12

Active Vision and Virtual Reality

Brian C. Madden[1]
Hany Farid[1]

ABSTRACT The experienced quality of virtual reality is currently limited by available computational resources, and will be for some time to come. It is essential that these limited resources not be wasted on the acquisition and processing of data that do not contribute significantly to the final percept. Efficient construction of the intermediate views that are the basis of some forms of virtual reality (telepresence) depends on the proper selection of the acquired views. A fixed array of sensors that afford adequate resolution over the entire scene can present a prohibitive cost in bandwidth and computation while complete sensor mobility is technically difficult to achieve without becoming unacceptably intrusive. Active vision provides a mechanism of effective and efficient resource allocation in the transformation of real scenes into virtual ones. The use of electronic cameras and lenses mounted on positionable platforms (pan and tilt) that can track objects of interest, maintain sharp focus on salient surfaces and zoom in on regions of interest can provide the functional equivalence of a much larger array of fixed sensors.

12.1 Introduction

Scientific advancement is often limited by technology, whether it is the need for a better microelectrode or a faster microchip. Virtual reality is currently in such an elastic region of advancement. Qualitative improvements in performance mirror quantitative improvements in the hardware. While it is certainly true that computational resources will increase and pixel size will decrease, improvements can also be obtained through better utilization of current resources.

In this chapter, we describe a system for acquiring virtual views of a remote real environment using an array of active cameras (see Fig. 12.1). Depending on the position of the participant, the content of the scene and the task at hand, the configuration of the cameras can be changed to improve the utility of the provided views. In the first section, we survey the range of systems that have been applied to this problem and discuss the compromises they incorporate to accommodate current hardware limitations. We then briefly discuss the advantages of the active vision paradigm and apply it to the telepresence view acquisition problem. In the next section, we describe the collection of photometric and range data, its integration

[1]GRASP Lab, Department of Computer and Information Science, The University of Pennsylvania

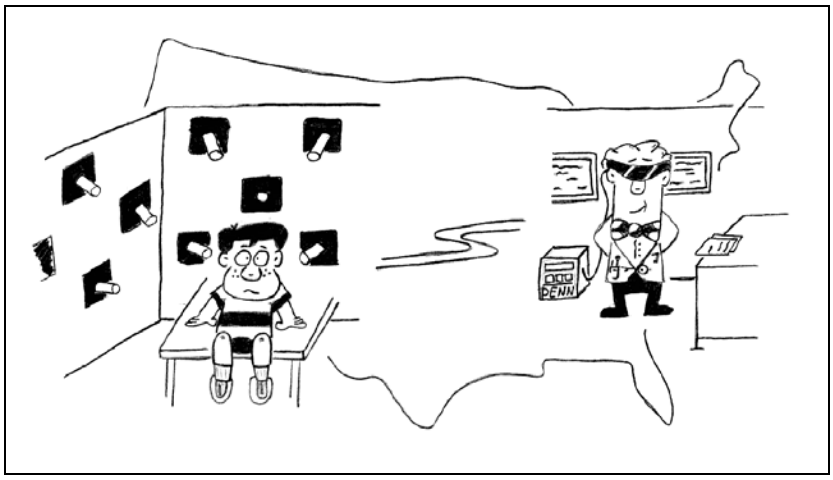

FIGURE 12.1. Active Telepresence Example. Using this system, a clinician wearing a head mounted display could view a patient thousands of miles away through the appropriate deployment of an array of positionable cameras. The combination of photometric and range information from the different cameras in the array in a world coordinate system would provide a representation that could be sampled from an appropriate perspective to provide the desired view.

into a single representation and the creation of an arbitrary intermediate view. In the following section, we apply the proposed techniques to both graphical simulations and digitized images. In the final section, we discuss some of the strengths and weaknesses of this approach and the direction of future work.

12.1.1 Virtual reality and telepresence

> "Imagine a wraparound television with three-dimensional programs, including three-dimensional sound, and solid objects that you can pick up and manipulate, even feel with your fingers and hands. Imagine immersing yourself in an artificial world and actively exploring it, rather than peering in at it from a fixed perspective through a flat screen in a movie theater, on a television set, or on a computer display. Imagine that you are the creator as well as the consumer of your artificial experience, with the power to use a gesture or word to remold the world you see and hear and feel. That part is not fiction. The head-mounted displays (HMDs) and three-dimensional computer graphics, input/output devices and computer models that constitute a VR (virtual reality) system make it possible, today, to immerse yourself in an artificial world and to reach in and reshape it (Rheingold, 1991)."

The basic principle of such a virtual reality system was introduced by Ivan Sutherland (1965). He proposed a system that supported real time spatial and temporal interactions with created worlds instead of numeric or symbolic simulations.

Since that time, the contributions of numerous researchers have made "virtual worlds" increasingly life-like. There is, however, a considerable distance to go before rendering and reality become indiscriminable. By approaching the problem in a different way, a new vision of telepresence has emerged (Fuchs & Neumann, 1993; Minsky, 1979), where *graphical worlds* are replaced with *real worlds*. The telepresence worlds are real in the sense that physical relations present in an actual environment serve to constrain the displayed scene whereas virtual reality scenes in general can be made up out of whole cloth and need not suffer from any external constraints. This distinction becomes especially important when the real environment being emulated is dynamic.

Telepresence is a simulation system where participants wear a head-mounted display to look around an actual, but physically remote, environment. The surface geometries of objects in the participant's line of sight are continuously sensed and updated as they maneuver through the remote environment. The challenge is to create for the participant an arbitrary 2-D image of the 3-D scene from the collection of available 2-D images currently being acquired. This creation must be done in real time (and potentially separately for each eye) to provide the views appropriate for the observer's current orientation in the scene and maintain the illusion of immersion within the remote environment.

A further distinction can be made within the domain of telepresence between systems that represent real worlds by manipulating digital images of the environment or by graphically rendering simulacra. The limited number of polygons that can be created in real time by graphics engines results in an assembly of subsampled surfaces that discards most of the minute variations that make up real objects. While it is true that the visual system itself discards most of these variations in the process of extracting and estimating parameters of the environment, their absence disrupts the feel of realism. Ultimately, as graphics *verité* improves, these distinctions will disappear. Until that time, however, digital images of dynamic complex scenes will be the best source of lifelike surface patterns and, at the very least, the source of the dynamic updating of the position of objects in the scene used for rendering.

The question is how to efficiently acquire these digitized views. Consider a camera that can be instantaneously repositioned with six degrees of translational (left/right, up/down and toward/away) and rotational (pitch, roll and yaw) freedom and can transmit images with no delay. With such a universal sensor, a telepresence system could be realized easily by simply displaying to each of the participant's eyes the appropriate camera views. Lacking such a device, researchers have applied various compromises to bridge the gaps in acquisition technology. Most approaches to telepresence are variations or combinations of the following methods.

Impersonal viewpoint: In the 1970s, Artificial Reality was the term applied by Myron Krueger (1991) to allocentric[2] systems that used nonintrusive tracking of

[2]Viewpoints are usually either first person (yours) or third person (someone else's). The application

participants based on pattern recognition techniques. The recovered position is used to update the patterns on a projection display that shows a representation of the participant's body in the virtual environment. This *second person* approach avoids many technical difficulties by restricting itself to capturing only the relation between the participant's body and the environment and not the associated personal point of view. Personalizing the view would require tracking of the six degrees of freedom of head position plus the direction of gaze and even the state of accommodation.

Flat view: Another technology that falls in the periphery of virtual reality is the large field of view displays used in some remote meeting applications. These displays afford percepts that are flat, with no stereo or motion parallax cues. Even for fields of view exceeding 60°, the participant gets no feeling of immersion at the remote site due to the lack of depth. Virtual gaze and movement are restricted and the participant becomes an observer. These devices fall more within the domain of multimedia and teleconferencing than in that of virtual reality.

Stereo view: Large field of view displays can take advantage of devices such as liquid crystal shutters mounted on glasses that alternate in synchrony with slightly disparate scenes on the display to induce stereopsis. The screen width acts as an aperture into a remote site. Ohya, Kitamura, Takemura, Kishino and Terashima (1993) used a 70-inch stereoscopic display as the vehicle to present wire-frame 3-D models of remote participants. Markers on the participants' faces were monitored by two cameras mounted on a helmet and magnetic markers were placed on the head and body to facilitate tracking of position. Articulation of the hands and fingers was measured by data gloves. The paucity of presentation detail allowed the wire-frame renderings to be presented in near real time (e.g., a 1000 node head model could be displayed at 6 frames per second). This rate is more than a factor of 3 away from that required to generate the illusion of smooth motion. In addition, latencies integral to tracking caused delays in the updating of the display. Initially distracting, these latencies can ultimately induce severe nausea. Although not incorporated into many systems yet, predictive techniques such as Kalman filtering can be used to compensate for much of the tracking lag (Azarbayejani, Starner, Horowitz & Pentland, 1993). Stereo systems bring with them another burden. By presenting images on a screen with a common objective depth, the participant is forced to dissociate accommodation from accommodative convergence. While human observers have some plasticity in this domain, the dissociation brings with it both fatigue and adaptation. Long term immersion could result in lowered performance both in the virtual world and subsequently in the real world.

Enveloping views: The CAVE[3] system (Cruz-Neira, Sandin, DeFanti, Kenyon & Hart, 1992) extends the large screen stereo approach to envelop the partic-

---

here uses the unusual perspective of viewing yourself remotely, hence the term allocentric, or *other-centered*.

[3]CAVE is a recursive acronym for CAVE Automatic Virtual Environment.

ipant in a cube of imaging surfaces. The advantage of this configuration over head-mounted displays is that only translation of the head need be tracked. Rotations and eye movements are unrestricted due to the simultaneous presentation of much of the remote scene on the facets of the cube. The apparent advantage of the panoramic presentation may be muted by the limited aperture of some stereo glasses. If the virtual objects are not too proximal, the tracking/display lags are not very intrusive since the large majority of perspective changes are small for minor head translations and mostly vanish for virtual distance exceeding 2 m. There can be, however, some distortion when objects overlap facet boundaries. It is difficult to compensate disparities in more than one plane for tangent distortion for all points of view. Other difficulties occur with this type of display when real objects such as hands improperly occlude virtual objects. Currently, the only way to avoid this conflict is to keep the virtual world behind the screen (*fish tank* VR).

Model-based: A different approach takes the form of a type of predictive coding wherein a model at the receiver is animated by key parameters extracted at the remote site. The savings in bandwidth are paid for with the increased computational load of reconstruction. Caudell, Janin and Johnson (1993) trained a neural net to reconstruct the silhouette of a face during speech given the location of a set of features (selected manually). They suggested that this technique can be extended to 2-D using polygons in the place of line segments. Terzopoulos and Waters (1993) applied a multilayer anatomical model that used the physical properties of the tissue, muscle and bone to generate deformations of the skin. The collection of these facial deformations structured expressions in a realistic manner. Complete range and photometric maps were obtained for each individual and were used to customize an epidermal mesh that overlaid the physical-anatomical model. The photometric information was texture-mapped onto the mesh to enhance realism. Extraction of key physical features (artificially enhanced) was used to control the deformation process from frame to frame. Extraction of facial features can also be used to track positional changes and allow the virtual head to be embedded in the remote environment (Azarbayejani et al., 1993). While the commercial potential of teleconferencing has resulted in much of the work being concentrated in the presentation of talking heads, these ideas may be extended to other classes of objects.

