Rendering Stereoscopic Augmented Reality Scenes with Occlusions using Depth from Stereo and Texture Mapping

Joseph Louis¹, Julio Martinez²

¹ PhD Student, School of Civil Engineering, Purdue University, 550 Stadium Mall Dr. West Lafayette, IN 47907, Phone 765.586.9353, Fax 765.494.0644, jlouis@purdue.edu
² Associate Professor, School of Civil Engineering, Purdue University, 550 Stadium Mall Dr., West Lafayette, IN 47907-2051, USA, Phone 540.392.1975, FAX 765.494.0644, julio@purdue.edu.

Abstract

Augmented Reality (AR) is commonly used in the construction industry to visualize proposed structures and/or operations in their real world environment as it greatly enhances the realism of the viewing experience without having to construct 3D models of complex real world scenes. In order to achieve convincing scenes in AR applications, the virtual objects should be correctly occluded by the real objects present in the scene, which is a non trivial problem that involves specialized equipment to capture depth information about the real scene. This paper presents a novel methodology that allows for the stereoscopic rendering of AR scenes with the correct occlusion of the virtual objects. Two pictures of the real scene that are captured simultaneously from viewpoints with a horizontal baseline separation of 1m are used to obtain both the stereoscopic images as well as the depth information of the occluding objects in the real scene. Potential applications of this methodology include its use in the creating and rendering engaging representations of the AR scene being visualized, by affording the user the perception of depth in the scene and as novel means to visualize 3D CAD and BIM models in their proposed/ existing settings in the real world.

Keywords: Augment Reality, Construction, Occlusions, Stereoscopy

1. INTRODUCTION

A key step in the development of Augmented Reality (AR) scenes in Building Information Modeling (BIM) is the seamless integration of real world pictures and computer-generated graphics. Tracking the user’s position in the scene, calculating the extrinsic parameters (position and orientation) of the user’s viewpoint, and correctly registering the virtual camera to match the real camera are important steps in ensuring that the virtual scene is correctly overlaid on the picture of the real world. However, in order to ensure that the virtual objects are correctly placed in the 3-D world, spatial information about the objects in the real world scene is required. This information is represented by a depth value assigned to each pixel in the photograph of the real world picture, which is a measure of the distance of the object from the camera at the time of image capture. Correct occlusion of the virtual objects is enabled with this depth information.

Capturing the depth information is not a trivial task as cameras usually capture only color data. Several methods have been proposed to calculate the depth buffer of scenes. These methods range from the use of stereo-cameras, obtaining graphical models of the
scene (static occlusion) to the use of specialized hardware such as time of flight range sensors and Laser Detection and Ranging (LADAR) devices.

This paper addresses the issue of occlusions in AR scenes with the use of a stereo camera setup and a having the user select objects in the real scene that are potential occluders of the virtual objects in the AR scene. This allows for a great flexibility in selecting objects that are of interest to the user and also eliminates errors in due to the aperture problem (Sietz and Dyer 1995) that arises due to uniformly colored surfaces in the scene. This paper also uses the two input images to create an intermediate image that can be used along with one of the input images to create a stereoscopic image. This creates an illusion of depth in the image to the viewer by providing two separate images to the left and right eyes, which have been offset by an Inter-Pupillary distance of 6 cm. The stereoscopic effect is also used in the entertainment media to create and provide depth cues for the viewer, thus enhancing the level of immersion into the AR Scene.

The objectives mentioned above, namely, rendering a stereoscopic AR scene with occlusion is achieved by using triangulation from stereo and texture mapping. An important assumption that has been is that the objects in the scene that are not sufficiently close to the viewer (so that they are not potential occluders of virtual objects) are infinitely far away from the viewer. While this method does require the user to select points of interest in the scene, it precludes any kind of scene setup such as placement of visual markers of fiducials in the scene. The methodology developed was tested on two scenes, an indoor scene and an outdoor scene. The assumption that the background objects are far away from the viewer did not hold true for the indoor scene and this gave rise to unwanted visual artifacts in the scene. The effect of the artifacts was much less pronounced in the outdoor scene. This has been described in the case study section in the paper. The findings of this research could lead to the easy development of stereoscopic AR scenes without the use of specialized hardware. In the construction domain, it would help in the creation of convincing and engaging visualizations of proposed buildings and structures in their planned real-world environment.

