Visual Mapping and Multi-modal Localisation for *Anywhere* AR Authoring

Andrew P. Gee, Andrew Calway and Walterio Mayol-Cuevas

Dept. of Computer Science, University of Bristol, UK

Abstract. This paper presents an Augmented Reality system that combines a range of localisation technologies that include GPS, UWB, user input and Visual SLAM to enable both retrieval and creation of annotations in most places. The system works for multiple users and enables sharing and visualizations of annotations with a control centre. The process is divided into two main steps i) global localisation and ii) 6D local mapping. For the case of visual relocalisation we develop and evaluate a method to rank local maps which improves performance over previous art. We demonstrate the system working over a wide area and for a range of environments.

1 Anywhere Authoring

Most Augmented Reality (AR) systems to date can be categorized by either having high levels of accuracy in small scale spaces, as provided by 3D visual simultaneous localisation and mapping (SLAM), or systems covering larger areas but resorting to approximate location, as offered by GPS. The former systems are capable of delivering accurate 3D object registration in unprepared environments and the latter well suited to deliver, for example, audio AR outdoors.

The vast majority of systems have also concentrated on the *retrieval* rather than the *input* of content, and therefore an AR application is often described solely as a system where annotations are visualized when the user is at the right location. To differentiate an AR system's ability to both retrieve and input content in any area, we use the term *anywhere authoring*. This is an ability needed in applications that aim to take AR to the next level of impact e.g. a fine-grained city maintenance system, worldwide AR encyclopedias or wide area forensics.

To combine GPS and local visual mapping may appear to be sufficient for anywhere authoring. Unfortunately this is not the case, in part because users spend most of the time indoors where GPS positioning is unreliable at best. This seriously hampers AR for most of the places that can be annotated and places high requirements on the visual mapping that can work indoors. In order to offer truly wide and robust anywhere authoring it appears likely that a range of localisation technologies from GPS to indoor positioning systems jointly with visual mapping have to operate seamlessly as the user moves in and out of areas. This, combined with an adequate framework for the propagation of both

existing and newly created content, are crucial for enabling fluid AR interactions anywhere.

To our knowledge, a system that seamlessly combines these many levels and modalities of localisation accuracies and the ability to enable users to retrieve and input AR content anywhere has not been presented before.

2 Related Work

The combination of global and local sensing has been explored in the related field of ubiquitous computing for some time. As an example using computer vision, the works in [1, 2] use visual feature descriptors to provide accurate object detection while GPS helps in the gating of the objects' database based on location. In both examples, the objects of interest are buildings whose facades are usually distinctive, relatively large, and less prone to perspective and occlusion problems. To extend the area of operation for AR outdoors, GPS was also the natural choice and this was the case for early systems e.g. [3]. The further addition of inertial sensors, markerless visual tracking and aerial photographs to GPS in [4] has achieved higher accuracy annotation of large, outdoor scenes. More recently, in [5] GPS combined with inertial sensors is shown to be able to deliver relatively good visualization of underground pipes outdoors despite not using visual methods.

For wide area indoor AR, systems have used localisation methods that include ultrasonic positioning [6] or odometry recovered from the user's steps [7], as well as visual tags from the ARToolkit or similar to provide well localized annotations [8–10]. Another recent alternative indoors is Ultra Wide Band (UWB) which in [11] has been combined with fiducial markers to provide extended indoor operation. The combination of inertial sensors and visual markers has been used in [12]. In the case of [13] ultrasound and GPS are combined with visual SLAM and demonstrated in a small scale environment.

The use of a global reference provided by any of the above methods helps to improve the localisation results and prepares the scene for integration of technologies with different accuracy granularities. When the global frame of reference is not built-in, the extreme alternative is to use the visual appearance of each area of interest as the way to position the user. This is the case in [14] where a visual SLAM system creates small submaps that are kept disjointed and that are compared against an input image to detect that the user is in the same area once again. Assuming that no area looks exactly the same, this is a viable possibility, however the scalability of a system based on purely visual (even when combined with geometric) appearance, and disregarding any global reference, appears unrealistic. Furthermore, a system that can deliver true anywhere authoring is likely to encounter areas where no global reference either from indoor positioning or GPS is available and this demands an alternative referencing method.

