Virtual View Synthesis using Visual Hulls

Nicholas F. Maunder * Gerhard de Jager Fred Nicolls

Digital Image Processing Group
Department of Electrical Engineering
University of Cape Town
Private Bag, Rondebosch, 7701

Abstract

Virtual view synthesis refers to the process of generating a novel view of a scene, or object, from a set of reference views. The novel view represents what would be seen from a particular viewpoint that does not coincide with the reference cameras. This paper discusses the implementation and evaluation of two approaches to virtual view synthesis. They are both based on the concept of the visual hull of an object and are therefore suited to generating novel views of objects rather than of whole scenes. The evaluation of the implementations is based exclusively on the visual quality of the synthesized view. The novel views that are generated are compared to additional views of the scene that were not used in the synthesis process. For this purpose a measure of error is formulated to quantify the differences between the rendered virtual views and the additional views.

1 Introduction

Virtual or novel view synthesis refers to the process of generating a virtual view of a scene, or object, from a set of reference views. These reference views are the images obtained from a camera, or a number of separate cameras, positioned at different viewpoints around the scene. A virtual view of the scene represents what would be seen by a camera if it were positioned at a point not coinciding with the original reference cameras but having a common field of view. Relevant information therefore needs to be extracted from the original reference views in order to render the image corresponding to the virtual viewpoint.

There are a number of practical applications for virtual view synthesis. Synthesised views of a real scene or object can be used to enhance the experience of computer generated environments in the field of virtual reality [13]. Similarly, in the field of augmented reality being able to synthesise novel views from the images of a real object allows for the correct visualization of real-life objects that have been artificially placed in an observed scene. Interest is also being shown in the entertainment industry in areas such as film making and video gaming [8].

*The financial support of the De Beers Technology Group (GTS) is greatly appreciated. The financial assistance of the National Research Foundation (NRF) towards this research is also hereby acknowledged. Opinions expressed and conclusions arrived at are those of the author and are not necessarily to be attributed to the NRF.

Surveying the relevant literature reveals that the current techniques can be divided into two groups, namely those that first reconstruct a three dimensional geometric model of the observed object using the reference views and then render a novel view, and those that generate the new view directly from the reference views.

The methods belonging to the first group are referred to as *geometry-based* rendering systems, because the new image is formed by rendering the reconstructed geometric model. The methods belonging to the latter group are termed *image-based* rendering systems, as the new view is rendered directly from the reference images [11].

In this work one technique from each of these groups was implemented and evaluated using a number of different data sets. These techniques are both built on the concept of the visual hull of an object and are therefore suited to generating novel views of objects rather than of whole scenes. An approximation to the photographed object's visual hull can be computed from its silhouettes via the process known as volume intersection [9]. For this purpose the calibration parameters of the reference views must be known. The colour reference images are then used to assign textures to the visual hull and with this information the novel view can be generated.

Section 2 presents a review of some of the many virtual view synthesis techniques. Section 3 discusses the concept of the visual hull of an observed object. The focus then moves to the two techniques that were implemented, described in sections 4 and 5. This is followed by a discussion of the results of the work and the conclusions.

2 A Review of Virtual View Synthesis Techniques

2.1 Geometry-based Rendering Techniques

Geometry-based systems make use of geometric descriptions of the surfaces of objects or volumetric data to model a scene and then render novel views [11]. Such systems first have to generate the geometric model using the reference images.

Constructing an approximate geometric model of a scene, or object, from a set of reference images can be done using image matching techniques and triangulation [8, 5]. An alternate approach is that of *volumetric scene modelling*. The basic principle behind this approach is that volumes that are consistent with the

given reference images are constructed in three dimensional space, thus reconstructing the scene. These volumes of space that are occupied by an object in the world can be represented as a regular tessellation of cubes, which are called *voxels* [5].

Two common categories of voxel-based reconstruction algorithms can be identified [5]. The first class includes those algorithms that make use of volume intersection to recover the approximate *visual hull* of the photographed object. The second class of algorithms performs a *colour consistency* test to distinguish between voxels that are part of the scene objects and those that are not. In a recent paper [17], Wong and Cipolla use uncalibrated images captured under circular motion to construct a voxel model of an object. An initial model of the visual hull is constructed using the silhouettes obtained from the captured images. This model is then refined by adding new silhouettes, captured from arbitrary viewpoints, to the original silhouettes.

