# Viewpoint Selection – An Autonomous Robotic System for Virtual Environment Creation*

Eric Bourque  Gregory Dudek

Centre for Intelligent Machines
McGill University
3480 University Street, Montréal, Canada H3A 2A7
{ericb,dudek}@cim.mcgill.ca

Proceedings of the IEEE International Conference on Intelligent Robotic Systems (IROS), 1998, Volume 1, pages 526-532.

**Abstract**

*This paper describes an integrated system for the automatic construction of image-based virtual realities to describe a real environment. A mobile robot autonomously navigates through the environment and uses a camera to make observations. At locations that are deemed sufficiently interesting, panoramic images are collected that are used to construct a multi-node VR movie.*

*Images of the environment are classified in terms of two features related to human attention: edge element density and edge orientation. The system deems locations interesting if they are sufficiently different from the surrounding environment. The parameterization of the surrounding environment is computed either in a pre-computation pass, or on-line using a technique termed alpha-backtracking. The panoramic images that describe the environment are automatically joined together in a navigable movie that simulates motion in the real environment.*

## 1 Introduction

It is often valuable to construct an archival record of a large-scale environment in pictorial form. In this paper, we describe an automated system that moves about and collects images that allow it to reconstruct a photo-realistic virtual reality for later examination.

Most people have collected vacation snapshots and then presented them to a friend later to illustrate a remote location. More prosaically, there are environments where regular cursory visual inspection by a person is important, but sending a person on-site is not desirable on a routine basis. One example might be the cooling pipes of a nuclear reactor; in some Canadian reactors regular inspections must be conducted by an operator at some cost in terms of both radiation exposure and down-time. Another example is the use of photographs to illustrate historical or aesthetic characteristics of a building.

We are developing a mobile robot system that can autonomously explore an unknown environment, automatically select viewpoints of interest, and construct an image-based virtual reality that records the appearance of the environment. This VR model is composed of a collection of panoramic images through which the user can navigate. In prior work [1, 2], we have outlined our approach to the selection of viewpoints of interest. In this paper we describe the overall system architecture which automatically selects and uses such viewpoints; we will also comment on an elaboration of our viewpoint selection algorithm.

## 2 Background

The present work involves a synthesis of techniques for autonomous robot exploration, navigation, image registration and virtual reality. In particular, we have developed a formal description of interesting views that we use to drive image acquisition. In this paper, our scope will be limited to considering the overall system architecture and the image selection process.

Work on human visual attention suggests that a key attribute of the loci of attention is that they are

*This work was supported by the Canadian National Centres for Excellence IRIS program and a NSERC research grant to G. Dudek

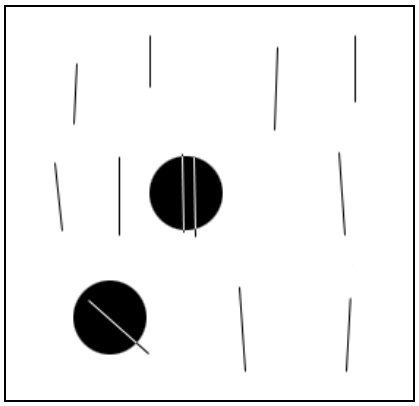

Figure 1: Example of the density and orientation operators on a simple 2D image. The black circles represent the locations chosen by the operators. The density operator chose the location containing the 2 parallel lines whereas the orientation operator chose the location containing the slanted line. This stimulus and the results resemble those used and observed in tests of human attention.

different from their surrounding context [3, 4, 5]. For short-term attention, several featural dimensions have been identified that lead to pre-attentive "pop-out" and, presumably, serve to drive attentional processing [6]. Likely human attention maps include those for color, edge orientation and edge density, among others. In this paper we will concentrate on edge density and orientation, extending our prior results using edge density alone. Figure 1 illustrates the results of using our operator, whose behavior resembles that predicted by the psychophysical literature.

Panoramic images have been used for documentary purposes from even before photographs were developed[1]. By moving about and selecting a *set* of panoramic images, we can capture much of the appearance of an environment. In principle, a suitable selection of panoramic images can serve to capture the *light ray manifold* in a scene [7], and perhaps even permit its reconstruction. The Lumigraph, the Light Field, and the plenoptic array are related constructions that couple the reconstruction of the view in a scene to the capture and sampling of it's light rays. [8, 9]. Scene visualization based on such methods is referred to as *image based rendering*. Our approach to visualization is based on collecting sample panoramic images at locations selected by the attentional operator.

