while there are many technical challenges remaining to make mobile AR development easier, we also acknowledge that because of the nature of AR, it is often difficult to assess the experience that mobile AR will provide to users without resorting to complex system implementations. We address this by presenting a method of creating low-fidelity prototypes for mobile AR entertainment systems.

Categories and Subject Descriptors: H5.1 [**Information Systems**]: Information Interfaces and Presentation (e.g., HCI)—*Artificial, augmented, and virtual realities*

General Terms: Algorithms, Design, Experimentation, Performance

Additional Key Words and Phrases: Mobile, Augmented Reality, Map Initialization, Sensor Fusion, Markerless

ACM Reference Format:
Klen Čopič Pucihar and Paul Coulton. 2014. Exploring the evolution of mobile augmented reality for future entertainment systems. ACM Comput. Entertain. 11, 2, Article 1 (December 2014), 16 pages.
DOI: http://dx.doi.org/10.1145/2582179.2633427

1. INTRODUCTION

Mobile phones are now the most ubiquitous consumer electronic device, and the continual advancement of their capabilities creates significant opportunities for the wide-scale deployment and evaluation of innovative new entertainment services. One such area is augmented reality (AR), in which advances in mobile phones' processing power; faster, higher-quality cameras; graphical accelerators; high-resolution displays; and

Author address: K. Č. Pucihar, P. Coulton, School of Computing and Communications, Infolab21, Lancaster University, Lancaster, LA1 4WA, UK.

DOI: http://dx.doi.org/10.1145/2582179.2633427

additional sensing capabilities (i.e., accelerometers, magnetometers, and gyroscopes), make them a compelling choice as an AR platform.

In handheld AR, virtual objects are projected into our real world, thus adding additional content to our environment when seen through the viewfinder of a handheld device. When combined with the ubiquity of a mobile phone, AR opens countless possibilities for seamless interaction with contextual data that surrounds us in our everyday lives.

Despite the significant advances of mobile phone capabilities, the direct mapping of AR systems developed for desktop PCs to a mobile phone is problematic due to the general processing speed of mobile phone hardware, which remains ∼15 to 30 times slower compared with a desktop PC [Klein and Murray 2009]; smaller form factor; and the different nature of the interface modalities. As this speed limitation is not expected to be overcome in the near future, the tuning of algorithms for mobile devices is crucial [Wagner and Schmalstieg 2009] and, coupled with the different nature of the device, justifies the differentiation between desktop and mobile AR.

One of the main driving forces of AR is entertainment, where the first adoptions of new technologies often emerge. AR technologies are primarily represented by sensor-based, marker-based, and markerless solutions. In the following section, the challenges and opportunities of each of the areas will be analyzed through a lens of existing and future entertainment services. A multi-user operation is now a vital component of many entertainment services; thus, there will be an additional focus on discussing the challenges of sharing an augmented space between multiple users.

Through evaluation of current state of the art, a broad spectrum of challenges for enabling wide-scale adoptions of mobile AR are identified, covering both the technical and user experience design. To illustrate this, we consider two specific challenges from opposite ends of spectrum and offer potential solutions. First, with a goal to decrease complexity of online markerless AR systems, we present a novel, fully automated, model-based single-step planar map initialization based on sensor data, through which natural features are un-projected to form a 3D map of a planar scene. Comparison with alternative move-matching techniques showed it is far simpler to implement, faster, less computationally demanding, and potentially more robust for users. Second, to support the evaluation of new mobile AR entertainment experiences, we present simple-to-implement approaches for the evaluation of potential AR entertainment services without the need for implementing time-consuming and costly prototypes.

2. TECHNOLOGY OVERVIEW

2.1. Sensor-based AR Solutions

The most widespread mobile AR solutions utilize common phone-sensing capabilities, such as GPS, accelerometers, magnetometers, and gyroscopes, to estimate the camera pose. Sensor-based camera-pose registration is relatively easy to implement, as the sensor data is normally accessible to the programmer through the phone's programming environment. However, due to the crude accuracy of sensory information, this method creates jerky augmentation with long response times to camera-pose changes, is limited to outdoor environments, and has problems with ensuring the augmentation with correct context. One such contextual problem is that sensor-based systems are not able to detect depth of view. This results in contextually inconsistent overlaying of information; for example, the projected information may belong to a point of interest located behind the one the user is looking at. Such contextual problems limit the possibilities for seamless interaction between physical and digital world as they break the conceptual connection between the object and the data for the user.

Nevertheless, due to the ease of implementation, there are a number of commercial, sensor-based systems that mainly focus on overlaying geo-tagged information on live

Fig. 1. Sensor-based AR solution Layar showing InfoLab21 as a Foursquare[1] venue.

video streams. Two successful commercial examples are Layar[2] (Figure 1) and Wikitude[3]. When designing services using sensor-based solutions, the main challenge lies in providing high-quality data sources. In cases where this data is not readily available, crowd sourcing may be an option.

In addition to overlaying points of interest, more complex applications have started to emerge. One such application is Wikitude Drive[4], an Augmented Reality Navigation System where route directions are projected into the real worldview with the intention of making instruction interpretation less abstract compared with a map representation. However, research by Mulloni et al. [2011] shows that the importance of providing correct information is even more crucial than in map representations, as the affordance of AR increased expectations on the accuracy of the visualized information. This suggests that quality expectations, coupled with reduced interpretation complexity of AR navigation, make it harder for the user to identify system faults.

