Robotic Sightseeing – A Method for Automatically Creating Virtual Environments

Eric Bourque, Gregory Dudek, Philippe Ciaravola
Centre for Intelligent Machines
McGill University
3480 University Street, Montreal, Canada H3A 2A7

Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 1998, Volume 3, pages 3186-3191.

Abstract—
This paper describes the fully automatic creation of an environment's description using an image-based representation. This representation is a collection of cylindrical sample images combined into an "image-based virtual reality". The locations at which the environment will be sampled are chosen automatically using an operator inspired by models of human visual attention and saccadic motion.

The image acquisition is performed by a mobile robot. The selection of vantage points is based on an analysis of the edge structure of sampled panoramic images. In order to trade off the optimality of the generated description of the navigation effort required in solving the on-line problem, a concept referred to as alpha-backtracking is introduced. The paper illustrates sample data acquired by the procedure.

Keywords— **mobile robotics, virtual reality, robotics**

I. Introduction

IN this paper, we present an approach to the automatic creation of environmental representations for human interaction. We use a mobile robot to collect image data that is used to construct a pseudo-realistic user experience. The technique is based on a model of visual attention.

A. Motivation

Graphical models of an environment can be used for applications such as architectural studies, environmental inspection, and telerobotic control. Such models that provide a realistic visual experience are frequently referred to as virtual reality (VR) models. The standard approach to VR modelling consists of using an *a priori* manually-constructed 3D model of the environment for real-time graphic rendering from a desired viewpoint. One factor limiting the utility of this type of VR modelling is that the construction of a realistic synthetic environmental model can be extremely labour intensive; the modelling and texturing of a single object can take months[1]. In addition, the computational burden involved in rendering scenes for model-based VR can be substantial, especially when we consider that when motion is involved, many successive frames of the environment must be rendered from slightly different viewpoints. Finally, obtaining a truly realistic result for an arbitrary environment remains exceedingly challenging.

Image-based virtual reality refers to the use of real image data (photographs) of an existing environment or model to create a VR environment. The image-based VR interface we use [3] allows a user to look in arbitrary directions from a given viewpoint, or to jump between pre-computed viewing locations. Although observer motion is constrained, image-based VR permits extremely realistic scenes to be displayed and manipulated in real time using commonplace computing hardware.

Image-based VR addresses the shortcomings of limited realism and/or high computational load imposed by conventional model-based VR. Unfortunately, it only partially alleviates the intensive effort needed to create a VR world model; the acquisition of the requisite images to construct an image-based VR model still entails effort and expertise. Furthermore, selecting suitable vantage points to produce an evocative and complete VR model is in itself an important issue. This paper deals with the automated acquisition and construction of image-based VR models by having a robotic system select and acquire images from different vantage points. The objective is to provide a fully or partially automated system for both the *selection* and *acquisition* of the image data required.

B. Applications

Image-based VR modelling appears promising in several contexts in addition to entertainment. In particular, these include task domains where the scene to be examined is either too remote, too dangerous, or inconvenient for a human operator to visit directly. As such, the potential application contexts overlap those for teleoperated robotics. This range of applications includes museum previews by computer (for example on the world wide web), security or surveillance applications where an environment must be inspected periodically, or the exploration of remote locations such as undersea or on another planet.

C. Overview

The key issues in producing an image-based VR representation are: (1) the selection of suitable vantage points to cover the interesting aspects of the environment, (2) the

This work is supported by a grant to G. Dudek from the National Sciences and Engineering Research Council of Canada.

[1]The construction of a single complete textured model of an airplane in the film "Con Air" took two months [1], and the rendering alone in the Disney Production "Toy Story" took 800 000 machine hours [2].

acquisition of suitably calibrated images. This is followed by post-processing of the image data to provide the VR model. When an image-based VR model consists of a collection of viewpoints between which the user can move, it is referred to as a *multi-node model*. The selected viewing locations form the nodes of a graph that determines possible (discontinuous) motions that a user may experience when using the model. In this paper, we describe an approach to the fully automated creation of image based VR models of a finite environment with essentially no human intervention.