The approaches that incorporate high level knowledge of the remote environment overcome bandwidth limitations that often hobble systems that depend on real time transmission of digital images. However, the cost is great in terms of flexibility. The same *a priori* knowledge that allows efficient and detailed rendering also precludes the introduction of arbitrary (nonmodeled) participants, an implicit presumption of static world composition. In addition, while the local generation of images is quite advanced in some cases, the remote extraction of features in real time continues to be a stumbling block. Many of the actual facial deformations result in only the slightest of luminance gradients. In addition, the very extraction of the facial features of participants who themselves are wearing head-mounted displays is problematic. The advancement of hardware will both increase the advantage of encoding by improving the quality of rendered images, and decrease it

as the bandwidth available to transmit images from remote sites increases.

Mobile sensor: The ability to move a boom-mounted stereo camera pair that is controlled by the head movements of a remote observer through a remote environment (or a mockup of one) is a solution at the other extreme in the trade between remote sensor positioning and local image manipulation. Some of the earliest flight simulators that incorporated visual feedback did so by passing a camera over a scaled landscape model. As the trainee altered the controls of the simulated aircraft and viewed the *out-the-window* displays, the path of the camera would be altered proportionately. The advantage of boom-mounted cameras is that the dynamic changes in the remote environment are independent of its complexity. For many applications the use of remote sensors *in loco ocularis* can provide a relatively inexpensive solution while providing a good deal of functionality in a dynamic environment.

Pan and tilt as well as boom-mounted sensors exhibit positioning latency problems. The acquisition latencies caused by delays inherent in moving the platforms are often far below the formidable rates that the participants exhibit when changing their position and especially their gaze. The use of predictive systems here as well may reduce errors due to tracking latency to an acceptable level. Some systems that don't track the participant's position require skilled operation of the cameras or, at the very least, occupy the hands of the participant to move the sensors (Cruz-Neira et al., 1992). This control requirement reduces the ability of the participant to interact with the environment, both locally and virtually. Furthermore, the booms can be intrusive and sometimes even dangerous for certain applications (e.g. tele-surgery, Adam, 1994; Fuchs & Neumann, 1993).

Local dome: An intermediate approach that employs a remote camera on a pan and tilt platform to acquire a series of images from the remote site can be used to texture-map images onto a virtual dome. Hirose, Yokoyama and Sato (1993) presented observers with a series of views on a head-mounted display that reflected the sensed position of their heads. The local updating eliminates acquisition latencies (transmission limitations and delays in camera positioning) that could disrupt the illusion of remote presence and eventually induce nausea. Additional preprocessing that yields the relative parallax of objects at different depths by rotation about a point other than the nodal point of the lens was used to deform the virtual dome and to provide enhanced realism through parallax distortion of near (within 2 m) objects. Although new photometric information could be acquired for each new viewing direction, the system as proposed was designed to allow the observer only rotational degrees of freedom in a static world. As with other systems that rely on local computation based on models, this approach is limited in its ability to represent novel events in the remote world.

Camera array: One way to improve the speed of image acquisition over that available with a mobile system is to preposition a fixed array of cameras about the perimeter of the remote environment. Fuchs et al. (1994) proposed using a *sea of cameras* to acquire both photometric and depth information. They used a multibaseline stereo algorithm to compute the disparity of local image patches and then used the disparities to construct a polygonal model for a given camera position.

The photometric information was texture-mapped onto the model and the result was used as the basis of display while the participant was in a viewing position near that of the selected camera. As the participant moved about the remote environment, the process was repeated for the cameras in the array most appropriate for the current point of view.

An array of cameras mounted along the ceiling and walls eliminates the positioning problem but incurs a host of problems inherent with static sensors. It is difficult to adjust the camera parameters to be appropriate for image acquisition over a large volume of a remote environment. Very often the most appropriate camera will be obstructed. In addition, calculating accurate, dense, high resolution depth maps from a pair of images offset in space is far from being a solved problem. In particular, specularities disrupt accurate depth estimation even when problems in obtaining correlations due to occlusions do not occur. Beyond the issue of depth acquisition, the computational cost of constructing a model from stereo data of the remote environment at each instance in time is formidable.

Retinal projection: The virtual retinal display (Kollin, 1994) proposes to project a high resolution, panoramic, chromatic image directly on the retina of the eye. Using Maxwellian view optics and by tracking the pupil of the eye, they intend to pass a 140° unoccluded image through the entrance pupil of each eye. They also propose to control the divergence of the beam thereby inducing a stimulus to accommodation. If successful, this technological leap would go a long way toward eliminating many flaws currently present in virtual displays. Even this device, however, will not solve the problem of how to obtain the views that are projected.

Although some of these methods exhibit a considerable amount of telepresence, each also inflicts considerable limitations on the illusion of remote presence. These limitations are the consequence of the formidable task of determining the viewing state of the observer, obtaining the appropriate views (or the information necessary to recreate the views) and presenting the views. To match the available resources, compromises are made in the degree to which the participant can interact with the remote environment, the level of detail available, and the flexibility of the approaches in coping with dynamic or novel events. As the presentation technology improves, some of the compromises will be relaxed; however, many of the difficulties arise from a need to acquire remote views in a versatile, timely and economical fashion. Perhaps, then, in a manner similar to the way that a different approach (telepresence) was needed to improve upon the original concept of virtual reality to overcome hardware limitations, an extension of the concept of telepresence itself would now be useful.

12.1.2 Active vision

As long as bandwidth limitations exist in the real time presentation of remote scenes, there will be a need for strategies that maximize the percentage of transmitted information that is relevant to the requirements of the observer. The essential

canon of active vision[4] (Aloimonos, Weiss & Bandopadhay, 1987; Bajcsy, 1988; Ballard, 1991) is based on feedback from the current image altering future processing and acquisition. Active vision is a collection of sensor control strategies wherein the cost of mobility and calibration is more than offset by the increase in the density of information appropriate for the task, information that is selected from the staggering amount available. Passive (static) sensing of the environment often results in attempts to reconstruct the world that are underconstrained. There are limits as to how detailed a representation of a given scene can be obtained by simply pounding on a set of pixels, and even if theoretically soluble, limits exist on the stability of solutions in the face of acquisition noise. Beyond these considerations, a complete and detailed recreation of the entire surroundings is not required for most tasks (recognition does not require total reconstruction, Aloimonos, 1990).

Active vision techniques use the control of the sensors' pose relative to the objects' surfaces both to select manageable portions of the environment relevant to the current task as well as to simplify the extraction of the required parameters from those selected views. Active vision is not just a change in viewing parameters, but is an adaptive response to the demands of the task given the environment. The question arises, at what level are sensors active (e.g., at the level of pixels, algorithms, positioning mechanisms, filter shapes, goals or priorities) and what is the difference between being active and merely adaptive? Clearly, circumnavigating an obstacle to obtain information previously obscured on a remote object is active; but is stopping down an aperture active as well? Yes, if the manipulation extracts a needed new dimension from the scene not previously available (e.g., depth), and no, if only the signal to noise ratio is being improved. As Pahlavan, Uhlin and Eklundh (1993) proposed, to be more than just trivially active, a system must be making a choice between at least two alternatives. These choices can be implemented in many ways, on many levels.

Perhaps the most basic operation of active vision is tracking (the ability to acquire and to maintain acquisition of a region of interest). Stabilization of a moving target within a sensor window offers several advantages (Madden & Cahn von Seelen, 1995). A constant target size facilitates the use of object-centered coordinate systems. Multiple estimates of minimally varying targets can be obtained. The motion blur of a moving target can be reduced (or eliminated) while the induced motion blur of the background can be used to help segment objects of interest. Since the dynamic range of target motion need not fit into a static window, the resolution of the region of interest can be higher. All of these consequences of actively tracking targets serve to reduce the complexity of higher-level processing.

In cases where the Lambertian reflectance component can be determined, object motion can be derived simply by servoing on the location of a fixed intensity pattern (Aloimonos et al., 1987). The motion of the object is evident from the camera

[4] Active is used in the sense of exploring the environment, not to distinguish sensors that emit energy (laser rangefinders) versus passive collectors (ccd cameras).

movement required to maintain fixation rather than from some optic flow calculation. In cases where optic flow is useful, the flow patterns can be optimized by controlling the motion of the sensor to extract the desired parameters. Even a static scene can yield flow information relevant to depth if the sensor moves through it.

If the responses of two sensors are combined to derive stereo information, active cameras can be verged to improve disparity resolution for a fixed baseline or, by moving the sensor, larger effective baselines can be created by incrementally combining the ease of correspondence in a short baseline to obtain the disparity resolution available in the larger. The technique of continually improving pose to obtain more independent estimates or to eliminate obstructions to aid underconstrained computations yields more accurate estimates of the target because no regularization is required (Aloimonos et al., 1987). By actively varying the focus of the camera lens, depth estimates can be obtained with a single sensor, thereby avoiding the correspondence problem altogether (Krotkov, 1989).

The advantages of active vision go beyond these (considerable) aids to parameter extraction. Some researchers have expanded the concept that began as a sensory acquisition strategy to cover virtually all cognitive tasks that rely on visual input (Aloimonos, 1990; Ballard, 1991). Active vision can be shown to facilitate the execution of complex tasks. Many problems cannot be solved from a static perspective. Visual problem solving often requires a sequence of views and the scanning patterns that are optimal will vary with the task. Conversely, visual properties are often dependent on the behaviors (and views) that generated them.

The proper design of such techniques and tools is an open question, but one thing is clear: The observer is more than just a recorder. Variables at many levels will contribute to optimal performance. Interaction with the environment depends on the task at hand. The active pursuit of information that will allow determination of one set of parameters also uncovers future questions that need to be asked. Active visual processing is inextricably part of cognitive processing and problem solving.

12.1.3 Active telepresence

What problems need to be solved in a telepresence application? The essential service that must be provided is the delivery of views appropriate for the participant's state relative to that of the remote environment. The nature of the supplied views and the weighting of errors in timing, geometry and photometry will vary with the application. Certainly the veridicality of views in a clinical diagnostic setting will have a higher weight than in an entertainment application where accuracy may be sacrificed to smooth over deficits and preserve the suspension of disbelief. Lacking the universal sensor described above, arbitrary viewpoints must be interpolated from available images. We propose to do this by combining the photometric information with a depth map obtained for the same view. Both maps are then supersampled using bicubic interpolation and the result is projected into a world coordinate system. Forward ray tracing is used to determine the projection of the points onto the virtual sensor appropriate for the desired point of view. Points distal

to the sensor along a common ray are eliminated as obstructed and the remaining projections are mapped to the nearest pixel locations. This procedure is repeated for each remote sensor and the virtual view is a weighted composite using a registration algorithm and coarse camera calibration to align the components. (The method of interpolating views will be described in detail in Section 12.2.) While this technique provides a required service, it is not active in the sense developed above.

In what ways can the image acquisition techniques of active vision enhance the performance of telepresence systems? While the physical sensor support for active telepresence could take many forms, we have developed a scenario wherein an array of active sensors (pan/tilt/focus/zoom) arranged about the periphery of an environment can unobtrusively acquire information from that environment. Our approach builds on ideas contained in the *sea of cameras* configuration (Fuchs et al., 1994) as well as some of the dynamic aspects of the *local dome* system (Hirose et al., 1993). This configuration is a balance between intrusion (interference with the activities of the remote environment) and functionality (the ability to provide intermediate views of dynamic and novel scenes containing concavities and interpositions). The active component comes from using sensor input to reconfigure the array so as to improve the view presented to the participant.