Sensor-based AR solutions are also targeting game development, although the crude accuracy of augmentation has significant impact on the game experience, particularly in urban environments. On the positive side, the fact that the camera-pose registration is absolute, along with the presented geo-tagged information, means the sharing of the same augmented space between multiple devices does not present a problem [Chehimi and Coulton 2008] unless users are in a situation whereby they can directly compare the augmentation with each other and sensor inaccuracies become apparent. The error of positioning by GPS is especially problematic in game scenarios, where the augmented objects are to be represented in close proximity to the user (i.e., below 10 meters) as errors become more readily apparent to the user, thus making the positioning accuracy of the GPS (best case, up to two-meter accuracy [Rashid et al.]) a significant problem for achieving a meaningful augmentation. In scenarios where augmented objects are far away from the user, the augmentation quality mainly suffers from camera tilt and orientation errors, which can be greatly improved by better hardware capabilities, such as gyroscopes.

Besides visual augmentation, audible augmentation is an equally compelling option for creating new entertainment services. Spatial sound augmentation could be utilized to create a more realistic variation of existing audible location-based games [Gustafsson

[1]http://www.foursquare.com.
[2]http://www.layar.com.
[3]http://www.wikitude.org.
[4]http://www.wikitude.com/drive.

et al. 2006]. Although audio-only games, such as Papa Sangre[5] and NightJar[6], have been developed for mobile, they don't currently utilize any sensor information. Further, visually impaired communities would benefit from such applications, as they would be able to play games on an equal footing with sighted players.

Despite all of the previously presented implementations, the accuracy of augmentation is highly problematic for the majority of envisioned entertainment services. The only way to achieve high precision of augmentation in handheld AR is to use computer vision algorithms [Kurz and Benhimane 2011], which will be explored in the following section.

2.2. Marker-Based AR Solutions

While there has been considerable progress in vision-based mobile AR over recent years, it has principally been related to either marker-based or offline markerless systems, in which 3D maps of environment are known in advance. Such systems are not appropriate for wide-scale deployment, as they require either mapping of our physical environment or augmentation of the world with fiducial markers. Furthermore, as distribution of mobile applications is now primarily done through electronic publishing channels, commonly referred to as app stores, the requirement to distribute markers becomes additionally problematic as studies of user behavior show that most applications are generally run once immediately after downloading, and an additional burden of having to print a marker is likely to severely limit the uptake of such services [Coulton and Bamford to be submitted].

In marker-based systems, features extracted from a fiducial marker are used for relative camera-pose registration. The accuracy of such camera-pose registration, when compared with sensor-based solutions, is significantly higher with achievable accuracies of less than 1 centimeter and 1 degree, which is generally considered the benchmark for mobile AR solutions. In comparison with natural feature-based solutions, marker-based solutions are less computationally expensive, less complex to implement, as well as more robust. However, the system works only while the marker is visible in the scene of view, and is limited to planar AR workspaces. Due to the limitation of detecting markers, no other objects within the real environments can be detected, which can lead to flawed augmentation. For example, the augmented object partially overlaps another object in the scene close to the marker. Furthermore, the presence of a marker within each captured camera frame limits the effective size of AR workspace, although this can be increased by adding support for multiple marker tracking.

Regardless of these limitations, the entertainment, publishing, and marketing industries successfully adopted marker-based technology. For example, Nintedo 3DS[7] and Sony Vita[8] allow users to augment their favorite characters and play AR games on flat surfaces using AR game cards (Figure 2). Both systems support multiple-markers tracking in order to enable the expansion of AR workspace. It is obvious that in a gaming scenario, requiring users to continually carry markers and obtain new markers limits the use and adoption of this technology. However, this is not a problem for use cases where marker distribution is less problematic or the experience is designed to be short-lived—for example, interactive books, where markers within the book are used to augment interactive content or illustrate the story, or magazines advertising a particular product or service. However, such markers can be obtrusive and meaningless to the reader until used for augmentation.

[5]http://www.papasangre.com/.
[6]http://www.the5experience.co.uk/.
[7]http://www.nintendo.com/3ds.
[8]http://www.playstation.com/psvita/.

Fig. 2. Nintendo 3DS AR card game Target Shooting, where users try to defeat the dragon (a), Sony VITA AR card game Cliff Diving, where diver Dan dives into your desk swimming pool (b).

An alternative to visible markers is unobtrusive infrared markers [Burnett and Coulton 2011]. Such markers are invisible, as infrared light cannot be seen by a human eye but is detectible by most phone cameras. For example, an interactive installation has been created at the National History Museum in London[9] using infrared markers to position tablets that visitors can use to view animated dinosaurs. Despite the fact that camera-pose registration of marker-based systems is relative, there is no need for sharing map information among users because the map information is shared through the predefined marker points. Thus, in order to achieve multiple device collaboration, only the information about the augmentation needs to be shared among users. The same principle is also applied by Nintendo 3DS and Sony Vita in order to support multi-player gaming within the same augmented space.

In summary, even though for some use cases the distribution of markers does not present a significant problem, there are many other cases where the need for markers limits AR use. This leads to the alternative markerless systems that will be discussed in the following section.

2.3. Markerless AR Solutions

There are three main types of markerless AR systems: offline systems where 3D maps of environment are known in advance; extensible tracking systems where 3D maps are initially known in advance and later extended to previously unknown areas; and online systems where maps of previously unknown environments are created in real time. In cases when 3D maps are known in advance and limited to planar scenes, offline markerless systems can be considered analogous to marker-based solutions, where instead of features from fiducial markers, existing natural features from environment are used to obtain camera pose. Thus any surface with enough texture can become a marker. This creates the potential to generate a bank of predefined AR surface images, stored locally or in the cloud, which can later be used for identification and camera-pose registration in an offline AR scenario. However, this results in a limitation for wide-scale deployment similar to that of marker-based systems, although such markers are far less obtrusive and easier to obtain. For example, marketing agencies such as Blippar[10] have adopted offline markerless technology for company branding purposes by utilizing company logos or other branding materials as a marker. As

[9]http://www.nhm.ac.uk/visit-us/darwin-centre-visitors/attenborough-studio/interactive-film/index.html.
[10]http://blippar.com/.

company-branding materials are widely available in magazines, newspapers, etc., it makes it easier for the user to obtain a copy and try out the application.

