SINGLE IMAGE AUGMENTED REALITY USING PLANAR STRUCTURES IN URBAN ENVIRONMENTS

Eric McClean, Yanpeng Cao, John McDonald

Department of Computer Science,
National University of Ireland Maynooth,
Maynooth, Co. Kildare, Ireland.

***Abstract*—In this paper, we present an effective method for integrating 3D augmented reality graphics into single images taken in urban environments. Building facades usually contain a large number of parallel lines aligned along several principal directions. We make use of the images of these 3D parallel lines and their corresponding vanishing points to recover a number of 3D planes from single 2D images of urban environments and then use them to represent the spatial layout of a scene. 3D objects are then aligned with respect to these recovered planes to achieve realistic augmented reality effects. In experiments we applied the proposed method to implement augmented reality in images from a benchmark image dataset of urban environments.**

I. INTRODUCTION

In this paper we present an approach to augmented reality based on single images. The motivation of our work is to develop vision techniques for image based navigation. In particular we aim to deploy the resulting technology in locality scale regions (e.g. 1km2) of general urban environments. Here the user should be able to capture an arbitrary image within the environment and have the system graphically augment it with relevant navigation and point-of-interest (POI) information. The underlying approach is to automatically relate the unknown input image to a database of stored landmark building images, and from the resulting match, compute the user's camera pose.

In such a system robust alignment of the real world and virtual objects is a crucial step, in which we need to know the mapping relation between a photo and its imaged 3D scene. Previously, a number of techniques [1, 2, 3, 4, 5, 6] have been proposed to solve this problem. These methods compute the full 3D camera motion based on geometric constraints between feature correspondences within multiple images (e.g. via the fundamental matrix [7]). Furthermore, these methods allow recovering the underlying 3D structure of a scene given a number of collected photographs. This procedure is usually referred to as Structure-from-Motion (SfM) [1, 2, 3, 7]. The main drawback of such methods is that they require a set of related images of a scene taken at similar viewpoints to establish frame-to-frame correspondences. Moreover these methods require batch-processing of the entire image sequence in a bundle adjustment operation [7] to achieve global optimum. This procedure is computationally expensive for interactive applications.

Research presented in this paper was funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan. The authors gratefully acknowledge this support.

In our previous work [8] we demonstrated an effective method for extracting dominant planar surfaces from single images taken in urban environments, and showed how they could be used to represent the 3D spatial layout of the imaged scenes. In man-made environments, where planar structures are usually visible somewhere in the scene, the resulting piecewise planar model yields good approximation. Building facades usually contain a large number of parallel lines aligned along several principal directions. The images of these 3D parallel lines and their corresponding vanishing points provide important cues for single-view based 3D reconstruction. Both the distribution of line segments and possible shape of building structure are taken into account to obtain reasonable 3D understanding of the imaged scene. The contributions of this paper are (i) to provide full details of the planar extraction algorithm mentioned above, and, (ii) to show how the recovered planar structures provide robust world reference frames to integrate 3D augmented reality graphics. In particular, we demonstrate this by using the homography between a planar building surface and its image to robustly superimpose 3D virtual contents onto images taken in urban environments.

The remaining sections of the paper are organized as follows: Section 2 reviews some existing approaches for photo-based camera pose estimation, 3D reconstruction and augmented reality. Single-view based 3D plane extraction is presented in Section 3. In Section 4, we introduce how to use the extracted planar surfaces for robust 3D augmented reality integration in urban environments. The performance of the proposed method is experimentally evaluated in Section 5. Finally, the conclusion and future work are discussed in Section 6.