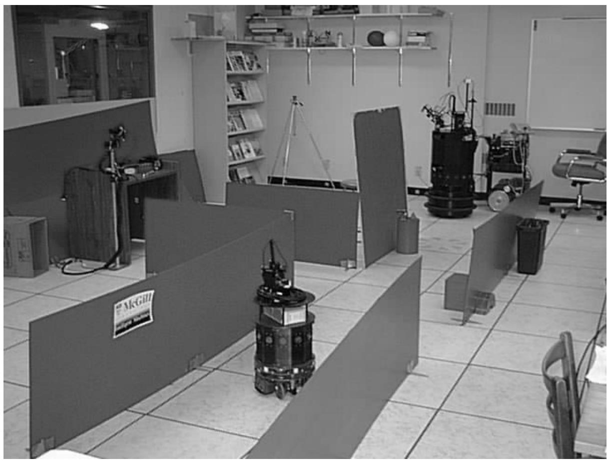


Fig. 1. Laboratory setup for the VR experiment. The RWI mobile robot is shown in the bottom portion of the picture, while the Nomad (not running in this particular experiment) is shown above.

Our approach is based on using a mobile robot (shown in Fig. 1) to navigate within an environment and collect the image data of interest. These images are in fact 360° cylindrical environment maps [4]. Our technique for selecting the appropriate sample locations uses a computational mechanism inspired by models of human attention. We assume that an autonomous exploration algorithm is available. While several such algorithms have been developed and implemented in our lab and elsewhere [5], [6], [7], [8], [9], their details are outside the scope of this paper. The current work simply presupposes that the robot travels along a particular trajectory through the environment, and that it can estimate its current position at any time.

Since our objective is to construct a virtual environment that appears subjectively realistic to human observers, our approach is inspired by models of human visual environment exploration. It has been demonstrated that human exploration of either an environment or an image is driven by a shifting attentional "spotlight" [10]. In building models of human attention, substantial research has been devoted to the computational mechanism involved [11], [12]. We concentrate here, rather, on the *locations* to which attention is driven. One class of attentional processing is characterised by visual saccades to areas of high curvature, or sharp angles [13]. More generally, things which are "different", or inconsistent with their surroundings, tend to attract visual attention. Thus, our approach is to compute a map over an image (perhaps a 3-dimensional image) of how much each point attracts attention. The extrema of this map provide a set of attentional features.

Our computational procedure for defining features is dependent on finding edges in the image [14], [15]. Edge structure has been used extensively in computational vision for several reasons including the apparent psychophysical relevance of edge structures, and because there is evidence that visual edges tend to be highly correlated with the projection onto the image of real physical events in the world (eg. object boundaries, markings) [16]. Higher level features such as large-scale edges, curves, circles, or corners, are difficult to detect robustly, and organising edge elements (*edgels*) produced by an edge operator into semantic tokens is a notoriously difficult problem. While the edge elements themselves are strongly suggestive of actual geometric structures in the environment, their distribution has the advantage to being robust to variations in illumination.

It is with this in mind that we have formulated a metric for visual attention based on edgel density. For example, areas with rich geometric content will have a high edgel density. To focus attention at locales that are notable, our attention mechanism is driven to locations where the local edgel density differs substantially from the mean edgel density in the environment.

II. Approach

The set of all possible views or images obtainable from a fixed location in the environment can be described as a *viewing sphere* or spherical image. More specifically, for every ray projected from a location in R^3 – in a direction along the unit sphere S^2 – we get an associated intensity from the environment. This transformation can be expressed as:

$$M_{3D} : R^3 \oplus S^2 \longrightarrow R \tag{1}$$

or

$$M_{3D}(x, y, z, \phi, \theta) = i \tag{2}$$

where (x, y, z) are spatial coordinates, (ϕ, θ) refer to the orientation of a light ray, and i is the intensity observed. This parameterisation of light rays is related to the *light ray manifold* defined by Langer and Zucker [17] and the Lumigraph [18].

