Markerless Augmented Reality using Image Mosaics

Pietro Azzari, Luigi Di Stefano, Federico Tombari, Stefano Mattoccia

ARCES - DEIS, University of Bologna,
viale Risorgimento 2, 40125 Bologna, Italy
{pazzari,ldistefano,ftombari,smattoccia}@deis.unibo.it
http://www.vision.deis.unibo.it/

Abstract. Augmented reality is a powerful tool for delivering spatially coherent information to a user moving in a known environment. Accurate and reliable pose estimation is the key to success. Many approaches track reference objects into the scene but as the environment grows larger more objects need to be tracked leading to computationally intensive methods. Instead, we propose a practical approach that is suitable for environment where big planar structures are present. All the objects laying on the structure are composed into a large reference object using image mosaicing techniques, so that the problem is reduced to that of finding the pose from a single plane. Experimental results show the effectiveness of this approach on two interesting case studies such as aeronautical servicing and cultural heritage.

Key words: Mosaicing, Augmented reality, Pose estimation, Markerless

1 Introduction

Augmented reality techniques convey information that is both semantically and spatially coherent with the observed scene. Information is shown by augmenting the scene captured through a camera with graphical objects that are properly aligned with the world 3D structure and often contextually close to the user needs. In this paper we mainly focus on structural coherence, nonetheless a simple demonstration of contextual awareness is given in the experimental results section.

The capability to deliver spatially coherent information to a user moving in a known environment is enabled by accurate and reliable pose estimation algorithms. Such algorithms try to compute the pose of the observer with respect to the world he is moving in by establishing correspondences among objects detected in the scene. Using those correspondences both the information to be displayed and the structure of the scene is estimated.

Most of the algorithms described in literature can be thought of in terms of a binary taxonomy: those that rely on absolute information [1, 2], such as known models, and those based on chained transformations [3, 4]. The former seek to find camera poses that correctly reproject some fixed features of a given $3D$

model into the $2D$ image. They do not drift but they often lack precision which results in jitter. The latter do not exploit a priori information but match interest points between images. Since the correspondences between adjacent frames are usually located very precisely, these algorithms do not jitter but suffer from drift or even loss of track.

Pose estimation algorithms usually represent the world as a collection of reference objects, modeled as $3D$ meshes, associated with appearance models, such as collection of key frames or image patches related to each vertex. Navigation of large environments is handled using several objects so that many of them are visible even though the user moves widely across the environment. Many algorithms are known to estimate the pose very quickly using a single object and a single image [1, 2]. However, in presence of several objects, the pose of the observer is optimized together with the relative position of the visible objects typically using temporal coherence constraints, i.e. objects projections in different images are expected to suggest the same act of motion. As the environment grows larger so does the number of required objects, thus yielding to computationally intensive algorithms.

To reduce the complexity Simon et al. [3] and Uematsu et al. [4] considers only planar reference objects. In this settings they can exploit both temporal and spatial coherence in the estimation, i.e. homographies between planes can be computed independently and added as additional constraints. This involves constructing at each frames a unified projective space and mapping all the planes to that space according to the computed homographies. The pose is subsequently computed using correspondences from that space and the image projections.

Nonetheless when several planar reference objects are also coplanar, the unified projective space can be profitably built in advance using image mosaicing techniques. As the cluster of objects becomes larger, using a mosaic as appearance model instead of a single shot, taken from larger distance or with shorter focal length, becomes more and more useful. In fact, the mosaic approach allows to maintain plenty of details that a single shot would miss.

We propose a practical approach that is suitable for environment where big planar structures are present. By mosaicing several planar objects during a training stage we shift off-line most part of the computation of the pose. At run-time, the algorithm simply determines the pose with respect to a single big reference object using approaches, such as [1, 2, 5], that are known to be fast and robust. This notably diminishes the on-line computational requirements and increases the accuracy of the estimated pose.

2 Methodology

The method is split up in two distinct stages. The first can be regarded as a training phase and it is performed off-line. It deals with the definition of a big planar reference object together with the construction of its appearance model, i.e. a mosaic of images that portray the planar structure. Several keypoints are extracted from the appearance model using the well known SIFT features

detector [6]. Metric measurements can be easily introduced in this framework by specifying the real world position of at least four non collinear points within the planar objects and computing the metric to projective homography accordingly.