II. RELATED WORKS

To integrate augmented reality in unprepared environments, an essential requirement is accurate and robust tracking of camera pose. The most popular class of methods for solving this problem is based on Structure-from-Motion (SfM) [7] or the related online approach of visual SLAM [6]. Such methods simultaneously estimate camera position as well the 3D structure of the imaged scene. These systems can deal with uncalibrated image sequences acquired with a hand-held camera. Broadly speaking the principal processing involved in these techniques can be divided into two components: features detection and matching (cf. [9] for a detailed review of recent

978-0-7695-4629-2/11 $26.00 © 2011 IEEE
DOI 10.1109/IMVIP.2011.10

approaches), and, (ii) structure and motion estimation. These SfM-based techniques permit accurate 3D registration and pose estimation in unstructured environments. However, they require multiple images of a same scene as input. Also, in order to establish robust frame-to-frame correspondences, even state-of-the art techniques require that images be taken at similar viewpoints.

Previously a number of techniques have been proposed for single-view based 3D reconstruction. Hoiem and his research group estimated the coarse geometric properties of a scene by learning appearance-based models of geometric classes [10, 11, 12]. In [13], a supervised learning approach was proposed for 3D depth estimation via the use of Markov Random Fields. It is noted that the man-made environments are highly constrained and their images usually contain regular structures including parallel linear edges, sharp corners, and rectangular planes. Such image patterns can be used for effective 3D reconstruction from single-view images. Kosecka and Wei [5] developed a method to recover vanishing points and camera parameters by using line segments found in Manhattan structures. Using the recovered vanishing points, rectangular surfaces aligned with major orientations were detected. In [14] authors worked on finding rectangles aligned with vanishing points from line segments. The detected rectangular structures give strong indications of the existence of co-planar 3D points. In [15], the authors used the normals of building facades to represent their shapes. Linear constraints based on connectivity, parallelism, orthogonality, and perspective symmetry, were imposed on the object shape formulation and the optimal solution was obtained for 3D reconstruction.

In augmented reality, marker based systems locate artificially inserted features in the scene. The ARtoolkit [19] is a popular example of this and uses square box markers. These square box markers are then located and then the Iterative Closest Point (ICP) algorithm is used to calculate pose and augment objects into the scene. Other systems like ARtag [20] use a similar technique. Further comparisons of square maker based methods are described by Zhang et al. [21]. Another focus of investigation in augmented reality is the calculation of pose information from naturally occurring features in the scene such as planes, edges, or corner points. Estimation of the camera pose using such features has been demonstrated in [22, 23]. These features have also been built upon to construct models for model based augmented reality. This approach uses models of the features such as 2D templates or CAD models to interpret a visual scene and augment into it. Examples of this can be seen in [24, 25].

III. 3D PLANE EXTRACTION FROM AN IMAGE

A. Overview

Given images taken in urban environments, we present an effective method to divide a single monocular image into several vertical strips, with each strip corresponding to a separate 3D plane (a side of the building surface) in the world. The method contains four major steps with each step feeding into the next: (1) camera tilt rectification; (2) parallel line grouping; (3) generating 3D layout candidates, (4) evaluating 3D layout candidates. The details of each step are presented in following subsections.

B. Tilt rectification

First we applied the approach described in [16] for line extraction. We assume that the buildings have vivid enough vertical outlines and their images are captured using a nearly upright camera thus the vertical direction can be robustly detected. Next we rectify the original image such that vertical boundaries of a building in 3D world become vertical in the rectified image. The advantage of tilt rectification is that building boundaries will appear vertical in the rectified image, making building structure more evident, and more importantly it removes two degrees of freedom from the subsequent processing. We divide the image into several vertical strips where each strip delineates a single 3D plane of the building. To do this, we select approximately vertical lines in the image (lines within $\pm\pi/6$ radians of the vertical image column) and compute a 3×3 rotation matrix to transform them to become vertical in the image. The same transformation is applied to the original image to create a rectified view where camera tilt effect is removed. We evaluated our method on the building images from the ZuBud dataset [17]. Fig. 1 shows some results of such tilt rectification. The tilt effects are very obvious in the original images (a rectangular structure will appear trapezoidal which is wider at the bottom in case of camera pitching up). After rectification, the images of vertical world lines become parallel to the image columns.

Figure 1: Some example results of camera tilt rectification (top row: original images, bottom row: rectified images). The vertical edges are highlighted to demonstrate the effect.