In addition to the advantages of better availability and less obtrusiveness, images also hold more meaning, thus enabling the intuitive merge between printed and digital media. In the research domain, Mobile Augmented Reality for identifying books on a shelf is one such example, where book spines are identified and later used to present augmented prices of identified books in different online stores [Chen et al. 2011]. Likewise, the commercially available Metaio Mobile SDK[11] library enables the distribution (if published through Junaio Augmented Reality Browser[12]), identification, and augmentation of additional digital content (e.g., 3D models, animations, and video clips) on top of any printed media. Recent examples of applications adopting Metaio technology are: Auto Zeitung DriveView[13], which merges digital media content with a printed version of the magazine; AR Puzzle[14], which brings augmented 3D puzzles to life; and AugmentedTeeShirt[15], which enables users to view animations and take pictures with animated content on their T-shirts.

Support for similar cloud-based banks of identifiable images has also been announced by Qualcomm through its Vuforia SDK[16] [Qualcomm 2012]. The number of identifiable images in a single data set is limited on both systems; however, while the information regarding the Metaio library is not available, the future release of Vuforia system is forecasted to support more than one million identifiable target images, thus creating numerous possibilities for novel entertainment services, for example, the American Apparel[17] shopping application demonstrated in the pre-release version of Vuforia library [Qualcomm 2012].

Despite the considerable opportunities created by offline markerless technology, mass-market adoption of the technology has still not occurred. There are many cases in which the size of supported data sets, as well as the need for markers, limits scenarios of where and when AR can be used. Additionally, the application works only when a specific marker is visible within each camera frame, thus limiting the size of AR workspace. Nevertheless, the mass adoption is expected to speed up by future advances in data set sizes, which will also enable camera-pose registration.

In order to create bigger AR workspaces, extensible tracking systems have been proposed and developed in the domain of desktop computers. Such systems are able to expand existing 3D maps to unknown environments, thus enabling the creation of bigger AR workspaces [Park et al. 1998]. However, our research has not identified any implementations of extensible tracking systems on handheld devices.

The alternative is online systems, in which no apriori information of the environment is required, thus making them more flexible and highly appropriate for wide-scale deployment. However, creating maps on the fly introduces the challenge of map initialization where the environment and camera pose are unknown. At each system initialization, a new 3D map with a different coordinate system is generated. Factoring in that camera pose registration is relative to the initialized 3D map means that map alignment is required in order to support sharing of AR workspace if the context of augmentation is to be maintained for all participating users. In the map alignment process, the coordinate systems are calibrated, map scales are initialized to the same

[11]http://www.metaio.com/software/mobile-sdk/.
[12]http://www.junaio.com/.
[13]http://www.autozeitung.de/driveview.
[14]http://www.ravensburger.com/uk/puzzles/augmented-reality/.
[15]http://www.augmenteeshirt.com/.
[16]http://www.qualcomm.com/solutions/augmented-reality.
[17]http://www.americanapparel.net/.

unit, and base planes are matched. All ambiguities can be resolved by sharing a minimum of three reference points—two, if the workspace is limited to planar scenes [Čopič and Coulton 2011].

Additionally, the map scale of online systems is unknown regardless of the map initialization method employed, as it is impossible to determine the scale of the scene based on a sequence of images alone [Hartley and Zisserman 2004]. This results in reduced contextual sensitivity of augmentation when compared with marker-based solutions. To date, the only method for fully autonomous scale initialization in online markerless AR systems, for devices with no additional measuring mechanisms (i.e., sonar and laser-depth sensors), is to utilize the autofocusing capability of a camera phone. The method is better known as Depth From Focus (DFF) technique [Čopič and Coulton 2011] and has been successfully implemented to introduce scale in online markerless AR system [Čopič et al. to be submitted]. By introducing scale, the context of augmentation can match that of marker-based systems.

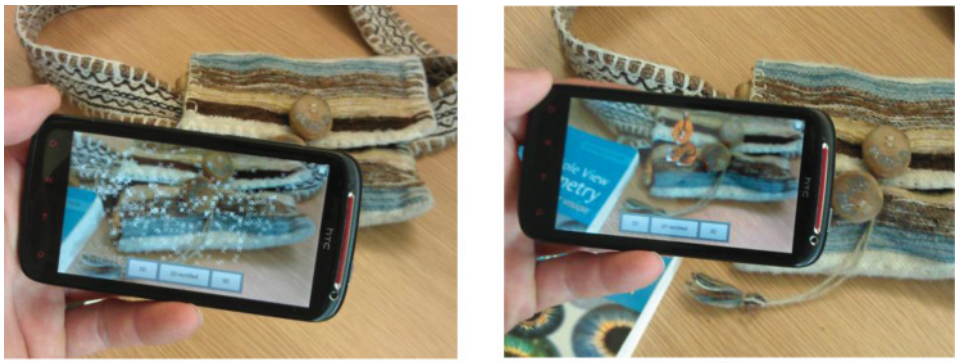
Online markerless AR methods can be divided into model-based [Hagbi et al. 2009; Wagner et al. 2008] and move-matching techniques [Klein and Murray 2009]. In case of model-based techniques, the camera pose is always estimated by correspondences between features of the initial key frame and the current camera frame. As the system is online, the 3D map needs to be initialized. In case of planar objects, map initialization can be achieved by capturing the initial key frame from a front-to-parallel view (Figure 4) of the object plane through which the location of extracted features in the image can be used as a 3D map model of a plane. An alternative is to computationally un-project extracted features from the initial key frame using additional sensor information or user interaction input (e.g., the manual selection of four points forming a square, selection of two perpendicular edges in the scene, etc.). This removes perspective distortions of camera projection by which the 3D map is initialized. Such map initialization is limited to planar scenes but avoids computationally expensive 3D reconstruction of the scene.