Whereas a fixed array of cameras must distribute its coverage evenly over the environment (or cope with blind spots), positionable sensors can dynamically alter the sampling density and bring resources to bear where they are currently needed. In addition, differences in the location of the camera platforms allow a redundancy in the coverage of any given area to also contain a range of available pose. For example, high resolution zoom images of critical regions could be merged with data from coarser wide field of view images. In contrast to fixed sensor systems, targets of interest would not be restricted to a particular sweet spot but could be mobile. This concentration of resources could follow targets of interest around the environment. Alternatively, some sensors could be dedicated to the acquisition of fine-grained depth information without incurring the penalty of the associated photometric distortion. This specialization can be carried further through the use of different types of sensors in the array. Cameras with megapixel sensors could be mixed with color cameras or, as they become available, sensors designed for real time focus ranging (e.g., Krishnan & Ahuja, 1993). This sensor diversity does not necessarily require multiple arrays since most sensors can provide multiple functionality. It is just the ordering of their performance on these tasks that varies.

The information that can be extracted from a scene will vary with the scene dynamics. While we propose that the information required for the construction of arbitrary views can be extracted from the acquired images and does not need *a priori* models, this does not mean that all information must be extracted anew from each frame. The bottom-up models are indeed *disposable* in that new models can be constructed from the projection of the photometric data of subsequent images into world coordinates. However, it is also true that in many applications not much displacement occurs in 1/30th of a second. As is often true, in this case the faster (more frequently) a process is executed, the simpler it becomes. In scenes where

not much is changing, small deformations detected in the new data can be accommodated in the world projection of the old data at a lower cost than recomputation from scratch. At the very least, the previous results can be used as the seeds for calculations based on the new data. The savings in time accrued by this reuse will allow the representation to be augmented in terms of both resolution and dimensionality. If the changes are occurring apace, then the acquisition parameters must be ready to handle a much larger range. To accommodate variations in demand, we propose a coarse-to-fine presentation of photometric and depth values. In cases where this is unacceptable in terms of perceived quality, an alternative would be to display the closest acquisition view and then transform it into the desired view as sufficient data are acquired. While sensor lags in positioning do occur and could potentially limit performance, the consequences are less intrusive since the interpolated views only have to be positioned somewhere in the space spanned by the available basis views[5]. There need be no delay, therefore, as would be the case with a one-to-one acquisition/presentation methodology.

Another dynamic that must be accommodated is the often severe spatial gradients brought on by perspective, interposition and differences in viewpoint. The quality of surface representation depends on the pose of the virtual view relative to that of the acquisition views. If the projection from the virtual view is much closer to the surface normal than that of an acquisition view, the sampling will be coarse. Estimates of the surface normals can be obtained from the depth gradients and coverage by the array could be altered to improve the sampling. The closest camera does not always afford the best view. Cameras more lateral in the array could be panned to supplement the sampling of a particular surface. Sensors with optical axes 90° or more away from that of the desired view, depending on the orientation of the surfaces in the scene, can contribute to the final composite. Determining the optimal allocation of pose among the cameras in an array is an open question in active vision which we hope to address with this configuration. The actual coverage will depend on the array spacing which, in turn, will depend on the application. However, two or three rows of positionable sensors spaced about 1 m apart should provide adequate coverage for simple viewing tasks over a working volume 2 m high and several meters in depth for the length of the array.

The distribution of coverage by the array can also be altered to minimize the extent of hidden surfaces (Fig. 12.2). There are at least two conditions where boundaries of missing data can be detected. If an object (a continuous region of depth more proximal than the background) lies along a boundary of an acquisition image and that boundary does not project inside of one of the other acquisition images, it is likely that that edge marks a region of missing data. In addition, if the pixels on the edge of a continuous region of similar depth values (bounded by depth discontinuities) do not project into a continuous region of continuous

[5]The acquisition views supplied by the camera array do not necessarily afford a complete representation of all possible surfaces within their field of view. The determination of the set of interpolated views completely covered by a set of acquisition views given a particular scene is not a solved problem.

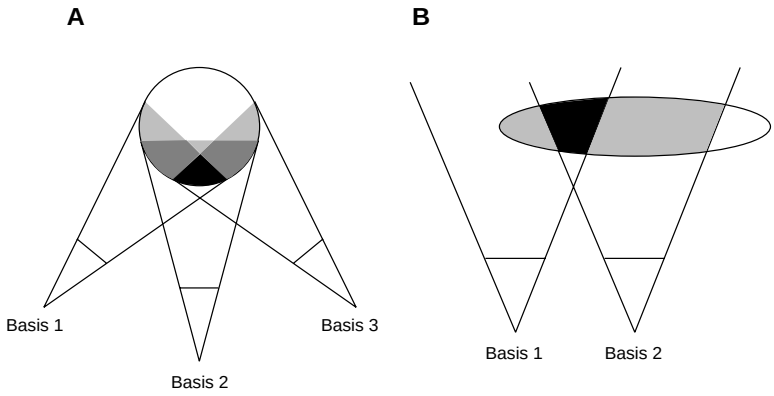

FIGURE 12.2. Hidden surfaces. In a given scene, not every surface will necessarily be completely within the span of the sensors; however, an occlusion boundary does not always indicate the edge of missing surface information. In A, each basis view is seen to cover a region of the sphere's surface that is slightly less than a hemisphere. The overlapping coverage often results in an occlusion boundary projecting to the middle of another view. It is only the boundaries that are not visible in other views that delimit the regions of missing data (unshaded). Similarly in B, while the ellipsoid is truncated by both sides of basis 2, only the right limit is a boundary of missing data. The minimization of such boundaries is one measure of the goodness of a given camera configuration.

depth values in another image, then they likely mark regions of missing data as well. Again, cameras in the array could be positioned to minimize the missing data boundaries or to eliminate them in a particular critical location.

In a manner analogous to the way the notions of purposive and animate vision extended the concept of active vision to more complex cognitive activities, it is possible to envision similar enhancements augmenting telepresence. A recent National Academy of Sciences report on virtual reality concluded that there are many applications where the enhancement of special features may be more important than realism (Durlach, 1994). Using data extracted from previous views, a participant could be guided to an appropriate pose for a given task. Work on using virtual fixtures has been shown to be a valuable tool for performing tasks remotely (Sayers & Paul, 1993, 1995). Fixtures don't just react to the state of the participants, they strive to actively assist their actions. The fixtures are always present in the environment but are activated only under the appropriate conditions. They channel the activity in a particular direction, but do not compel compliance. For example, a fixture could be used to maintain the position of a clinician relative to a freely moving remote patient but would still allow sufficiently large or rapid deviations in position or gaze to break off the attachment. Over time, use of such a *visual toolbox* would become second nature.

In summary, then, the extension we propose to the concept of telepresence is the assimilation of the techniques of active vision into the acquisition of images for remote viewing — an active telepresence. The essence of this proposal is that

a universal sensor can be approximated and even improved upon by the application of sensor strategies that interact with the remote environment.

In the following section, we will detail how the continuum of intermediate views that are necessary for an observer to experience a remote presence can be generated from a limited number of sensors. For images obtained from an array of active cameras that are adjusted appropriately in focus, zoom and direction, we will demonstrate that, barring any narrow concavities, acquired photometric data can be brought into registration, merged and warped into a pattern that approximates any viewpoint in the span of the camera positions (and, at increasing risk, may extend beyond that span).

12.2 Generating Views

The remote environment is a collection of point sources. Each point on an object acts as a radiator, throwing off a near continuum of spherical waves. With the exception of some special circumstances where these wavefronts exhibit phase coherence and produce interference patterns, the waves simply expand and superimpose on one another. A manifestation of the particle nature of light occurs at low intensity levels in the statistical distribution of photons. An active sensor acts as a sampler of this quantized superposition of diverging light. At a basic level, vision is the adjustment of the position and optical parameters of a sensor so as to select the appropriate portions of the wavefronts necessary to disambiguate the superposition with a criterion level of reliability. Beyond the contributions of the degrees of freedom inherent in the scene geometry, this disambiguation is rendered somewhat more complex by the superposition of specularities that not only retain varying degrees of the spectral composition of the illuminants but the associated wavefront curvature as well.

The fundamental resource of active telepresence is a sensor or sensors that can alter their properties to obtain the necessary estimates of surface reflectance in response to the state of the participant. A small array of active cameras can provide dense coverage of an environment by adjusting the fields of view to abut or overlap in a manner that supports the interpolation of intermediate views. With control of zoom, abutting fields of even frontoparallel cameras can be formed at arbitrary depth planes — an approximation to translation in depth. With control of pan and tilt, cameras can be brought to bear on regions throughout a wide range and, therefore, subjects need not be restricted to a special focal area. Each acquired view can be thought of as a *basis* vector in the multidimensional space that is the remote environment. The task is to obtain a sensor configuration that forms a basis set that spans that portion of the space containing the desired views. A major advantage of using such a basis set to construct virtual views is the decoupling of the acquisition frustum (the 3-D pyramidal view of the camera), from the participant's field of view, thereby easing restrictions on both.

There are as many potential sensor configurations as there are applications. The

scenario developed in this paper is based on an array of cameras arranged about the periphery of a scene. The proposed array is composed of three rows of sensors spaced 1 m apart horizontally and vertically. The center row is approximately at head height. This configuration provides good coverage (a good basis set) over a large volume (a variation is also being examined that shifts the center row laterally by half the sensor spacing to improve the acquisition of concave surfaces). The cameras are mounted on motorized platforms that allow a range of pan and tilt on an axis through the nodal point of the lens. The lenses are motorized and allow control of zoom and focus. The following sections detail the physical and algorithmic components that generate the interpolated views.

12.2.1 Camera calibration

A standard camera calibration method based on the method of Tsai (1987) is implemented in our system. This method calibrates both the extrinsic (rigid-body transformation between camera and world coordinate systems) and intrinsic camera parameters (focal length, lens distortion, and scaling parameters). The calibration technique requires a *calibration fixture* with at least seven, noncoplanar, calibration points with precisely known positions relative to an arbitrary world coordinate system. Each basis view may be calibrated by first determining the position of each of the calibration points in image coordinates. With calibration points in hand, the extrinsic camera parameters may be recovered by solving a set of, possibly overconstrained (i.e., more than seven calibration points), linear equations. Intrinsic parameters are recovered through a gradient descent, or comparable minimization (see Tsai, 1987, for more details). Although the mathematical principles underlying this calibration technique are sound and well understood, highly accurate camera calibration has proven to be quite elusive in practice. Problems exist in determining the *ground truth* of targets as well as locating these calibrated features in the digitized image with sufficient precision. In addition, nonlinear minimization is susceptible to being trapped by local minima during the selection of solutions for the camera parameters. As such, we propose to use the results of this camera calibration as the initial conditions for the image registration stage described below. To model the response of the lenses over the available range of focus and zoom, a number of static calibrations need to be performed (Willson, 1994). These results will form a lookup table that can be interpolated for any arbitrary condition of a lens.

12.2.2 Digitization

For the purposes of experimentation and validation, the array of active sensors will be approximated by positionable electronic cameras (a TRC BiSight 2-axis camera platform) mounted on a robotic arm (PUMA 560). Additional static cameras may be added to provide a wider range of simultaneous views. The combined eight degrees of positioning freedom can be used both to simulate the pan and tilt of the active sensor and the position of the sensors in the array. Two SONY

XC77-RR CCD cameras can be asymmetrically verged by the TRC platform with an inter-pan axis baseline of 25 cm. The pan axes are capable of a peak velocity of 1000 deg/sec and a peak acceleration of 12,000 deg/sec^2 while maintaining a precision of $2'$.

