Interactive Insertion of Virtual Objects in Photos and Videos

R. Nóbrega†[1] and N. Correia[1]

[1]CITI, Departamento de Informática, Faculdade de Ciências e Tecnologia, FCT, Universidade Nova de Lisboa

Abstract

The introduction of virtual objects in photos or videos has been the focus of many Augmented Reality applications. This paper proposes a framework using image analysis methods to automatically detect scene features to introduce virtual 3D geometry objects. The current high-level features include surfaces, depth and scene orientation automatic detection. The main use of this technology is in AR tools, games or applications, which require the user to introduce an object that blends with a photographed or filmed scene. The main advantage of the proposed approach is that it can work only with one or two photos without prior knowledge, being ideal for mobile applications with camera or to be used with photo albums. Additional sensors can be added to increase reliability such as depth sensors or accelerometers but they are not essential. The main algorithms are intended to create a scene model in a few seconds and allow an interactive behaviour of the augmented objects.

Categories and Subject Descriptors (according to ACM CCS): H.5.1 [Information interfaces and presentation]: Multimedia Information Systems—Artificial, augmented, and virtual realities

1. Introduction

Combining computer graphics with real-world images is the main problem addressed by Augmented Reality. For objects to blend seamlessly in interactive applications their geometry must blend with the lines, textures and properties of the photographed world. The problem addressed in this paper is how to introduce virtual objects in an interactive scene, which react to detected features of an image. In this paper a framework is proposed which analyses one or two images in order to create an internal model of the photographed scene. Using the internal model it is possible to introduce augmented objects that interactively react to properties of the scene. The current considered properties are: point descriptors (SURF or SIFT), surface registration, depth detection and scene orientation. The main goal is to create a system that allows building interactive applications for mobile devices with cameras or to use user-generated photos as input.

The first augmented reality applications included the creation of virtual objects that followed a marker on the scene (i.e. ARToolKit). More advanced applications tie the object to a set of keypoints in the scene [KM07], as seen in Figure 1. Some applications use pre-built models of the scene to help the introduction of virtual reality [Hal]. Most of these techniques require a pre-configuration of the chosen scenario, as all the marks and scene models must be known *a priori*. Although prepared to use marker-style AR, the main advantage of the proposed framework is that it only uses the images as input for interactive applications. Some other projects use additional sensors to perform similar tasks in interactive applications such as depth sensing cameras (i.e. KinectFusion) or accelerometers to detect the direction of the floor. In the current paper, the main focus was given to the creation of an internal model of the scene using single image analysis [LHK09] [SSN09] and stereo vision (from photos and videos [IAH95]).

2. Interactive Augmented Reality

The main novelty of this work is that by analysing the images in search for low-level features (interesting points, image displacement, line segments) several high-level features (homographies, depth, surface normals, vanishing points, scene orientation) are extracted. The input of the problem is generally two images from one scene. The initial processing of the images includes extracting keypoint descriptors (FAST, SURF or SIFT), lines and comparing keypoints between images.There are some additional steps for each high-level fea-

† Work funded by FCT/MEC and CITI/FCT/UNL, through grants SFRH/BD/47511/2008 and PEst-OE/EEI/UI0527/2011

DOI: 10.2312/conf/EG2012/posters/007-008

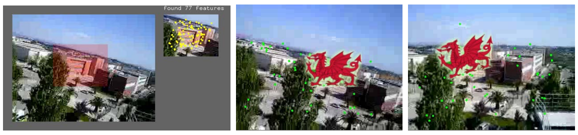

Figure 1: *Augmented reality in videos based on SURF descriptors. The user selects an area in one frame (left image). The object, the dragon, is superimposed on that area (middle image). The object is then replicated to the other frames of the video in the same position and rotation as it was when placed (example frame in right image).*

ture. In Figure 1, an example of how to use keypoints and homographies to edit a video is shown. Here a virtual object can be anchored by the user to a specific place inside the video. This particular example was implemented with SURF features. This follows the traditional marker technique where the object is superimposed on top of the mark but here the mark is represented by a set of SURF descriptors. The object can be replicated to the entire video using the homography matrix between frames.