While some of the above systems combine a few localisation techniques, none seems to have the seamless interaction over the different areas that we are after.

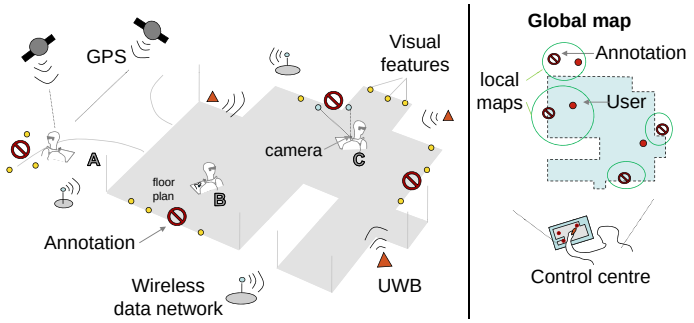

Fig. 1. System overview showing multiple users authoring a scene with AR annotations and using different localisation methods. See text for detailed explanation.

Importantly, none of them appear to be built with a multi-user and robust communications infrastructure for the input of annotations, as needed for anywhere authoring.

3 Operational Overview

Figure 1 shows an overview of the overall system in operation featuring three different modes of localisation: A) GPS, B) floor plan maps and C) UWB. The insertion and retrieval of annotations is made locally accurate by using visual SLAM. Figure 1 shows the SLAM features represented by yellow circles which serve as anchor points for the annotations. A communications infrastructure (in our case using WiFi and TETRA [15]), links users and allows visualization in a global map at a control centre. Local SLAM maps (green circles in global map) are positioned in the global reference with different accuracies depending on the positioning method at the time of authoring. However, visible annotations will always be displayed with local accuracy relative to the camera thanks to the automatic SLAM relocalisation, even if the location of the annotation in the global map is metrically inaccurate. The global map is used primarily as a topological representation for gating and rough navigational guidance.

4 Multi-modal Positioning

In this paper we divide the overall operation of the system into two main steps: i) locate the user in 3D space and ii) use a 6D referencing method to position accurately local AR annotations. The positioning of the user helps to establish a frame of reference that can later be used to provide only the relevant information for the immediate environment. This is the idea of location-based gating mentioned before. User positioning needs to be achieved on indoor and outdoor areas before we can combine it with an accurate local frame of reference.

GPS and UWB As with other systems, we employ GPS, when available, to provide an accepted alignment with an absolute frame of reference. Our GPS uses a Teseo GPS chipset to provide 3D positioning accurate to 2m with a 50%

confidence limit. For the indoors case we employ a UWB positioning system composed of multiple transponders [16]. These can be located indoors or outdoors and self-calibrate once they are active. In a typical indoor environment, the UWB system provides 3D positioning to at least metre-level accuracy, enabling the visualization of paths and places that users have visited. Accuracy varies according to the coverage of the UWB base units, which is affected by obstructions in the lines of sight between units and reflective surfaces in the environment that add multipath effects.

A rigid transformation can be found to align the UWB transponders with a reference from GPS, however when a global map is not required this alignment is not necessary. This is because even if these two references (GPS and UWB) are kept separate, the system can still determine at any instance if there is coverage by one or the other system and a decision can be made as to which reference will be used with priority (in our case it is UWB). Recall that an external reference is sought only for the task of gating which annotations should be near the user. This does not require an absolute or aligned set of frames of reference. In our system, switching between the UWB and GPS is transparent to users.

Interactive input In contrast to previous systems for wide-area AR, we employ user interactivity as a bridge to operate between the areas covered by GPS and UWB. For the case of a system designed for people, user input is a sensible alternative for positioning almost anywhere. When the user wishes to create an annotation, and when neither UWB nor GPS are available, the system prompts the user to refine location on a 2D map shown centered on the last trusted position fix. The user can then simply select an approximate location in this map. Our system uses street maps showing only the outlines of buildings (Fig. 7), but nothing prevents the use of more detailed map representations. The maps can also potentially be extended to include architectural floor plans if available.

By combining automatic referencing with the interactive user input we are able in principle to perform authoring anywhere.