2. BACKGROUND

A lot of work has been done on occlusion handling since the late 1990’s. Breen et al. (1996) utilized two different methods (model-based and depth-based method) to resolve the real occluding virtual problem. In the model-based approach, the geometric models of real objects should be registered to their real-world counterparts. In the depth based approach, the depth of the map of the real world is calculated and the occlusion is handled by comparing the depth of the virtual objects and the real objects.

2.1 OCCLUSION HANDLING IN AR SCENES

Berger (1997) proposed an occlusion solving method which is an alternative to 3D reconstruction. The utility area of this method is limited to static scenes. To make the occlusion resolve applicable to image sequence, Ong et al. (1998) proposed a framework the occlusion problem without assuming a prior computer model from the input scene. The disadvantage of this method is that the 3D model will change in the intermediate frames.
due to the motion of the viewpoint. Lepetit and Berger (2000) made an improvement to this method. They only compute the 3D boundary from two consecutive key views.

Schmidt et al. (2002) presented a real time technique for computing dense disparity maps from a binocular stereo camera system. They successfully used this method to solve the occlusion problem in static scenes. Hayashi et al. (2005) proposed a method for real-time stereo matching using a contour-based approach to acquire depth of the boundary of the real objects. Fortin and Herbert (2006) proposed a depth based method which supports a mobile viewpoint. But a set of targets should be placed in the scene in order to deal with the camera motion. An automatic occlusion handling method proposed by Tian et al. (2010) obtains the proper spatial relationship between virtual and real objects in real-time by first constructing the disparity map of the scene and then handling the occlusions by segmenting the scene into different regions based upon their pixel depth values.

The method presented by Fischer et al. (2003) for detecting the occlusion of virtual objects by natural occluders in a scene is based on a graphical model of static backgrounds in the natural environment which have to be previously acquired. This graphical model consists of a set of textured polygons, in front of which occlusion can be detected. These polygons have to correspond to surfaces in the real world, so each surface is described by a number of textured coplanar polygons and is called dynamic occlusion background (DOB). The background is assumed to be occluded whenever a difference is found between the camera image and the appearance that is expected so, if there are no differences or the differences are minimal, it is possible to see the background without any occlusions. Four major drawbacks with this approach are acknowledged: 1) the algorithm can’t deliver the actual depth information of the scene so it always assumes that a real object detected in front of a DOB is actually in front of all virtual objects, 2) the algorithm performance doesn’t seem to be good enough for real-time applications, 3) the way in which point correspondences are detected can be problematic because very large marker tracking inaccuracies prevents correspondences from being found at all. 4) The pixel comparison criteria don’t always produce a flawless occlusion mask.