C. Line grouping

Man-made environments usually contain a large number of parallel building lines. The images of a group of 3D parallel lines will intersect into a vanishing point. We propose an effective approach for clustering parallel lines by referring to their associated vanishing points. In this step, we use the tilt rectified images instead of the original images to improve performance. We equally divide the rectified image into a number of vertical strips. For each strip which corresponds to a single wall, we used the RANSAC technique [18] to find a dominant vanishing point for the lines contained within it. Given a potential vanishing point (i.e. the intersection of two line segments), we compute its supporting score as follows:

$$vote(V_i) = \sum_{all\ accpted\ L\ of\ V_i} |L|.dist(V_i, l) \qquad (1)$$

The score depends on both the length of line segment |l| and the distances between the line and its corresponding vanishing

point $dist(V_i, l)$. The vanishing point with the highest supporting score is chosen and its corresponding inlier lines are kept for further grouping. Dividing the entire collection of line segments into many small subsets, it becomes easier to identify the true vanishing points and to remove outliers. The method can robustly identify parallel building edges in the presence of large amount of clutter as shown in Fig. 2 (a). Also the method can successfully find a vanishing point for the lines on a small plane (Fig. 2 (b)) and can even handle some curved building facades (Fig. 2 (c)) by approximating them as piecewise-planar. Finally, we refine the result of line grouping using the Expectation Maximization algorithm [16]. EM iteratively estimates the coordinates of vanishing points as well as the probability of an individual line segment belonging to a particular vanishing direction.

(a) (b) (c)

Figure 2: Example results of line grouping. The color coding corresponds to the membership assignment of the individual line segments.

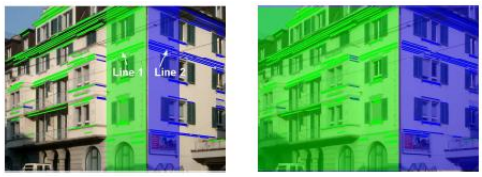
D. Generating 3D layout

Assuming the buildings have vivid enough vertical boundaries, we use the vertical lines extracted on the tilt rectified image to generate its 3D layout models in a cascade manner. First we choose the leftmost and rightmost vertical lines to generate the model containing one single dominant 3D plane. Then we select another vertical line and add it into an existing model to generate the models containing two planes. By repeatedly adding more vertical lines to the existing structure, we can create the piecewise planar model for a scene containing multiple 3D planes, as demonstrated in Fig. 3.

Figure 3: The process of generating piecewise planar 3D building layout candidates by adding more vertical lines into the existing models.

E. Evaluating 3D layout

The line segment from a horizontal vanishing direction provides a strong indication of the existence of a 3D plane in its direction. In Fig. 4 (a) Line 1 defines a vertical strip to support a 3D plane in its corresponding direction, while Line 2 suggests another plane in a different direction. After generating a number of 3D layout models based on the extracted vertical lines, we iterate through and evaluate how well each potential 3D layout fits the collection of horizontal line segments. The best fitting model is chosen to describe the underlying 3D geometry of the imaged scene.



(a) (b)

Figure 4: The line segments from horizontal vanishing directions provide important cues for 3D understanding of the scene.

Consider L_x is the candidate model which contains x planes. Accordingly the image will be divided into x vertical image strips $S = \{s_1, s_2, .. s_x\}$. For a strip s_k, the supporting score for it belonging to a vanishing direction V_i is computed as:

$$f(V_i, s_k) = \frac{\sum_{l_j \in C(V_i, s_k)} |l_j|}{\sum_{l_j \in C(s_k)} |l_j|} \quad (2)$$

where $C(s_k)$is the set of line segments contained within the strip s_k, $C(V_i, s_k)$ is the set of lines belongs to the vanishing direction V_i within the strip s_k, and $|l_j|$ denotes the length of the line l_j . The direction V_k^* with the maximum supporting score will be assigned to this strip. Then the fitting score for this layout candidate is computed as:

$$F(L_x) = \sum_{s_k \in S} f(V_k^*, s_k) * AREA(s_k) \quad (3)$$

where, $AREA(s_k)$ is the area percentage of vertical strip s_k in the image. The model which produces the highest fitting score will be chosen to describe the 3D layout of the scene. In practice, if the fitting score does not increase significantly (0.1 in our implementation) after adding more planes, we use the model of fewer planes for better efficiency. Fig. 4(b) shows the final result of image segmentation, in which each color-coded vertical strip corresponds to a different 3D plane.