Offering three additional degrees of freedom per camera are the Fujinon motorized lenses (H10x11E-MPX31). The lenses have motorized focus and zoom (focal length: 11 to 110 mm). The 10 to 1 change in magnification has a corresponding effect on the field of view (approximately 4 to 40°). Light levels can be controlled either by a motorized iris in the lens or by an adjustable flux integration time (electronic shutter) in the camera. The monochromatic NTSC camera output is digitized (Data Translation DT1451; 512 by 480 pixels) and read into a SPARCStation (Sun Microsystems) for analysis.

12.2.3 Active estimation of surface depth

As pointed out by Pahlavan et al. (1993), the projection of the world onto a camera sensor is a 3-D to 3-D mapping. Although the sensor records the impinging wavefronts in a 2-D array of pixels, the distribution of the light from each point in the environment is distributed in a manner characteristic of its distance and the optics of the lens (Hopkins, 1955). This curvature of the optical wavefronts is commonly neglected in the formulation of the plenoptic function of the elemental sources of visual information (e.g., Adelson & Bergen, 1991). Along each ray from the nodal point of the lens, there is a distance at which a point radiator will produce the smallest blur circle in the sensor plane. The collection of these points form a (sometimes complicated) surface that will have the sharpest image. As a point being imaged moves along the ray between it and the nodal point in either direction, the distribution of the light from that point is enlarged. For incoherent light, a good approximation to the changing point image is a disk of varying diameter. As the point moves further from the surface that is conjugate with the sensor, the increasing disk diameters act as omnidirectional lowpass filters with ever lower cutoff frequencies.

Krotkov (1987) developed a technique to use these changes to assign depth values to local patches of the image. His focus ranging algorithm was based on determining the focus setting of a calibrated variable focus lens that produced the image with the greatest amplitudes in the high spatial frequency range. The Tennengrad metric of local image content was computed by convolving each small (10 to 20 pixels square) region of the image with 3 by 3 vertical and horizontal Sobel edge operators and combining them to form the oriented edge gradient at each pixel position. The sum of these high frequency bandpass filters over the local patch was tracked. Then, by knowing the relation between the focus setting and the distance of the surface that was conjugate with the sensor, a depth value could be assigned to a local image patch that corresponded to the peak of the Tennengrad metric. Depth resolution was found to be 1% of distance over a range of 1 to 3 m.

It is somewhat surprising that more effort has not gone into developing this technique given the advantage of not having to solve the correspondence prob-

lem. One reason focus ranging has been relegated to the backwaters of computer vision may be the lack, until recently, of a range of commercially available motorized lenses. Advances in technique should increase its application. Xiong and Shafer (1993) improved depth resolution by an order of magnitude by replacing the peaks found by a Fibonacci search method with the mean of a Gaussian fit to a range of focus settings about the maximum found by the Fibonacci search. Equally large advances could be made with better hardware. Much of the optics used today is designed to have a large depth of field. It is considered advantageous to have as much of the field of view in focus as the light level and sensor sensitivity will allow. Focus ranging, on the other hand, would benefit from a very small depth of field. In addition, many of the commercial motorized lenses are designed for surveillance applications where precision positioning of the optical components is not a requirement.

Advantage can be taken of redundancies and parallelism in the current application. Portions of the scene are viewed by more than one camera. The overlap can not only be used to consistently merge the different views into a world coordinate system, but can provide multiple opportunities for focus ranging. These overlapping views may have very different optical parameters, allowing very precise regions of depth measurements to be combined with larger regions of coarse values. Parallelism exists in the form of the hundreds of small patches in each acquisition image that are used to determine the Tennengrad metric. While it is possible that each patch is independently displaced in depth between frames, it is far more likely that in the 30th of a second frame interval the motion of the objects in the scene relative to the camera can be well modeled by a rigid (or at least a piecewise rigid) transformation. Depending on the reliability of the depth measures, significant deviations from rigidity can be incorporated into the projection of the data into the world view. The parallelism of focus metric computation is itself redundant in the regions of view overlap. In addition to range from focus (Krotkov, 1987) and range from defocus (Pentland, 1987), which requires detailed knowledge of the optics and two distinct focus settings, the method proposed here forms a third way to use focus information to constrain depth estimates — slight perturbations of hundreds or even thousands of depth computations that in the aggregate capture the approximately rigid change in the camera to object transformation that occurred between the acquisition of the previous and current frames.

12.2.4 Registration

Knowledge of the position of the basis cameras in a common world coordinate system is essential to successful interpolation. As was stated in Section 12.2.1, recovering accurate camera position (extrinsic parameters) from standard camera calibration techniques (Tsai, 1987) has proven to be a formidable task. As such, we propose to use the results of camera calibration as initial conditions for an image registration algorithm. The formulation of image registration developed by Besl and McKay (1992) starts with a model shape in a model coordinate system and aligns 3-D data taken from a sensor coordinate system by determining the

rotation and translation that minimizes the distance between the shapes using a mean-square distance metric. In our application, there is no model. We are interested in bringing two or more data sets, whose location in the world coordinate system is known only approximately, into mutual registration by aligning their overlapping regions. The data from the camera location nearest to the point of view being interpolated are taken as the basis to which the other shapes are registered. Using the results of camera calibration as initial conditions, we apply the the iterative closest point (ICP) algorithm of Besl and McKay (1992) to compute the accurate position of the basis views relative to one another. For views that have no overlap, registration into the world coordinate system is based solely on the camera calibration of the acquiring sensor.

12.2.5 Warping

A standard *forward warping* technique (Wolberg, 1990) is used to *warp* or *project* each of the basis views to the desired virtual view. Both the intensity image and depth map, as discussed above, are required for this procedure. The warping is performed by first projecting each sensor element in the basis view into a common world coordinate system, as specified by the associated depth map. This point is then reprojected into the virtual sensor plane (Fig. 12.3(A)). The intensity of the basis sensor element is assigned to the virtual sensor element whose region was intersected by the projection. Although this warping technique is straightforward, several technical difficulties arise. It is very common that a single sensor element in the basis view will not project into a single sensor element in the virtual view (Fig. 12.3(B)). As such, we currently perform a nearest neighbor interpolation and assign the intensity to the nearest sensor element. This interpolation is necessary for both the generation of the final digital image and for the hidden surface computation. That is, if two points project into the same sensor position in the virtual view after nearest neighbor interpolation, their depth values are used to determine which point should "shadow" the other. To avoid aliasing artifacts due to the nearest neighbor interpolation, the original basis views are upsampled by a factor of two before warping. The virtual view is subsequently subsampled accordingly.

Although the upsampling eliminates some of the aliasing, several artifacts still remain. We are currently investigating several approaches for eliminating these artifacts. With each of the basis views warped to the virtual position, a complete view is generated by integrating across each of these views.

12.2.6 Integration of views

The integration of warped basis views is currently based on a somewhat *ad hoc* weighting scheme (Fig. 12.4). In particular, for each pixel in the virtual view, the contribution of each warped basis view is given by $e^{-0.05\theta}$, where θ is the angle between the rays connecting the object in 3-space and the virtual and basis views. This weighting function was chosen so as to have a value of 1 when $\theta = 0$ (i.e., the virtual view is coincident with a basis view) and fall off gradually as θ increases

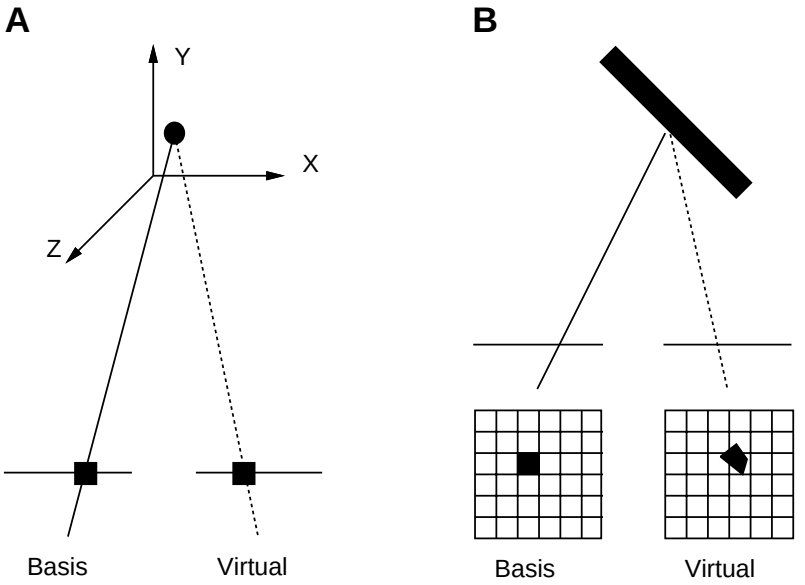

FIGURE 12.3. Warping of a pixel. A basis view is three dimensional. Each photometric map that is acquired must have a corresponding depth map. Each depth value is associated with a frustum delimited by the square or rectangular region that is the area of a pixel as seen from the nodal point. When projected into the world coordinate system and then back to the virtual view, the projection of the pixel area undergoes a perspective distortion. The resulting region that corresponds to a given basis pixel may cover (all or part of) several virtual pixels, or it may only fill a small portion of a single pixel.

(i.e., the virtual view moves away from the basis view). More sophisticated methods are currently being investigated which weight the contribution of each basis view by its position relative to the surface normal. In particular, the finest sampling of a surface (and hence the richest representation) is not necessarily obtained from the closest view, rather it is obtained when the imaging sensor is parallel to the surface being imaged (i.e., perpendicular to the surface normal). Estimates of the surface normals can be obtained from the depth gradients and the contribution of each basis view can then be specified as a function of the surface normal.

12.2.7 Distortions

The interpolated views contain two sources of error, intensive and positional. It is helpful to separate out the two kinds of error when analyzing the quality of the generated views. Comparisons are ideally made between the actual view that would be obtained at an intermediate position and the interpolated view for corresponding surface points in the two images. Comparisons based on a local measure obtained using differences between pixels at the same locations in the two images are corrupted by the configuration of the nearby patches.