[1] One instance of panoramic imagery that predates photography is the art of Hendrik Willem Mesdag and his associates. An example is a cylindrical room adorned by a panoramic painting painted c. 1880, on exhibit at Museum Panorama Mesdag in The Hague.

## 3 Approach

We can model the sampling of the intensity of light rays projected to a fixed point in an environment as follows:

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

or

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

In other words, from a fixed location $(x, y, z)$ in the environment looking in a direction $(\phi, \theta)$ along the unit sphere, we can sample the intensity of exactly one light ray. In practice, we have a mobile robot whose trajectory lies in the plane, and whose intensity sensor (CCD camera) is able to pan and tilt. This idealized (in terms of the robot) transformation can be expressed as:

$$M_{Camera} : R^2 \oplus S^2 \longrightarrow R^n \tag{3}$$

or

$$M_{Camera}(x, y, \phi, \theta) = \mathbf{I} \tag{4}$$

Where $n$ denotes the number of pixels in the image $\mathbf{I}$ taken by the camera. This leads to a parameterization of a set of images $\mathbf{I}_{x,y}$ whose intensities (pixel values) are denoted by $\mathbf{I}_{x,y}(\phi, \theta)$. It is this parameterization of intensities which we will use in our computational model of visual attention.

## 4 Computing Attention

Since both 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 (or set of images) differ from the mean as a metric for attention. In order to quantify the distinctiveness, we concentrate on the information available from the edge structure of the region under consideration, namely the location of edges, and their orientation. The edge detection process returns two maps – an edge map composed of edge elements and their associated intensities, as well as an orientation map containing the orientations corresponding to the edge elements. The extraction of information from these two maps will be discussed separately in the following sections. We will then demonstrate a method for combining the information available from these maps to make appropriate selections in our environment.

### 4.1 Density

Our first metric for computing visual attention is based on edge element density. Each element in the edge map $E(\mathbf{I})$ of image $\mathbf{I}$ has an intensity associated with the strength of the edge to which it belongs. We compute a density map $D(i,j)$ over the entire image by convolving a Gaussian[2] windowing operator of size $A \times B$ with the edge map. Each point in the map is divided by the total possible number of edgels, giving the following measure of density:

$$D(i,j) = \alpha \int_{j-\frac{B}{2}}^{j+\frac{B}{2}} \int_{i-\frac{A}{2}}^{i+\frac{A}{2}} e^{-\frac{(\phi-i)^2+(\theta-j)^2}{2\sigma^2}} E(I_{x,y}(\phi,\theta)) d\phi d\theta \tag{5}$$

with

$$\alpha = \frac{1}{\int_{-\frac{A}{2}}^{\frac{A}{2}} \int_{-\frac{B}{2}}^{\frac{B}{2}} e^{-\frac{l^2+m^2}{2\sigma^2}} dl dm} \tag{6}$$

Since we are interested in unusual locations, we define the interest, $\Gamma$, as the deviation from the mean $\hat{D}$ over the entire image:

$$\Gamma(i,j) = |D(i,j) - \hat{D}| \tag{7}$$

We then find the extrema of this map, that is, the locations with the highest deviations from the mean and define those as the most interesting locations, based on edge density alone. This involves an implicit assumption that the edgel density distribution is uni-modal, since otherwise we may occasionally obtain non-intuitive results. The extrema of this operator will typically be associated with edge junctions and other geometric "events" in typical indoor images, although they can also be associated with empty regions in textured images. See section 6.2 for examples of the use of the density metric.

### 4.2 Orientation

The second operator we use for computing attention is edgel orientation. Each entry in the orientation map $\Theta(\mathbf{I})$ is the orientation of the corresponding edgel in the edge map. We compute a local orientation signature $O(i,j)$ similar to the density map defined above, as follows. In order to select orientations that are maximally different from the typical orientation structure in the scene, we make a noise-insensitive estimate of the most likely orientation: a robust maximum.

[2]A Gaussian operator has desirable properties in terms of localization in both space and frequency space.

Given a function,

$$\Phi(k,i,j) \quad k \in [0,\pi) \tag{8}$$

which returns the number of edgels with orientation $k$ in the local neighborhood of $(i,j)$, we can compute the robust maximum orientation as follows:

$$\Phi^*(k,i,j) = \Phi(k \bmod \pi, i, j) \quad k \in R \tag{9}$$

$$O(i,j) = \max_{k\in[0,\pi)} \int_{q-\frac{\omega}{2}}^{q+\frac{\omega}{2}} \Phi^*(k,i,j) dk \quad \omega \in (0, \frac{\pi}{2}) \tag{10}$$

where $\omega$ is the width of the subsection of the orientation distribution we wish to consider. In practice, $\Phi(k,i,j)$ is also convolved with a Gaussian windowing operator. Again, we are interested in unusual locations with respect to orientation, so we define interest as the deviation from the overall robust maximum orientation $\hat{O}$:

$$\Omega(i,j) = |O(i,j) - \hat{O}| \tag{11}$$

We then find the extrema of this map which will be the local neighborhoods with the highest deviation from the maximum orientation, and define them as the most interesting locations based on orientation alone. Section 6.2 demonstrates the behavior of the orientation operator.