To create virtual applications that use as input a space from a user supplied photo, a spatial model of that space has to be extracted from the images. Figures 2 and 3 give some examples of this. Figures 2 shows an example of how can depth and surfaces be extracted quickly from image keypoints. This means that a user can take two pictures of a scene and the detected model can be used to insert 3D geometry in context. To simplify the process of inserting objects, they could be aligned with the scene direction and react to a virtual plane representing the floor. To do this the line geometry that composes the images are analysed as seen in Figure 2. Our detection algorithm analyses the lines using different luminance thresholds to rule out cluttering and detect the main vanishing points. Using the vanishing point, the horizon is extracted and the virtual scene is rotated accordingly so that the virtual object assumes an initial position oriented with the scene. The current framework is implemented in C++, with some structures and algorithm implementations from OpenCV 2.3 (namely SURF, SIFT, Hough Line detection and Delaunay triangulation). It runs on standard dual core computers with 4GB of RAM and the proof-of-concept examples were built in OpenGL with the support of the openFrameworks library.

3. Conclusions and Future Work

The framework shows promising results, especially in Manhattan images. The final goal will be to implement and integrate several techniques that will help in the development of interactive augmented reality applications using computer vision input.

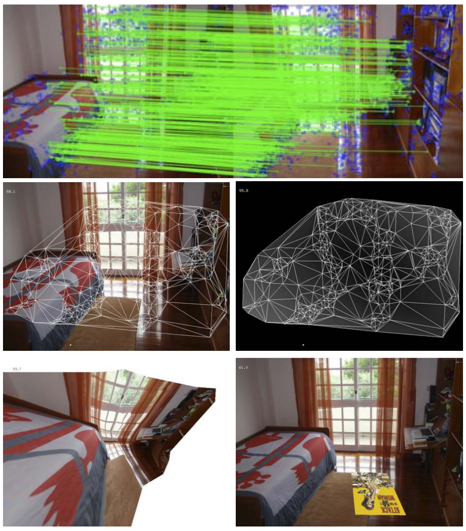

Figure 2: *Interactive application for introduction of virtual objects attached to surfaces. Using two images and their SIFT descriptors, the matching points are used to create a 3D model using Delaunay triangulation. The depth of each point is inferred based on the camera parameters and the displacement of each point between images. Using the normal vector of each polygon created with the triangulation the virtual object (yellow poster) can be attached to surfaces.*

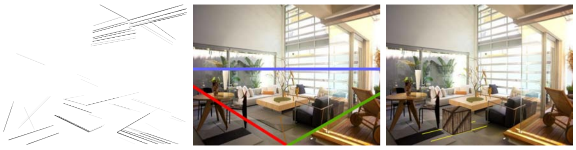

Figure 3: *Interactive application for introduction of virtual objects oriented as the room and attached to the floor. Using line detection and vanishing point detection, the horizon is calculated and the main line directions detected (lower left image, horizon in blue, main directions red and green). In the end, the virtual box can be dynamically introduced in the floor, with the room orientation (lower right image).*

References

[Hal] Haladova Z.: Reconstruction of Cultural Heritage Object Utilizing its Paper Model for Augmented Reality. pp. 7–8. 1

[IAH95] Irani M., Anandan P., Hsu S.: Mosaic based representations of video sequences and their applications. In *Proceedings of IEEE International Conference on Computer Vision* (1995), IEEE Comput. Soc. Press, pp. 605–611. 1

[KM07] Klein G., Murray D.: Parallel Tracking and Mapping for Small AR Workspaces. *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality 07* (2007), 1–10. 1

[LHK09] Lee D. C., Hebert M., Kanade T.: Geometric reasoning for single image structure recovery. *IEEE Conference on Computer Vision and Pattern Recognition (2009) 0*, June (2009), 2136–2143. 1

[SSN09] Saxena A., Sun M., Ng A. Y.: Make3D: learning 3D scene structure from a single still image. *IEEE Transactions on Pattern Analysis and Machine Intelligence 31*, 5 (2009), 824–40. 1