2.2 Image-based Rendering Techniques

Image-based rendering systems generate a virtual view of a scene directly from the photometric data contained in the reference views. Classification of the various methods into distinct categories is not straightforward. A previous survey [16] prefers to view the different approaches as a "continuum" of image-based rendering techniques ranging from those that make use of *no* geometrical information to those that make use of *implicit* or *explicit* geometrical information.

View interpolation refers to the process of producing intermediate views of a scene from the images of two reference views [15]. The process usually involves establishing point correspondences between the images, followed by an interpolation of the displacement between, and colour values of, the related points. Chen and Williams [2] use a depth map and the relative pose between cameras to easily find matching image points. They point out that if the transformation between the three cameras is restricted to a translation that is parallel to the image plane then the result of the interpolation will be perspectively correct. Seitz and Dyer [15], furthermore, demonstrate that by first rectifying the reference images a valid intermediate view can also be synthesised.

The geometric constraints that exist between multiple views of the same scene can be utilized to synthesize a new view. Such an approach would be through the manipulation of the epipolar geometry that exists between pairs of images, as was investigated by Laveau and Faugeras [10]. Avidan and Shashua [1] make use of trilinear tensors to generate novel views of a scene from two or three reference images. Their method requires a dense correspondence between two reference images but they do not recover the full camera calibration parameters.

An algorithm developed by Matusik et al. [13] renders the image of a textured visual hull of an object without having to first reconstruct the geometric model—hence they call the approach *image-based visual hulls*. Their method exploits the epipolar geometry that exists between the virtual view and each of the reference views.

3 Visual Hulls

The closest geometric approximation of an object that can be reconstructed using only its silhouette images is referred to as its *visual hull* [9]. The visual hull can therefore be viewed as the largest shape (in terms of volume) that can be substituted for the original object while still producing the same silhouettes. Obtaining the visual hull is accomplished through the technique known as *volume intersection* [9].

Given a number of views of an object, the silhouettes are usually obtained by segmenting the input images into binary images. A pixel marked as part of the silhouette indicates that its associated line of sight, or visual ray, from the camera centre meets the observed object [5]. All the intersecting visual rays for a particular image form a visual cone, and the intersection of the individual cones from all the input images gives the approximate visual hull (see figure 1).

It is only an approximation because the actual visual hull is described by Laurentini [9] to be the intersection of the cones corresponding to silhouettes obtained from *all* possible viewpoints exterior to the object's convex hull[1]. Increasing the number of input images will thus improve the accuracy of this approximation.

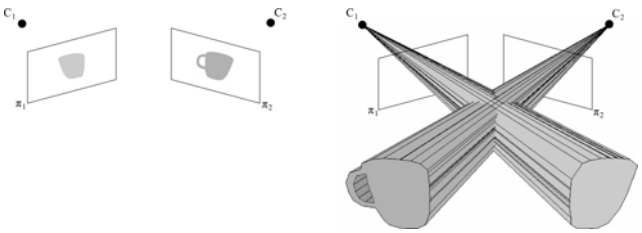

Figure 1: The intersection of the viewing cones defined by an object's silhouettes gives its approximate visual hull.

4 Voxel Reconstruction and Texturing

The first approach to virtual view synthesis that was implemented reconstructs an explicit three dimensional model of the observed object. This model is then rendered using scan-conversion algorithms [6] giving the desired virtual view. The approach is therefore classified as a geometry-based rendering technique.

Recovery of the object's geometrical structure is achieved via a technique known as volume intersection thus giving an approximation to its visual hull as discussed in section 3. The volume of space occupied by the object is modelled using voxels. To facilitate the storage and creation of the voxel model an *octree* data structure is utilized.

The volume occupied by the observed object is at first represented by a single all encompassing voxel. This voxel is then repeatedly subdivided until the model is formed. The subdivision happens as follows. The initial voxel, and every voxel that is subsequently processed, is projected into each of the reference camera

[1]Laurentini also introduces the idea of an *internal* visual hull. He does, however, point out that the principal case is the approximation of the visual hull from viewpoints exterior to the object's convex hull.

views. Comparing these projections to the associated silhouette images can have one of the following outcomes [17]:

- The projections lie within the boundaries of every one of the silhouettes. The voxel is therefore recorded as being part of the model.
- One or more of the projections lie outside the boundaries of the silhouettes. The voxel is not part of the model and is therefore removed.
- The projections straddle the boundaries of the silhouettes. The voxel is sub-divided into eight smaller voxels and the process is repeated for each one.