In our particular case, we have a camera mounted on a pan and tilt unit at a fixed location on a mobile robot. For the purposes of this paper, let us assume that the robot is constrained to a flat floor, and thus we restrict the camera to a plane. This constrains the origin of the ray to R^2, and we have the idealised 2-D observer in a 3-D world:

$$M_{2D} : R^2 \oplus S^2 \longrightarrow i \tag{3}$$

or

$$M_{2D}(x, y, \phi, \theta) = i. \tag{4}$$

A minor variation is the case of an idealised camera which only pans, which is common in most image-based VR. Since we are now dealing with a camera, as opposed to a single ray, the result of the transformation is an image or a set of intensity measurements given by a cone about the camera direction:

$$M_C : R^2 \oplus S \longrightarrow R^n \tag{5}$$

or

$$M_C(x, y, \phi) = \mathbf{I} \tag{6}$$

where n is the number of pixels in $\mathbf{I}$, a pixel-indexed image $I(a, b)$ implicitly dependent on the field of view of the camera. Each pixel is, of course, also specified by Eq. 4. An entire *spherical* panoramic image $I(x, y)$ where each pixel is a ray corresponding to Eq. 4 is given by

$$M_S : R^2 \longrightarrow R^n \tag{7}$$

where n is the number of pixels in the image, thus leading to a parameterisation of a *set* of images $\mathbf{I}_{x,y}(a, b)$ whose individual pixels implicitly depend on the viewing parameters of the camera.

It will prove convenient to consider the problem of directing attention to a sub-image of a large image. In this case, we specify the sub-image pixels as $I_{x,y}(\phi, \theta)$ so that it notationally resembles the specification of a pixel from a set of panoramas in Eq. 7:

$$I_{x,y}(\phi, \theta) = I(x + \phi, y + \theta) \tag{8}$$

An image-based VR model is founded on approximating a continuous set of spherical images given by Eq. 7 from a (discrete) set of representative points in the environment. In practice, image based VR allows a user to move between specific locations and look in (almost) any viewing direction from any of these locations.

To construct a navigable environment, several such nodes must be created, and a method must be defined for internodal movement. In practice, one can define *hot-spots* within the images to create such links in the nodal graph. The desired result is to obtain a graph composed of such nodes which encompass all the distinctive regions in the environment, as well as a means of navigating smoothly between them. That is, if two nodes are chosen which have no overlapping visual information, it would be desirable to have a node in between which would allow a smooth transition. It is the automated selection of the nodal positions $\mathbf{P}_i$ which we will now develop further.

III. Methodology

To encode an environment using image-based VR nodes, we must first determine which locations in the environment, that is, which viewing cylinders from the set of all those possible, are most worthy of retention. We accomplish this by establishing which viewing spheres are most *distinctive*, where interest is measured by the extent to which a location attracts visual attention.

To construct our attention operator, we will consider the case of simple two dimensional images. In principle, we would like to exploit image geometry and semantics. Work in human psychophysics suggests that various types of geometric structures – such as line endings, oriented line segments, or curves – "pop out" of an image when they are different from the rest of the scene [11]. Since edge linking and segmentation remain open problems in a generic context, we settle, instead, on exploiting the variations in the *distribution* of edge *elements* as cues to attentional fixation. Because psychophysics and intuition suggest that we wish to concentrate on regions that are unusual or distinctive, we can evaluate the extent to which regions of an image differ from the mean. We begin with an image $I_{x,y}(a, b)$ and compute the binary edge map $E(I_{x,y}(a, b))$. A generic and computable metric for image content is local edge-element density. We compute this by convolving the image with a windowing operator to obtain the local edgel density $D(i, j)$. While in principle a Gaussian windowing function is suitable, in the interest of real-time performance we use a square-wave kernel of size AB:

$$D(i, j) = \frac{1}{AB} \int_{j-\frac{B}{2}}^{j+\frac{B}{2}} \int_{i-\frac{A}{2}}^{i+\frac{A}{2}} E(I_{x,y}(a, b))\ da\ db \tag{9}$$

Note that this can be computed in the context of images from either Eq. 7 or 8.

The interest value of a point is then given by the absolute deviation from the mean local edgel density $\hat{D}$:

$$\mathcal{D}(i, j) = |\hat{D} - D(i, j)| \tag{10}$$

We then sort these points $\mathbf{P_i}$ based on their absolute deviation $\mathcal{D}$ from the mean to provide a list of the K most-interesting locations for which nodes are created. In practice, additional constraints – such as assuring no two points are too close together – are desirable. For the purposes of the present synopsis, we will simply assure that no two regions on the list of interesting places are permitted to overlap. If they do, we evaluate pairwise combinations and delete the less interesting of the two.

IV. Implementation

Our approach to attention described above assumes that the statistics of the edge distribution of the environment are fully available when decisions are made. Such a paradigm is sometimes referred to as an *off-line algorithm*. In this context, it involves an analysis of image data from every point in the environment, followed by a selection of the best few locations for which panoramic image nodes are subsequently created and interconnected. Examples of the performance of this approach are presented in Section V.