To date, the only commercial library supporting online model-based markerless tracking for planar surfaces is the Metaio mobile SDK. The library supports tracker configuration at run time, in addition to the definition of tracking targets by raw images with no pre-processing requirements, thus enabling developers to define new tracing targets on the fly.

The alternative to model-based camera tracking is move-matching, in which the information is extracted based on the frame-to-frame movement of tracked features. Techniques used for solving systems where no apriori knowledge of the environment is available in advance are better known as Simultaneous Localization and Mapping (SLAM) techniques, where camera pose as well as 3D maps are simultaneously initialized and updated. These approaches can typically be divided into incremental and long-term mapping systems. However, regardless of system type, the main challenge is initializing the map from a sequence of images alone.

The first implementation of six degrees of freedom (6-DOF), SLAM is a highly modified variation of PTAM [Klein and Murray 2009] running on the iPhone. PTAM is a long-term mapping system, where the map is initialized in a fully autonomous, two-stage procedure that requires the user to select only the initial key frame [Klein and Murray 2009]. Autonomous map initialization makes the system less prone to user error, because the selection of the second key frame that forms a valid stereo pair for map initialization is done automatically.

Generally, the map-initialization problem can be classified as the structure-from-motion problem. This problem can be solved by decomposing the camera movement from frame-to-frame, after which the map points can be initialized using



(a) (b)

Fig. 3. Online 3D map initialization of Metaio SALM system (a). After providing sufficient movement around the scene whilst continuously tracking initial feature points, map is initialized and the lion is augment onto the dominant base plane of AR workspace (b).

the stereovision principle. Compared with model-based map initialization, such an approach has the advantage of being able to initialize in a non-planar scene. Thus, the challenge of correct map initialization is how to select the two appropriate key frames where the camera movement can be sufficiently defined and will meet the stereovision baseline requirement for map initialization. Although such a process can be implemented in a fully autonomous way, they are complex, computationally more expensive, harder to implement, and potentially less robust compared with the previously discussed model-based map-initialization method.

Besides PTAM, there have been two other implementations of SLAM on handheld devices, namely: Smart AR[18] implemented on Sony Vita and the previously mentioned Metaio system for mobile phones. At the time of writing Smart AR has not been released to the public, making Metaio the first commercially available SLAM implementation for handheld devices (Figure 3).

Despite the fact that AR workspaces of markerless systems are not limited to planar scenes, as in the case of marker-based systems, no examples of online systems with higher complexity of AR workspaces have yet been demonstrated. Improving reconstruction models of 3D environments would make for a more realistic user experience and gain a crucial advantage over marker-based systems.

To sum up, this section delivered an extensive overview of augmented reality technology. In case of sensor-based systems, the crucial problem was the quality of augmentation, whereas the main limitation of marker-based systems was the requirement for obtrusive fiducial-markers, which subsequently limits mass adaptation of the technology. When moving to offline markerless systems, where any sufficiently textured surface could become a marker, the problem of distributing markers becomes less as they are easier to find in our every day lives (i.e. a company logo, a picture in the newspaper, product packaging, etc.). In addition to availability, pictures hold more meaning than randomly generated fiducial markers thus enabling the intuitive merging between the printed and digital media. While the size of datasets of identifiable images is limited these are expected to grow in the future, thus creating an opportunity for wider adoption.

Even though, online markerless solutions have the greatest potential for wide scale deployment there have been a limited number of commercial use-cases. Arguably this

[18]http://www.engadget.com/2012/03/09/sony-shows-off-playstation-vitas-augmented-reality-chops-at-gdc/.

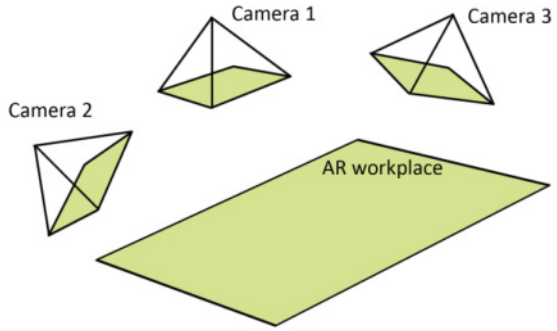


Fig. 4. Front-to-parallel view is a view where the AR workspace is viewed from the front and is parallel to the camera plane as in the case of camera 1 view.

is due to higher complexity of such systems where contrary to offline system, 3D map initialization is required. This increases the computational cost of system initialization and camera pose tracking. In addition to that, ensuring the robustness of operation in unknown environments remains to be a problem. Therefore, in order to decrease complexity and reduce computational cost of online markerless system initialization the following section introduces an alternative sensor based planar map initialization.