### 4.3 Combining Density and Orientation

A suitable function is needed to combine the information from the density and orientation operators such that we achieve results which are stronger than the results which each operator can provide independently. To produce a compound interest operator, we combine the individual interest ratings due to the individual measurements using and $L_n$ metric:

$$C(i,j) = \sqrt[n]{(\gamma\Gamma(i,j))^n + ((1-\gamma)\Omega(i,j))^n} \tag{12}$$

where $n$ is a constant and the value chosen for $\gamma$ depends on the type of environment being sampled. By using a large value of $n$ we obtain a behavior that resembles a winner-take-all scheme, while smaller values of $n$ exploit combinations of features. We have also considered the use of multiple attention maps at multiple spatial scales, leading to feature detections $\mathcal{M}_f^s(i,j)$ where $s$ and $f$ are indices that specify the scale and feature type. In this case, we combine these maps using:

$$C(i,j) = \sqrt[n]{\Sigma_s \Sigma_f \gamma_f^s \mathcal{M}_f^s(i,j))^n}. \tag{13}$$

Section 6.2 demonstrates the effectiveness of the combined density-orientation operator with $n = 1$.

## 5 System Architecture

Our hardware system is composed of a Nomadic Technologies Nomad 200 mobile robot with an onboard computer, and a CCD camera mounted in a 2-DOF pan and tilt unit on top of the robot. There are several software components to the entire viewpoint selection system including robot exploration, image acquisition, attention processing, image registration, and graph creation. The specifics of these subsystems are outlined below.

The final component in the system is a software package from Apple Computer Inc. which combines the stitched panoramic photographs and the topological representation from the nodal graph component to form a multi-node *QuickTime VR* movie.

In principle any exploration algorithm may be used with the viewpoint selection system so long as the environment is fully sampled; to exemplify this independence the software was designed using a "plug-in" architecture. In practice, we have used[3] a simple algorithm akin to a bouncing ball: the robot travels in a straight path until it is obstructed at which point it rotates by a random angle until it can once again move forward. Although this algorithm does not exploit the layout of the environment it still manages to cover the free space quite well [10]. The motivation for the plug-in exploration architecture was based on the potential variability in environments one will encounter and the fact that they might mandate environment-specific strategies. For example, in an office environment, exploration based on covering the Voronoi diagram might be more appropriate [11, 12]. Similarly, an open environment would most likely need to be sampled more evenly, perhaps using a grazing exploration method [13].

As the robot explores the environment, video images are collected using the CCD camera. In order to minimize warping effects during stitching, we rotate the camera about its optical center or nodal point.

The first phase of the image analysis process performs edge detection on the images acquired using the Canny operator [14, 15], which returns an edge map, and an orientation map. The image analysis process implements the attention metrics outlined in section 4 on all of the images acquired. These are then statistically analyzed so that the top $n$ extrema of the density/orientation map(s) can be chosen as the locations which will be part of the final graph.

To produce a panoramic image at each location, adjacent images taken with the same $(x, y)$ but different orientations must be fused together to produce a single cylindrical or spherical image. To solve this "mosaicing" problem we use cross correlation to find the best correspondence; observe that the problem is simplified by the fact that the images are acquired using only rotations about a fixed nodal point [1, 16]. Once this overlap is found, the intensities of the two adjacent images are blended (averaged) to remove any seam which may be present [17].

Because we wish to create a multi-node VR *scene*, the relationship between the panoramic photographs must be established, and a facility provided for the user to move between these photographs. As mentioned above, QuickTime VR provides a facility for moving between nodes called *hot-spots*. These are encoded using a mask over the panoramic image, with the value of the mask at the current location of the user's pointer determining which node will next be visited should the user to decide to move. In order to automate the creation of these masks, we construct a fully connected graph representing the virtual environment. Because each image is associated with a known pose $(x, y, \theta)$ in the plane[4] we are able to determine the arc lengths and positions in the mask which correspond to other nodes in the graph. Although localization errors will, even with correction techniques, corrupt these pose estimates, we only require approximate positions to construct the topological representation. Consider figure 2: assuming the radius of each