The occlusion problem in AR scenes has also been attempted to be solved by the means for special hardware as opposed to the use of computational methods to calculate the depth of scene. Behzadan and Kamat (2008) presents a methodology for producing correct occlusion effects in real-time in unprepared outdoor environments, such as construction sites, which was what it was originally intended for. The main contribution of this research is that it takes into account the dynamics of the real world in which the AR animation takes place. Also, unlike other algorithms, no prior assumptions about the shape and positions of objects in the real world scene are made. Such assumptions do not support the dynamic nature of the objects in the scene and the change in viewpoint from which the scene is viewed in the AR application. The algorithm developed uses depth information of the real world scene, recorded using a flash LADAR. The flash LADAR obtains the depth information by casting laser beams into the real world scene and recording the time that it takes to bounce of objects and return to the source. Color information about the real world scene is captured to a video camera that ensures a live video stream of the real world. Both the camera and the flash LADAR are attached to the user, so that the information present is dependent upon the user’s position and orientation in the real world. This information is
obtained with the use of a GPS and a tracker, which are also mounted upon the user which capture the position and orientation of the user respectively. The AR scene is displayed to the user on a wearable Head Mounted Display (HMD). In the research conducted by Fischer et al. (2008), a depth of flight range sensor is utilized to obtain a 2D map of distances to real objects in the environment, which is then registered with high resolution color images obtained with a digital video camera. The ARToolkit library is used along with optical markers to obtain the pose of the camera. While rendering the AR scene, the depth map obtained is used in a specialized shader program running on the Graphics Processing Unit (GPU) while comparing if the pixels belonging to the virtual object are behind or in front of those belong to the real world scene. The GPU then draws the unoccluded pixels in order to represent the correct occlusion of objects in the AR scene.

2.2 OBTAINING STEREOSCOPIC IMAGES

In order to obtain a stereoscopic image, it is necessary to take two photographs, offset the human IPD, which is approximately 6 cm. It is not possible to take two simultaneous photographs using conventional DSLR cameras due to space restrictions. It is required to take the pictures simultaneously if the scene being shot has dynamic elements in it. If the scene is static, a single camera with two exposures would suffice for capturing the stereo images. Extant literature relating to the creation of stereo scenes of AR with occlusions was not found during a review of the state of art.

2.3 MOTIVATION FOR THIS RESEARCH

The main motivation for performing this research was the need for a methodology that enabled the creation and sharing the viewing experience of convincing AR scenes of 3D CAD models in their real world environment amongst multiple users. Research into the application of AR in construction (Behzadan and Kamat, 2008) has focused on delivering real time dynamic AR scenes to user(s) who is present on the site. This restricts the viewership of the AR scene to just the single user who is controlling the equipment, as the scene displayed would depend on his position and orientation in the scene.

Also the use of specialized hardware, such as Time-of-Flight sensor and the flash LADAR as described in the previous sections, would discourage the adoption of AR by users who would otherwise look to exploiting its benefits as a visualization tool for their projects. Another aspect of AR that was found to be lacking in the state of the art was the development of stereoscopic scenes. The use of stereoscopic scenes would greatly enhance the realism of the viewing experience by providing additional depth cues to the scene.

Thus, this research was done with the twin goals of delivering realistic stereoscopic AR scenes of buildings and construction sites without the use of special hardware. This research also ensures that the viewing experience of the AR scenes can be shared by multiple users after a one time creation process.

3. METHODOLOGY PROPOSED

The methodology that is presented in this paper allows for handling of occlusions in the AR scene by using 2 cameras in stereo mode with only a horizontal translation between their centers of projection. The methodology proposed also allows for the creation of a
stereoscopic view of the image, by generating a new image that can be used in conjunction with the left input image to create the stereoscopic pair of images. The first section of the methodology explains the camera setup and the process used by the user to select objects in the foreground used to acquire the depth information about the scene, while the second section discusses the generation of the stereoscopic image. From here on, the images will be referred to as left and right image.

3.1 Use of Stereo to obtain Depth Buffer of Scene

As was mentioned in the section on the state of the art on methods in occlusion handling, depth from stereo is a method that does not require any specialized hardware to measure the depth of the scene, nor does it require any prior assumptions to be made about the scene. The two major disadvantages of using stereo are: 1) processing time is considerably higher than other methods, 2) it is a function of the image resolution and 3) automating the finding of correspondences between the images is not a trivial task. In order to overcome these limitations, it was decided to allow the user to select correspondences between if he so desired. If not, correspondences were made for each point selected on the left image by using a color matching algorithm to find the best pixel along the point epipolar line in the right image.