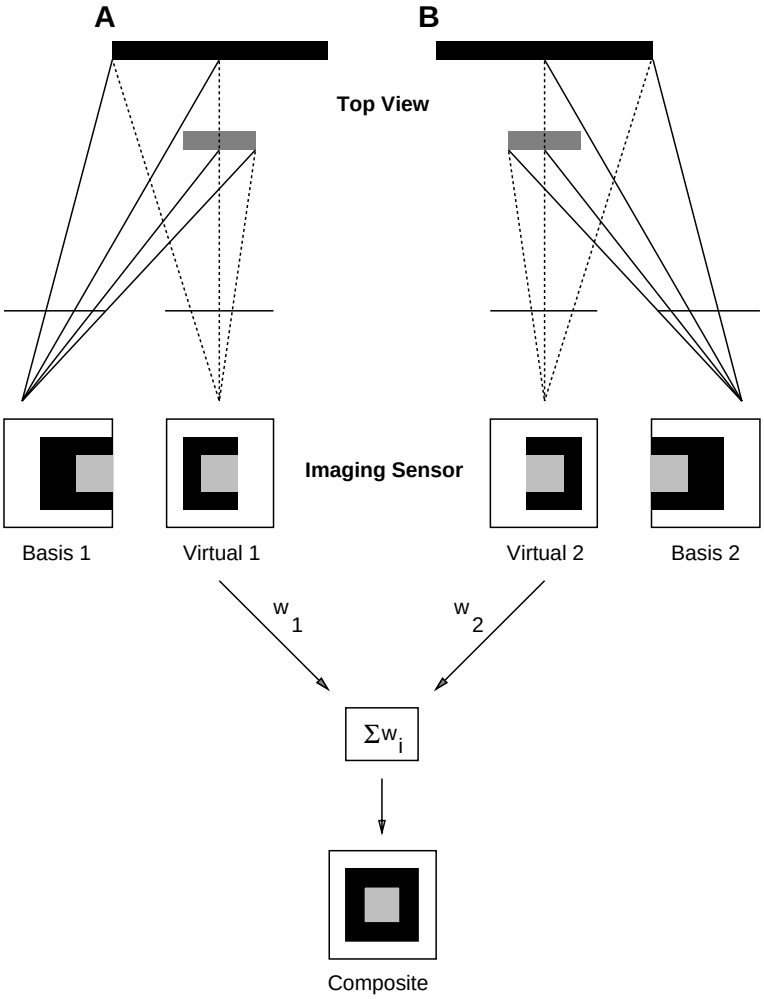

FIGURE 12.4. Integration of views. In general, due to a limited field of view or by interposition, a surface needed for a virtual view may not be completely visible from any position of any sensor in the array. By combining the multiple acquisition views in a single world coordinate system, it is possible to form composite surfaces. A ray is computed between each point projected from a given basis view to the nodal point of the virtual view. The virtual projection of the point occurs where the ray intersects the virtual sensor. Each projection is labeled with its distance from the virtual nodal point. Multiple projections to the same sensor region with different depth values indicate an occlusion (e.g., the dotted vertical line in (A) and (B)). The distal point is removed. The remaining projections are combined to form the regular sampling grid of the virtual image. Each partial virtual image is then weighted by the angular difference between the basis and virtual views. The weighted sum forms the composite image.

Photometric errors

The measure of photometric error compares the value of the reflectance of a particular surface patch when actually viewed from the intermediate position to the value obtained *for the corresponding patch of surface* wherever it is located in the interpolated view. Given the nature of surface reflectance, there are two separate components that need to be modeled, Lambertian and specular reflectance (Torrance & Sparrow, 1967). The Lambertian component is a diffuse reflectance that ideally appears equally bright in all directions. The specular reflectance is restricted to a narrow cone located symmetrically opposite the rays from the light source (with the angles referenced to the surface normal). While this distinction will be maintained here for purposes of analysis, it should be noted that few real surfaces appear equally bright in all directions and few specularities mirror the source of illumination without some diffusion.

Photometric errors can occur when there are changes in surface sampling. Changes in sampling density occur when the viewing angle of the interpolated image relative to a surface is different than that of the acquisition view. If the interpolated view is closer to the surface normal, the surface will be dimmer than the actual view. It is possible (but not included in the current system) to use the gradient of depth values to obtain an estimate of the surface normal. This estimate could be used to correct brightness for differences in surface slant. Pattern information on a relative slanted surface suffers from such subsampling. This distortion is not improved by simply scaling by differences in slant and can only be corrected by acquiring an image with better pose.

Specular error is linked to the surface normal in a different way. Specularities appear at positions on a surface where the surface normal bisects the angle formed between the point, the source and the nodal point of the view. In general, as the view changes, so will the distribution of specularities. Several methods exist to remove the specular component of an image (Klinker, Shafer & Kanade, 1987; Lee & Bajcsy, 1992; Madden, 1993). It is possible to actively acquire the location of the light source(s) and to obtain multiple views of the same surfaces from different orientations. Together with an estimate of the distribution of surface normals, this information could be used to locate, extract and reposition specularities so that the specular component of the interpolated view is correct. These techniques are also not incorporated in the current system. In practice, an uncorrected interpolation will give the incorrect impression that the light source has shifted or that there are multiple sources (a specular diplopia).

Geometric errors

The positional error measure compares the location of a surface patch when actually viewed from the intermediate position to the position of the same surface patch in the interpolated view. Very few of the incorrect positional displacements are due to sensor distortion. With the quality lenses available today, what little distortion there is can be quantified by calibration and corrected by a radial lens model. By far the largest component of the geometrical error is in the depth estimates. Impact

of these errors increases with the angle between the acquisition and interpolation views. When the desired view matches one of the basis views, there is no need at all for depth measures. The distortion caused either by absolute errors in the projection of points into the world coordinate system or by their quantization reaches a maximum when the interpolated view is 90° away from the acquisition view. The depth errors are brought about in part by the presence of steep gradients. The Tennengrad metric is computed over a region of several hundred pixels. If the region has point images that form at very different depths, the metric will be degraded. One of the measures of a good sensor configuration is how close the pose of each sensor is to the surface normals of the region of interest.

12.3 Results

By increasing the density of an array of active sensors, the errors in interpolated views can be kept arbitrarily small. The cost in hardware and computation cannot be said to do the same. We need to determine how much of the look and feel of actual camera motion between two positions can be obtained simply through appropriate merging of available information in the displayed scene. In this section we will present images corresponding to interpolated views created using both graphically generated and digitized basis views. Examples are presented that show the quality of images within the span of the basis views as well as what happens when the desired views extend beyond this. It should be noted that these static images give only a partial indication of the appropriateness of the interpolated views. Full judgment should be reserved for assessment of the images presented at frame rates in phase with movements of the participant.

The simulations are 3-D arrangements of spheres and ellipsoids rendered using varying surface reflectance models (both Lambertian and specular) and a single light source. The rendered images are created to correspond to simulated camera positions. Each rendered image is matched with a depth map. Together with the location of the nodal point and image plane of the sensor, the depth map can be used to locate each pixel in the simulation at a position in the world coordinate system. Specification of the nodal point and image plane of the view to be interpolated is all that is required to map (warp) the pixels in the world coordinate system into the desired view. Relatively simple antialiasing and defocus routines are applied to the warped image. The individual warped virtual views are blended to form the final image.

12.3.1 Simulations

We built a graphics-based test-bed of our proposed telepresence system. This system allowed us to generate arbitrary views of a graphics world through standard ray tracing techniques. Virtual views from arbitrary positions were interpolated from a small set of ray-traced basis views. The errors in the virtual view were an-

alyzed by comparing the virtual view to a ray-traced image of the scene from the same position as the virtual camera.

Our graphics world consisted of an arbitrary number of quadric surfaces. The general form of a quadric surface, $f(x, y, z)$, is given by the following equation:

$$ax^2 + by^2 + cz^2 + 2dxy + 2eyz + 2fxz + 2gx + 2hy + 2jz + k = 0. \quad (12.1)$$

This quadric equation is a 3-dimensional extension of the general conic equation. Spheres, cones, cylinders, ellipsoids, and paraboloids and hyperboloids of revolution are among the shapes that may be generated from a quadric equation. This representation is desirable since quadric objects have surface normals that can be computed analytically. Objects in the world had a Lambertian surface property with specular highlights. Lighting consisted of a point light source plus ambient light. When objects were positioned between the point light source and a surface, shadows were cast accordingly. Arbitrary views of the world could be generated by specifying the position and viewing direction of a camera, along with its focal length and aperture.

A standard ray tracing algorithm (Foley, van Dam, Feiner & Hughes, 1992) was employed to generate views of the graphics world. A ray was projected from the camera's nodal point (a pin-hole camera model was assumed) through each sensor element. Calculations were done to determine whether the ray struck any of the quadric surfaces in the world. Since all objects in the world were analytic, it was possible to compute the intersection between a line in the world coordinate system (the ray) and any quadric surface. If the ray intersected more than one surface, the closest surface was assumed to obstruct the other surfaces (i.e., no transparent surfaces were allowed). At obstruction boundaries where depth discontinuities occurred, the quantization of depth in the hidden surface calculation led to a certain amount of aliasing. For the rays that were not determined to be obstructed, the intensity (I) reflected by a point on the surface was given by:

$$I = \alpha(I_{background} + I_{point}(\beta \cos(\theta_1) + \gamma \cos^n(\theta_2))/\sqrt{d}), \quad (12.2)$$

where the surface reflectance (α), the diffuse-reflection coefficient (β), the specular-reflection coefficient (γ), and the specular-reflection exponent (n) are user-defined parameters that specify both the Lambertian reflectance (the intrinsic absorptance of the surface) and the specularity (an image of the source of illumination warped by both the macroscopic and microscopic characteristics of the surface geometry). $I_{background}$ and I_{point} are the intensity of the background ambient and point light source, respectively. θ_1 is the angle between the surface normal at the point of intersection and the point light source and θ_2 is the angle between the viewing direction and the point light source mirrored about the surface normal. d is the distance from the nodal point of the camera to the object.

Shadows from the point light source were determined using the same principles as the ray tracing algorithm. A point, P, on a quadric surface, S_1, was in the shadow of another surface, S_2, if the ray from the point light source to P intersected S_2. The intensity of each point determined to be in the shadow of another object was

given the value $\alpha I_{background}$. That is, the contribution from the point light source in Eq. 12.2 was set to zero.

Until now, a pin-hole camera model has been adopted. Although chosen for its simplicity, this camera model has the undesirable property that all objects in its field of view are in focus. Thus, an otherwise strong depth cue, blurring due to depth of field, is lost. As such, we simulated a finite aperture by blurring each pixel in the rendered view by a Gaussian with a standard deviation that increased as a function of the distance from a user-specified depth of field plane.

Since they were analytic and synthetic creations, precise knowledge of the depth values existed for each of the basis views. A variety of depth recovery algorithms may be simulated by perturbing the exact depth values with noise followed by quantization of the depth values. In the current system, both the noise and quantization level were scaled linearly with viewing distance. For example, at a distance of 1 m, a standard deviation of the noise distribution of 0.1 cm and 1.0 cm quantization resolution corresponded at 2 m to 0.2 and 2.0 cm, respectively. As more detailed data on the depth acquisition processes applied to digitized images are acquired, this model will be improved. With intensity and depth estimates in hand, arbitrary (i.e., virtual) views of the world can be generated by warping each of the basis views to the virtual camera position and then integrating.

