COMBINING ABSOLUTE POSITIONING AND VISION FOR WIDE AREA AUGMENTED REALITY

Tom Banwell, Andrew Calway
Department of Computer Science, University of Bristol, Bristol, U.K.
{*tom,andrew*}*@cs.bris.ac.uk*

Keywords: Visual SLAM, Localisation, Mapping, Absolute Positioning, Augmented Reality, Sensor Fusion.

Abstract: One of the major limitations of vision based mapping and localisation is its inability to scale and operate over wide areas. This restricts its use in applications such as Augmented Reality. In this paper we demonstrate that the integration of a second absolute positioning sensor addresses this problem, allowing independent local maps to be combined within a global coordinate frame. This is achieved by aligning trajectories from the two sensors which enables estimation of the relative position, orientation and scale of each local map. The second sensor also provides the additional benefit of reducing the search space required for efficient relocalisation. Results illustrate the method working for an indoor environment using an ultrasound position sensor, building and combining a large number of local maps and successfully relocalising as users move arbitrarily within the map. To show the generality of the proposed method we also demonstrate the system building and aligning local maps in an outdoor environment using GPS as the position sensor.

1 INTRODUCTION

A fundamental requirement for Augmented Reality (AR) applications is to be able to localise the pose of a mobile device with respect to the physical environment. In the past, work in this area has focused primarily on localisation based on known structure in the form of calibrated targets (Piekarski et al., 2004) or models (Park et al., 2008; Pupilli and Calway, 2006). However, more recent work has attempted to overcome the limitations of this by employing techniques able to operate in previously unseen environments. Of particular interest has been the significant advances made in vision based simultaneous localisation and mapping (SLAM) systems which have their roots in the Robotics literature, see for example the monocular systems described in (Davison et al., 2007; Chekhlov et al., 2006; Klein and Murray, 2007). These systems have now reached a level of robustness where they can be reliably used in a variety of AR applications (Chekhlov et al., 2007; Castle et al., 2008).

Although the robustness and reliability of these visual SLAM systems is impressive, the challenge of building very large maps over wide areas still remains. One limiting factor is computational effort, which for most algorithms increases quadratically with the number of features in the map. This can be addressed by adopting sub-mapping techniques to build consistent maps over relative wide areas (Clemente et al., 2007; Pinies and Tardos, 2008). Clemente et al (Clemente et al., 2007) were the first to apply the sub-mapping technique to a monocular-camera system. They demonstrated impressive results in an outdoor courtyard, building sub-maps of limited size before initialising a new map referenced to the current pose. This allowed sub-maps to be treated as statistically independent. Remapping common features in two sub-maps enables loop closure and the sub-maps to be joined into a single global map. Pinies and Tardos (Pinies and Tardos, 2008) extended the sub-mapping method to be able to share common information in the form of state components between sub-maps. This reduces the loss of information caused by map initialisation. Although these systems enable a wider area to be mapped, they assume continuous texture between the sub-maps and cannot handle situations where there is no texture or when the camera has been kidnapped.

To overcome these limitations Castle et al (Castle et al., 2008) developed a system that assumes there is not always continuous texture between sub-maps. Their system enables a user to build sub-maps in areas of interest and relocalise in those maps once they return to them. This allows users to build sub-maps of spatially separated areas. However, the relative location and orientation of the sub-maps are not known and relocalisation may become an issue once there are a large number of maps.

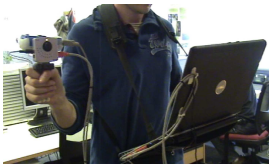


Figure 1: User with Mobile Device and Laptop Desk.

A kidnapped camera can provide no information about its current position relative to a previous map until it reobserves features from that map. One method to recover such information would be to seek support from an additional sensor. Pinies et al (Pinies et al., 2007) combine their monocular-camera system with an Inertial Measuring Unit, this improves their estimated trajectory and map. Newman et al (Newman et al., 2001) developed a building-wide AR system by combining ultrasound position estimates and rotation estimates from an inertial tracker. This allows the wearer of a head mounted display to be positioned, but does not provide detailed information about the environment they are in. The main focus of these systems has been improving the overall system robustness or improving the accuracy of the estimate. In this work we overcome these limitations by using a second sensor, an absolute positioning system. During areas of texture we simultaneously estimate our global trajectory using the absolute positioning system and our local trajectories using a monocular-camera system. Using a least-squares approach we estimate the transformation between the global and local trajectories and build a single scalable consistent global map. During periods of loss of visual information we use the absolute positioning system to estimate 3-D position within the global map. This allows the system to efficiently relocalise, when returning to a previously built local map, even when there are a very large number of maps.