3.2 Setup of the Equipment

Capturing depth from stereo requires that two images need to be taken of the same scene from locations that are known relative to each other. This can be done by using markers or fiducials in the scene being photographed or by setting up the stereo cameras so that their relative positions and orientations are known to the user. The authors used the latter option presented above as it precluded the need for setting up markers in the scene. The cameras were fixed on brackets, which were then mounted on an optical rail and this was fixed on a tripod as shown in Figure 1. The optical rail was mounted on a ball and socket joint, which allowed it 3 degrees of freedom, whereas the camera brackets had one degree of freedom of yaw and could be translated along the rail.

![Figure 1: Stereo Camera Setup.](image)
During the setup of the cameras, it was ensured that both the cameras were oriented with the same yaw, roll and pitch rotation angles. This was confirmed by reading the calibrations of the camera mounts and the bubble level placed on the cameras. The horizontal separation was measured along the optical rail that the cameras were placed on.

3.3 Defining Occluders using Stereo

This section describes the steps followed in order to calculate the depth of the scene from the two stereo images obtained. It is assumed that both the cameras’ intrinsic parameters (width and height of the image, horizontal field of view of the camera) and the extrinsic parameters (in this case, only the horizontal distance between the two cameras) are known to the user. A larger field of view is desirable as it allows for a greater overlap of the scene captured in the left and right images.

In order to calculate the depth or distance of the occluder from the view position, it is required that correspondences be made between the common elements of the occluder in the left and right images. In this research, the user has the ability of making these correspondences manually or by letting the program choose it for points of interest in the two images. This task is greatly simplified by searching only along the epi-polar line of the pixel in the right image. The average values of the RGB (Red-Green-Blue) channels of an area of size 5x5 pixels is calculated around the pixel in the left image and along each of the pixels on the epi-polar line in the right image till an acceptable match has been found. However, this method is not robust as it greatly depends upon the composition and lighting of the scene, which could render the scene differently in the two images, depending upon the center of projection of the image. Since the user is only required to select the edges of the occluding shape that is of interest (as opposed to finding correspondences of each pixel in the left image), it is possible to perform this task manually, which is an option provided to the user. This method greatly improves the accuracy of the correspondence matching.

Once the correspondences have been made, triangulation is used to obtain the 3d position of the point whose projection was found on both the pictures. Once this has been done for all the edges of the user-selected occlude, a triangle mesh is constructed using the points obtained. This is a partial model of the occluder that is correctly placed with respect to the position of the cameras (It is assumed that the origin of the coordinate system lies at the center of projection of the left camera). Furthermore, the triangular mesh is enhanced with a texture map to obtain a textured model of the occluder that is correctly placed in the real scene. The texture coordinates are calculated by dividing the x and y pixel coordinates of the projection of the vertices of the triangular mesh by the width and height of the image. The texture used is the portion of the picture taken of the occluder in the left image.

Since the model of the occluder is correctly placed with respect to the coordinate system we have defined at the COP of the left camera, it is possible to obtain the depth value of each pixel within the occlude. As mentioned earlier, it is assumed that the rest of the scene that is not contained in the area selected by the user is at an infinite distance (very far away) from the view point. The above process can be repeated in the same scene for multiple occlude objects in the scene.
3.4 Obtaining stereoscopic photographs

Stereoscopic images are offset images that are viewed separately by the left and right eyes and fused in the brain to provide an illusion of depth in the scene. In order to obtain such a set of images, it is required that the they (left and right image) be captured of the same scene at the exact same instant of time such that the COP of images are separated by the inter-pupillary distance(IPD) of the human eyes which is approximately 6 cm (Dodgeson 2004). While this task is easily achieved for computer generated 3d models and scenes, it requires special stereo cameras to capture such images of the real world.