A. On-line Viewpoint Selection with α-backtracking

In practice, as the robot moves through the environment, it would be highly advantageous to make decisions when locations are encountered so that there is no need to either acquire and store immense amounts of data, or to backtrack to selected locations to obtain the panoramic images. To do this, nodes must be selected based only on partial information of the statistical distribution of image content over the environment, giving rise to an *on-line algorithm*. Assuming that the off-line algorithm performs well, we seek an on-line algorithm whose performance is a good approximation of that obtained with the off-line method.

We can assure that the on-line algorithm exhibits arbitrarily good performance, as compared to the ideal of the off-line algorithm, by permitting the robot to backtrack. We can define the *forward interest* of a point from partial information as

$$\mathcal{D}_t(i,j) = |\hat{D}_t - D(i,j)| \tag{11}$$

where the subscript t denotes statistics computed from the initial fraction $t \in (0,1]$ of the entire data set. We define *on-line viewpoint selection with α-backtracking* as a variant of the off-line algorithm such that the best K non-overlapping points are selected as the exploration proceeds. As each point is selected, a corresponding panoramic node is constructed. Density values are also stored for all other points visited. As the exploration proceeds, t increases and the forward interest values of previously visited locations may evolve. If a prior unselected point – *which is no further back than a fraction α of the current trajectory length* – becomes more interesting than one of the K selected points, the robot backtracks and uses it instead of the point it replaces. Clearly, the performance (in terms of the points selected) of this algorithm approaches the ideal as α approaches one.

B. Environmental VR

Our approach to environmental VR is based on having a mobile robot traverse the 2-dimensional environment to be mapped[2]. It is independent of the traversal strategy, although it assumes that the topology of the trajectory is known so that the multi-node model can be constructed. In addition, in order to avoid closely-spaced sample nodes, an approximate *local* estimate of distance is desirable (eg. from odometry). Our technique has been implemented to function with robots from both RWI and Nomadic Technologies, both of which provide odometry data whose accuracy is far in excess of our requirements.

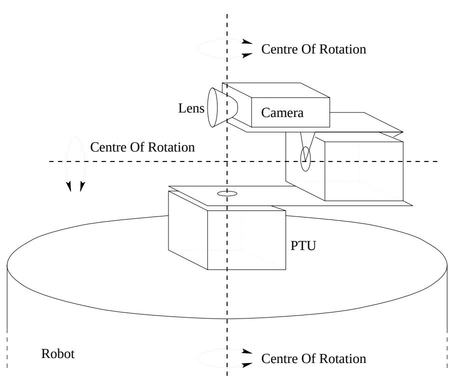


Fig. 2. Configuration of robot and camera used to acquire the images.

The robot we used has a digital camera mounted on a pan and tilt unit, and is capable of translating and rotating, thus giving a the final camera position a total of 4 degrees of freedom (DOF) as shown in Fig. 2. In practice, we have fixed the tilt of the camera, and therefore define an (x, y, θ) location for each image taken in the environment. During the exploration, the robot periodically stops and gathers a set of images in one or more orientations, so that they may be evaluated for candidacy within the set of salient points. The most complete sampling of the environment demands a set of images which collectively cover the view in every direction; that is the data presented in this paper.

The difference between the off-line and on-line implementations is apparent at this point. In the case of the on-line algorithm, the robot must immediately decide whether or not to create a VR node and obtain additional data if necessary. If the current location is not selected, the image data acquired can be discarded. In the off-line case, on the other hand, the robot must retain all of the data. The final decision regarding node selection is made after the exploration is complete. In the off-line case, the robot must acquire all the additional data for a panorama at each location, or return to the selected locations after selection is complete. These nodes are then incorporated together to form a complete image-based VR representation of the previously unknown environment.