Figure 5: Some examples of 3D reconstruction. The color-coded vertical strips correspond to 3D planes in different directions

We have tested this method on 100 images selected from the ZuBUD building dataset [17]. We manually divided the images into several vertical strips and labeled the ground truth for each one. In total, 51 images had less than 10% misclassified pixels and 89 images had less than 20% misclassified pixels. On average, 84% of the image areas were correctly labeled using the proposed method. Some example results are shown in Fig. 5. The output of our approach consists of a number of detected 3D planes. In the following section we provide details of how these planes can be used for 3D augmented reality.

IV. 3D Augmented Reality Integration

A. Camera Pose Estimation

In 3D vision-based augmented reality applications, accurate estimation of camera pose relative to the real world is essential for robust registration of virtual objects. We set the recovered 3D planes from the last step as z=0 planes and use them as tracking references. Within each extracted image strip which corresponds to a 3D plane, we choose four line segments (two from the vertical direction $\overline{P_{11}P_{12}}$, $\overline{P_{13}P_{14}}$ and two from a horizontal direction $\overline{P_{11}P_{14}}$, $\overline{P_{12}P_{13}}$) and compute their points of intersection to construct a quadrilateral $(P_{11}, P_{12}, P_{13}, P_{14})$ (see Fig. 6). Four corners of the defined rectangle in the 3D world have the forms as follows:

$$\mathbf{X} = \begin{bmatrix} 0 & 0 & s.h & s.h \\ 0 & h & h & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad (4)$$

where h is the height of 3D rectangle and s is the ratio between the width and the height. Their projections into the image are defined as follows:

$$\mathbf{x} = P\mathbf{X} = K[R_1 R_2 R_3 t] \begin{bmatrix} 0 & 0 & s.h & s.h \\ 0 & h & h & 0 \\ 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

$$= K[s.R_1 \;\; R_2 \;\; t] \begin{bmatrix} 0 & 0 & h & h \\ 0 & h & h & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} = H_s \begin{bmatrix} 0 & 0 & h & h \\ 0 & h & h & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix} \quad (5)$$

where P is the 3×4 camera projection matrix and H_s is the 3×3 planar homography which transforms a 3D square to a quadrilateral patch in the 2D image. The image coordinates of the four corners of the quadrilateral are measured, therefore H_s can be solved in a closed form. Since R_1 and R_2 are columns of a rotation matrix and should have unit normal, the aspect ratio s can be recovered as follows:

$$s = \frac{\|H_s^1\|}{\|H_s^2\|} \quad (6)$$

where H_s^1 and H_s^2 are the first and second columns of matrix $K^{-1}H_s$. Then the six-degree of freedom camera pose including three degrees of freedom for translation and three for orientation can be estimated as follows:

$$R_1 = \frac{H_s^1}{s}, R_2 = H_s^2, R_3 = R_1 \times R_2, t = H_s^3 \quad (7)$$

Note the camera pose with respect to a 3D scene plane is automatically computed given a single-view image. As well we can compute the camera position for other image frames, using only the frame-to-frame homographies between multiple-views.

Figure 6: Four line segments are chosen to construct a quadrilateral within each extracted vertical strip.

B. AR Content Authoring

One of the challenges in AR is the development of tools to assist the process of content authoring. That is, the creation of content that will be displayed in the AR view. Typically this is achieved through the use of a 3D model of the environment, into which the augmented 3D graphics are added. Here the 3D model acts a frame of reference for the author in that it allows them to position content relative to the world.