Our results show that by using a second sensor we can overcome these limitations of scalability and the need for continuous texture, and that it is possible to build many local maps that are known relative to one another. This provides us with a novel AR capability, to be able to track in a local map and see AR in other local maps that are not joined to the current local map by continuous texture.

The remainder of the paper is organised as follows. The next section describes the general framework for combining a mapping system with an absolute positioning system. The third section describes the implementation we have developed based on this framework. Conclusions are drawn in the final section.

2 POSITIONING AND VISION

This section describes the core method underlying our work described in a general framework. One of the advantages of our method is that it is directly applicable to any camera mapping system, for example; probabilistic and Structure from Motion approaches and any absolute positioning system, for example; Ultrasound, Ultra-Wideband and GPS.

Another advantage of this work is that we can create a very large number of local maps at an arbitrary position, orientation and scale and combine all of these maps into a single coordinate frame to provide a single scalable global map. Crucially, the estimated map and trajectory from the local mapping system are locally correct, but a transformation out, when compared to the estimate in the global coordinate frame. By simultaneously estimating this local trajectory and the global trajectory we are able to estimate this alignment transformation and combine the local maps into the global coordinate frame.

2.1 Estimating Transformations

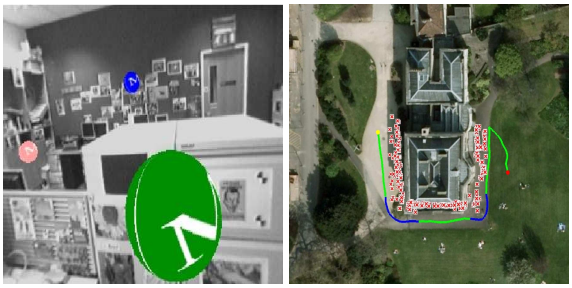
We begin by estimating the motion of a mobile device using two sensors; a monocular-camera and an absolute positioning system. The two sensors are rigidly attached at a known offset. Each sensor estimates the mobile's trajectory in their own coordinate frame; our goal is to recover the transformation between the two coordinate frames.

To estimate the transformation we need to estimate two trajectories, propagated to a common time. After each measurement update a new position is estimated and this is stored in the trajectory set. This provides us with two sets of data, or trajectories, X for the vision system and Y for the absolute positioning system.

Now we have the two trajectories we estimate the transformation between them. We find it best to estimate the transformation parameters over the full trajectories once the local map has been 'finished'. To estimate the desired transformation we use a least-squares approach introduced by Umeyama (Umeyama, 1991). This method is based on Singular Value Decomposition, which is known for its numerical stability. We now give a brief overiew of the method. First we find the means and variances of the two trajectories and then the covariance Σ_{xy}, between the two. We then determine the singular value decomposition of Σ_{xy} which gives UDV^T. From these matrices we follow the steps described in Umeyama (Umeyama, 1991) to estimate the rotation R, translation t, and scale s, of the transformation. These are the parameters of the transformation required to convert the local map into the global map.

2.2 Building the Global Map

Once the transformation parameters s_j, R_j and t_j have been estimated for local map j, they can be used to



(a) Relocalised Camera (b) Mapping with GPS

Figure 2: (a) Relocalised within a Local Map: The mean-spheres in neighbouring maps can be seen. (b) Local maps aligned with a truly global coordinate frame.

transform the local map into the global coordinate frame. The trajectory X_{lj} of local map j, is transformed into the global coordinate system X_{gj} as follows:

$$X_{gj} = s_j R_j X_{lj} + t_j \quad (1)$$

Each local-feature f_{li}, in local map j, can be transformed to the global coordinate frame f_{gi} as follows:

$$f_{gi} = s_j R_j f_{li} + t_j \quad (2)$$

3 EXPERIMENTAL RESULTS

To demonstrate our method we conducted experiments both indoors, using an ultrasound system, and outdoors using GPS.