To construct an image-based model, we must first gather a set of images from each point $\mathbf{P}_i = (x_i, y_i)$ in the environment we wish to model. These images are then tiled into a mosaic which can be subsequently mapped onto a viewing volume [19], [20]. In practice, the mosaic is produced by "stitching" or fusing all of the images from one sample location into one composite image [19]. To achieve this, the camera rotates about the lens' optical centre, resulting in consecutive images that are related to one another by a pure horizontal translation [19]. In order to find the translation matrix M which relates one image to the other, it is necessary to identify the relative corresponding regions that the two images have in common [19]. The strategy used finds the intensity difference for all possible overlapping configurations between the two images. This is computed using *cross-correlation*. The transformation which results in the highest correlation is then chosen as the best "match" between the images. In this implementation, we assume that only 2-D translation is needed (not rotation or scaling). The strategy used to find M minimises $\hat{I}$, defined as:

$$\hat{I} = \frac{\sum_{i=1}^{N}[I(x_i', y_i') - I(x_i, y_i)]}{N} = \frac{\sum_{i=1}^{N} \hat{\imath}_i}{N} \tag{12}$$

where N is the size of the overlapping region, and $I(x, y)$ is defined as the intensity of the pixel at position (x, y). As the images are "stitched", the two contributing images are blended together to reduce visible artifacts on the resulting image, according to the weighting function $w_t \in [0, 1]$:

$$I(x'', y'') = I(x_i', y_i')w_t + I(x_i, y_i)(1 - w_t) \tag{13}$$

For any viewing vector $\mathbf{v} = (r, \phi, \theta)$ where r represents the zoom factor, and ϕ, θ the Euler angles, we can then map the appropriate field of view onto a planar surface for

[2]While there is no reason that the methods we have discussed could not be extended to R^3, it is simply outside of the scope of this paper.

display [20]. The sampling location $\mathbf{P}_i$, defined as a *node* in the environment, now encompasses all possible viewing directions, within the constraints of the cylindrical map.

V. Experiments

In order to create a cylindrical image, either a special panoramic camera must be used, or multiple consecutive images taken by a traditional camera must be registered. Our approach uses a traditional video camera to acquire the images. The reason for this choice is that special panoramic cameras are costly, and that such cameras which use a wide angle lens tend to distort the image, whereas those which use a parabolic mirror often have lower pixel resolution per steradian than normal cameras [21]. In contrast, normal cameras are inexpensive and they generally give better results than panoramic cameras.

We have examined the performance of our environmental sampling technique in the context of a small, fully-controlled test environment that can be manipulated at will. In the configuration shown in Fig. 1, the environment resembled a simulated office space, or a section of a maze. We used a small mobile robot with a top-mounted camera to navigate this environment[3]. The field of view of the camera was such that the tops of the images were just slightly below the tops of the walls of the test maze. This was to ensure that no spurious image data was brought into the experiment. Two pictures were mounted on the inside surface of the walls at different locations, and a few small objects were dispersed throughout the environment. At each junction there was a view into the "open world" which was considerably different from the somewhat constrained internal environment.

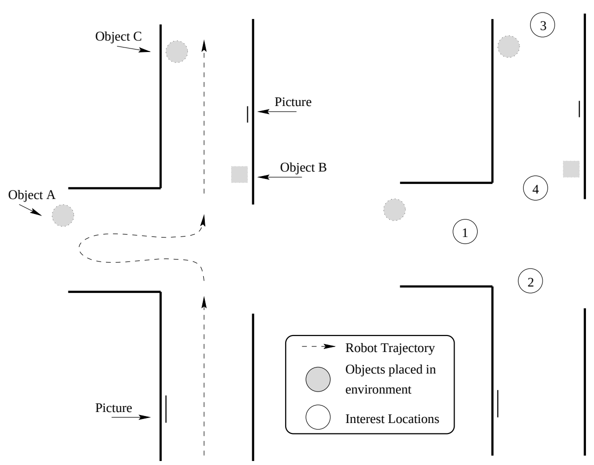


Fig. 3. Left: Trajectory taken by the RWI through the unknown environment. Right: Node locations chosen by the multiple image, off-line implementation. The numbers inside the circles denote order of distinctiveness.

The robot followed the path shown in Fig. 3 and collected 35 images from 12 vantage points for a total of 420 images gathered within the environment. In this experiment, the robot used a simple path-planner to guide it through the entire short trajectory. The actual image data was collected with a control system that automatically directed the camera at the candidate interest points [22]. The robot was equipped with sonar sensors that were used for navigation and collision avoidance. Raw data from these sensors was used to illustrate the layout of the environment in approximate form.

The image data was then analysed and the images were sorted by their descending absolute deviation as described in section III. Fig. 4 shows various images and their rank in the resultant sorted collection of images.

