

# Main approaches in realistic virtual view synthesis

Livatino, Salvatore

*Published in:*
CVonline : the Evolving, Distributed, Non-Proprietary, On-Line Compendium of Computer Vision

*Publication date:*
2003

*Document Version*
Publisher's PDF, also known as Version of record

# Main Approaches in Realistic Virtual View Synthesis

Salvatore Livatino

Email: sl@cvmt.dk

A brief overview of main reference approaches in the field of realistic virtual view synthesis is presented. This review is intended for a reader interested in knowing about main approaches in image- and model-based rendering approaches, but not as an exhaustive survey of the research field.

The synthesis of realistic virtual views has received increasing attention in the last decade, mainly due the increased popularity of virtual reality and the spread of its applications. Hence, to the demand of increasing realism into generated sceneries while simplifying the modeling process.

A solution to the problem of providing visual realism to computer generated images is searched into the possibility of re-creating naturally occurring physical phenomena by real world observations. The "road" mainly investigated has been to capture the occurring phenomena through photographs, hence, directly or "indirectly" transfer them into novel generated views.

The direct way refers to warping algorithm which do not take into consideration any geometric information associated to the observed scene. These approaches usually require a dense set of reference-views. The "indirect" way instead refers to the case when reference image-views are supported by associated knowledge, (pixel correspondences, depth maps), or 3D models.

The achievement of such ambitious goal, a realistic synthesis, has mainly be attempted through two different rendering approaches: model-based and image-based.

The model-based approach represents the traditional way to generate virtual views of an object or scene. This approach is usually referred as *Model-Based Rendering* because it relies on a geometrical 3D model of the object or scene wished to be rendered. In this context the research is focused on improving model fidelity by using image-based modeling. In particular: geometric model extraction and representations for rendering purposes, object-texture extraction and mapping to geometric models, illumination effects recovery and rendering.

At a high level a model-based rendering approach involves three processes. First, an event or a scene must be recorded, then, a 3D model of the environment has to be extracted using computer vision techniques, and at the end, the obtained 3D model is rendered from the view of a virtual camera.

The image-based approach represents instead an alternative to model-based rendering,

and a competing means of creating virtual views, primarily relying on real images taken as reference in place of a geometric 3D model. This approach is then referred as *Image-Based Rendering*. In order to produce novel views, reference images are usually interpolated or re-projected from source to target image.

This rendering approach is less generic than model-based rendering since utilized techniques often depend on the applications, thus, type of environment and required rendered field of view. The common characteristic is a rendering time independent from scene complexity and no need in principle for reconstruction of geometric models. Image-based rendering techniques often require an additional knowledge to input reference images, such as image-correspondences, depth information, epipolar relations, etc. This additional knowledge often is extracted from same input images or it is provided a priori.

Image-based rendering is usually applied to static environments whereas model-based rendering is often proposed for dynamic scene visualization. Authors have also proposed both the two approaches for the same application context, (Blanc, Livatino and Mohr [3], [5]).

A growing interest, which could also be considered as an "evolution" of image and model based rendering is towards hybrid methods, so that in the last years authors have presented methods which lie in-between image- and model-based rendering. A successful example is represented by the work of Debevec, Taylor and Malik [15], which proposes generation of novel views based on a reconstructed geometric model, where textures in novel views are mapped view-dependently.

Survey papers have also started with classifying works in realistic virtual view synthesis as a "continuum" of representations, (which might include model-reconstruction and model-based rendering), based on the tradeoff of many aspects. Among them: number of required input images, motion assumed for the virtual-camera, knowledge about the scene geometry, depths and correspondences, the way pixel are transferred etc. Previous work on the field is usually classified by the authors depending on which is the aspect they would like to focus on in their contribution.

In their classification of image-based rendering H. Shum et al., [44], propose three categories according to how much geometric information is used: no-geometry, implicit geometry (i.e. correspondences), and explicit geometry. D. Forsyth and J. Ponce, [19], also propose three categories but based on the type of approach: volumetric reconstruction, points transfer, and light-fields. L. McMillan, [32], proposes to distinguish approaches based on the way images have supplemented the image generation process: images to represent approximations of scene geometry, images in a database to represent different environment locations, and images as reference scene models from which to synthesize new views. S. Kang [26] proposes a categorization primarily based on the nature of the scheme for pixel indexing or transfer: non-physically based image mapping, mosaicking, interpolation from dense samples, and geometrically-valid pixel reprojection.

# 1 Image-Based Rendering

As above mentioned, the image-based approach relies on real images taken as reference in place of a geometric 3D model. However, input reference images often do not suffice for the purpose of rendering novel views, so that many of the proposed systems require additional knowledge, such as image-correspondences, depth information, epipolar relations, etc.

The advantage represented by avoiding model reconstruction may also mean software rendering and no exploitation of graphic hardware, (unless the system has been designed for the exploitation of some graphic functions, e.g. projective texture mapping).

Image-based rendering methods often mentioned in current literature, were developed in the half of the nineties, (Chen [11], Levoy and Hanrhan [28], Gortler et al. [21], McMillan and Bishop [33], Seitz and Dyer [40], Chen and Williams [10], Shashua and Werman [43], Chang and Zakhor [8], Leveau and Faugeras [27], Regan and Pose [38]). Research is still very active in the field and new techniques have also been proposed for investigation. For example, the new projection model based on the *two-slit camera*, (Granum et al. [22]).

In this section summaries of representative works in Image-Based Rendering are presented. The authors name at the top of each summary identifies presented approach together with a "pioneer" reference publication. Figure 1 represents typical computational steps involved in Image-Based Rendering. presented works.