Our technique permits a simplified approach whereby rather than requiring a detailed global 3D model, we use input images to directly add content to the environment. For example, given the image in Fig. 7(a), by processing it with our planar extraction technique we produce a set of front-parallel views, with one view for each of the facades of the building.

(a) (b)

(c) (d)

(e)

Figure 7: Process of image authoring on ZuBUD database

Authoring then consists of adding content relative to these planes. This process is shown for 2D content in Fig.7 (b) where the technique has extracted two planes and where an AR label has been added to one of these planes. Figures 7(c) and 7(d) show this label reprojected into the original view and unseen view, respectively.

We have also used this approach to perform authoring for 3D content whereby the planar facade is taken as corresponding to the XY-plane of a 3D frame with the Z direction orthogonal to that plane. Figure 7(e) shows an example of 3D content attached at the same point as the 2D tag.

We have found that this approach to authoring is very intuitive and, although it does not give the complete freedom of augmenting a global 3D model, it is suitable for many applications. For example, in the future we plan to use this approach to allow users to author content directly within image space using standard smartphone and tablet type interfaces, without the need for complex 3D input modalities.

V. Experimental Results

The methods described above have been implemented in Matlab and C++ using OpenGL and GLUT libraries. The 3D planar extraction is performed in Matlab and computes camera pose and outputs the projection matrix and model view matrix needed as input to OpenGL to perform the 3D augmentations. In order to demonstrate this method these 3D augmentations were placed into various scenes of different complexity. Some examples of applying the proposed method to the images from ZuBUD are shown in Fig. 8.

Figure 8: Buildings from ZuBUD database augmented with direction arrows

Dominant 3D planar structures are extracted in the original images to provide world reference for AR integration. It is noted that there is no perceptible drift or jitter in the resulting images. We also applied the technique to images captured in the campus of the National University of Ireland, Maynooth. Fig. 9 shows a set of images of the John Hume Building captured by a hand-held camera from different viewpoints. Here, the same 3D cube (i.e. relative to the building façade) is rendered in each image using our technique. As can be seen from the figure, the technique provides an accurate rendering over considerable change in viewpoint.

Figure 9: Cube augmented to the side of a building varying at different angles and distances

The system works well on planar structures however an issue arises on buildings that have only partially planar facades or completely lack any planar facades. Any curved area of the building will be treated by the technique as if it had a planar profile. The system will not be able to augment correctly onto non planer areas of the building however it will work for any planer parts of the facade. This may be resolved in future using 3D model data of the building to assist with augmentation. Another issue that can occur with the software is when calculating the line grouping, it can also sometimes confuse lines on the ground with being part of the building. Even though these lines are outliers, if the building does not have a strong enough collection of line segments these line segments can sometimes be considered when identifying the true vanishing points. This will affect any augmentation onto the building.

VI. Conclusions

In this paper we have presented a single image based augmented reality technique. The system works by identifying and extracting the planar facades of buildings, and then uses those facades as reference frames for augmenting the view. Results of the application of this technique to real-world images of general urban environments have been shown. These results included augmenting scenes with both 2D and 3D graphical elements.

In the future we aim to use this approach for the development of a vision based AR navigation system that will operate over a ~1km^2 area of an urban environment. Here the user should be able to capture an arbitrary image within the environment and have the system graphically augment it with relevant navigation and point-of-interest (POI) information. The target platforms for this system will include smartphone and tablet based devices. Due to the computational limitation of these devices we would use a client-server approach which will transmit an image to a server for processing. While there would be network latency issues associated with this approach we would try to precompute as much as possible to lower the amount of data that needs to be transmitted.

References

[1] Cornelis, N., Leibe, B., Cornelis, K., Gool, L, "3D urban scene modeling integrating recognition and reconstruction," Int. J. Comput. Vision 78 (2008) 121-141.

[2] Pollefeys, M., Van Gool, L., Vergauwen, M., Verbiest, F., Cornelis, K., Tops, J., Koch, R, "Visual modeling with a hand-held camera," Int. J. Comput. Vision 59 (2004) 207-232.