5 Visual Mapping and Relocalisation

The requirement for working in unprepared, untagged environments has favoured the use of visual SLAM methods. Indeed, it was the construction of a local map for an AR scenario that was the first application of real-time visual SLAM. That system was based around an EKF process [17]. The PTAM system [18], a more recent take on the problem, uses bundle adjustment and splits the tasks of mapping and tracking to make gains from parallelization while delivering impressive results.

While the framework for mapping is important to the achievable accuracy, it is the way in which the system will re-localise in a previously visited area which is more critical for the application we are considering in this paper. Anywhere authoring demands a method that is able to work with efficiency over many local submaps while providing unambiguous camera pose recovery. This is important because although location-based gating helps to reduce ambiguity, a truly robust

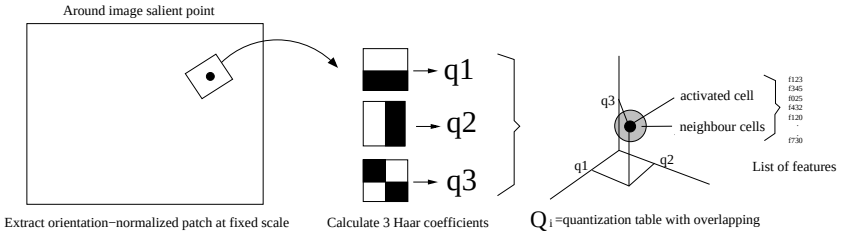

Fig. 2. The process of relocalisation from an input image in a single local map is based on the computation of three appearance coefficients per saliency point to approximate a nearest neighbor search using a quantization table.

system should be able to work when there is large uncertainty in the location of the user, perhaps when entering an area annotated using interactive input as described above, or if one of the other positioning systems fail.

In [19] a method is presented for visual SLAM relocalisation that uses randomized trees for re-detecting features, combined with a RANSAC verification step for pose estimation. Randomized trees are generated offline and use relatively large storage space — about 1.3MB per map point [14]. The PTAM system [18] uses a relocalisation method based on low resolution keyframes which has been used in the work of [14] for localisation over multiple maps. This approach is better from the point of view of data storage, however, in our experience, keyframe based localisation is prone to false positives, in particular when operating in roughly similar areas.

Another popular alternative is to use visual codebooks as used in [20] to match image frames. Visual codebooks are usually found after an optimization process of clustering and are therefore not easily updated on the fly, something which is corrected in [21].

In this paper we use the method for relocalisation and mapping described in [22]. This method uses robust visual descriptors and geometry consistency checks. The relocalisation is based on a quantization table which is small in comparison to other description approaches (e.g. using randomized trees) and can be updated on the fly. The method described in [22] was designed to work on a single map but in this work we extend that approach to work more efficiently with multiple maps as needed here and as described in Sec. 5.2. Furthermore, our method differs from the previous multiple map relocalisation work in [21] both in the smaller size of the descriptors used and in our use of a relatively small quantization table created only from the 3D features in our SLAM maps.

5.1 Single Map Relocalisation

Relocalisation assumes that a map M_i of features has been built previously and the 3D geometry of features together with their visual descriptors is available. To attempt to relocalize, a saliency detector is run on the input image. Around all image areas above a saliency threshold, a fixed-size window is used to obtain a rough estimate of local orientation. This local orientation allows extraction of a fixed-size patch from which three Haar coefficients are computed. These

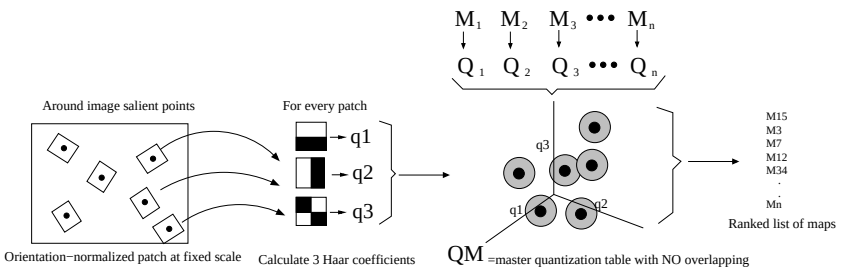

Fig. 3. When multiple maps have to be searched to attempt relocalisation, a master quantization table QM assists in the ranking of the maps to speed up the process.

coefficients encode the rough appearance of that patch in x, y, and xy. These numbers are used to index a quantization table Q_i where descriptors of other similar patches have been stored jointly with their 3D position, i.e. a cell c_{ij} in Q_i contains a list of features $F = \{f_k, \ldots, f_m\}$ generated by visual SLAM at the time M_i was created. In relocalisation, only the descriptors in c_{ij} and neighboring cells are compared with the input patch's descriptor. The process is illustrated in Fig. 2. The use of a fixed size patch here does not prevent working at different scales since the system builds a multi-scale stack of descriptors for every feature in a background process [22], and these are indexed too via Q_i.