Interpolation with full depth

To demonstrate view interpolation, we selected the simple interposition of a sphere and an ellipsoid (Fig. 12.5). A sphere 13 cm in diameter was located at the center of the world coordinate system with a horizontally elongated ellipsoid (13 by 40 cm) centered about a point 25 cm behind, 75 cm to the left and 5 cm above the origin. The surface reflectance properties of the sphere (ellipsoid) were: the specular-reflection exponent n, 10 (4); the specular-reflection coefficient γ, 0.25 (0.25); the diffuse-reflection coefficient β, 1.0 (1.0); and the surface reflectance α, 0.3 (0.3). The point light source was located 150 cm in front of the sphere, 10 cm to the right of the Z-axis in the XZ-plane. Since only the quadric surfaces were actually ray-traced, the background was approximated as a uniformly radiant dome surrounding the objects. The background ambient illumination was set to 1/10th of the level of the point light source. As a computational short cut, the ambient illumination was applied uniformly to all surface points. Shadows due to the point light source can be distinguished from the background by the attenuation of the surface reflectance coefficient (α). The sensor array was in a plane parallel to the XY-plane, 1 m in front of the center of the sphere. The cameras were located in three rows in a Cartesian grid 1 m apart. The camera sensor array was a 2.54 cm square with 0.2 mm/pixel resolution (128 by 128). The simulated array was smaller by a factor of 16 than that commonly found in CCD cameras. This reduction was maintained during system development to speed up the time required for rendering. The focal length of the lens was 64 mm and the equivalent aperture was approximately 50 mm. Object points 1 m from the nodal point were at best focus for each of the cameras and the blur circle diameter was increased linearly

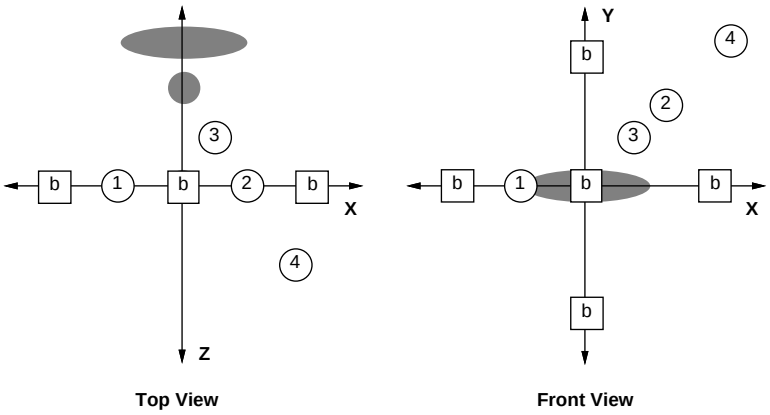

FIGURE 12.5. Configuration of simulated objects and sensors. Two quadric surfaces were positioned within the span of an array of five sensors. A sphere was centered at the world coordinate system with an ellipsoid positioned behind, above and to the left. The center camera was along the Z-axis, 1 m in front of the sphere. The other four cameras were arranged in a diamond about the center camera, each 1 m distant. Virtual views were computed for four positions: (1) midway between the center and left cameras; (2) midway between the top and right; (3) the same view as (2) but translated 0.5 m toward the sphere; and (4) the same view as (2) but translated 1.0 m away from the sphere.

with distance from best focus. Correction of tangent errors will be incorporated in future versions so that the conjugate surface is indeed a plane. For the simulations, basis views were generated for a camera in the middle row and its four nearest neighbors. The selected views demonstrated an interpolation well within the span (view 1) as well as others on the edge of the range of views covered by the basis set (views 2-4). Fig. 12.6 illustrates the five basis views.

The virtual, actual and difference maps from a virtual camera position well within the span of the basis set (midway between the center and left acquisition images) are illustrated in Fig. 12.7. The virtual view is very close to the actual view in the position, the shape and the shading of the objects. The differences that do appear are largely due to aliasing effects in the rendering and quantization in the warping. These distortions follow from simplifications incorporated in these stages that can readily be improved upon. Another source of error comes from the camera parameters selected for this simulation. Although all cameras were verged on the center of the sphere, they were all focused 1 m in front of their nodal points. While the sphere was in sharp focus in the center image, the other four images were somewhat blurred. The differential defocus is a corruption of the photometric data which when combined with a hidden surface algorithm that tried to assign a single depth value to each location in the virtual view, resulted in the inappropriate removal of some projections near depth discontinuities. A distortion that will be somewhat more difficult to correct is the smearing of the specularity in the virtual image due to the summation of the different specular components in the basis set

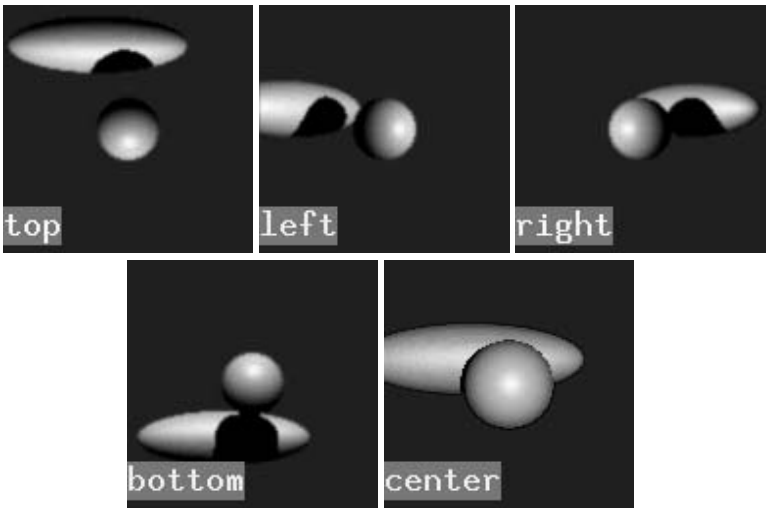

FIGURE 12.6. Simulation basis views. Views are presented that are (along the top row) above, to the left and to the right; and (along the bottom row) from below, and directly down the Z-axis toward the sphere. The center image is slightly closer (by $1/\sqrt{2}$) than the others due to the layout of the planar sensor array. The light source is slightly behind and to the right of the center camera. The effects of the light source can be seen in the movement of the specularity on the sphere as the viewpoint changes as well as the shadow of the sphere cast on the ellipsoid.

FIGURE 12.7. Virtual view within the span. A virtual view positioned midway between the center and left basis images is well within the span of the basis set. In this figure (and the series of figures that follow) both the rendering of the desired view (left) and the view interpolated from the basis set (center) are presented. On the right is the absolute value of difference between the two images scaled to the range 0 to 255. The combined errors from rendering, warping and defocus are generally small and are largest at interposition boundaries. Problems are concentrated at depth discontinuities because of aliasing in the rendering, relatively large changes in the surface sampling of steep gradients as seen from the virtual view, and the spread of blur circles across object boundaries even though the surfaces lie at different depths.

FIGURE 12.8. Virtual view on the edge of the span. When a virtual view is positioned midway between the top and right basis views, it is at one of the limits of the span supported by the five acquisition views. The weighting of the contributions from each basis view is different than those used in Fig. 12.7. Slight mismatches brought on by the simplicity of the current blending scheme can be seen in the relative brightness of the lower left portion of the ellipsoid. Diffuse specularities occur over large regions of the ellipsoid due to both low curvature and a somewhat more matte surface (than the sphere). The sharp boundaries of the different virtual view components adds to the visibility of the differences.

as they shifted with the point of view. The interpolated surface appears much more matte but is not geometrically distorted.

Another set of virtual, actual and difference maps is presented in Fig. 12.8. These images represent a view right on the edge of the viewing space spanned by the basis set. While the span of the basis views is not a simple Cartesian metric (the slant of the surfaces relative to the acquisition views as well as the spacing of multiple surfaces can both introduce gaps within and extend coverage beyond the convex hull of the sensors), the quadric surfaces insure that positioning beyond the outer boundary defined by the sensors will incur distortions due to missing data. This view in the periphery of the coverage gives a good example of the distortions that can occur because the partial virtual views are warped separately and then combined. Variations in the weighting of the different virtual components occur because the relative contribution from each basis view depends on its proximity to the virtual view. The diffuse specularity of the ellipsoid can cause large regions to change in apparent brightness with changes in viewpoint. Even though the pixel differences involved are small, their spatial configuration is such that a human observer is very sensitive to them. Currently there is no smooth blending at the boundaries of the different virtual patches. This source of error should be eliminated when all the views are projected to the virtual sensor plane before the application of the nonlinear operations that are needed to combine them into the regular spacing of the virtual sensor grid.

The span of virtual views covered by a given basis set is, in general, three-dimensional. Even though, in our example, all of the basis views lie in a plane, it is possible to construct views that are closer or further away from the objects being viewed. Both the change of scale and the perspective distortion are accommodated by projecting the acquisition data into a common world coordinate system and then projecting it appropriately onto the virtual sensor array. Adjustment of this sort must be done in any case for views within the plane of cameras since an arbi-

FIGURE 12.9. Virtual view in front of the sensor plane. As the point of view is shifted from position (2) (Fig. 12.5) 0.5 m toward the center of the world coordinate system (position (3)), the magnification, the perspective distortion and the field of view change appropriately. In consequence, as the point of view moves toward a set of objects, it is possible to leave the span of the basis set because of the visibility of previously obscured surfaces.

FIGURE 12.10. Virtual view behind the sensor plane. When the shift in view point is 1.0 m away from the sensor plane (position (4)), changes in perspective, magnification and field of view are appropriate for the increased distance. Motion away from objects can render previously obscured surfaces visible as well.

trary object will be at a nonuniform distance from the different nodal points. One factor that is not currently adjusted for is blur due to defocus. As is the case with the basis views of the sphere, overlapping views of a given surface patch may not all be at best focus or even have the same degree of defocus. It is possible (but also not done) to artificially add varying amounts of blur to points as they are projected into the virtual image, and it is even possible (though much more difficult) to selectively deblur portions of the surfaces through knowledge of their positions relative to the acquisition and virtual nodal points and the parameters of the two cameras. Nonetheless, moderate excursions both in front of (Fig. 12.9) and behind (Fig. 12.10) the sensor plane produce interpolated views with smooth contours that have appropriate shape, shading and geometry. The combination of appropriate scale and perspective changes dwarfs the effect of defocus errors under a wide range of conditions. Boundaries on the span of the set of basis views exist for these types of view changes as well. It is possible to see behind objects insofar as at least one oblique camera view allows you to see those surfaces (and even somewhat beyond with objects such as these quadrics, since it is difficult for participants to detect missing data errors on severe obliques).

The previous simulations demonstrated the quality of interpolated views possible within the span of the basis set. What kind of failures occur when the basis set does not provide adequate coverage of the environment? To demonstrate inadequate coverage, we have replaced the five basis views used in the previous simulations with a single view offset to the right (Fig. 12.11). As might be ex-

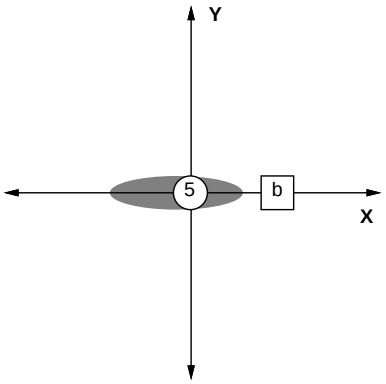

FIGURE 12.11. Inadequate sensor coverage configuration. To demonstrate the errors incurred by not having sufficient basis views to span the position of the desired view, a single acquisition image replaces the array of basis views used in previous examples (cf. Fig. 12.5). The single camera is positioned 0.5 m to the right of the Z-axis in the XZ-plane. The virtual view (5) is located where the center sensor used to be, 1 m in front of the sphere.

pected, surface data is missing from the back of the sphere (Fig. 12.12). With a

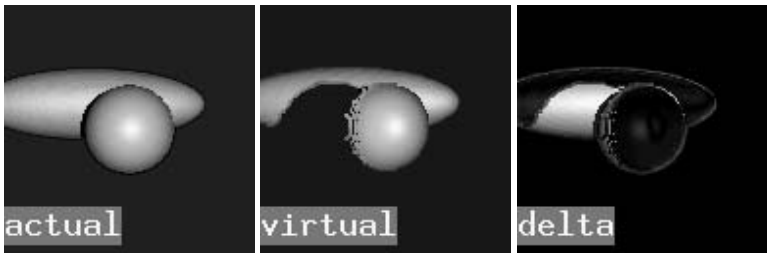

FIGURE 12.12. Consequences of inadequate sensor coverage. When there are not enough cameras, or they are not appropriately positioned, some surfaces will not be available for view. The loss of information can be due to self-occlusion (the back of the sphere) or by interposition of another object (the notch missing from the ellipsoid).

single camera, it is not possible to see more than a hemisphere and the 26° shift in viewing position about the center of the sphere is more than can be masked by the steep gradient at the edge of the sphere. In addition, the acquisition "camera shadow" of the sphere leaves a notch in the ellipsoid for which there are no data.