Since two stereo images have already been obtained, albeit at a distance much greater than the IPD, a methodology to generate an intermediate image is proposed in this paper. The methodology assumes that the background images are sufficiently far away so as to remain unchanged when only translation is applied to the viewpoint. With this assumption, it follows that the only objects that have been selected by the user as being potential occluders would translate along its epipolar line when the viewpoint is translated. Hence, the generation of the right-eye image from the previously captured photographs was done using the following three steps: a) Render the background of the left image as the background of the right image (since it is assumed that the background is sufficiently far away); b) Remove the occluding objects as seen from the left eye from the background; c) Project and render the occluders and the virtual objects as seen from the right eye.

While steps a) and c) are trivial and easily achieved since we have the textured model of the occluders. It is removing the occluder that is a non-trivial issue. This is done by placing the color data inside the silhouette of the with the color data from the right image. This is done in order to see more of the background that was hidden in the left image. Figure 2 shows the left image, left image with the occluder hidden using data from the right image and the final right eye image with the occluder as viewed from the right eye.

![Figure 2: Obtaining stereoscopic image](image)

4. CASE STUDY

The proposed methodology was tested on rendering AR scenes. The first scene consisted of a black box kept on a table top indoors while the second scene consists of a pile of dirt outdoors. In both cases, the cameras used were a pair of Nikon D300s fitted with a 20 mm lens. It was ensured that both the cameras had the same settings with regard to shutter speed, ISO value, aperture and white balance. Also, the pictures were shot simultaneously in manual focus mode using a single wireless remote clicker.
Case Study 1: Indoor scene

Figure 3 shows the set of input images of the scene. Figures 4 shows the images with the correspondences marked on them by the user. The third set of images shown in Figure 5 form a stereoscopic view of the AR scene. Furthermore, a computer generated teapot has been added to the scene and is shown partially occluded by the black box.
Visual artifacts can be noticed in the right-eye image towards the right of the black box. This is due to the fact that the assumption of the background being sufficiently far away does not hold good in this case. The stereo nature of the images can be tested by staring cross-eyed at the images in Figure 5 until they fuse to form a single stereoscopic image.

Case Study 2: Outdoor scene

The outdoor scene selected for the study consisted of a heap of mud, as may be noticed in an earthmoving site. A cube texture mapped with the image of a truck was added to create an AR scene from the pictures. Figure 6 show the set of input images of the scene. Figure 7 shows the images that form the stereo pair of the scene.

![Figure 6: Left and Right Input images for outdoor scene](image)

![Figure 7: Left-Eye and Right-Eye Images for Stereoscopic AR scene with occlusion](image)

The artifacts that can be seen here are not due to the failure of the assumption, but rather due to the coarse triangle mesh generated formed by only 6 vertices that were triangulated from as many corresponding points in the two images.

5. CONCLUSION

This paper presents a novel methodology that uses two stereo images to generate a stereoscopic AR scene with occlusions. The user is required to select the edges of potential occluders in one of the images. The corresponding point is the other image can be identified by the program by performing a color matching algorithm along the pixels on the point’s epipolar line or it can be selected by the user. The latter method avoids errors that may be caused due to the scenes composition and lighting.
A very important assumption that has been made in the methodology is that the objects that outside the perimeter of the occluder(s) as defined by the user are sufficiently far away from the viewer so that they do not occlude any of the virtual objects in the scene and that they it remains intransient to any translation of the viewpoint. The effect of this simplifying assumption was pronounced in the right-eye image generated for the stereoscopic image of the indoor scene and the background was not in fact far away from the viewer. Another kind of visual artifact is produced due to the coarseness of the triangle mesh of the occluder generated. However, this is entirely dependent upon the number of points that the user selects in the scene. This effect was very pronounced in the outdoor scene, wherein the occluder was of an irregular shape.

Future extensions to this work could include the use of edge detection algorithms to detect the edges of occluders without the user having to select them manually. This method could possibly preclude the need to make the assumption of the background being infinitely far away as it would be possible to generate a triangle mesh of the entire scene and without extracting much effort from the user.

6. REFERENCES