After candidate matches are found with this procedure, a RANSAC method attempts to compute a consistent camera pose. If successful, and if an annotation linked to M_i is visible in the current frame, it will be displayed as an AR object.

In our tests, this approach uses only about 3% of the comparisions needed by an exhaustive search. The whole process is also fast, usually relocalizing within $50 - 300ms$.

5.2 Multiple Maps Relocalisation

When considering many local maps, the naive approach would be to run the above process in every M_i individually, perhaps gated by location. When the number of maps in a vicinity is small, that process may be sufficient but in general we would need to be prepared to run relocalisation on many maps to ensure robustness. To this end, we developed a system of map ranking based on the single map method described in Sec. 5.1 by combining the information of the individual Q_is as follows.

We create a master quantization table QM based on all the quantization tables Q_i from the local maps. This QM uses the same input as needed in the single map relocalisation. The process therefore starts with three Haar coefficients extracted around every salient point in the input image but in this case the coefficients are first used to index cells in QM. Every cell in QM keeps a list of the index i of all the maps M that have features in that cell. Therefore if a cell in QM is activated by an input patch, a list of all possible M_is that have to be searched is obtained. In addition, each cell is weighted by the **tf-idf** measure in a similar way as introduced in [20] to reflect the uniqueness of a cell. In this way cells that activate for every map will have a lower weight than

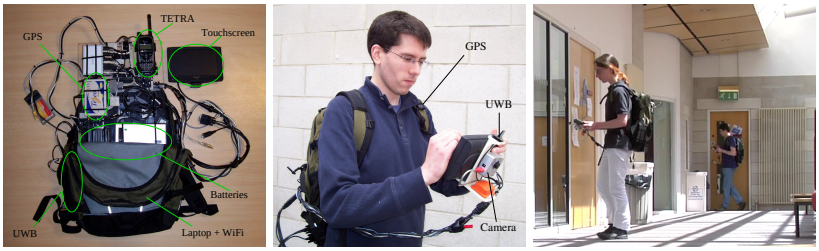

Fig. 4. Hardware components and multiple users exploring and annotating an area.

those that activate for fewer maps. By combining the weighted lists generated for every patch on the image it is possible to rank all maps according to the cosine similarity score between the **tf-idf** vectors for each map and the current image.

The process is illustrated in Fig. 3 and is very fast as we only need to look at the weighted frequency of i indices and rank them. The rank establishes the order in which relocalisation in the individual maps is to be attempted as per Sec. 5.1. When the first relocalisation is successful the process stops and switches to AR visualization, since in our experience the method does not produce false positives in real applications.

6 Experiments with Multiple Maps Relocalisation

Each hardware unit integrates components around a dual core Centrino laptop worn on a backpack as shown in Fig. 4. The interface with the user is displayed on a handheld touchscreen which has a firewire camera with a horizontal FOV of 80° rigidly attached to a 3D orientation sensor (which is not used in this work). The touchscreen also has the UWB antenna attached to it so that the most accurate sensors are close together. The GPS antenna is worn on the backpack's shoulder strap to enhance reception strength.

We performed experiments on the performance of the relocalisation in multiple maps. For this we assume the worst case where no location based gating is available. Experiments were conducted for an indoor scenario with 20 maps and an outdoor scenario with 5 maps, as shown in Figs. 5 and 6 respectively. We do not need to consider more maps than this, since the location based gating in the real system will always place a relatively low bound on the number of maps that need to be checked.

The performance of the map ranking was evaluated using camera tracking and exhaustive single-map relocalisation to provide a ground-truth estimate of the correct map for each frame. This was matched against the multiple map relocalisation ranking computed at each frame and used to plot the cumulative distribution function of the ranking. The results of the ranking method were then compared against the baseline case of a randomized sort of the maps.