The subdivision of voxels can continue until every voxel projects into the silhouettes. It is, however, more practical to limit the level of subdivision, thereby putting an upper bound on the resolution of the final model [17].

The texture information for the voxel model is obtained from the original reference images. Depending on the observations, each surface voxel is assigned a colour. A *view independent* texturing strategy was implemented in that for a particular surface voxel the relevant colour values obtained from each of the reference are averaged to determine its value. Unlike the *view dependent* texture mapping approach described by Debevec, Yu, and Borshukov [4] the illumination effects that are unique to a particular camera view will *not* be reproduced in any of the synthesized virtual views.

To determine which of the pixels in the reference images map to a particular voxel all of the surface voxels are rendered onto each of the reference image planes using a *scan-conversion* algorithm [6]. It is important that the visibility of these voxels in each of the reference views is established. Including the projection of an occluded voxel in the colour analysis distorts the assignment of values because the observed colours actually belong to the occluding voxels. A common method for resolving the visibility issues related to model rendering is the implementation of a *z-buffer*, also known as a *depth-buffer* [6]. Thus for a particular reference view, the visibility of the surface voxels can be established by rendering them all with the aid of a z-buffer [3]. Instead of storing a colour value at each pixel the ID of the current voxel being processed is stored. The result is then an image map with the value at each pixel identifying which surface voxel is visible along that particular line of sight.

During the colour computation for a surface voxel the algorithm first needs to scan the image map for all pixels with the correct voxel ID. The colour values of the corresponding pixels in the original reference image are then averaged to determine the contribution from the associated view. This process is repeated for each reference view with the colours again being averaged and the resulting value is assigned to the voxel in question.

5 Image-based Visual Hulls

A viewpoint-dependent representation of an object's visual hull can be computed without actually reconstructing an explicit geometric model. The result will take the form of a depth map relative to a particular viewpoint—each pixel in the image gives an indication of the distance to the surface point of the visual hull along that particular line of sight. These depth values can then be used to extract colour information from the reference images, thereby generating the novel view. The algorithm discussed in this section is based on the approach to virtual view synthesis entitled *image-based visual hulls* [13].

The computation of the observed object's visual hull is performed in the following manner: for each pixel in the virtual image the three dimensional point in the world where the pixel's line of sight meets the visual hull of the object must be calculated. The information necessary for this computation can be extracted directly from the silhouettes of the object by making use of the epipolar geometry that exists between the virtual view and each of the reference views [13].

The visual ray associated with a particular pixel in the virtual image is projected into each of the silhouette images. For a given silhouette image the ray projection can be found using the fundamental matrix that relates points in one image to epipolar lines in another. A search is now performed on the line in order to determine whether it overlaps the actual silhouette. Only the visible portion of the image plane, namely the silhouette image, is searched. Overlapping segments are projected back into three dimensional space giving the corresponding line segments along the visual ray in question. Figure 2 illustrates the back projection of the segments. This reprojection can be performed as follows. The line segments are each specified by two points and each point defines a three dimensional line, or visual ray, passing through that point and the reference camera's centre of projection. The intersection of these new visual rays with the original visual ray from the virtual camera is then found. Corresponding pairs of three dimensional points, marking the intersections, represent the back projections of the two dimensional line segments from the silhouette image [14].

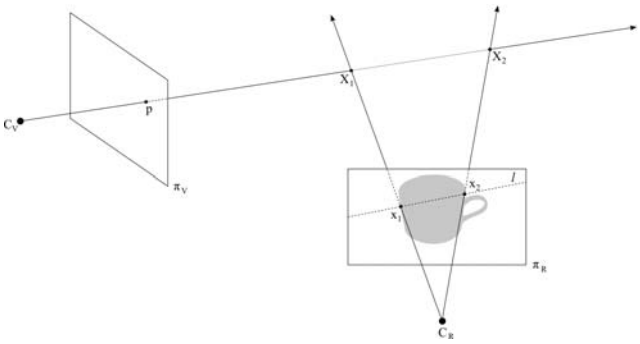

Figure 2: Back projecting the two dimensional silhouette intersections into three dimensional space. The points (x_1 and x_2) at which the epipolar line (l) intersect the silhouette image (π_R) are recorded, thereby specifying the line segment that overlaps the silhouette. These points define visual rays stemming from the reference camera (C_R). The three dimensional points (X_1 and X_2) where these rays meet the visual ray of the virtual camera (C_V) mark the back projection of the overlapping line segment ($\overrightarrow{x_1 x_2}$) in the image.