3.1 Indoor Office Environment

The hardware was setup as shown in Figure 1. The user has a handheld mobile device that consists of two sensors that are rigidly attached at a known offset. The two sensors are; a calibrated handheld camera with 320x240 pixels and wide-angled lens, and an ultrasound receiver. To enable mobility a laptop desk is used to carry the laptop. An ultrasound positioning system is used to provide estimates of 3-D position (Randell and Muller, 2001). To provide the local maps and trajectories we use the visual-SLAM system developed by Chekhlov et al (Chekhlov et al., 2006).

3.1.1 Building and Correcting Local Maps

Testing was performed in an indoor environment. A sequence of steps can be seen in Figure 3 (the full experiment can be seen in the attached video). The ultrasound system estimates the 3-D position of the mobile device in the global coordinate frame. Although in practice one would want to start a new map after a loss of visual track, for ease of testing we allow the user to control the building of local maps. The user provides input to start building a local map. Once the user has decided they have finished building the local map they again provide input and the system stops building the current local map. At this point the system estimates the transformation to align the local trajectory with the global trajectory and applies that transformation to the local estimate.

3.1.2 Viewing Into and Across Local-Maps

One of the major advantages our method offers for AR applications is the ability to see across disjoint maps. To demonstrate this we placed rotating spheres at the mean of aligned maps, as can be seen in Figure 3. After aligning six local maps the user entered the relocalisation phase. The mobile returned to and relocalised in the second map (green sphere). As this local map had been globally aligned the camera's global position could be estimated. All global graphics can be seen in the camera view allowing it to 'look-across' maps and see the contents of those maps. This is shown in Figure 2a. This novel contribution is not achievable with other current single-camera systems.

3.1.3 Efficient Relocalisation

To demonstrate the second contribution of our work we improve the relocalisation method developed by Chekhlov et al (Chekhlov et al., 2008). Between building local maps, if the user returns to a previously mapped area, we seek to relocalise the camera. Once there are a very large number of maps, it is unrealistic to attempt to relocalise in all maps. Our contribution comes from the fact that our method provides us with an estimate of the global position of the mobile and all local maps. We use this information to decide which local map(s) we should attempt to relocalise in. Once the current local map has been 'finished' the system enters a relocalisation phase. A map is selected as a potential for relocalisation if it falls within a sphere centered about the mobile's position. The sphere defines a region where there is a significantly high chance of being able to detect features from the map, it's radius is estimated based on the distance between the mobile's current position and the mean of the local map.

3.2 Combining GPS and Vision

To demonstrate the generality of our method we have applied it using an alternative sensor, GPS. Using GPS we have the potential to scale across very wide areas and the entire global. This enables users to share maps and combine them with applications such as Google Earth. Using the same mapping system as described in section 3.1 and the method of section 2 we tested the system outside by walking around a building. Each wall of the building was locally mapped using the vision system, once finished the local map was aligned with the global coordinate frame to update the global map. This can be seen in Figure 2b.