Five different cell sizes for QM were tested. Although average performance was better than the baseline in all cases, the results showed that no single cell size gave good results for all maps. Sorting the maps by their mean rank over the five

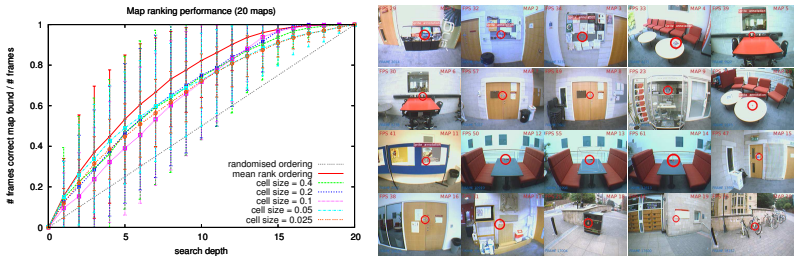

Fig. 5. Twenty maps were generated over a large indoor space incorporating many similar areas (several tables with red chairs). The cumulative distribution function of the ranking shows the improvement in performance achieved by the proposed method.

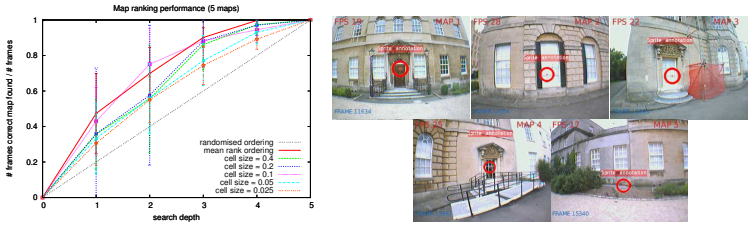

Fig. 6. Five maps were generated over a local outdoor area within a 10m radius representative of GPS accuracy. The cumulative distribution function of the ranking shows the improvement in performance achieved by the proposed method

different cell sizes improved the average performance and reduced the number of individual maps that performed worse than than the baseline. Alternative methods of combining the ranks from the different cell sizes, such as the median, minimum or maximum rank, were also considered but provided less performance improvement than the mean rank method.

In all cases, exhaustive relocalisation over all maps provided just a single positive match to the correct submap. This is despite the fact that the test sequences contain several instances of maps with very similar appearance. This supports the claim that the single map relocalisation method produces very low false positive rates in real scenes.

7 Demonstration

The performance of the system was demonstrated by building multiple maps over a 0.1km^2 area containing a mixture of indoor and outdoor locations. The scenario mimics a maintenance task where users label multiple objects to be revisited by other users at a later time. In some indoor locations a UWB positioning system was available to provide absolute position. The full set of 16 maps is shown in Fig. 7.

In areas with UWB coverage, a 2m distance threshold was used and the separation of the constructed maps was such that a maximum of one candidate map was returned for relocalisation. In one of the maps (map 3), the UWB accuracy was degraded by the surrounding furniture, producing position mea-

Fig. 7. Sixteen maps were created over an area containing a mixture of indoor and outdoor locations and with a mixture of GPS, UWB and User Input positioning. The 20m search radius reflects that the user is currently in an area using the interactive input positioning.

surements outside the expected distance threshold and preventing automatic relocalisation. However, single map relocalisation was successful when the map was selected manually from the user interface.

Areas with GPS coverage used a 10m distance threshold and returned a maximum of two candidate maps for relocalisation. In areas requiring interactive input to define absolute position, the 20m distance threshold returned between two and six candidate maps. The multiple map relocalisation method found the correct map within the first two maps tested on each of the six occasions it was used.

8 Conclusions

This paper has presented a novel system that combines a range of positioning technologies with local visual SLAM to enable the retrieval and creation of AR annotations. We have developed and evaluated a method for the efficient ranking of visual maps to improve performance and demonstrated the system operating over various areas in a maintenance-like scenario where multiple users cover an area finding and labelling objects practically anywhere in the environment.

Acknowledgement This work was funded by the UK Technology Strategy Board and the UK Engineering and Physical Sciences Research Council. The authors wish to thank all partners in the ViewNet project for their discussions and participation in this work. Ordnance Survey mapping © Crown copyright.