As was expected, the selected regions were those which encompassed the edge information at the extrema of the edgel density distribution. All junctions in the environment revealed information which was quite different from that contained within, and were therefore suitable candidates for selection. Other potential candidates from the environment were the objects and pictures placed on the inside surface of the walls. The selection algorithm performed very well in this regard – the top selection was one which contained two objects: a gold samovar[4], and a picture on the wall. Given the intended applications of this research, these results are very promising.

Also of interest are the image samples which ranked in other areas of the order. Since the order is a direct function of image distinctiveness, we would hope that the samples become decreasingly interesting as their sorted list is traversed. In our experiments, the samples from the middle and bottom of the order revealed a diminishing amount of disparity.

Our primary goal is to provide a complete nodal graph through the distinctive parts of the environment. The right-most portion of Fig. 3 exemplifies this accomplishment.

VI. Summary and Discussion

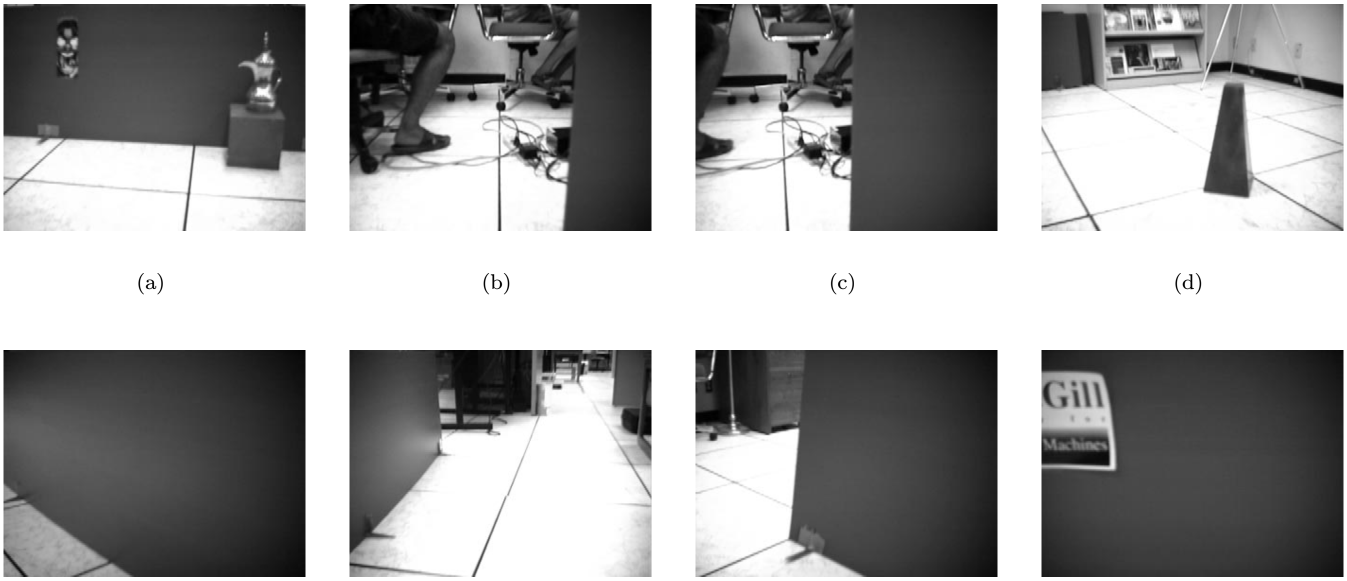
In this paper, we have outlined an approach to the selection of representative views to convey the appearance of an unknown environment. The method assumes that a set of spherical or cylindrical images is the medium used to store and convey the appearance; this is presented using image-based virtual reality.

Our secondary objectives are to limit the number of spherical images acquired and used, and to minimise the trajectory length needed to acquire them. In the experiments presented here, image acquisition is accomplished using a mobile robot. By using an attention mechanism, the robot acquires images only at "interesting" locations.

The image samples we obtain seem to effectively capture many important aspects of the scene being observed. For some applications, it may be desirable to explicitly specify certain views of interest *a priori*, for example for an art gallery it may be important to have frontal-parallel views of the pictures. Such views could be specified either manually (using map information) or procedurally (using a task-specific set of criteria). Exactly how to facilitate the spec-

[3] An alternative, larger, robot is used for conventional environments.