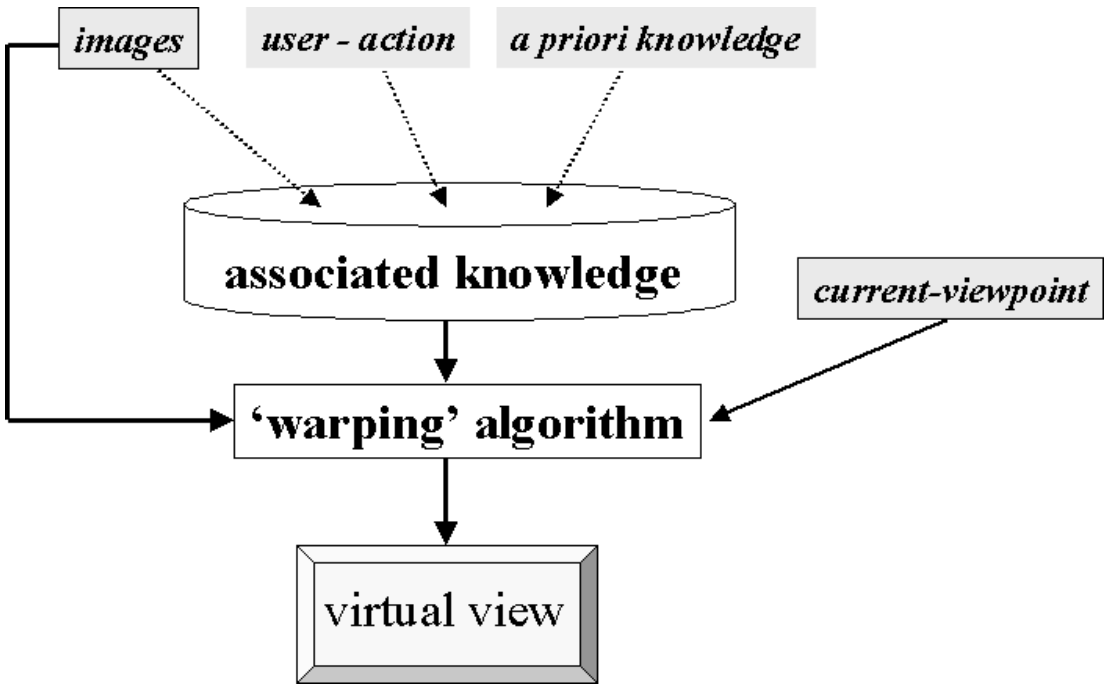


Figure 1: Image-Based Rendering: typical computational-steps.

## Chen [11]

S. E. Chen proposes an image-based rendering system called *QuickTimeVR* developed by Apple Computer.

The paper presents a way for a computer to systematically deal with movies. It includes an algorithm for storing moving pictures and play them back as fast as possible without using any extra hardware. The system includes a "panoramic movie" technology which enables users to explore spaces, and an "object movie" technology which enables users to examine objects interactively. This system is used on the world wide web to display 3D objects from various viewpoints.

The scene is represented by a set of cylindrical images created at key locations. Based on these images the system is able to synthesize new planar views in response to user input by warping one of these cylindrical images. The user is so able to navigate "discretely" from location to location, and while at each location continuously change the viewing direction. Translation of viewing position can instead only be approximated by selecting reference cylindrical images closest in viewpoint to current viewing position. The above is achieved at interactive rates (greater than 20 frames per second).

The speed of the processor largely determines the quality of visualized movies. If the system can not process all frames in the movie, QuickTimeVR drops some frames. The system can be applied to exchange video on the Internet, virtual navigation of real environment, (useful for architecture planning, museum tours), etc.

Among the advantages: the method represents a practical way of exchanging video on the Internet, virtual navigation of real environment, allows for immersive navigation of visual environment. No need for considering the viewing angle when selecting a reference images, (references are cylindrical), no need for specialized hardware, high-quality images, distortion corrector, multimedia possibilities.

Among the disadvantages: visualized scenes must be static, only playback visualization, several photographs are required and properly registered.

Class of Approaches: no-geometry rendering [44], light-fields [19], mosaicking [26].

## Levoy-Hanrahan [28]

This paper describes *Light Field Rendering*, a simple and robust method for generating new views from arbitrary camera positions without depth information or feature matching, simply by combining and re-sampling the available images.

The major idea behind the technique is a representation of the light field, the radiance, as a function of position and direction, in region free of occluders. In these regions the "light field" is a 4D parameterization of viewing position and direction. An image is a

two dimensional slice of the 4D light field. Creating a light field from a set of images corresponds to inserting each 2D slice into a 4D light field representation. Similarly, generating new views corresponds to extracting and re-sampling a slice. Once a light field has been created new views may be constructed in real time by extracting slices in appropriate directions. The desired ray can be looked up in the light field database of rays using the 4D parameterization of viewing position and direction.

Image generation using light fields is inherently a database querying process, much like the movie map image-based approach of Chen [11]. The interpolation scheme used by the authors approximates the re-sampling process by simply interpolating the 4D function from the nearest samples. The authors have investigated the effect of using nearest neighbor, a bilinear interpolation, and full 4D quadrilinear interpolation.

Since the success of the method depends on having a high sample rate, the authors describe a compression system that is able to compress the generated light fields by more than a factor of 100:1 with very little loss of fidelity. A vector quantization scheme is used to reduce the amount of data used in light field rendering, yet achieving random access and selective decoding. The authors have also addressed the issues of anti-aliasing during creation, and re-sampling during slice extraction. In particular, to reduce aliasing effect, the light field is pre-filtered before rendering.

Among the advantages: real-time display of new views by extracting slices in appropriate directions, high freedom in the range of possible views, no model information such as depth-values or image-correspondences is needed to extract the image values, image generation involves only re-sampling (a simple linear process), simple compression schemes can be applied (because of the 3D structure of the light field), re-sampling process simpler than depth or correspondence -based, image-based rendering approaches.

Among the disadvantages: large amount of data that may be required, (but possible high compression by the proposed method), long time for image acquisition (reference images are acquired by scanning a camera along a plane using a motion platform), the flow of light is completely characterizes only through unobstructed space in a static scene with fixed illumination, sampling density must be high (to avoid excessive blurriness).

Class of Approaches: no-geometry rendering [44], light-fields [19], interpolation from dense matching [26].

## Gortler-Grzeszczuk-Szeliski-Choen [21]

This paper discusses *The Lumigraph*, a new computational method for capturing the complete appearance of both synthetic and real world objects and scenes, representing this information, and then using this representation to render images of the object from new camera positions.

The Lumigraph as in the case of light-field rendering is a ray-database query algorithm. The lumigraph uses a 4D parameterization of viewing position and direction, (a 4D parameterization of rays passing through a pair of planes with fixed orientation). The lumigraph, unlike the light field, considers the geometry of the underlying models when reconstructing desired views. The geometric information is used to control the blending of the images. The lumigraph, as well as the light field, is defined as data intensive rendering process. However, the lumigraph can tolerate a lower sampling density since the available geometric information.

Among the advantages: fast scene rendering, arbitrary camera poses are used to construct the database of visible rays, high freedom in the range of possible views.

Among the disadvantages: the preparation of the database requires considerable preprocessing, large amount of data that may be required, the flow of light is completely characterizes only through unobstructed space in a static scene with fixed illumination, sampling density must be high.

Class of Approaches: no-geometry rendering [44], light-fields [19], interpolation from dense matching [26].

## McMillan-Bishop [33]

L. McMillan and G. Bishop propose *Plenoptic Modeling* as a consistent framework for the evaluation of image-based rendering systems. The authors give a concise problem definition and propose an image-based rendering system in light of the Plenoptic framework.