With no positive data to the contrary, the warping in the region of the notch extends clear to the background. The actual shadow from the point light source (at the left edge of the sphere-ellipsoid boundary of the ideal view in Fig. 12.12) is also missing. Note that the specularity is not blurred (there is only one basis view) but that it is in the wrong position. The warping acts as if the specularity is anchored to the surface.

Interpolation with corrupted depth

In the simulations presented above, the depth map associated with a given acquisition image was known exactly (within floating point precision). In practice, however, the available depth map will have limited resolution, perhaps severely so. The focusranging results of Krotkov (1987) achieved a resolution of 1% over a distance of 1 to 3 m. Xiong and Shafer (1993) were able to improve resolution by an order of magnitude by fitting a curve to closely spaced responses about the peak of the distance metric. In Fig. 12.13, depth is quantized to 1 cm (at the 1 m

FIGURE 12.13. Virtual view with 1% depth quantization. When the depth map was quantized to the approximate resolution achieved by Krotkov (1987), additional distortions brought about by the quantization were not perceptually salient. The approximately 1 cm steps in depth did not add any noticeable photometric or geometric errors. The reduction in the specular highlights resulting in a relatively matte finish was due to the movement of those contributions in the basis views and could be seen even with full depth resolution.

viewing distance to the center of the sphere). As can be seen by comparing this figure with Fig. 12.8, the interpolated view shows very little increase in degradation due to the quantization. However, when the quantization error is increased fourfold, the distortions become apparent (Fig. 12.14). Distorted by the larger depth errors, the smooth surface takes on a mottled appearance and flanges appear about the perimeter of the objects. Note that the mottled appearance of the surface is due to very small luminance differences, generally less than 10%. Depth errors of this magnitude will require a more sophisticated combination algorithm than the simple warping now employed. When the views are degraded further by noise, any new warping algorithm will likely involve a smoothing that will result in the loss of spatial detail — small contours, edge sharpness and spaces between nearby objects.

FIGURE 12.14. Virtual view with 4% depth quantization. The large 4 cm gaps between successive depth planes in the different basis views now form noticeable striations in the composite. In addition, when the depth quantization cuts through a steep gradient and the result is viewed at an angle appreciably away from that acquisition view, flanges appear extending out from the body of the object.

FIGURE 12.15. Mannequin face. This is an oblique view of the mannequin used in the digital image example.

12.3.2 Digitized Image Interpolation

To test our active telepresence method on actual images, we digitized views of the face of a mannequin (Fig. 12.15). Due to the subtle interpretations humans make of facial expressions, this is a good test example. Since we do not yet have a real time implementation, it is also a good test example in that she doesn't move. The basis views were selected to bracket the mannequin's face (Fig. 12.16). Closeups of the eyes and mouth were obtained with an overlapping region in the middle of the nose (Fig. 12.17). Depth maps were obtained for each of the basis views (Fig. 12.18) and were brought into registration (Fig. 12.19). Six virtual views were then computed (Fig. 12.20) in two rows of three views. Three of the views were at the same vertical height as the bottom basis view (the center view coincided with the basis view). The other three virtual views were displaced vertically. All virtual views were oriented to the mannequin's mouth. This configuration resulted in virtual views that were oriented between 20° and 27° with respect to the Z-axis. The virtual views display both the subtle change in the features as well as the larger effects (such as foreshortening in the lower views) that are associated with viewing a three dimensional object. The views also exhibit the not so subtle degradations

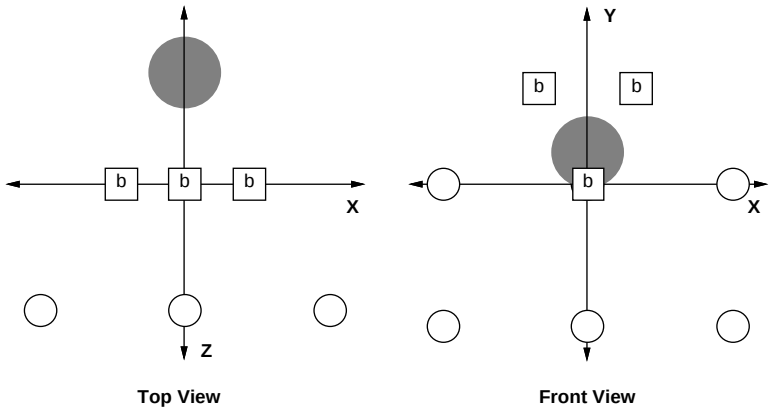

FIGURE 12.16. Layout of basis and virtual views. The origin of the world coordinate system is at the mannequin's mouth. The plane containing the basis and virtual views is parallel to the XY-plane and is translated 35 cm in the Z direction. The basis views of the eyes are 8 cm above and 3 cm to the side of the basis view of the mouth. The virtual views are in a grid, each separated by 20 cm.

FIGURE 12.17. Photometric basis images. Closeups of the two eyes and the mouth were obtained of the mannequin's face to serve as the basis for the construction of interpolated views. The images of the eyes were displaced vertically by 8 cm and laterally by 3 cm from the view of the mouth.

from the current implementation of our merging techniques. There was also no attempt made to correct the photometric errors.

12.4 Discussion

The rationale behind the system we propose is that for many virtual reality applications, specifically for those involving telepresence in real environments, computing resources are better spent on acquiring an appropriate set of remote images that will allow the warping of surface reflectance into a desired point of view than on the extraction of parameters needed to generate a graphical reconstruction local to the viewer. On the other hand, we propose that it is not necessary to design

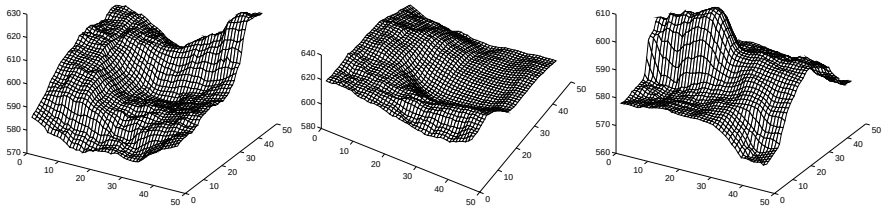

FIGURE 12.18. Basis depth images. A series of images with varying focus were obtained for each basis view and those images were used to estimate depth using the focus ranging algorithm. To display the depth variations, the orientations were not made to match those of the previous image.

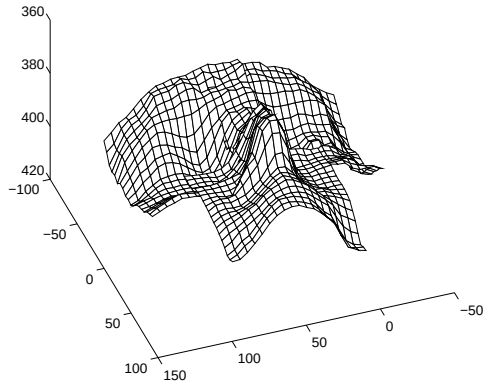

FIGURE 12.19. Registration of basis depth views. The separate depth maps were brought into registration and merged in the world coordinate system.

a positioning system that will rapidly and accurately mimic the position of the participant's eyes in the remote environment. We base our belief partially on an examination of the recent efforts to establish telepresence. Although the intrusive latencies associated with tracking of the participant and with positioning the sensors can be reduced with the use of predictive filters, the compromises that remain are severe. Some attempts reduced the quality of views to wire-frame representations, while others that increased the quality of the rendering also required the availability of much specialized hardware. In addition, the increased quality of rendering required much off-line authoring (creation of models). The extraction of the parameters needed to animate graphical models is often difficult (e.g., facial expression) and the range of a given model is limited — novel events or objects cannot be accommodated. In the system we propose, the creation of models is data-driven. Photometric data are positioned in a world coordinate system through the use of associated depth information and then reprojected into a desired point of

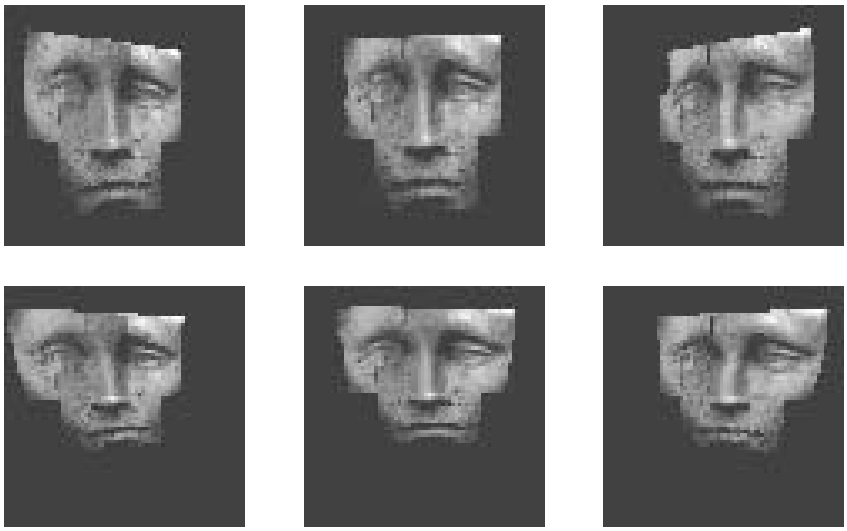

FIGURE 12.20. Virtual views. Six views of the mannequin's face are presented from different points of view. The views vary from direct frontal to 20° below and 27° to the side.

view. The computational cost of warping the information in world coordinates to the desired view is much less than rendering an image of equivalent detail. All that is required of such a system is to retain sufficient accuracy in the distribution of photometric data in the world coordinate system so that the information isn't distorted in an intrusive way. All the realistic grit of actual surfaces comes for free. For many environments and for many tasks, the proposed peripherally-positioned array of sensors can provide views that adequately approximate the data obtained from a remote pair of cameras that move in synchrony with the participant's gaze.

We also base our belief on an examination of what active vision offers in the acquisition of appropriate basis images. Apart from all of the optical limitations of focus and sensitivity, a fixed array of sensors is encumbered by the need to cover the environment so that each region is in the field of view of at least one camera. This coverage is difficult to achieve without either sacrificing the resolution needed for realism or providing a very dense array of sensors. In addition, unless the cameras can be configured in advance for a specific state of the environment, some surfaces will be viewed very obliquely, and thus sampled relatively poorly. The improvement of the acquired sampling of a surface through the adjustment of pan, tilt, focus and zoom can provide the equivalent coverage of a very large number of fixed cameras. Active vision, however, can do more than provide an efficient way to collect views of surfaces. It is not just the application of computer vision algorithms to disentangle highlights from surface properties in a desired view by the application of *shape-from* models of illumination and reflectance. Active vision is not optimizing what you have, it is getting what you need. Applied to telepresence, active vision is the movement of sensors about the remote scene to acquire

the location and composition of the sources directly. It is acquisition of several views of a given surface patch from different angles to measure specular changes. It is the acquisition of depth gradients through the use of focus ranging and stereo to approximate surface normals as one of a multiplicity of sources of information that constrains the behavior of highlights. It is the interactive refinement of these estimates by successively removing contaminants and other obstacles to the performance or the perceptual experience of the participant. It is the provision and appropriate allocation of flexible resources (such as visual fixtures) to accomplish the goals of the participant. Active vision does not simply attempt to reconstruct the environment, it provides a mechanism to interact with the environment on both photometric and functional levels.