[4] A Russian urn used to boil water for tea.



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 4. Selected viewpoints from environmental VR experiment: (a)-(d) were chosen as most interesting, (e)-(h) were evaluated to be substantially less interesting.

ification of domain specific viewing constraints remains a subject for future investigation.

Even when the desired views are chosen manually, the generic attention mechanism is useful for selecting additional views of the environment to provide continuity and completeness in the VR viewing experience. In this case, the objective is to complement the manually specified views. In addition to supplementing these views, this work provides a mechanism for regularly and effortlessly updating the VR model; this update ability is an advantage even if *all* the views are selected manually.

We are currently developing alternative attentional functions to specify where the views should be selected. In addition, we are forming a quantitative framework to measure the quality of the VR model generated, using both psychophysical criteria and image reconstruction paradigms.

References

[1] J. Rygiel, "Digital effects in motion pictures," *Invited talk, Computer Vision and Pattern Recognition*, June 1997.

[2] A. Thompson, "Toy wonder," *Entertainment Weekly*, vol. 304, December 1995.

[3] A. Computer, *QuickTime VR 2.0 Authoring Tools Suite*. Apple Computer, 1997.

[4] N. Greene, "Environment mapping and other applications of world projections," *IEEE Computer Graphics and Applications*, vol. 6, pp. 21–29, November 1986.

[5] N. Roy, "Multi-agent exploration and rendezvous," Master's thesis, McGill University, 1997.

[6] G. Dudek, M. Jenkin, E. Milios, and D. Wilkes, "Robotic exploration as graph construction," *IEEE Transactions on Robotics and Automation*, vol. 7, no. 6, pp. 859–865, 1991.

[7] G. Dudek, P. Freedman, and I. M. Rekleitis, "Just-in-time sensing: efficiently combining sonar and laser range data for exploring unknown worlds," in *Proceedings of the IEEE International Conference on Robotics and Automation*, vol. 1, (Minneapolis, MN), pp. 667–671, April 1996.

[8] B. Kuipers and T. Levitt, "Navigation and mapping in large-scale space," *AI Magazine*, pp. 25–43, summer 1988.

[9] J. Borenstein, H. R. Everett, and L. Feng, "Where am I? Sensors and methods for mobile robot positioning," tech. rep., University of Michigan, April 1996.

[10] R. M. Shiffrin and W. Schneider, "Controlled and automatic human information processing: II. General theory," *Psychological Review*, vol. 84, pp. 127–190, March 1977.

[11] A. M. Triesman, "Perceptual grouping and attention in visual search for features and objects," *Journal of Experimental Psychology: Human Perception and Performance*, vol. 8, no. 2, pp. 194–214, 1982.

[12] J. K. Tsotsos, "Analysing vision at the complexity level," *Behavioral and Brain Sciences*, vol. 13, no. 3, pp. 423–496, 1990.

[13] D. Noton and L. Stark, "Eye movements and visual perception," *Scientific American*, vol. 224, pp. 35–43, June 1971.

[14] J. F. Canny, "A computational approach to edge detection," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, pp. 679–698, November 1986.

[15] R. Deriche, "Using canny's criteria to derive a recursively implemented optimal edge detector," *International Journal of Computer Vision*, vol. 1, May 1987.

[16] D. Marr, *Vision*. San Francisco: W.H. Freeman, 1981.

[17] M. Langer, "Diffuse shading, visibility fields, and the geometry of ambient light," in *Proceedings of the Fourth International Conference on Computer Vision*, pp. 138–147, May 1993.

[18] S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen, "The lumigraph," in *Proceedings of the ACM SIGGRAPH*, pp. 43–54, August 1996.

[19] R. Szeliski, "Video mosaics for virtual environments," *IEEE Computer Graphics and Applications*, vol. 13, pp. 22–30, March-April 1996.

[20] S. E. Chen, "QuickTime VR – An image based approach to virtual environment navigation," in *Proceedings of the ACM SIGGRAPH*, (New York), pp. 29–38, ACM, 1995.

[21] H. E. Malde, "Panoramic photographs," *American Scientist*, vol. 71, pp. 132–140, March-April 1983.

[22] P. Ciaravola, "An automated robotic system for synthesis of image-based virtual reality," Tech. Rep. CIM-TR-97-12, Centre for Intelligent Machines, McGill University, 1997.