The paper introduces the use of the 5D plenoptic function, $P_5(V_x, V_y, V_z, \theta, \phi)$, defined as the intensity of light rays passing through the camera center at every space locations $(V_x, V_y, V_z)$ at every possible angle $(\theta, \phi)$. The original 7D plenoptic function was presented by Adelson and Bergen [1]. The simplest plenoptic function is a 2D panorama, cylindrical or spherical, when the viewpoint is fixed.

Within the proposed plenoptic modeling the goal of image-based rendering is to generate a continuous representation of the Plenoptic function. The authors claim that all image-based rendering techniques can in fact be casted as attempts to reconstruct the Plenoptic function from a sample set of that function. They believe there are significant insights to be gained from this characterization, so they propose their system in light of this Plenoptic framework.

The samples used are cylindrical panoramas. The "angular disparity" of each pixel in stereo pairs of cylindrical panoramas is computed and used for generating new plenoptic function samples. The authors also introduces a geometric invariant for cylindrical projections that is equivalent to the epipolar constraint defined for planar projections. The original samples, cylindrical panoramic images, can so be used to reconstruct new virtual

views from arbitrary locations. The reconstructed views are also capable of describing perspective effects and occlusions. In particular, the authors introduce a novel visible surface algorithm which guarantees back-to-front ordering.

Among the advantages: real-time display of visually rich environments (both indoor and outdoor) is possible without the need for special graphic hardware, the method allows for acquisition and exploitation of compact sample images, realistic visualization of complex sceneries where perspective effects and occlusion are correctly modeled, real-time display and the use of commonly available equipment.

Among the disadvantages: visualized scenes must be static and with fixed lighting conditions, reference images should be acquired close to each other, reconstructed views should be generated close to sample images.

Class of Approaches: no-geometry rendering [44], light-fields [19], mosaicking [26], geometrically-valid pixel reprojection [26].

## Seitz-Dyer [40] [41]

S.M. Seitz and C.R. Dyer propose *View Morphing*, a way to generate new views of a scene from two basis views. This can be applied to both calibrated and uncalibrated images. At minimum, two basis views and their fundamental matrix are needed.

A scan-line algorithm for making image interpolation is presented that require only four user provided feature correspondences to produce valid orthographic views. The paper describes a simple image rectification procedure which guarantees that interpolation does in fact produce valid views, under generic assumptions about visibility and projection process.

The proposed technique uses basic principles of projective geometry, and introduces an extension to image morphing that correctly handles 3D projective camera and scene transformations. The authors propose to exploit monotonicity along epipolar lines to compose physically valid intermediate views without the need for full correspondence information. Under the assumption of monotonicity, it is shown that the problem is theoretically well-posed.

This result is significant in light of the fact that is not possible to fully recover the structure of the scene due to the aperture problem [1]. Moreover, they demonstrate that for a particular range of views, the problem of view synthesis is in fact well-posed and does not require a full correspondence, that is, images interpolation is a physically valid mechanism for view interpolation. Views can consequently be generated by linear interpolation of

[1] The Aperture problem arises due to uniformly colored surfaces in the scene. In the absence of strong lighting effects, a uniform surface in the scene appears nearly uniform in projection. It is then impossible to determine correspondences within these regions.

the basis images, (if the basis images are first rectified).

Among the advantages: the method represents a practical and simple way of generating new views of a scene (under monotonicity assumptions), view synthesis does not suffer from the aperture problem, the technique may be applied to photographs as well as rendered scene, ability to synthesize changes both in viewpoint and image structure, interesting 3D effects via simple image transitions, applicable to both calibrated and uncalibrated images, suitable for application in entertainment industry and for limited bandwidth teleconferencing.

Among the disadvantages: the method requires multiple image re-sampling (loss of quality), local blurring when monotonicity assumption is violated, artifacts arising from errors in correspondence, it is only suitable for static scenes, the method needs four user provided feature correspondences, visualized regions need to be free of occluders.

Class of Approaches: implicit-geometry rendering [44], volumetric reconstruction [19], geometric-valid pixel reprojection [26].

## Chen-Williams [10]

This paper presents *View Interpolation* an image interpolation approach to synthesize 3D scenes, where input images are a structured set of views of a 3D object or scene.

In order to reconstruct desired views several reference images are used along with image correspondence information. The view synthesis is based on linear interpolation of corresponding image points using range data to obtain correspondences, (as in view-morphing [40]).

Intermediate frames are used to approximate intermediate 3D transformations of the object or scene. The authors have investigated smooth interpolation between images by modeling the motion of pixels (i.e. optical flow) as one moves from one camera position to another. They have investigated special situations in which interpolation produces valid perspective views. They conclude that interpolated images do not in general correspond to exact perspective views. They point out and suggest solution for determining the visible surfaces. Like image morphing, View Interpolation uses photometric information as well as local derivative information in its reconstruction process.

Among the advantages: the proposed method can be performed at interactive rates, suitable for virtual holograms, walk-trough in virtual environments, incremental rendering, motion blur acceleration, and soft shadows cast (by area light sources) acceleration, the approach works well when generated views share a common gaze direction and the synthesized view-points are within 90 degrees of this gaze angle.

Among the disadvantages: problems in the generated images for points which are not mutually visible on both reference images (difficult to establish the flow field informa-

tion), view approximation when the change in viewing position is not slight, static scene, problems may arise when the generated views do not share a common gaze direction, and when the synthesized view-points do not stay within 90 degrees of the gaze angle.

Class of Approaches: implicit-geometry rendering [44], volumetric reconstruction [19], interpolation from dense matching [26].

## Shashua-Werman [43]

This paper based on the existence of certain trilinear functions of three views, (with a corresponding tensor of 27 intrinsic coefficients), [42], derives connections between the trilinear function invariants across three views and intrinsic structures and invariants of 3D space.

The result shows that the tensor of coefficients determined by three views replaces entirely the role of the fundamental matrix (and associated intrinsic structures of two views) in 3D tasks. In other words, the projective structure of the scene follows directly from the tensor without the need to recover any intrinsic structure associated with two views.

In addition the tensor encompass 2-view structures in the sense that the fundamental matrix is readily expressed as a solution of a linear system determined by the tensor, the rotational component of camera motion is expressible in closed form by the tensor, and a variety of means exist for recovering the epipoles from the tensor.

The major result is that exists a decomposition of the tensor into three matrices that corresponds to three intrinsic homography matrices of the three distinct planes. The planes are associated with the camera coordinate frame of the third view and provide a reference basis for reconstruction of invariants. This provides a geometric intrinsic structure of three views.

The author claims that the tensor offers a host of a new algorithms for recovering 3D information from 2D views, cuts through the epipolar geometry, makes room for statistics, and generally exploits the information available from measurement across views in a more efficient manner than any technique based on 2-view geometry.