3. SENSOR FUSION FOR AUTONOMOUS MAP INITIALIZATION

In model-based tracking systems the 3D map of a planar object can be initialized by acquiring front-to-parallel (Figure 4) view of the plane. The front-to-parallel view is used because such images have no perspective distortions. In case of undistorted images, extracted feature locations can directly be used as map points. The distortions are caused by the nature of image formation of a pinhole camera model where the objects farther away from the camera are rendered smaller than object closer to the camera. In case when all object points are located at the same distance from the image plane the perspective distortion is not present. An alternative to providing a front-to-parallel view is to computationally un-project the image by using accelerometers of a camera phone in a way that will maintain a constant distance of all points on the object pane from the image plane. Such map initialization avoids going through any reconstruction of the 3D scene simplifying map initialization.

The first attempt to merge sensor and vision data on a mobile phone in order to support map initialization of online markerless system is a mobile specific implementation of Gepard [Lee et al. to be submitted] where the system initializes a planar patch by removing distortions caused by pitch rotation (on Figure 5 denoted as Rx) of the phone, leaving the map alignment of the other rotation to the user. In the initialization phase the system rectifies the whole camera image from which a single patch is selected for tracking purposes.

Another project relating to sensor fusion on handheld devices focused on improving feature-descriptor-based tracking systems. Work presented by Kurtz et al. [2011] introduced Gravity Aligned Future Descriptors (GAFD) for close-to-vertical surfaces. Compared with standard approaches that gain feature orientation from the intensities of neighboring pixels, GAFD future descriptors achieve orientation invariance by aligning features with gravitational vector, resulting in better and computationally less complex feature descriptors [Kurz and Benhimane 2011]. As each feature has been aligned with gravity vector, current gravitational vectors of camera pose can be used to assist frame-to-frame feature matching. Thus the method does not focus on

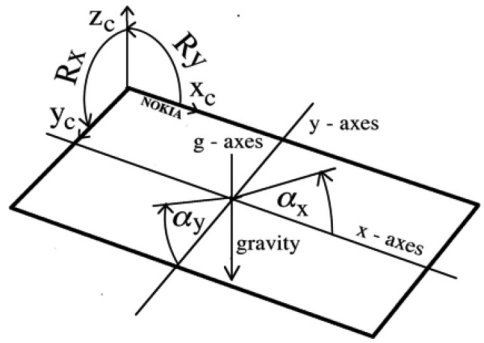


Fig. 5. Sensor and camera coordinate system in case of horizontal plane initialization.

rectifying the image in order to remove perspective distortions from the capture video stream, but instead is used to define the orientation of the feature.

Following GAFD, the Scene Derived Mobile Augmented Reality (SID-MAR) library was the first model-based markerless AR system in which sensors were used to correct for perspective distortions caused by both pitch and roll rotations (on Figure 5 denoted as Rx and Ry), thus enabling a fully atomized map initialization of online markerless systems on horizontal surfaces. The initial concept and preliminary implementation of SID-MAR sensor fusion was previously demonstrated by Čopič et al. [to be submitted]. In the following section, a detailed description of the proposed sensor fusion will be discussed and evaluated.

Gravity Rectified Feature Descriptor (GREFD) is the latest advance of sensor fusion on handheld devices; each camera image is gravity-rectified before extracting feature descriptors that are later used for feature marching. Gravity-rectified camera images create a virtual front-to-parallel view in a way that is similar to Gepard and SID-MAR library. However, in contrast to Gepard and GREFD, SID-MAR rectifies only extracted feature points, thus reducing the computational cost of the operation [Kurz and Benhimane 2011].

3.1. SID-MAR Design

Accelerometers provide information about the orientation of the phone by utilizing gravity; the acquired rotations are relative to the gravitational axis orientation and are thus detectable only around two axes (on Figure 5 marked as x and y) Gilbertson and Coulton [to be submitted]. When initializing on a horizontal plane this is not a problem, as the rotation around the gravity axes does not create perspective distortions, making it possible to correct for all angles of rotation that create perspective distortions. However, this is not the case when initializing on vertical surfaces.

Fully autonomous map initialization is thus limited to horizontal surfaces unless a calibration of the surface tilt (similar to the ones used in games with accelerometer interfaces, where users calibrate the orientation of their phone prior to starting the game [Chehimi and Coulton 2008]) is introduced. The calibration can be achieved by aligning the phone with the plane, as in the case of capturing a front-to-parallel view. Note that this would be necessary only when the orientation of the workplace surface changed. To support initialization of planes where rotation around the gravity access causes perspective distortions, a gyroscope or compass is required. However, it is important to note that in majority of use cases, the AR workspace is expected to be a horizontal surface, making the proposed method usable in many cases without any calibration.

In order to un-project image points x, the system needs to define the camera projection matrix P. The most common camera model used in AR systems is the finite pinhole camera model [Hartley and Zisserman 2014], which is described by a 3x4 projection matrix P (1). The projection matrix P is defined by intrinsic camera parameters K, rotation matrix R and a translation vector t. In cases where the camera has been calibrated, the intrinsic parameters K are known; therefore, there are only six degrees of freedom left (three for rotation and three for translation) in projection matrix P (1).

$$\begin{aligned} x &= P \cdot X \Rightarrow X = (X, Y, 0, 1) \quad x = (x, y, 1) \\ P &= K \cdot [R \,|\, t] = K \cdot [r_1 r_2 r_3 t] \end{aligned} \tag{1}$$

In cases of planar scenes, the object coordinate system can be chosen in a way to fix the plane at $Z = 0$. By doing this the last column of the rotation matrix R can be removed, and the projection matrix P becomes a square 3×3 H matrix, also known as homography (2). As the projection matrix is now a square matrix, it is possible to calculate the inverse homography H^{-1}.