The second stage performs on-line and addresses the estimation of the pose of the observer at a given instant using a set of points correspondences between the visible scene and the constructed appearance model. This stage is composed of a feature tracker that finds point matches and any chosen pose estimation algorithm based on point correspondences. The projection of virtual objects is easily accomplished once the pose is known.

2.1 Construction of the appearance model

The first stage concerns the construction of the big reference object and its appearance model from a collection of pictures using a mosaicing algorithm. The idea of using mosaics in augmented reality applications is not a novelty in itself. For instance, Dehais et al. [7] use mosaics to augment the scene with virtual objects. However, with their system the user is allowed to rotate only and both the training and the testing sequence must be captured from the same vantage point. The approach proposed by Liu et al. [8] is also based on image mosaicing, but it requires fiducial markers and the viewpoint is again allowed to rotate only. Instead, our method relies on natural markers present in the scene and allows for every kind of motion as long as a portion of the model is visible to the observer.

During a training stage we construct the appearance model using several views of a roughly planar structure in the scene. The transformations among the views are homographies as long as the observed subject is planar. The algorithm we use to mosaic images can be regarded as an iterative version of the pairwise DLT method described in [9]. From each pair of views we compute a set of point correspondences and fit the best homography in the least square sense. Then we repeat this procedure for all the pairs and concatenate the homographies. This can be seen as the common projective space computed by [4] when all the patterns are coplanar.

Instead of building a mosaic one might also capture the whole planar structure with a single shot taken from a larger distance or with a shorter focal length and then use such a shot as the appearance model. Indeed, this choice is potentially preferable when, given the resolution of the acquisition device, objects are as small as they can be captured by a single shot without losing too much information. In fact, in such a case objects are already registered with respect to each other and taking a picture is quicker than building a mosaic. However, in any application scenario the more appropriate approach should be identified carefully. In the experimental results section we propose a comparison between the two approaches considering two different case studies.

Finally, given the appearance model, the SIFT feature detector extracts a set of keypoints p_i from it. Extracted features that appear in the model but do not belong to the planar reference object are discarded using outlier removal techniques such as Ransac.

2.2 Pose estimation and augmentation

Pose estimation from point correspondences for calibrated perspective projection cameras has been extensively studied in literature. To demonstrate the effectiveness of our proposal we choose two well known algorithms that address the problem from very diverse points of view.

The pose estimation problem can be stated as that of estimating the rigid transformation, made up of a rotation matrix R and a translation vector t, that relates a set of noncollinear 3D coordinates of known reference points p_i with their corresponding normalized projections (u_i, v_i) so that:

$$\begin{aligned} u_i &= \frac{R^1 p_i + t_x}{R^3 p_i + t_z} \\ v_i &= \frac{R^2 p_i + t_y}{R^3 p_i + t_z} \end{aligned} \tag{1}$$

where $p_i = (x_i, y_i, z_i)$ are expressed in an object-centered frame, R is 3×3 orthonormal matrix and t is a 3×1 vector.

In these settings, the first algorithm we consider is that illustrated by Simon et al. [5], which has been considered for long the classical photogrammetric formulation. In practice they solve for the unknown pose by optimizing the following objective function:

$$\sum_i^N \| \left(\widehat{u_i} - \frac{R^1 p_i + t_x}{R^3 p_i + t_z} \right), \left(\widehat{v_i} - \frac{R^2 p_i + t_y}{R^3 p_i + t_z} \right) \|^2 \tag{2}$$

where $\widehat{u_i}, \widehat{v_i}$ are measured image points. This computation minimizes the error distance among projections in the image space. In place of the sequential estimation proposed in their paper, we compute the pose of each frame with respect to our appearance model thus avoiding potential estimation drift issues.

From a theoretical viewpoint, an equivalent reformulation of the problem consists in estimating (R, t) that relates the known reference points p_i with the corresponding q_i so that:

$$q_i = Rp + t \tag{3}$$

where $p_i = (x_i, y_i, z_i)$ and $q_i = \left(x_i^{'}, y_i^{'}, z_i^{'} \right)$ are expressed in an object-centered and camera-centered reference frame respectively. Based on this viewpoint, the second algorithm, proposed by Schweighofer et al. [1], aims at minimizing an object space distance error by means of the line-of-sight projection matrix $\widehat{V_i}$ (for further details refer to [1]). This algorithm yields the best results according to a recent analysis of the state-of-the-art carried out in [2].