Among the advantages: new algorithms for recovering 3D information from 2D views, an order of magnitude improvement compared to conventional techniques that rely on epipolar geometry (when synthesizing novel views from a pair of model views), applications in virtual reality, 3D television, recognition, fast rendering, 2-views structures (fundamental matrix, epipoles) are recoverable (linearly) from a tensor.

Among the disadvantages: static scene, some fiducial points are needed.

Class of Approaches: implicit-geometry rendering [44].

## Avidan-Shashua [2]

This paper proposes a method where views are reconstructed directly without first estimate the depth, by exploiting certain invariants in the geometry of the problem.

Input consists of 3 images from which it is possible to compute a trilinear tensor who will provide a correct way to generate virtual views of the observed object. In particular, the trilinear tensor is computed from the point correspondences between reference images. In case of only two images, one of the images is replicated and regarded as third image. If the camera intrinsic parameters are known, then a new trilinear tensor can be computed from the known pose change with respect to the third camera location. The new view can subsequently be generated using the point correspondences from the first two images and the new trilinear tensor.

The authors claim that the trilinear tensor gives user wider perspective transformation possibilities than other methods in literature. Texture is achieved by an interpolation of reference images. A realistic effect is achievable with this technique, however, image rendering might not be real-time because of the dense matching and tensor computation.

Among the advantages: realistic effect, the use of tensor (recovering 3D information from 2D views, no epipolar geometry, etc), efficient synthesis of novel views and wide visualization range, applications in virtual reality, 3D television, recognition, fast rendering, 2-views structures (fundamental matrix, epipoles) are recoverable (linearly) from a tensor.

Among the disadvantages: this approach does not correctly reconstruct points that become occluded.

Class of Approaches: implicit-geometry rendering [44], points transfer [19].

## Laveau-Faugeras [27]

The authors propose a system where views are reconstructed directly without first estimate the depth. Under the assumption that a complete pixel-wise correspondence is available, it is possible to predict a broad range of views. The use of epipolar geometries between images restricts the image flow field in such a way that it can be parameterized by a single disparity value and a fundamental matrix which represents the epipolar relationship. The authors also provide a two-dimensional ray-tracing-like solution to the visibility problem which does not require an underlying geometry description. Their method does, however, require establishing correspondence for each image point along the ray's path.

Class of Approaches: implicit-geometry rendering [44], points transfer [19].

## Chang-Zakhor [8], [9]

This paper presents a method to generate arbitrary views of three dimensional scene by means of an intensity-depth representation.

By using an uncalibrated camera which scans a stationary scene under approximately known camera trajectories, and then by transforming points on camera image planes onto the plane of the virtual view, the proposed system derives dense depth-maps at several preselected viewpoints.

The authors propose an adaptive matching algorithm which assigns various confident levels at various regions. Once the depth maps are computed at preselected viewpoints, the intensity and the depth at these locations are estimated using a stereo algorithm and used to reconstruct arbitrary views of the 3D scene.

Among the advantages: fast and flexible image acquisition (hand held cam-corder, uncalibrated cameras, unknown camera position), well estimate depths, image-quality good for the most part, well-reconstructed horizontal edges, few errors concerning occluded regions.

Among the disadvantages: artifacts due to specularities of the surface, image matching performed poorly for background regions which are seen through holes of foreground regions, static scene (stationary 3D objects), only horizontal motion.

Class of Approaches: geometric-valid pixel reprojection [26].

## Rousso-Peleg-Finci [39]

This paper concerns with stitching together images from adjacent viewpoints in order to generate a realistic panoramic virtual view of an observed environment. The authors propose an algorithm based on the method proposed in [37], to solve the main problem of panoramic mosaicing which is related to the forward camera motion (e.g. zooming). Pictures are segmented in vertical strips which are aligned by a "stretching" technique. In this way distortions appear greatly reduced.

## CohenOr [12]

This paper presents a way to exploit projective texture-mapping to render adjacent views of reference images. The authors called these views *Extrapolated Views*. The aim was to improve time-performance of a walk-through in remote virtual environment.

## Hirose [23]

This paper proposes the use of a camera with position sensors in order to make a interactive walk-through, based on pre-recorded sequence of images which are stored in a

database. The use of image interpolation greatly reduces the required number of pre-recorded images while the known image-position allows to the system to recover reference images of interest from the database.

### Regan-Pose [38]

This paper describes a hybrid system in which plenoptic samples are generated on the fly by a geometric-based rendering system at available rendering rates, while interactive rendering is provided by the image-based subsystem. At any instant, a user interacts with a single plenoptic sample. The authors also discuss local reconstruction approximations due to the changes in the viewing position. Local reconstruction approximations amount to treating the objects in the scene as being placed at infinity, resulting a loss of kinetic depth effects.

## 2 Model-Based Rendering

Model-based Rendering usually recovers the geometry of the real scene and then render it from desired virtual view points. Methods for the automatic construction of 3D models have found applications in many field, including Mobile Robotics, Virtual Reality and Entertainment. These methods generally fall into two categories, active and passive methods.

Active methods often require laser technology and structured lights or video, which might result in very expensive equipments. However, new technologies have extended the range of possible applications, (Levoy et al. [29], Hogg et al. [24], Fisher et al. [18]), and new algorithms have improved the ability to cope with problems inherent to laser scanning, (Castellani, Livatino and Fisher [7], [6], Stulp [46], Davis et al. [14]).

Passive methods usually concerns the task of generating a 3D model given multiple 2D photographs of a scene. In general they do not require a very expensive equipment, but quite often a specialized set-up, (e.g. Kanade et al. [25], Fuch et al. [20], Tseng and Anastassious [49]). Passive methods are commonly employed by Model-Based Rendering techniques.

Some of the research contributions in the field have proposed fully working systems for specific applications, some other have instead mostly focused on one or some of the involved aspects but provided a general application context. For example, Moezzi et al. [34], [35], propose an entire specific system for image-acquisition, model-construction and play-back interactive rendering, while Ofek et al. [36], mostly focus on extraction of textures from a generic video sequence for high-fidelity model-based texture mapping.