The process of finding the three dimensional line segments for a visual ray from the virtual image is repeated for each silhouette image. The intersection of each of these line segments is then calculated and the point that is closest to the image plane of the virtual camera marks the surface point of the visual hull that is visible [13].

Unfortunately, the pose of the virtual camera in relation to any one of the reference cameras can cause certain parts of the epipolar line in that reference view to be invalid. Therefore only the visible epipolar line segment that represents the visual ray extending from the virtual camera's centre of projection and passing through its image plane should be searched [14]. Determining the appropriate range of the visible epipolar line is dependent upon the position of the reference camera with respect to the virtual camera's image plane, and also on whether the line segment from the reference camera's centre of projection to the vanishing point of the visual ray from the virtual camera intersects the reference camera's image plane. In this regard, approximately four separate cases can be identified. More details on this topic can be found in [12] and [14].

A *view-dependent* texture mapping of the computed visual hull is performed by assigning a colour value to each pixel in the depth map thus completing the synthesis of the novel view. For a particular pixel in the virtual image the three dimensional point on the surface of the object's visual hull is calculated. This calculation is only performed for those pixels which correspond with pixels in the depth map that image the surface of the visual hull. The surface point can then be projected into the appropriate reference view using its camera projection matrix and the corresponding colour value can be read. Since the texture mapping is to be view-dependent the reference camera that should be selected is the one that has the closest viewpoint to that of the virtual camera [4].

The most appropriate view can be determined by considering the angle between the vector linking the virtual camera's centre of projection to the surface point of the visual hull, and the vector linking the reference camera's centre of projection to the same surface point. The reference view associated with the smallest angle is the view that must be used.

A problem that arises is that although a camera may have a favourable viewing angle it may not have an unoccluded view of the surface point. Hence, to improve the quality of the texturing process the visibility of the surface points for each reference view must first be determined. Matusik, et al. [13] proposes an approach that compares points lying in the same epipolar plane. When testing the visibility of a surface point for a particular reference camera the only points that might occlude it from view will lie in the epipolar plane formed by itself, the virtual camera, and the actual reference camera. Due to the discrete nature of an image the algorithm will, however, only produce an approximation to the actual visibility [14].

Once the visibility information has been computed the surface point is projected into the reference view with the most appropriate viewpoint, and from which it is also visible. The colour value is then calculated by performing a *bilinear interpolation* on the four pixels surrounding the projected point.

6 Experimental Results

The implementations were evaluated using a number of different input data sets. These data sets consist of reference views of either images of real-life objects or images of artificial objects generated on a computer. The data sets consisting of real images were acquired using a single digital camera and, in some instances, a turntable. The computer generated images were created with three dimensional modelling software. With all the data sets the viewpoints were positioned at approximately the same height above the ground plane and were arranged in a circular pattern around the object.

The silhouettes of the objects were obtained by manually segmenting the reference images using image editing software. In the case of the computer generated images the silhouettes were rendered separately through the manipulation of the modelling software.

The algorithms receive a set of colour images and a set of binary images as input, representing the reference views and the corresponding silhouettes of an observed object. Apart from the input images the algorithms also require the calibration parameters of the reference cameras, as well as the calibration parameters of the virtual camera for the desired view.

6.1 Establishing a Measure of Performance

The evaluation of the implementations is based exclusively on the visual quality of the synthesized view. The measure of performance thus involves a comparison between the newly rendered images and additional reference images of the object that were not used in the synthesis process. For the real data sets these "additional reference images" are actual images of the object acquired using a digital camera. In the case of the computer generated data sets these images are rendered using the calibration parameters of the virtual viewpoint.

Comparing any two images is done at a pixel level. Since a view of the object does not occupy every pixel in the image only a subset of the image pixels are processed, thereby limiting the number of background pixels included in the comparison. The region of interest is defined as the smallest rectangular area that will enclose all the foreground pixels in both the additional reference image and the virtual image. When evaluating a particular series of data sets, for instance the data sets of the ceramic cat (table 1), the region of interest is kept constant and is chosen so that it encloses all the foreground pixels of each virtual image that is processed.