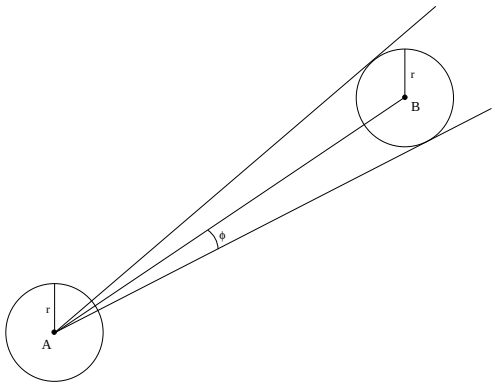


Figure 2: Calculating arc length on viewpoint A.

node is fixed[5] we can compute the intersection range

[3] We have also developed and tested additional exploration algorithms but they are outside the scope of this paper.

[4] We assume planar environments, although our approach could be readily extended to 3D environments.

[5] In practice the radii are fixed to a certain amount of vertical lines (pixels).

$I_{AB}$ of the panorama $B$ on the panorama $A$ (in $A$'s local orientation frame):

$$\phi = \tan^{-1}\left(\frac{\sqrt{(A_x - B_x)^2 + (A_y - B_y)^2}}{r}\right) \quad (14)$$

$$\gamma = \tan^{-1}\left(\frac{B_y - A_y}{B_x - A_x}\right) \quad (15)$$

$$I_{AB} = [\gamma - \phi - A_\theta, \gamma + \phi - A_\theta] \quad (16)$$

where $r$ is the radius of the panoramas. A compli-

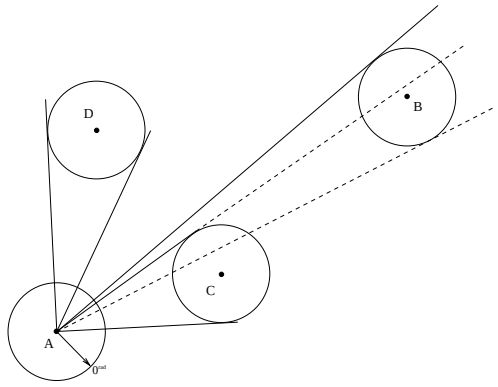


Figure 3: Viewpoint A's view of location B is obscured by intermediate location C.

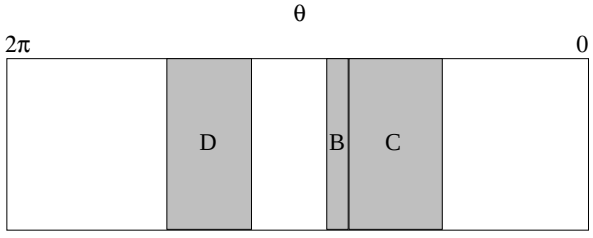


Figure 4: The connectivity mask associated with viewpoint A as depicted in figure 3.

cation arises if we have a situation similar to that of figure 3. This suggests an algorithm for creating the masks – process adjacent nodes in order of their increasing distance from the source node. If the arcs of two nodes overlap, only keep the non-overlapping remainder of the arc for the occluded node. The resultant mask can be seen in figure 4.

This pre-computation provides a fully connected graph; however, since we do not build a map of the environment in the current exploration model, some adjacent nodes may be occluded by the environment itself. Although we do not consider this to be a limitation, edges in the graph could be easily removed by the user.

# 6 A Visit to the Art Gallery

## 6.1 Overview

The experimental data discussed in this paper was obtained at the Canadian Centre for Architecture, located in Montréal. Due to the sensitive nature of the environment, the exploration was carried out manually.

The gallery floor contained pedestals holding exhibits encased in glass, which were scattered about the environment. The walls contained many exhibits, usually of uniform size, spread around a room.

The path followed by the robot covered several rooms of the gallery, and included 3 rooms which were not part of the exhibit. The robot acquired 36 images at each of the 17 different locations along the trajectory. These images were then edge detected, analyzed, and registered as outlined in section 5.

## 6.2 Results

The individual locations chosen in the panoramic images according to the density and orientation operators are shown in figure 5. The top three photographs demonstrate the effectiveness of the density operator. The images whose edgel density variance is highest show the two rooms in the gallery which were substantially different from the remainder of the gallery. Furthermore, the candidate with the lowest variance shows a region of a wall which contains little more than a brick-like texture. The effectiveness of the edgel orientation operator is displayed in the next three images. Here, the candidates containing multiple curves such as the pattern in the painting, the frosted glass pattern in the door, or the marble fireplace, have the highest variance. This is due to the fact that the remainder of the gallery is mostly rectilinear as shown by the candidate whose variance was lowest. The latter shows a series of photographs in frames, which are all organized adjacently.