There can well be different ways of generating 3D models from photographs, from sim-

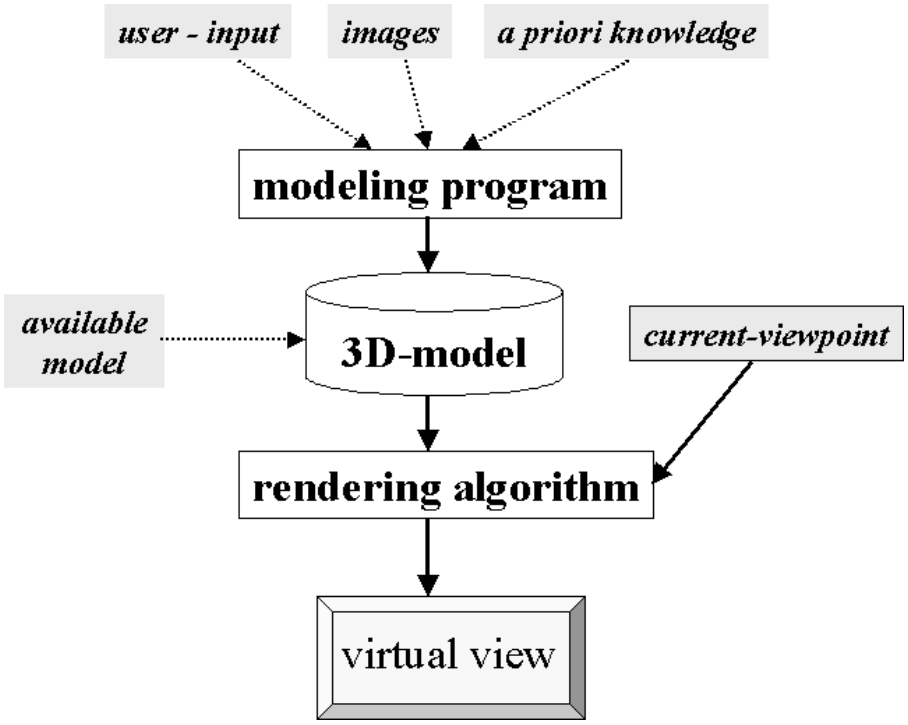


Figure 2: Model-Based Rendering: typical computational-steps.

ple 3D silhouette models dynamically cut-out and texture-mapped from video-sequences, (Livatino and Hogg, [31]), to polyhedral visual hulls generated by multiple-view silhouettes [2] and estimated stereo disparities, (Li, Schirmacher and Seidel, [30]). A survey of image-based volumetric scene reconstruction can be found in the works of Slabaugh et al. [45], Dyer [17], and Forsyth and Ponce [19].

In this section summaries of some representative works in Model-Based Rendering are presented. The authors name at the top of each summary identifies presented approach together with a "pioneer" reference publication.

Figure 2 represents typical computational steps in presented works.

## Kanade-Narayanan-Rander [25]

Kanade et al. coin the term *Virtualized Reality* to characterize a system that is able to capture dynamic scenes and render them from different virtual viewpoints. This in order to immerse viewers in a virtual reconstruction of real-world events.

[2] The visual hull represents a conservative shell that envelops the true geometry of the objects, consisting in the shape obtained from silhouette image data. The visual hull techniques require that foreground objects in the input images can be segmented from the background.

A visual event such as an actor motion is captured using many cameras (from six cameras to many more) placed all around a hemispherical dome 5 meters in diameter that cover the action from all sides. Consequently, several real-images of the scene are acquired.

The 3D structure of the event, aligned with the pixels of the image, is computed from a few selected directions using a stereo technique. In particular, depth information are recovered by stereo matching and the combination of depth/color data converted into a triangle mesh model on graphics rendering engine.

The authors use a multi baseline stereo algorithm to compute time-varying 2.5-D depth maps representing the scene geometry. Stereo-based depths are aligned with the pixels of their corresponding images.

Based on the viewer's position, the depth map from the closest camera is used to render the scene. Triangulation and texture mapping enable the placement of a "soft-camera" to reconstruct the event from any new viewpoint. Virtualized reality allows a viewer to move freely in the scene, independently from the angles used to record the scene.

Among the advantages: the system provides 3D structure of events for virtual reality applications, safe training, user guided visualization of events (for entertainment).

Among the disadvantages: the system does not allow for an on-line processing, (but only play-back) due to the high computational cost.

Class of Approaches: volumetric reconstruction [19].

## Fuchs-Bishop-Arthur-McMillan-Bajcsy-Lee-Farid-Kanade [20]

This paper proposes the use of image data acquired by many stationary cameras installed around a small environment such as a conference room and the use of stereo methods to compute time-varying 2.5-D depth maps representing the scene geometry.

The authors propose to reconstruct the real world scene from a large amount of fixed cameras by applying a correlation based depth from stereo. Wide baseline stereo is used to extract depths maps, which are updated, maintained, and then combined to create a virtual scene from viewer current position and orientation. The data can be acquired in a remote site while the viewer position and orientation is local.

In shown results a depth image of a human subject is calculated from 11 closely spaced video camera positions. The user is wearing a head-mounted display and walks around the 3D data that has been inserted into a 3D model of a simple room.

Among the advantages: the system is suitable for teleconferencing applications, provide 3D structure of events for virtual reality applications, safe training, user guided visual-

ization of events (for entertainment).

Among the disadvantages: only play-back is allowed, (the speed of the stereo algorithm was much limited by the poor machine performance).

## Tseng-Anastassiou [49]

This paper proposes the use of images captured simultaneously by a set of equi-distant cameras with parallel axis, in vertical and horizontal lineups and the use of stereo methods to compute time-varying 2.5-D depth maps representing the scene's geometry. Virtual images are generated by interpolating real views scanline-by-scanline based on disparity information.

Within the MPEG standardization the transmission of a stereoscopic (left and right views) sequence is possible by utilizing the proposed high profile double layer structure of temporal scalable coding. The left stereo sequence is coded on the lower layer and provides the basic non-stereoscopic signal. The right stereo bitstream is then transmitted on the enhancement layer and when combined with the left view results in the full stereoscopic video. After decoding the two extreme views, an "intelligent" scheme is proposed to interpolate the intermediate views.

Among the advantages: MPEG can be applicable to two sequences of stereoscopic signals through the use of spatial and temporal scalability extensions, easy and direct implementation convenience, graceful stereo image degradation, and high SNR reconstructions, compression, and improvement in image reconstruction.

## Moezzi-Tai-Gerard [35]

This paper proposes to recreate the original dynamic scene in 3D, the system allows photo-realistic interactive playback from arbitrary viewpoints using video streams of a given scene from multiple perspectives.

The idea is to capture multiple images of an object and then construct a 3D textured model of an object and use view-dependent texture mapping for rendering (from any view angle). The authors use 17 cameras surrounding a stage area to record various performances. The 3D model is extracted by an accurate recovery of the 3D shapes of dynamic or foreground objects by means of a volume occupancy method.

This work is based on Moezzi et al. [34] who construct a visual hull, i.e. a conservative shell that envelops the true geometry of the objects, consisting in the shape obtained from silhouette image data. The visual hull is constructed using voxels in a off-line processing system. The shape of the visual hull can be determined from object silhouettes in multiple images taken from different viewpoints, (the silhouette information is obtained by background subtraction).

In order to render the obtained model from a virtual camera point of view, a true 3D model is created with fine polygons, each separately colored. There is no need for texture rendering support and the viewing position plays no role in the modeling process.