Once the pose is retrieved it is then possible to project $3D$ models in the image according to (R, t) and the known camera intrinsics.

3 Experimental results

This section reports the performance of the considered algorithms in two different case studies. Performance are measured in terms of estimation steadiness and

smoothness. Under this perspective the most stable the estimated pose over time the better the algorithm. In the following we plot the recovered position of the camera center's coordinates $O^C = \left(O_X^C, O_Y^C, O_Z^C\right)$ expressed in the object-centered frame. Both algorithms are run twice on each sequence with different appearance models, the first time using a single image (Fig. 1 top), the second time using a mosaic (Fig. 1 bottom). All the frames used to build the models do not belong to the test sequences.

The two test sequences have been acquired by a freely moving observer using a webcam (Logitech Quick Cam Sphere). Each sequence is about 600 frames long and the images have a resolution of 640×480 pixels.

Fig. 1. Small (top) and large (bottom) appearance models

3.1 Aeronautical servicing

The first case study is drawn from a collaborative research project addressing the application of Augmented Reality to the field of aeronautical servicing. The ultimate aim of the project is to equip engineers with see-through helmets by which a context-aware system will act as a virtual assistant providing information on the maintenance procedure in real-time using augmented reality. The sequence portraits the inside of a cockpit of a plane. Useful information in this context concerns the position of the most important switches and leverages as well as instructions on how to use them properly (refer to Fig.3 for some examples).

In the upper row of Fig.2 the position of O^C according to the pose estimated using a small appearance model is reported. While the pose is correct most of the time, the peaks in the plots denote that the estimation suffers from jitter. It is worth pointing out that both pose estimation methods are affected by these peaks approximately in the same way. Conversely, the plots in the lower row of

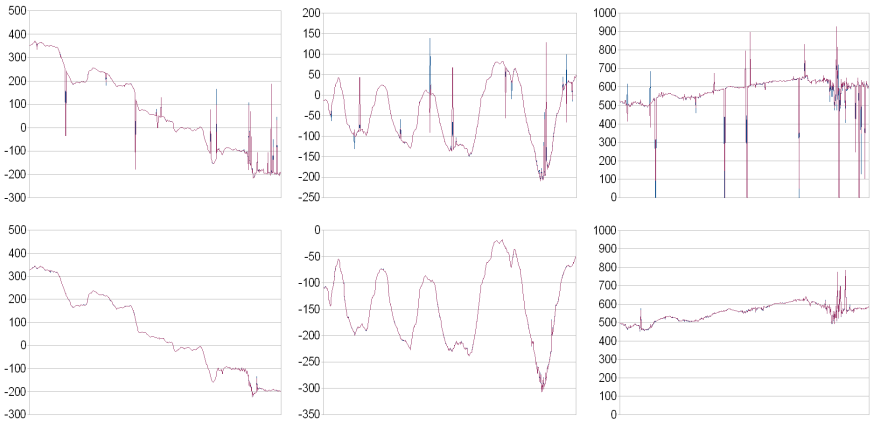

Fig. 2. Recovered camera center's coordinates using small (top) and large (bottom) appearance models: Schweighofer et al. (violet), Simon et al. (bordeaux). Left to right: O_X^C, O_Y^C, O_Z^C.

Fig.2 show that, when using the mosaic as appearance model, the estimated pose exhibits a much smoother behaviour and jitter is almost completely eliminated, with the exception of some creases regarding the z component. It is also worth noticing how the accuracy is not affected by the considerable lighting changes occurring in the scene, as shown by Fig.3.

Fig. 3. Augmented cockpit sequence samples.

3.2 Cultural heritage

The second case study concerns the field of advanced context-aware systems for delivering information to visitors of museums or archaeological sites. The considered sequence displays a showcase with Etruscan jewellery. Fig. 5 shows that the pose of the observer with respect to the showcase is accurately retrieved, as vouched by the coloured outlines superimposed on the borders of the shelves.

Besides, additional context aware information is conveyed by highlighting the object that is likely to be the most important for the user given his position and orientation.