[3] Snavely, N., Seitz, S.M., Szeliski, R, "Modeling the world from Internet photo collections," IJCV 80(2) (2008) 189-210

[4] D. Robertsone, R. Cipolla, "An image-based system for urban navigation," BMVC04 (2004) 819–828.

[5] J. Kosecka,W. Zhang, "Extraction, matching, and pose recovery based on dominant rectangular structures," Comput. Vis. Image Underst. 100 (3) (2005) 274–293.

[6] H. Durrant-Whyte and T. Bailey "Simultaneous localisation and mapping (SLAM): Part I the essential algorithms", Robot. Autom. Mag., vol. 13, p.99 , 2006.

[7] R. Hartley, A. Zisserman, "Multiple View Geometry in Computer Vision," Cambridge University Press, New York, NY, USA, 2003.

[8] Y. Cao, J. McDonald, "Viewpoint invariant features from single images using 3D geometry," WACV09 (2009) 1–6.

[9] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L. Van Gool, "A comparison of affine region detectors," Int. J. Comput. Vision 65 (1-2) (2005) 43–72.

[10] D. Hoiem, A. A. Efros, M. Hebert, "Automatic photo pop-up," ACM SIGGRAPH (2005) 577–584.

[11] D. Hoiem, A. A. Efros, M. Hebert, "Geometric context from a single image," IEEE International Conf. on Computer Vision (2005) 654 – 661.

[12] H. Derek, A. A. Efros, M. Hebert, "Putting objects in perspective," IEEE Conf. on Computer Vision and Pattern Recognition (2006) 3–15.

[13] A. Saxena, S. H. Chung, A. Y. Ng, "3-D depth reconstruction from a single still image," Int. J. Comput. Vision 76 (1) (2008) 53–69.

[14] G. Wang, Z. Hu, F. Wu, H.-T. Tsui, "Single view metrology from scene constraints," Image Vision Comput. 23 (9) (2005) 831–840.

[15] Z. Li, J. Liu, X. Tang, "Shape from regularities for interactive 3D reconstruction of piecewise planar objects from single images," MULTIMEDIA (2006) 85–88.

[16] J. Kosecka, W. Zhang, "Video compass," ECCV (2002) 476–490.

[17] T. S. H. Shao, L. Van Gool, "Zubud-zurich buildings database for image based recognition," Tech. Rep. 260, Swiss Federal Institute of Technology (2004).

[18] M. Fischler, R. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," CACM 24 (6) (1981) 381–395.

[19] H. Kato and M. Billinghurst. Marker tracking and HMD calibration for a video-based augmented reality onferencing system. In IWAR'99, pp. 85-94, 1999

[20] Mark Fiala, ARTag, a Fiducial Marker System Using Digital Techniques, Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2, p.590-596, June 20-26, 2005 [doi>10.1109/CVPR.2005.74]

[21] Xiang Zhang , Stephan Fronz , Nassir Navab, Visual Marker Detection and Decoding in AR Systems: A Comparative Study, Proceedings of the 1st International Symposium on Mixed and Augmented Reality, p.97, September 30-October 01, 2002

[22] L. Vacchetti, V. Lepetit and P. Fua. Combining edge and texture information for real-time accurate 3d camera tracking. In ISMAR '04, pp. 48-57, 2004.

[23] G. Simon, A. Fitzgibbon, and A. Zisserman, "Markerless Tracking Using Planar Structures in the Scene," Proc. IEEE/ACM Int'l Symp. Augmented Reality, pp. 120-128, Oct. 2000.

[24] P. Fua and V. Lepetit. Vision based 3D tracking and pose estimation for mixed reality. In M. Haller, M. Billinghurst, B. H. Thomas Eds, Emerging Technologies of Augmented Reality Interfaces and Design, pp. 43-63, Idea Group, Hershey, 2007.

[25] G. Reitmayr and T. Drummond. Going out: robust model-based tracking for outdoor augmented reality. In ISMAR '06, pp. 109-118, 2006.