The remaining question is how to define the homography matrix H, whose inverse can later be used to un-project image points. As has previously been discussed, the rotation around x and y axes are the only two factors that create the distortion when initializing on horizontal surfaces. Therefore the rotation matrix R needs to be defined in such a way that it takes into the account Rx and rotations detected by phone accelerometers. To support initialization of non-horizontal surfaces rotation around z-axes, R_Z would need to be added to the rotation matrix R. Using the angle of rotation, on Figure 5 denoted as a_x and a_y, it is easy to construct R_x and matrices which can then be concatenated into one rotation matrix R [Bradski and Kaehler 2008]. In order to keep the object coordinate system aligned with the camera coordinate system, the x and y coordinates of the translation vector need to be equal to 0. However, the z coordinate has effect only on the map scale, which was already lost in the process of image formulation. Nevertheless, it is important to choose the translation vector t in a way that will enable us to calculate the inverse matrix of H. The inverse matrix can be calculated if the determinant of H is not equal 0. Thus, the translation vector was chosen as t = (0, 0, 1).

$$\begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = K \cdot [r_1 r_2 r_3 t] \cdot \begin{pmatrix} X \\ Y \\ 0 \\ 1 \end{pmatrix} = K \cdot [r_1 r_2 t] \cdot \begin{pmatrix} X \\ Y \\ 1 \end{pmatrix} \tag{2}$$

Now that the inverse homography is defined, the image points can be un-projected to form a 3D map of the object plane.

3.2. Evaluation

The proposed map initialization can be implemented using any point-based tracking system and is thus not limited to a specific tracking technology. The map initialization task in object-based systems involves two stages, feature extraction and map creation. The majority of existing map-initialization methods use different feature-tracking systems and have been tested on a wide range of platforms, thus making direct performance comparisons difficult. In case of GREFD, the system has been evaluated using a desktop PC [Kurz and Benhimane 2011], limiting comparison of published results with Gepard and SID-MAR, as those have been evaluated on a mobile phone.

The proposed map initialization is implemented as part of SID-MAR library, which is currently supported for the maemo operating system. SID-MAR is designed as a point-based tracking system, where natural features are extracted from the image based on Shi and Tomasi good corner definition [Shi and Tomasi 1994]. The feature-tracking

Fig. 6. Screenshot of SID-MAR[19] library running on Nokia N900. The application is augmenting a Rubik's Cube on a surface of a randomly selected book. Map is initialized by combining sensor and vision data.

system is based on optical flow, which is calculated in the small window region (5x5 pixels) of selected points by the Pyramid Lucas and Kanade algorithm [Lucas and Kanade 1981].

The proposed map-initialization method has been tested on Nokia N900 with ARM Cortex-A8 600 MHz phone. Throughout testing, the number of extracted feature points was limited to 20. Map initialization was repeated 30 times under varying phone angles using the same horizontal surface. The mean map initialization time was 174 ms, where ∼99 percent of the execution time was used for feature extraction. Compared with system initialization times of Gepard running on the iPhone 3GS, SID-MAR initializes approximately 25 times faster. However, its important to note that Gepard uses a far more complex feature extraction procedure that involves generation of 225 patch views, later used for camera pose tracking and reinitialization. Compared with alternative move-matching techniques, such as map initialization of PTAM running on iPhone [Klein and Murray 2009], the proposed map initialization executes approximately 10 times faster.

4. DESIGNING AR SOLUTIONS

In the previous section the area of system implementation was addressed, while here we address the challenge from the opposite side of the spectrum: designing the user experience. Regardless of emerging commercial development tools for AR, such as Vuforia and Mataio, there is a general lack of development and design tools for evaluating future AR applications. This, coupled with the complexity of AR, makes prototype development a very time-consuming and costly operation. Therefore, alternative design methods are crucial for evaluating future AR entertainment services.

AR applications involve a high degree of visual representations, which can make conveying such ideas verbally a formidable task. In addition to that, without being able to create prototypes, it is very hard to integrate user perspective into the design process. The alternative solution to creating working prototypes is to use a well-known strategy of low-fidelity prototyping (here we use the term *low-fidelity prototyping* to describe applications that try to create an illusion of a functional application) that can be coupled by Wizard of Oz experiments in which the users believe in the autonomy of the system, when it is actually being fully or partially operated by an unseen human being [Kelley to be submitted].

[19]http://www.youtube.com/watch?v=nNMiUWF-QbU.

Fig. 7. Screenshot of Interactive TV low-fidelity prototype application running on Android Nexus One phone.

Low-fidelity prototypes are much cheaper and faster to develop compared to creating working prototypes, but create an equally powerful representation of the concept, especially when presented in a pre-staged environment or as a Wizard of Oz experiment. To bring in the concept of mobility coupled with user interface limitations of mobile phones, it is vital to have low-fidelity prototypes run on actual devices.

The development tools can be divided into content creation and mobile application implementation. The choice of tools for content creation depends on the type of AR applications. In case of creating 2-D overlays, any image-editing software coupled with video -editing software that allows adding image overlays can be used.

The second part requires low-fidelity prototypes to be implemented as applications with user interface on the actual device. The choice of tools depends on the target platform. Adobe AIR technology is a good candidate due to ease of development and flexibility of running developed applications on multiple platforms. However, playing a video and implementing user interfaces are well supported by most native software development kits (SDK). Further, using native development kits has the benefit of higher performance, coupled with the native look and feel of developed applications.

4.1. Designing an Interactive TV Application

To demonstrate the previously proposed method for designing AR solutions, a low-fidelity prototype of an interactive TV application was designed and is shown on Figure 7. The application enables users to use their mobile phone as a magic lens to identify individuals they can see on their TV screen.