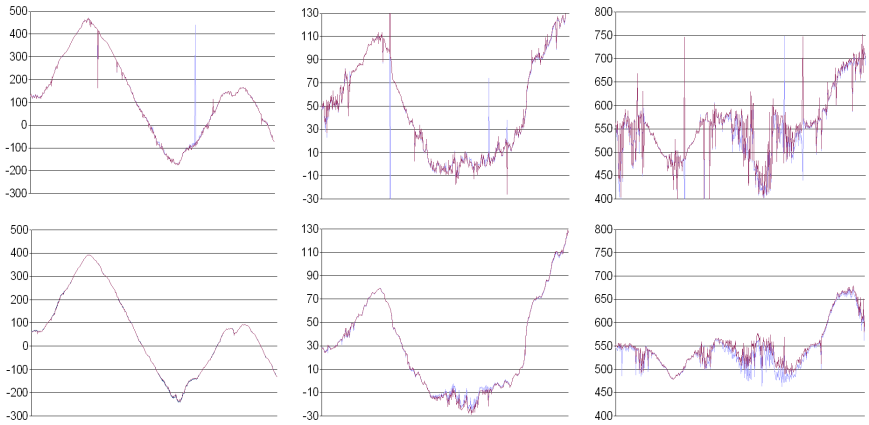

Fig. 4. Recovered camera center's coordinates using small (top) and large (bottom) appearance models: Schweighofer et al. (violet), Simon et al. (bordeaux). Left to right: O_X^C, O_Y^C, O_Z^C.

As before, the estimation using a small appearance model is quite good but suffers from jitter (as it can be seen in the upper row of Fig.4). When using the mosaic (lower row of Fig.4), jitter mostly disappears and, unlike previous experiment, the pose is smoother even when there are no macroscopic estimation error. Similarly to the cockpit sequence, the z component exhibits the worst reconstruction quality because the model is mostly observed from ahead with limited tilt angles, thus ill-conditioning the optimization procedure.

Fig. 5. Augmented jewel sequence samples.

4 Conclusions

In this paper we have presented a practical approach to augmented reality that is suitable for environments where big planar objects are present. Instead of modeling the reference objects using a single image or a set of independent images, we propose to build a mosaic by registering several detailed views. The pose is then estimated from the correspondences between the actual frame and the appearance model of the reference planar object using any chosen pose estimation algorithms. The experiments demonstrate that two very different pose estimation algorithms largely benefit from the proposed approach. In this sense our proposal can be thought as a preprocessing step able to improve the computational performance and accuracy of any pose estimation algorithms. The major limitation of the proposed approach is represented by the planar structure constraint.

References

1. Gerald Schweighofer and Axel Pinz. Robust pose estimation from a planar target. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(12):2024–2030, 2006.
2. F.Moreno-Noguer, V.Lepetit, and P.Fua. Accurate non-iterative o(n) solution to the pnp problem. In *IEEE Intl. Conf. on Computer Vision*, Rio de Janeiro, Brazil, October 2007.
3. Gilles Simon and Marie-Odile Berger. Real time registration of known or recovered multi-planar structures: application to ar. In *in Proc. of BMVC*, 2002.
4. Yuko Uematsu and Hideo Saito. Vision-based registration for augmented reality with integration of arbitrary multiple planes. In *Proc. of ICIAP*, pages 155–162, 2005.
5. Gilles Simon, Andrew W. Fitzgibbon, and Andrew Zisserman. Markerless tracking using planar structures in the scene. In *Proc. of ISAR*, pages 120–128, Munich, Germany, May–June 2000.
6. D. Lowe. Distinctive image features from scale-invariant keypoints. *Intl. J. of Computer Vision*, 60(2):91–110, November 2004.
7. Christophe Dehais, Matthijs Douze, Geraldine Morin, and Vincent Charvillat. Augmented reality through real-time tracking of video sequences using a panoramic view. In *Proc. of ICPR*, pages 995–998, 2004.
8. Peiran Liu, Xiaoyong Sun, Nicolas D. Georganas, and Eric Dubois. Augmented reality: A novel approach for navigating in panorama-based virtual. In *Proc. of HAVE*. unknown, 2003.
9. R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Second Edition, 2003.