The proposed approach can use standard object formats such as VRML delivered through the Internet and viewed with VRML browsers. Hence, the approach is suitable to the client-server scenario because real views do not need to be transferred to the client.

Among the advantages: accurate 3D model reconstruction, no texture rendering support needed, use of VRML format for browsing, suitable for transmission.

Among the Disadvantages: need for a off-line processing.

## R.Szeliski [48] [47]

R. Szeliski proposes different ways of computing image warping and he recovers a 3D model depending on the application: 2D planar image mosaicing, partial 3D model recovery and fully 3D model recovery. When recovering a full 3D model the utilized techniques are: volumetric description from silhouette or stereo matching from image sequence.

In volumetric description, the 3D model is recovered from a binary silhouette of an object against its background, local optical flow is computed and converted into sparse 3D point estimates, and the occluding contours of an object are tracked to generate 3D space-curves.

These techniques are suitable to reconstruct an isolated object undergoing known motion. Similar techniques can however be used to solve a more general 3D scene recovery problem where the camera motion is unknown. In particular, it is proposed the projective motion algorithm for determining an object motion based on recovering "projective depths".

Among the advantages: possibility for automatically creating large panorama images of arbitrary shape and detail.

Among the Disadvantages: limited 3D rendering.

## CohenOr-Rich-Lerner-Shenkar [13]

The paper proposes the use of a textured mapped voxel-based model to represent terrains and 3D objects.

The use 3D voxels-model is proposed because this fits better a high-detailed real-object, such as a real terrain, than a polygonal-model. Terrains are textured from b/w photographs and some objects (e.g. house buildings) are textured by a more detailed texture. The author uses b/w photographs because of the applications on Missile cameras. This leads to generate a less realistic rendering than using colors but it allows for real-time performance.

The system is based on a portable software rendering able to generate photo-realistic images in real-time (on a parallel machine). This performance is due to an innovative rendering algorithm based on discrete optimized ray-casting algorithm, accelerated by ray-coherence and multiresolution transversal.

Among the Advantages: real-time fly-through, portable software rendering, photo-realism.

As mentioned above, some of the research works have mostly focused on one or few aspects involved in realistic visualization of virtual-views. These works are also relevant for model-based rendering. For example, an automatic extraction of textures from a generic video sequence could represents a convenient approach to model-based texture-mapping, as proposed by the research work following summarized.

## Ofek-Shilat-Rappoport-Werman [36]

The proposed method focus on automatically deriving realistic 2D textures from video sequences for texture mapping purposes. The term realistic is here used to indicate textures free of disturbing effects such as highlights, reflections, shadows, etc.

The recorded scene is a video-sequence where the object of interest is viewed in different resolutions and different perspectives. A simple 3D model may also be provided by the user to improve system performance. The authors propose a model given by hand from five point at least. The authors also discuss an automatic model generation through a mask.

A multiresolution texture is proposed and for each pixel the color is computed by a weighted average from correspondent pixels in the video sequence. The multiresolution texture is proposed to be exploited for generation of virtual views of recorded scenes.

The proposed approach allows for identification of undesired features like highlights, reflections and shadows. The system is able to recognize the above features by looking at patches which contain very sharp step edges, since these are most likely to occur with depth discontinuities or reflective highlights. The system is then able to conveniently remove the disturbing features from the texture.

The image quality in resulting textures is as high as the original video-stills and so suitable to be used as reference-views. During visualization reference-views can be selected view-dependently based on current observation viewpoint, and the closer reference-view can be used for the mapping.

Among the advantages: suitable for mapping textures on 3D models from video-sequences, suitable for merging texture appearing in different resolutions, efficient storage of the resulting texture in a multiresolution data structure.

The following two methods are related to both model- and image- based rendering, but they represent different implementations. In the first work, the authors propose both the two approaches one at a time, for the same application contexts. In the second more complex approach, the system generates virtual views based on both, a reconstructed geometric model and an image-based texture-mapping (view-dependent).

## Blanc-Livatino-Mohr [3] [5]

The authors present a methods allowing for the exploration of a 3D scene based on triangular-mesh model recovered from 2D views. A comparison between the proposed method for virtual view-synthesis based on a sparse match and a view-synthesis based on a dense match, [4], is also proposed, (earlier in [3] and later in [5]).

The 2D views are photographs of a real scene and proposed as reference views. From these references either a projective model ([4]) or a triangular 3D mesh ([3]) is estimated. Hence, virtual views can be generated to allow a user to virtually navigate inside the scene and appreciate the tri-dimensional structure.

The method based on a dense match uses point reprojection. This method starts with a dense matching between the reference views and then each matched couple is reprojected using the trilinear relations to generate a new view from arbitrary viewpoints.

The method based on sparse matches uses model-based rendering (based on textured triangles). First, a sparse match (e.g. corner points) between the reference views is computed. Then, a textured filtered triangular mesh is calculated based on an initial Delaunay triangulation. Eventually, new views are synthesized by a model-based rendering.

The realistic effect in the first method is due to the fact that each pixel is directly reprojected from the real references views to the virtual view. In the second method the realistic effect is due to the fact that triangles are filled in with the texture from the reference views.

Among the advantages: no camera calibration, no 3D-model of the scene with the first method (a projective model is enough), fast synthesis, applications to high-rate video compression (only the references views and the displacement of the camera need to be transmitted and transmitted data does not depend on image size), fast sparse matching and no "holes" in the synthesized views with the second method as well as a mostly automated fast and realistic scene modeling.

Among the disadvantages: not real-time matching phase because of the computed dense-matching (first method), "holes" in the synthesized images arising from unmatched points and from adjacent pixels in the reference views which are not adjacent in the synthesized view (first method), more than two reference views are needed because of the trilinear tensor (first method), false matches are exacerbated if visualization is required from view-

points distant from the original viewpoint (second method).

## Debevec-Taylor-Malik [15]

The system developed by the authors use photographs and an approximate geometry to create and render realistic models of architectures.

The system requires only a small number of photographs, i.e. fews different views of the object, and a few indications to specify an approximate geometry and rough correspondences between the photographs. A method for photogrammetric modeling implemented on an interactive modeling program (called "Façade") is proposed for this purpose.

The model is then refined by means of proposed *model-based stereo*, which exploits estimated geometry (a coarse object model) and epipolar relations, to match stereo views on a wide baseline. In particular, by re-projecting one image of the stereo pair from the other image viewpoint. In this way, the foreshortening problem for the wide baseline stereo pair is eliminated and the stereo reconstruction can be done more robustly.

When an object is rendered, this is textured by proposed view-dependent texture-mapping technique which interpolates textures coming from photographs which are closer to current viewpoint. In particular, the interpolation is performed using a geometric model to determine which pixel from each input image corresponds to the desired ray in the output image. Among the corresponding rays, those that are closer in angle to the desired ray are weighted to make the greatest contribution to the interpolated result.