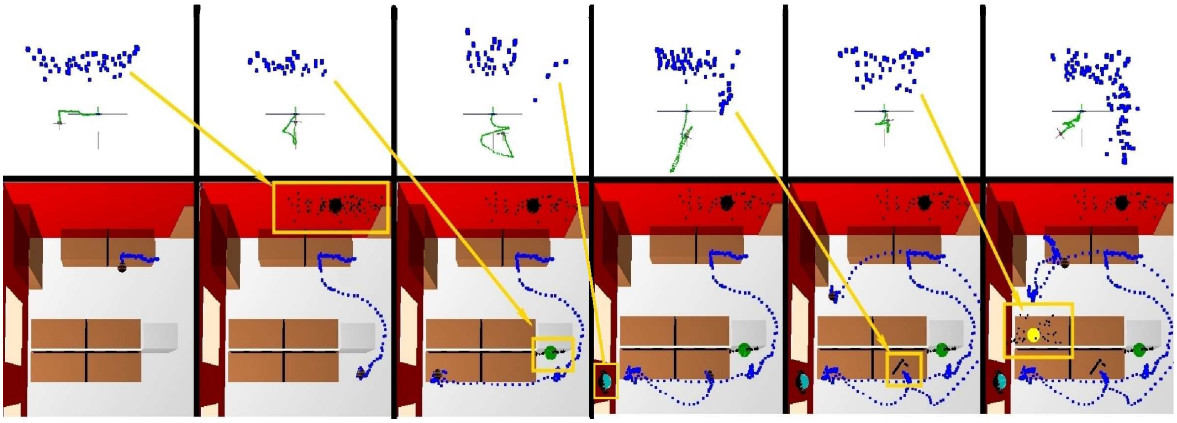


Figure 3: Screenshots from a map building sequence. The top row shows the building of local maps, with the locally estimated trajectory (green) and features (blue). The bottom row shows the same trajectory estimated in the global coordinate frame (blue dotted-line) and the correctly transformed local maps (black). The arrows (yellow) represent the correct transformation of the local maps. The spheres (black, green, cyan and yellow) represent the means of the local maps.

4 CONCLUSIONS

This work developed a new scalable mapping and localisation technique. It combines trajectories from an absolute positioning system and a local mapping system to produce a single map. The key contributions are the ability to map in areas where there is no continuous texture and the correct alignment of local maps into a single global coordinate system. The position and orientation of the local maps are known relative to each other. This provides the ability to view across disjoint maps and improves the efficiency of relocalisation. We demonstrated our method by building and transforming local maps to a global coordinate frame using different sensors both indoors and outdoors. Future work will focus on auto-calibrating the ultrasound system to reduce the installation cost.

REFERENCES

Castle, R. O., Klein, G., and Murray, D. W. (2008). Video-rate localization in multiple maps for wearable augmented reality. In *Proc 12th IEEE Int Symp on Wearable Computers 2008*, pages 15–22.

Chekhlov, D., Gee, A., Calway, A., and Mayol-Cuevas, W. (2007). Ninja on a plane: Automatic discovery of physical planes for augmented reality using visual slam. In *International Symposium on Mixed and Augmented Reality (ISMAR)*.

Chekhlov, D., Mayol-Cuevas, W., and Calway, A. (2008). Appearance based indexing for relocalisation in real-time visual slam. In *19th Bristish Machine Vision Conference*, pages 363–372.

Chekhlov, D., Pupilli, M., Mayol-Cuevas, W., and Calway, A. (2006). Real-time and robust monocular slam using predictive multi-resolution descriptors. In *2nd International Symposium on Visual Computing*.

Clemente, L., Davison, A., Reid, I., Neira, J., and Tardos, J. (2007). Mapping large loops with a single hand-held camera. In *Robotics: Science and Systems*.

Davison, A., Reid, I., Molton, N., and Stasse, O. (2007). MonoSLAM: Real-Time Single Camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1052–1067.

Klein, G. and Murray, D. (2007). Parallel tracking and mapping for small AR workspaces. In *Proc. Sixth IEEE and ACM International Symposium on Mixed and Augmented Reality (ISMAR'07)*, Nara, Japan.

Newman, J., Ingram, D., and Hopper, A. (2001). Augmented reality in a wide area sentient environment. In *Augmented Reality, 2001. Proceedings. IEEE and ACM International Symposium on*, pages 77–86.

Park, Y., Lepetit, V., and Woo, W. (2008). Multiple 3d object tracking for augmented reality. In *Proc. Seventh IEEE and ACM International Symposium on Mixed and Augmented Reality*, pages 117–120.

Piekarski, W., Avery, B., Thomas, B., and Malbezin, P. (2004). Integrated head and hand tracking for indoor and outdoor augmented reality. pages 11–276, Chicago, IL.

Pinies, P., Lupton, T., Sukkarieh, S., and Tardos, J. D. (2007). Inertial aiding of inverse depth slam using a monocular camera. In *IEEE International Conference on Robotics and Automation, Roma, Italy*.

Pinies, P. and Tardos, J. (2008). Large-scale slam building conditionally independent local maps: Application to monocular vision. *IEEE Transactions on Robotics*, 24(5):1094–1106.

Pupilli, M. and Calway, A. (2006). Real-time camera tracking using known 3d models and a particle filter. In *International Conference on Pattern Recognition*.

Randell, C. and Muller, H. (2001). Low cost indoor positioning system. In *Ubicomp 2001: Ubiquitous Computing*, pages 42–68.

Umeyama, S. (1991). Least-Squares Estimation of Transformation Parameters Between Two Point Patterns. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 13, pages 376–380.