To quantify the differences the distance in RGB colour space between the colour coordinate of a pixel in the additional reference image and the colour coordinate of the corresponding pixel in the virtual image is calculated [3]. This error measurement is computed for each pixel of interest using the following formula:

$$E = \sqrt{(R_v - R_{ref})^2 + (G_v - G_{ref})^2 + (B_v - B_{ref})^2}$$

where E is the distance error in RGB colour space, $[\ R_v \quad G_v \quad B_v\ ]^\mathsf{T}$ are the red, green, and blue colour values, respectively, of the pixel in the virtual image, and $[\ R_{ref} \quad G_{ref} \quad B_{ref}\ ]^\mathsf{T}$ are the red, green, and blue colour values, respectively, of the corresponding pixel in the additional reference image to which the virtual image is being compared. A better approach would be to consider the way humans perceive colour and hence make use of a colour space where the distance between colour coordinates is related to the difference in observed colour. These are referred to as perceptually *uniform* colour spaces [8]. Such an error measure was however not investigated.

The mean distance error for an image is calculated by averaging the error values for all the processed pixels, thereby giving an indication of the quality of the rendered output. This value is used to

compare techniques and also investigate how the number of reference views used in the synthesis process influences the quality of the novel view.

6.2 Quantitive Evaluation

One series of data sets given as input to the implementations consisted of reference views of a small ceramic cat. These real images were captured using a digital camera and a turntable. The calibration of the viewpoints was accomplished using silhouette consistency constraints and is derived from the concepts discussed in [7]. A similar approach to camera calibration has been proposed by Wong and Cipolla [17].

Table 1 gives the average error values calculated for both the geometry-based and the image-based techniques when using the cat data sets as input. In most cases, as the number of reference views increases so the average error calculated decreases. This is the expected behaviour because adding more views not only increases the amount of photometric information available but it also refines the approximation of the object's visual hull due to the increased number of silhouettes [9].

Figure 3(b) and 3(c) show the novel views generated by the geometry-based technique and the image-based technique, respectively. Since the geometry-based technique uses a view-independent texture mapping strategy the relative illumination levels across the surface of the cat in the desired view (figure 3(a)) are not reproduced in the new view.

Table 1: Average error calculated for the ceramic cat data sets.

Number of Views	Geometry-based Technique	Image-based Technique
5	33.49	31.66
8	27.54	27.59
10	28.27	27.14
16	28.16	26.45

Table 2 gives the average error values calculated for both the geometry-based and the image-based techniques when using the computer generated radio data sets as input. As with the ceramic cat sequence, using more than five reference views decreases the average error value. There is, however, an increase in the average error when using ten reference views as opposed to eight. A possible explanation for this behaviour is that it is related to the relative placement of the cameras, which were spaced equally around the radio.

The body of the radio is in the shape of a rectangular box with flat faces and slightly rounded edges. The difference between the camera configurations of the two data sets is that with eight cameras, four of the cameras are positioned parallel to the radio's faces while with ten cameras only two are parallel. The result is that the front face of the model constructed using ten views is curved, and not flat as it should be, thus causing its shading to appear warped which increases the associated error. The novel views of the radio are shown in figure 3.

In both tables the error values calculated for the image-based approach are generally less than the values calculated for the geometry-based approach. The reason for this is that the level of detail of the reference images that can be reproduced by the geometry-based method is restricted by the resolution of the voxel representation. The image-based approach does not have this limitation. It produces a more accurate sampling of the object's visual hull which is determined by the image resolution of the virtual image [13]. The geometry-based approach, however, creates a quantized sampling of the visual hull related to the dimensions of the voxels. Each voxel will, therefore, generally map to more than one pixel in the new image. In contrast, the implemented image-based method establishes a separate mapping of colour between a *single* pixel in the virtual image and one of the reference images.

Table 2: Average error calculated for the model radio data sets.

Number of Views	Geometry-based Technique	Image-based Technique
5	18.84	11.60
8	15.26	7.95
10	17.35	8.17
16	15.80	6.87

7 Conclusions

The work presented in this paper covers the implementation of two different approaches to virtual view synthesis. Both are based on the concept of the visual hull of an object and are thus more suited to synthesising novel views of objects as opposed to whole scenes.