We also looked at the candidates from the combined density-orientation operator. Here we see an interesting and positive result – the top candidate in the distribution is the image *just between* the top density candidate, and the second orientation candidate.

Equally appealing is the operator's choice representing minimal variance – the last image is convincingly boring!

## 7 Discussion

We have outlined the structure of an intelligent robotic system that can automatically construct an image-based model of an environment. Once the model is built, it can be inspected at leisure by a human. Such a model could be used for inspection, teleoperation, or tourism. The results obtained from our approach appear to be surprisingly well suited to such tasks, even without environment specific tuning. In practice, however, one might wish to augment the approach with either domain-specific rules on what views are most interesting, or specific locations that embody required views. For example, in constructing a tour of an art museum it might be desirable tune the system to acquire fronto-parallel views of rectangular objects (paintings) while in a nuclear reactor facility one might like to acquire an image looking at the cooling pipes no matter how boring they appear.

In the work described here, we have not not addressed how the required "mean" estimates of the local environment are computed: that is, our presentation presupposed that all the data is available (i.e. it is an off-line algorithm). In fact, it is possible to perform on-line estimation of the environmental parameters and use an on-line algorithm referred to as *alpha-backtracking* to select the viewpoints in real time [1]. While a detailed discussion of this is outside the scope of this paper, there is a natural tradeoff between the extent to which the final viewpoints match the theoretically optimal ones, and the computational and physical resources used in the task [18, 1, 2].

In ongoing work we are addressing the above issues. A related issue is the reconstruction of viewpoints that have never actually been sampled. This is an active research problem in it's own right, and may eventually provide a useful complement to the approach described here.

## References


[1] E. Bourque, G. Dudek, and P. Ciaravola, "Robotic sightseeing - a method for automatically creating virtual environments," in *Proceedings of the IEEE International Conference on Robotics and Automation*, vol. 4, (Leuven, Belgium), pp. 3186–3191, May 1998.

[2] E. Bourque and G. Dudek, "Automated creation of image-based virtual reality," in *Proc. SPIE Conference on Intelligent Systems and Manufacturing*, (Pittsburgh, PA), pp. 292–303, October 1997.

[3] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurbiology*, vol. 4, pp. 219–227, 1985.

[4] W. Schneider and R. M. Shiffrin, "Controlled and automatic human information processing: I. Detection, Search, and Attention," *Psychological Review*, vol. 84, pp. 1–66, January 1977.

[5] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modelling visual attention via selective tuning," *Artificial Intelligence*, vol. 78, no. 1-2, pp. 507–547, 1995.

[6] 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.

[7] M. Langer, G. Dudek, and S. W. Zucker, "Space occupancy using multiple shadowimages," in *Proceedings of the IEEE Conference on Intelligent Robotic Systems*, (Pittsburgh, PA), pp. 390–396, IEEE Press, August 1995.

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

[9] L. McMillan and G. Bishop, "Plenoptic modeling: An image-based rendering system," in *Proceedings of the ACM SIGGRAPH*, (Los Angeles, CA), pp. 39–46, ACM, August 1995.

[10] P. Beame, A. Borodin, P. Raghavan, W. L. Ruzzo, and M. Tompa, "Time-space tradeoffs for undirected graph traversal by graph automata," *Information and Computation*, vol. 130, pp. 101–129, November 1996.

[11] B. Kuipers and Y.-T. Byun, "A robot exploration and mapping strategy based on a semantic hierachy of spatial representations," *Robotics and Autonomous Systems*, vol. 8, pp. 46–63, 1991.

[12] H. Choset, K. Nagatani, and A. Rizzi, "Sensor based planning: Using a homing strategy and local map method to implement the generalized voronoi graph," in *Proc. SPIE Conference on Mobile Robotics*, (Pittsburgh, PA), 1997.

[13] T. Balch and R. C. Arkin, "Communication in reactive multiagent robotic systems," *Autonomous Robots*, vol. 1, no. 1, pp. 27–52, 1994.

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

[17] 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.

[18] E. Bourque and G. Dudek, "Automated image-based mapping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition–Workshop on Perception of Mobile Agents*, pp. 61–70, June 1998.

(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 5: Selections made by the viewpoint selection system. Images (a) and (b) were the top selections using the density operator, while (c) was the lowest. Images (d) and (e) were the top selections using the orientation operator, and (f) was the lowest. Images (g) and (h) were th top selections for the combined density-orientation operator, and (i) had the lowest variance.