In the current system, we concentrated on the application of focus ranging techniques to construct the models in the world coordinate system because of the availability of motorized optics and positioning systems. Using focus ranging to determine the depth of local surface patches relative to the cameras affords us the benefit of the acquisition of moderately fine-grained depth information without having to solve the correspondence problem. We explored the use of focus ranging information, not because it was intended as a replacement for other sources of information, but because it was relatively unexplored (as compared to stereo). We fully expect a working system to incorporate any and all sources of information in the refinement of the distribution of photometric data in the world coordinate system.

Finally, we base our belief on an examination of the creation of intermediate views from simulated data. Using five basis views obtained by verging widely-spaced cameras on an object of interest (the sphere), we were able to generate arbitrary views that correspond to a range of observer points of view. These views not only vary over the part of the sensor plane bounded by the cameras, but also toward and away. The quality of the interpolated views remains high through the registration and combination of the multiple views even though the pose of the cameras is not optimal for all the surfaces of interest.

What else is needed to make our belief more believable? We need to improve the graphical rendering so that ideas may be tested with the comfort that rendering errors are negligible (currently this is not the case). It is important to have a source of basis images for which ground truth is known. Graphical rendering of scenes can fill this need if the consequences of the quantization error brought about by digital representation and bounds on available computation can be kept small. Because it was important for us to have knowledge and control of the errors involved in rendering, we created our own simple test-bed in Matlab (The MathWorks). Now that we have a better feeling for the trade-offs involved, we will take advantage of one of the many packages available (both commercial and freeware). Similarly, we need to improve our warping algorithm to more seamlessly combine the contributions of the various acquisition views into a virtual view. The current warping method results in unnecessarily large errors, especially in the presence of steep gradients.

Apart from the generation of views, there is a need to assess their quality and

utility. A large body of psychophysical knowledge exists that should be applied to telepresence. With the new technology, reality is being distorted in a manner that is new to the participant. Photometric and geometric distortions that are subthreshold when viewed statically in a single image suddenly can become salient when viewed dynamically in sequence.

Traditional psychophysical measures of detection and discrimination alone are insufficient to measure the quality of virtual reality. New properties such as immersion need to be measured. Immersion is a measure of capture/involvement/reality of the presented views. It is a difficult measure to quantify. However, intrusion — the flip side of immersion — is somewhat easier to measure. An intrusive factor can be judged on how much it differs from reality or degrades performance. Intrusiveness is not limited to the quality of the photometric signal presented to the participant. Consequences of distortions other than appearance (such as stress to the oculomotor system) can limit the ultimate utility of this technology far more severely.

In summary, then, whether it is for an ophthamologist in an office in Bethesda, Maryland examining a patient in eastern Montana or a movie patron experiencing a car chase in Lethal Weapon VII, when both photometric and range information are available from a number of basis views, it is possible to construct a personalized viewpoint in real time that supports the perception of remote presence. In the former scenario, the clinician can be seen as tuning an instrument to obtain the necessary pose and resolution for an accurate diagnosis. In the latter application, a considerable amount of accuracy can be sacrificed to maintain the perceptually smooth continuity required for the maintained suspension of disbelief. In either case, the array of cameras on the remote clinic wall or in the stunt car can do more than simply provide multiple views.

- Image acquisition strategies based on active vision can be used effectively to obtain images for telepresence applications.
- A sparse array of active cameras (pan/tilt/zoom/focus) can be used to capture views of a remote environment equivalent to a much denser array of fixed sensors.
- The generation of views by warping digital images is computationally much less expensive than rendering.
- The combination of photometric and depth maps obtained in real time can eliminate the need to prepare models of the remote environment in advance and, at the same time, offer improved response to dynamic and novel events.
- Focus ranging provides a source of depth information that doesn't require solution of the correspondence problem.
- For many environments and applications, much of the appearance of remote views can be maintained with relatively coarse depth maps.

Acknowledgements

We would like to thank Norm Badler, Ruzena Bajcsy, Wendy Hunt, Visa Koivunen and Richard Paul for helpful comments and criticisms. The equipment used in this paper was supported by Navy Grant N00014-92-J-1647; Army/DAAL 03-89-C-0031PRI; NSF Grants CISECDA 88-22719, IRI 89-06770, ASC 91-08013, MSS-91-57156, CISECDA 90-2253, CDA91-21973 and GER93-55018; NATO Grant 0224/85; A. I. du Pont Institute, Barrett Technology Inc., duPont Corporation, General Motors, and The Preservation Hall Trust.

12.5 REFERENCES

Adam, J. A. (1994). Medical electronics. *IEEE Spectrum*, *31*, 70–73.

Adelson, E. H. & Bergen, J. R. (1991). The plenoptic function and the elements of early vision. In M. S. Landy & J. A. Movshon (Eds.), *Computational Models of Visual Processing* (pp. 3–20). Cambridge, Massachusetts: MIT Press.

Aloimonos, Y. (1990). Purposive and qualitative active vision. In *International Conference on Pattern Recognition* (pp. 346–360). Los Alamitos, California: IEEE Computer Society Press.

Aloimonos, Y., Weiss, I. & Bandopadhay, A. (1987). Active vision. In *DARPA Image Understanding Workshop* (pp. 552–573). San Mateo, California: Morgan Kaufmann Publishers.

Azarbayejani, A., Starner, T., Horowitz, B. & Pentland, A. P. (1993). Visually controlled graphics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*, 602–605.

Bajcsy, R. (1988). Active perception. *Proceedings of the IEEE*, *76*, 996–1005.

Ballard, D. H. (1991). Animate vision. *Artificial Intelligence*, *48*, 57–86.

Besl, P. J. & McKay, N. D. (1992). A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*, 239–256.

Caudell, T. P., Janin, A. L. & Johnson, S. K. (1993). Neural modeling of face animation for telecommuting in virtual reality. In *Proceedings of the First IEEE Virtual Reality Annual International Symposium* (pp. 478–485). Piscataway, New Jersey: IEEE Press.

Cruz-Neira, C., Sandin, D. J., DeFanti, T. A., Kenyon, R. & Hart, J. C. (1992). The CAVE – audio visual experience automatic virtual environment. *Communications of the ACM*, *35*, 64–72.

Durlach, N. (1994). *Virtual reality: scientific and technological challenges*. Washington, DC: National Academy Press.

Foley, J. D., van Dam, A., Feiner, S. K. & Hughes, J. F. (1992). *Computer Graphics: Principles and Practice* (2nd Ed.). Reading, Massachusetts: Addison-Wesley.

Fuchs, H., Bishop, G., Arthur, K., McMillan, L., Bajcsy, R., Lee, S. W., Farid, H. & Kanade, T. (1994). Virtual space teleconferencing using a sea of cameras. In *First International Symposium on Medical Robotics and Computer Assisted Surgery* (pp. 161–167). Pittsburgh, PA.

Fuchs, H. & Neumann, U. (1993). A vision of telepresence for medical consultation and other applications. In *Proceedings of the 6th International Symposium on Robotics Research*. Cambridge, Massachusetts: MIT Press.

Hirose, M., Yokoyama, K. & Sato, S. (1993). Transmission of realistic sensation: development of a virtual dome. In *Proceedings of the IEEE Virtual Reality Annual International Symposium* (pp. 125–131). Piscataway, New Jersey: IEEE Press.

Hopkins, H. H. (1955). The frequency response of a defocused optical system. *Proceedings of the Royal Society London A*, *231*, 91–103.

Klinker, G. F., Shafer, S. A. & Kanade, T. (1987). Using a color reflection model to separate highlights from object color. In *Proceedings of the International Conference on Computer Vision* (pp. 145–150). Los Alamitos, California: IEEE Computer Society Press.

Kollin, J. S. (1994). The virtual retinal display. Technical Report MS FJ-15, University of Washington, Human Interface Technology Laboratory.

Krishnan, A. & Ahuja, N. (1993). Range estimation from focus using a non-frontal imaging camera. In *Proceedings of the Eleventh National Conference on Artificial Intelligence* (pp. 830–835). Menlo Park, California: AAAI Press/The MIT Press.

Krotkov, E. P. (1987). *Exploratory visual sensing for determining spatial layout with an agile stereo camera system*. PhD thesis, The University of Pennsylvania.

Krotkov, E. P. (1989). *Active Computer Vision by Cooperative Focus and Stereo*. New York: Springer-Verlag.

Krueger, M. W. (1991). *Artificial reality II*. Reading, Massachusetts: Addison-Wesley.

Lee, S. W. & Bajcsy, R. (1992). Detection of specularity using colour and multiple views. *Image and Vision Computing*, *10*, 643–653.

Madden, B. C. (1993). Extended intensity range imaging. Technical Report MS-CS-93-96, Department of Computer and Information Science, University of Pennsylvania.

Madden, B. C. & Cahn von Seelen, U. (1995). 3-dimensional redundant tracking. Technical report, Department of Computer and Information Science, University of Pennsylvania.

Minsky, M. (1979). Toward a remotely-manned energy and production economy. Technical Report 554, AI Laboratory, Massachusetts Institute of Technology.

Ohya, J., Kitamura, Y., Takemura, H., Kishino, F. & Terashima, N. (1993). Real-time reproduction of 3-D human images in virtual space teleconferencing. In *Proceedings of the IEEE Virtual Reality Annual International Symposium* (pp. 408–414). Piscataway, New Jersey: IEEE Press.

Pahlavan, K., Uhlin, T. & Eklundh, J. (1993). Active vision as a methodology. In Y. Aloimonos (Ed.), *Active perception* (pp. 19–46). Hillsdale, New Jersey: Lawrence Erlbaum and Associates.

Pentland, A. P. (1987). A new sense for depth of field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *9*, 523–531.

Rheingold, H. (1991). *Virtual Reality*. New York: Summit Books.

Sayers, C. & Paul, R. (1993). Synthetic fixturing. In *Advances in Robotics, Mechatronics and Haptic Interfaces, ASME Winter Annual Meeting, New Orleans, USA* (pp. 37–46).

Sayers, C. & Paul, R. (1995). An operator interface for teleprogramming employing synthetic fixtures. *Presence*. In press.

Sutherland, I. (1965). The ultimate display. In *Proceedings of the IFIP Congress* (pp. 506–508). Amsterdam: North-Holland.

Terzopoulos, D. & Waters, K. (1993). Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *15*, 569–579.

Torrance, K. E. & Sparrow, E. M. (1967). Theory for off-specular reflection from roughened surfaces. *Journal of the Optical Society*, *57*, 1105–1114.

Tsai, R. Y. (1987). A versatile camera calibration technique for high-accuracy 3-D machine vision metrology using off-the-shelf TV cameras and lenses. *IEEE Journal of Robotics and Automation*, *RA-3*, 323–344.

Willson, R. G. (1994). *Modeling and calibration of automated zoom lenses*. PhD thesis, Carnegie Mellon University.

Wolberg, G. (1990). *Digital image warping*. Los Alamitos, California: IEEE Computer Society Press.

Xiong, Y. & Shafer, S. A. (1993). Depth from focusing and defocusing. In *DARPA Image Understanding Workshop* (pp. 967–976). Palo Alto, California: Morgan Kaufmann Publishers.