References

1. Fritz, G., Seifert, C., Paletta, L.: A mobile vision system for urban object detection with informative local descriptors. In: Int. Conf. on Computer Vision Systems. (2006)
2. Hutchings, R., Mayol-Cuevas, W.: Building recognition for mobile devices: incorporating positional information with visual features. Technical Report CSTR-06-017, Dept. of Computer Science, University of Bristol (2005)

3. Höllerer, T.: User Interfaces for Mobile Augmented Reality Systems. PhD thesis, Columbia University (2004)
4. Höllerer, T., Wither, J., Diverdi, S.: Anywhere augmentation: Towards mobile augmented reality in unprepared environments. In: Loc. Based Services and Tele-Cartography. (2007)
5. Schall, G., Mendez, E., Kruijff, E., Veas, E., Junghanns, S., Reitinger, B., Schmalstieg, D.: Handheld augmented reality for underground infrastructure visualization. Personal and Ubiquitous Computing **13** (2009)
6. Newman, J., Ingram, D., Hopper, A.: Augmented reality in a wide area sentient environment. In: Int. Symp. on Augmented Reality. (2001)
7. Kourogi, M., Sakata, N., Okuma, T., Kurata, T.: Indoor/outdoor pedestrian navigation with an embedded GPS/RFID/self-contained sensor system. In: Int. Conf. on Artificial Reality and Telexistence. (2006)
8. Wagner, M.: Building wide-area applications with the AR toolkit. In: Int. Augmented Reality Toolkit Workshop. (2002)
9. Reitmayr, G., Schmalstieg, D.: Location based applications for mobile augmented reality. In: Australasian User Interface Conference. (2003)
10. Nakazato, Y., Kanbara, M., Yokoya, N.: Localization system for large indoor environments using invisible markers. In: ACM Symp. on Virtual Reality Software and Tech. (2008)
11. Newman, J., Schall, G., Barakonyi, I., Andreas, S., Schmalstieg, D.: Wide-area tracking tools for augmented reality. In: Int. Conf. on Pervasive Computing. (2006)
12. Wormell, D., Foxlin, E., Katzman, P.: Advanced inertial-optical tracking system for wide area mixed and augmented reality systems. In: Int. Immersive Projection Tech. Workshop/Eurographics Workshop on Virtual Environments. (2007)
13. Banwell, T., Calway, A.: Combining absolute positioning and vision for wide area augmented reality. In: Int. Joint Conf. on Computer Vision, Imaging and Computer Graphics Theory and Applications. (2010)
14. Castle, R., Klein, G., Murray, D.: Video-rate localization in multiple maps for wearable augmented reality. In: Int. Symp. on Wearable Computers. (2008)
15. Efthymiou, C., Gormus, S., Fan, Z., Calway, A., Mayol-Cuevas, W., Doufexi, A.: Application of multiple-wireless to a visual localisation system for emergency services. In: IEEE Int. Symp. on Personal, Indoor and Mobile Radio Communications. (2010)
16. Harmer, D., Russell, M., Frazer, E., Bauge, T., Ingram, S., Schmidt, N., Kull, B., Yarovoy, A., Nezirović, A., Xia, L., Dizdarević, V., Witrisal, K.: EUROPCOM: emergency ultrawideband radio for positioning and communications. In: IEEE Conf. on Ultra-Wideband. (2008)
17. Davison, A., Mayol, W., Murray, D.: Real-time localisation and mapping with wearable active vision. In: Int. Symp. on Mixed and Augmented Reality. (2003)
18. Klein, G., Murray, D.: Parallel tracking and mapping for small AR workspaces. In: Int. Symp. on Mixed and Augmented Reality. (2007)
19. Williams, B., Klein, G., Reid, I.: Real-time SLAM relocalisation. In: Int. Conf. on Computer Vision. (2007)
20. Sivic, J., Zisserman, A.: Video google: A text retrieval approach to object matching in videos. In: Int. Conf. on Computer Vision. (2003)
21. Eade, E., Drummond, T.: Unified loop closing and recovery for real time monocular SLAM. In: British Machine Vision Conf. (2008)
22. Chekhlov, D., Mayol-Cuevas, W., Calway, A.: Appearance based indexing for relocalisation in real-time visual SLAM. In: British Machine Vision Conf. (2008)