A method for averaging textures of neighboring regions in case when regions are textured from different image-sources is also proposed, in order to avoids seams and abrupt transitions of textures.

A view-dependent texture mapping is later proposed in [16] to further reduce the computational cost and have smoother blending, by means of visibility processing, polygonal view-maps, and projective texture-mapping.

Among the Advantages: sparse set of reference images, wide-baseline stereo matching, realistic response.

Among the Disadvantages: projective texture mapping involves expensive computation, the visibility problem needs to be addressed, texture seams may arise.

Class of Approaches: explicit-geometry rendering [44], volumetric reconstruction [19].

# 3 Summary

Image-domain approaches emphasize the role of photographs or still-video in order to provide realism to computer generated images. This methodology uses 2D photographic images instead of 3D geometrical models. Realistic virtual views as they are seen from a virtual camera can be generated by an image-based rendering algorithm without any tedious reconstruction of 3D models. Views generated using this method have an advantage compared to those created using the geometrical model-based method, generating the same image quality is much easier. In addition, views generation is relatively easy following "preparation" of the 2D images.

In general image-domain approaches need less computation resource than 3D model-based approaches and the produced image quality is as good as conventional 2D media. However, interaction with the world is limited and they need larger amount of data space, because they have to handle redundant data. Huge amount of data space represents a trade-off for making application involving networks, since a high bandwidth is required to share an image-based virtual world. Also, image-based approaches limit supported virtual views to a "narrow range" and scene is constrained to convex and not occluded objects.

Model-based approaches, on the other hand, are very generic, capable of generating any world and object by using a geometrical model from the beginning. Users encounter no limitations in interacting with the world. Model-based approaches make larger the range of the possible virtual views, and faster the rendering process since they can exploit hardware rendering provided on nowadays graphics-workstations.

Depending on the details and fidelity of the recovered model, model-based methods can yield realistic images. In particular, a realistic synthesis depends on: accuracy of the geometric model of the objects, textures, object surface properties, illumination simulations, rendering algorithms, etc. Pre-acquired information, heuristics and additional effects can also be integrated. A summary of main advantages and disadvantages of the two classes of approaches is presented in figure 3.

As it comes out from the state of the art, a realistic image synthesis of virtual environments is a large field of applications not yet generally solved. The success of some approaches mainly depends on the application and application constraints. However, from the many different techniques proposed, it is possible to gain a general idea on what approach and on what technique may better fit the individual application dependencies. The parameters playing an important role are:

* real-time performance;
* static/dynamic environment;
* convexity/concavity of the scene objects;

| | Advantages | Disadvantages |
|---|---|---|
| Model-Based Rendering | • very generic approach, any world any object<br>• no restriction in virtual views<br>• no limitation in interacting with the world<br>• exploitation of progress in graphic hardware | • 3D-model reconstruction<br>• approximated 3D-models<br>• strongly dependent on CPU capability and special hardware<br>• approximate realism |
| Image-Based Rendering | • realistic visualization (image quality as conventional 2D media)<br>• no 3D-model reconstruction<br>• rendering time independent from scene complexity | • limited interaction with the world<br>• narrow range of possible virtual views<br>• software rendering only<br>• strongly dependent on memory capacity<br>• high bandwidth to transfer data<br>• scene dynamics difficult to achieve |

Figure 3: The table summarizes advantages and disadvantages of image- and model- based rendering techniques.

* object visibility and mutual occlusions;

* range of perspective transformations required;

* level of fidelity required.

Several years ago the image-based approaches were unthinkable to propose since memory devices and high speed data links costed so much. Then amazing advances in semiconductor technologies, including reduction of memory device cost, made possible to explore such methodology.

A remarkable progress has also been reached by computer graphics technology, and by 3D geometric modeling. We currently have many sophisticated 3D graphics tools such as 3D modeling systems, 3D scanning systems, that combined with hours of labor let us generate sophisticated graphics images as seen in movies.

# References

[1] E.H. Adelson and J.R. Bergen. *The Plenoptic Function And The Elements Of The Early Vision*, chapter 1. MIT Press, Cambridge, MA, 1991.

[2] S. Avidan and A. Shashua. Novel view synthesis in tensor space. In *Conference on computer vision and pattern recognition*, pages 1034–1040, San Juan, Puerto Rico, June 1997.

[3] J. Blanc, S. Livatino, and R. Mohr. Fast and realistic image synthesis for telemanipulation purposes. In *European Workshop on Hazardous Robotics*, pages 77–83, Barcelona, Spain, November 1996.

[4] J. Blanc and R. Mohr. From image sequence to virtual reality. In E.P. Baltsavias, editor, *ISPRS Intercommission Workshop: From Pixels to Sequences*, pages 144–149, Zurich, Switzerland, March 1995.

[5] J. Blanc and R. Mohr. Towards fast and realistic image synthesis from real views. In *The 10th Scandinavian Conference on Image Analysis (SCIA)*, pages 455–461, Lappeenranta, Finland, June 1997.

[6] U. Castellani and S. Livatino. Scene reconstruction: Occlusion understanding and recovery. In Robert B. Fisher, editor, *CVonline: The Evolving, Distributed, Non-Proprietary, On-Line Compendium of Computer Vision*. School of Informatics, University of Edinburgh, December 2001.

[7] U. Castellani, S. Livatino, and R.B. Fisher. Improving environment modelling by edge occlusion surface completion. In *1st International Symposium on 3D Data Processing Visualization and Transmission (3DPVT)*, Padova, Italy, June 2002.

[8] N.L. Chang and A. Zakhor. Arbitrary view generation for three-dimensional scenes from uncalibrated video cameras. *Speech and Signal Processing*, 1995.

[9] N.L. Chang and A. Zakhor. View generation for three-dimensional scenes from video sequences. *IEEE Transaction on Image Processing*, 6(4):584–598, April 1997.

[10] S. Chen and L. Williams. View interpolation for image synthesis. In *Computer Graphics (SIGGRAPH'93)*, pages 279–288, August 1993.

[11] S.E. Chen. Quicktime vr - an image-based approach to virtual environment navigation. In *Computer Graphics (SIGGRAPH'95)*, pages 29–38, August 1995.

[12] D. Cohen-Or. Model-based view extrapolation for interactive vr web-system. In *Computer Graphics International*, Hasselt, Belgium, 23-27 June 1997.

[13] D. Cohen-Or, E. Rich, U. Lerner, and V. Shenkar. Real-time photo-realistic visual flythrough. *IEEE Transactions on Visualization and Computer Graphics*, 2(3), September 1996.