Comparing the results obtained for the two solutions reveals that the image-based approach achieves lower average error values than the geometry-based approach. This suggests that the image-based approach produces a more accurate approximation of the desired virtual view.

The relative positioning of the cameras around the observed object can influence the accuracy of the approximated visual hull and thus the quality of the synthesised views. As was noted in the results section (section 6.2), a flat surface will not be computed as being flat unless it is observed by a camera that is positioned at a viewpoint parallel to that surface.

In general, increasing the number of reference views used to generate the novel view decreases the average error achieved. A possible reason for the observed anomalies, other than the cases related to the relative positioning of the cameras, could be camera calibration that is not sufficiently accurate.

References

[1] S. Avidan and A. Shashua, "Novel View Synthesis by Cascading Trilinear Tensors," *IEEE Transactions on Visualisation and Computer Graphics*, Vol. 4, No. 4, October—December 1998.

[2] S. E. Chen and L. Williams, "View Interpolation for Image Synthesis," *Proceedings of the International Conference on Computer Graphics and Interactive Techniques*, pages 279–288, 1993.

[3] W. B. Culbertson and T. Malzbender, "Generalized Voxel Coloring," *Proceedings of the International Workshop on Vision Algorithms: Theory and Practice*, Vol. 1883 of Lecture notes in Computer Science, pages 100–115, Springer-Verlag, 2000.

(a) Desired novel view (b) Geometry-based technique (c) Image-based technique

(d) Desired novel view (e) Geometry-based technique (f) Image-based technique

Figure 3: Figures (b)–(c) show novel views of ceramic cat generated using 10 reference views. Figures (e)–(d) show novel views of radio generated using 8 reference views.

[4] P. Debevec, Y. Yu, and G. Borshukov, "Efficient View-Dependent Image-Based Rendering with Projective Texture-Mapping," *Proceedings of the Eurographics Rendering Workshop*, June 1998.

[5] C. R. Dyer, "Volumetric Scene Reconstruction from Multiple Views," In L.S. Davis, editor, *Foundations of Image Understanding*, pages 469–489. Kluwer, Boston, 2001.

[6] J. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes, *Computer Graphics: Principles and Practice*, Addison-Wesley Publishing Company, Inc., 1990 with corrections 1997.

[7] K. Forbes, A. Voigt, and N. Bodika, "Using Silhouette Consistency Constraints to Build 3D Models," *Proceedings of the Fourteenth Annual South African Workshop on Pattern Recognition*, PRASA, 2003.

[8] D. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*. Englewood Cliffs, N.J.: Prentice-Hall, August 2002.

[9] A. Laurentini, "The Visual Hull concept for Silhouette-Based Image Understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 16, No. 2, February 1994.

[10] S. Laveau and O. Faugeras, "3-D Scene Representation as a collection of images," *Proceedings of the International Conference on Pattern Recognition*, Vol. 1, pages 689-691, 1994.

[11] L. McMillan and G. Bishop, "Plenoptic Modeling: An Image-Based Rendering System," *Proceedings of SIGGRAPH'95*, August 1995.

[12] L. McMillan, *An Image-Based approach to Three-Dimensional Computer Graphics*, Ph.D. Thesis, University of North Carolina, 1997.

[13] W. Matusik, C. Buehler, R. Raskar, S. J. Gortler, and L. McMillan, "Image-Based Visual Hulls," *Proceedings of the International Conference on Computer Graphics and Interactive Techniques*, pages 369–374, 2000.

[14] W. Matusik, *Image-Based Visual Hulls*, Master of Science Dissertation, Massachusetts Institute of Technology, 2001.

[15] S. M. Seitz and C. R. Dyer, "Physically-Valid View Synthesis by Image Interpolation," *Proceedings of the IEEE Workshop on Representations of Visual Scenes*, pages 18–25, 1995.

[16] H. Y. Shum and S. B. Kang, "A Review of Image-based Rendering Techniques," *IEEE/SPIE Visual Communications and Image Processing (VCIP) 2000*, pages 2–13, June 2000.

[17] K. K. Wong and R. Cipolla, "Reconstruction of sculpture from its profiles with unknown camera positions," *IEEE Transactions on Image Processing*, Vol. 13, No. 3, pages 381–389, 2004.