The selected video for prototype displays a football match. By looking at the screen through the phone, users see an augmentation displaying player names that follow each player as they move around the screen. The prototype creates an illusion of viewing a live video stream by operators synchronizing the video presented on the television and on the screen of the phone, as well as by placing the user at approximately the same position from where the video was originally captured. To make the prototype more immersive, the orientation of the phone is detected by onboard sensors. In cases when the phone is pointed away from the screen, the video disappears and becomes visible again only after aligning the phone with the television screen. The prototype was implemented and tested on Android Nexus One phone, using Android SDK[20], while for content creation, Adobe Flash[21] was used.

[20]http://developer.android.com/sdk/.
[21]http://www.adobe.com/products/flash.html.

By looking at the screenshot of the low-fidelity prototype application (Figure 7), one can easily identify the proposed system concept and verify the potential for representing new ideas to users. In addition to that, the prospect of using low-fidelity prototypes to engage potential users in the design process is also evident.

5. CONCLUSION

A review of mobile AR technologies from sensor-based to marker-based and markerless AR solutions has highlighted the potential reasons why mass adoption of AR is yet to occur. In case of sensor-based systems, the main limitations are the crude accuracy of sensory information, which results in rudimentary and jerky augmentation, as well as the limitation to outdoor environments. The accuracy of GPS sensors is especially problematic for game development, because in game scenarios the augmented objects are usually augmented in close proximity to the user. Nevertheless, due to the ease of implementation and flexibility to operate in unknown outdoor environments, many commercial use cases of mobile AR currently utilize sensor-based techniques.

Vision-based solutions can provide a more accurate camera-pose registration compared with sensor-based solutions. However, in case of fiducial marker-based systems, the main limitation is the requirement of providing an obtrusive marker. In the case of offline markerless systems, any sufficiently textured surface can become a marker, thus replacing obtrusive meaningless makers with images. Such technology creates the potential to generate a bank of predefined AR surface images, stored locally or in the cloud, which can later be used for identification and camera-pose registration in an offline AR scenario. The size of datasets of identifiable images is currently limited, although in the future these are likely to grow.

Regardless of new opportunities created by offline markerless technology, the marker requirement and the limited size of datasets restricts instant use of the application and places marker-based AR solutions in a place of severe shortcoming. This is because mobile applications are primarily distributed through app stores, where analysis of user behavior showed that most applications are run once immediately after downloading [Coulton and Bamford to be submitted]. However, there are certain use cases, such as interactive books and marketing campaigns, where the distributed content is encouraging application download, and the experiences are designed to be short-lived, thus eliminating the disadvantage of marker necessity.

It is evident that marker-based solutions are not optimal for gaming scenarios, as users have to continually carry markers or obtain new markers. Furthermore, such systems are limited to planar scenes and work only when the marker is visible, thus limiting the size and shape of AR workspaces. Even though multiple marker tracking systems to an extent address size restriction of AR workspaces, marker-based AR games currently appear be restricted to the realms of novelty.

The alternative vision-based systems are online markerless systems where no apriori information of the environment is required; they are therefore more appropriate for wide-scale deployment. However, as developer tools for such systems are still in their infancy, system development continues to require a high degree of skill both, from the developer and subsequently the user. This, coupled with complex camera-pose-tracking algorithms plagued with problems ensuring robustness, currently makes such technology too immature for mass-market adoption.

Challenges covering a wide spectrum of areas have been identified, covering both the requirement to improve the technology and help evaluate the user's experience during the design phase. The paper suggests two possible solutions within the field. First, to support improved usability and reduce system complexity of AR systems, this paper presents a method to simplify system initialization of online markerless AR solutions by incorporating use of accelerometers. The method clearly illustrates that sensor and

vision fusion offer a significant enhancement to markerless AR systems running on mobile devices and can be used to create a faster, more elegant 3D map initialization of a planar scene. In contrast to the move-matching map-initialization methods, the proposed model-based map initialization is limited to planar scenes. In cases where planes are non-horizontal, an additional ground plane calibration step is required for correct map initialization. However, the method has an advantage over the alternative move-matching initialization techniques, as it is much simpler to implement, potentially more robust, and computationally less expensive.

Second, to enable the evaluation of the experience of possible AR systems, a method utilizing low-fidelity prototypes coupled with pre-staged environments, or Wizard of Oz techniques, was proposed. The method demonstrated how low-fidelity prototypes can be used to engage potential users in the system-design process, as well as introduce system concepts without the need for implementing fully working prototypes.

To conclude, the potent mix of processing power, sensors, high-quality displays, good cameras, and accessibility makes mobile phones an exciting AR platform, particularly for future entertainment services. Merging these qualities creates thriving environments with great potential where enhancements, such as those presented in this paper, become indispensable.

6. ABOUT THE AUTHORS

Klen Čopič Pucihar is currently a Ph.D. student at Lancaster University, researching the Mobile Augmented Reality with emphasis on systems appropriate for wide-scale deployment that can operate in unknown environments. Klen has already gained substantial experience in mobile development and innovation, winning first prize of the worldwide coding competition WidSets Challenge 2008 and second prize at the Maemo Coding Competition in 2010.

Dr. Paul Coulton has more than 15 years' research experience in mobile and is senior lecturer at Lancaster University. He leads the Mobile Experiences Group for the Nokia Innovation Network, an elite group of 20 universities around the world selected by Nokia. Paul has published extensively in mobile both in terms of both academic papers was selected as one of 50 most talented mobile developers worldwide. The main focus of his research is innovative mobile experiences design, with an emphasis on mobile entertainment in the form of games and play.