[14] J. Davis, S.M. Marschner, M. Garr, and M. Levoy. Filling holes in complex surfaces using volumetric diffusion. In *1st International Symposium on 3D Data Processing Visualization and Transmission (3DPVT)*, Padova, Italy, June 2002.

[15] P.E. Debevec, C.J. Taylor, and J. Malik. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In *Computer Graphics (SIGGRAPH'96)*, pages 11–20, August 1996.

[16] P.E. Debevec, Y. Yu, and G. Borshukov. Efficient view-dependent image-based rendering with projective texture mapping. In *9th Eurographics Workshop on Rendering*, pages 105–116, 1998.

[17] C.R. Dyer. Volumetric scene reconstruction from multiple views. In L.S. Davis, editor, *Foundations of Image Understanding*, pages 469–489. Kluwer, 2001.

[18] Fisher R.B. (coordinator), Div. Informatics, University of Edinburgh. The camera project (cad modelling of built environments from range analysis), eu tmr-project. http://www.dai.ed.ac.uk/homes/rbf/CAMERA/camera.htm, 1998-2001.

[19] D. Forsyth and J. Ponce. *Computer Vision - A Modern Approach*, chapter 26, Application: Image-Based Rendering, pages 780–808. Alan Api, 2002.

[20] H. Fuchs, G. Bishop, K. Arthur, L. McMillan, R. Bajcsy, S. Lee, H. Farid, and T. Kanade. Virtual space teleconferencing using a sea of cameras. In *First International Symposium on Medical Robotics and Computer Assisted Surgery*, pages 161–167, 1994.

[21] S.J. Gortler, R. Grzeszczuk, R. Szeliski, and M.F. Choen. The lumigraph. In *Computer Graphics (SIGGRAPH'96)*, pages 43–54, New Orleans, August 1996.

[22] Granum E. (coordinator), CVMT lab., Aalborg University. Being there without going (benogo), eu fet-project. http://www.benogo.dk, 2002-2005.

[23] M. Hirose. Image-based virtual world generation. *IEEE Multimedia*, 4(1):27–32, March 1997.

[24] Hogg D. (coordinator), Dept. Computer Studies, University of Leeds. The resolv project (reconstruction using scanned laser and video), eu-project. http://www.scs.leeds.ac.uk/resolv, 1995-1999.

[25] T. Kanade, P. Narayanan, and P. Rander. Virtualized reality: Concepts and early results. In *IEEE Workshop on Representation of Visual Scenes*, pages 69–76, June 1995.

[26] S.B. Kang. A survey of image based rendering techniques. *VideoMetrics, SPIE*, 3641:2–16, 1999.

[27] S. Leveau and O. Faugeras. 3-d scene representation as as collection of images and fundamental matrix. Technical Report 2205, INRIA Sophia-Antipolis, February 1994.

[28] M. Levoy and P. Hanrahan. Light field rendering. In *Computer Graphics (SIGGRAPH'96)*, 1996.

[29] Levoy M. (coordinator), Dept. Computer Science, University of Standford. The digital michelangelo project: 3d scanning of large statues. http://graphics.stanford.edu/projects/mich, 2002.

[30] M. Li, H. Schirmacher, and H.P. Seidel. Combining stereo and visual hull for online reconstruction of dynamic scenes. In *IEEE Workshop on Multimedia and Signal Processing*, December 2002.

[31] S. Livatino and D. Hogg. Image synthesis for telepresence. In *European Workshop on Semi-Autonomous Monitoring and Robotics Technology*, Las Palmas, Canary Islands, Spain, 7-10 January 1999.

[32] L. McMillan. Image-based rendering using image-warping - motivation and background. In *Computer Graphics (SIGGRAPH'99), course n.39*, August 1999.

[33] L. McMillan and G. Bishop. Plenoptic modeling: an image-based rendering system. In *Computer Graphics (SIGGRAPH'95)*, pages 39–46, August 1995.

[34] S. Moezzi, A. Katkere, D. Kuramura, and R. Jain. Reality modeling and visualization from mulitple video sequences. *IEEE Computer Graphics and Applications*, 16(6):58–63, November 1996.

[35] S. Moezzi, L. Tai, and P. Gerard. Virtual view generation for 3d digital video. *IEEE Multimedia*, 4(1):18–26, Jan.-Mar. 1997.

[36] E. Ofek, E. Shilat, A. Rappoport, and M. Werman. Highlight and reflection-independent multiresolution textures from image sequences. *IEEE Computer Graphics and Applications*, 17(6), March-April 1997.

[37] S. Peleg and J. Herman. Panoramic mosaics by manifold projection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 338–343, San Juan, Puerto Rico, June 1997.

[38] M. Regan and R. Pose. Priority rendering with a virtual reality address recalculation pipeline. In *Computer Graphics (SIGGRAPH'94)*, 1994.

[39] B. Rousso, S. Peleg, and I. Finci. Mosaicing with generalized strips. In *DARPA Image Understanding Workshop*, pages 261–264, 1997.

[40] S.M. Seitz and C.R. Dyer. Phisically-valid view synthesis by image interpolation. In *Workshop on Representations of Visual Scenes*, 1995.

[41] S.M. Seitz and C.R. Dyer. View morphing. In *Computer Graphics (SIGGRAPH'96)*, pages 21–30, August 1996.

[42] A. Shashua. Algebraic functions for recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):779–789, 1995.

[43] A. Shashua and M. Werman. Fundamental tensor: On the geometry of three perspective views. In *IEEE International Conference on Computer Vision (ICCV)*, pages 920–925, 1995.

[44] H. Shum and S. Kang. A review of image-based rendering techniques. In *SPIE Int. Conf. on Visual Communication and Image processing*, pages 2–13, 2000.

[45] G. Slabaugh, B. Culbertson, T. Malzbender, and R. Schafer. A survey of volumetric scene reconstruction methods from photographs. In K. Mueller and A. Kaufman, editors, *Volume Graphics 2001, Proc. of Joint IEEE TCVG and Eurographics Workshop*, pages 81–100, Stony Brook, New York, USA, June 2001. Springer Computer Science.

[46] F. Stulp. *Completion of Occluded Surfaces.* PhD thesis, Rijksun Universiteit, Groningen, Holland, 2001.

[47] R. Szeliski. Image mosaicing for tele-reality applications. In *IEEE Workshop on Applications of Computer Vision*, pages 44–53, Los Alamitos, California, 1994. IEEE CS Press.

[48] R. Szeliski. Video mosaics for virtual environments. *IEEE Computer Graphics and Applications*, pages 22–30, March 1996.

[49] B. Tseng and D. Anastassiou. Compatible video coding of stereoscopic sequences using mpeg-2's scalability and interlaced structure. In *International Workshop on HDTV'94*, Torino, Italy, October 1994.