ACKNOWLEDGMENTS

The authors would like to thank Nokia and in particular Sean White for the software and hardware for the software and hardware that the Mobile Radicals research group has used to carry out this research.

REFERENCES

Bradski, G. and Kaehler, A., Learning Open CV: Computer Vision with the OpenCV Library. O'Reilly, 2008.

Burnett, D. and Coulton, P., Using Infra-Red Beacons as Unobtrusive Markers for Mobile Augmented Reality. *MobileHCI 2011 Workshop: Mobile Augmented Reality: Design Issues & Opportunities,* Stocholm, Sweeden, 2011.

Chehimi, F. and Coulton, P., Motion controlled mobile 3D multiplayer gaming. *Proceedings of the 2008 International Conference in Advances on Computer Entertainment Technology - ACE '08, New York*, ACM Press, 267–267.

Chehimi, F., Coulton, P. and Edwards, R., 3D Motion Control of Connected Augmented Virtuality on Mobile Phones. *Proceedings of 2008 International Symposium on Ubiquitous Virtual Reality*. 67–70.

Chen, D., Tsai, S., Hsu, C.-h. and Singh, J.P. Mobile Augmented Reality for Books on a Shelf. *Multimedia and Expo (ICME),* 2011.

Coulton, P., Bamford, W., Experimenting Through Mobile 'Apps' and 'App Stores'. *International Journal of Mobile Human Computer Interaction (IJMHCI)*, 3(4), 55–70.

Čopič Pucihar, K. and Coulton, P., Estimating Scale using Depth From Focus for Mobile Augmented Reality. *Procceedings of the ACM SIGCHI Symposium on Engineering Interactive Computing Systems 2011*, Pisa, Italy, (2011).

Čopič Pucihar, K., Coulton, P. and Hutchinson, D., Utilizing Sensor Fusion in Markerless Mobile Augmented Reality. *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Service*s. 663–666.

Čopič Pucihar, K., Coulton, P., Towards Collaboratively Mapped Multi- View Mobile Augmented Reality. *MobileHCI 2011 Workshop: Mobile Augmented Reality: Design Issues & Opportunities,* Stocholm, Sweeden, 2011.

Gilbertson, P., Coulton, P., Chehimi, F. and Vajk, T., Using tilt as an interface to contro no-button 3D mobile games. *ACM Computers in Entertainment*, *6* (3). 1–13.

Gustafsson, A., Bichard, J., Brunnberg, L., Juhlin, O. and Combetto, M., Believable environments: generating interactive storytelling in vast location-based pervasive games. *Proceedings of the 2006 ACM SIGCHI international conference on Advances in computer entertainment technology*, 2006.

Hagbi, N., Bergig, O., El-Sana, J. and Billinghurst, M., Shape recognition and pose estimation for mobile augmented reality. *Proceedings of the 8th IEEE International Symposium on Mixed and Augmented Reality*, (2009), IEEE Computer Society, 65–71.

Hartley, R. and Zisserman, A., *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2004.

Kelley, J.F., An iterative design methodology for user-friendly natural language office information applications. *ACM Transactions on Information Systems (TOIS)*, 2 (1). 26–41.

Klein, G. and Murray, D., Parallel Tracking and Mapping on a Camera Phone. *Proceedings of the 8th IEEE International Symposium on Mixed and Augmented Reality*, (Orlando, 2009).

Kurz, D., and Benhimane, S., Inertial sensor-aligned visual feature descriptors. *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

Kurz, D., and Benhimane, S., Gravity-Aware Handheld Augmented Reality. *Proceedings of the 8th IEEE International Symposium on Mixed and Augmented Reality*, (2011), pp. 111–120.

Lee, W., Park, Y., Lepetit, V. and Woo, W., Point-and-shoot for ubiquitous tagging on mobile phones. *Proceedings of the 9th IEEE International Symposium on Mixed and Augmented Reality*, 57–64.

Lucas, B. D., and Kanade, T., "An Iterative Image Registration Technique with an Application to Stereo Vision", in *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, 1981, pp. 674–679.

Mulloni, A., Seichter, H., and Schmalstieg, D., Enhancing Handheld Navigation Systems with Augmented Reality, *MobileHCI 2011 Workshop: Mobile Augmented Reality: Design Issues & Opportunities,* Stocholm, Sweeden, 2011.

J. Park, S. You, and U. Neumann. Natural feature tracking for ex- tendible robust augmented realities. *Proceedings International Workshop on Augmented Reality*, 1998.

Rashid, O., Mullins, I. and Coulton, P., Extending cyberspace: location based games using cellular phones. *Computers in Entertainment (CIE)*, 1–18.

Shi, J. and Tomasi, C., Good features to track, *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 1994, pp. 593–600.

Qualcomm, Vuforia Extends Image Recognition to the Cloud, http://www.qualcomm.com/media/releases/2012/06/27/vuforia-extends-image-recognition-cloud (accessed on: 11. Oct. 2012).

Wagner, D., Reitmayr, G., Mulloni, A., Drummond, T. and Schmalstieg, D., Pose tracking from natural features on mobile phones. *Proceedings of the 7th IEEE International Symposium on Mixed and Augmented Reality*, (Washington, DC, USA, 2008), 125–134.

Wagner, D. and Schmalstieg, D., History and Future of Tracking for Mobile Phone Augmented Reality. *Proceedings of the 2009 International Symposium on Ubiquitous Virtual Reality*, (GIST, Guangju, Korea, 2009), 7–10.