Graphics + Robotics + AI = Fast, 3D Scene Construction

Brian C. Mashburn and Douglas S. Blank
Department of Computer Science and Computer Engineering
University of Arkansas
Fayetteville, AR 72701
{bmashbu,dblank}@comp.uark.edu

Abstract

Scene construction is the process of building realistic, three-dimensional representations, or models, of real world environments, such as rooms, landsacpes or buildings. Because of the realistic quality of images being produced, current scene construction algorithms require manual processing by human experts. However, the benefits of having such 3D models are great. Consider a situation where a three-dimensional model of an evironment must be created in real-time. Existing scene construction algorithms will not suffice. Therefore we have outlined a new area of quick and dirty scene construction where usable, low resolution, three-dimensional models of real world environments can be created in real time. This paper describes the need for such a system, and provides a generalized approach for accomplishing the task.

Introduction

The process of constructing scenes, also known as "scene modeling" or "scene rendering" (Debevec , Taylor, & Milik 1996) (Szeliski & Johnson 1995) has been applied by computer-aided developers to many environments. Currently, the process is primarily dedicated to the development of photo-quality representations of the real world. Representations of photo-quality images are very time consuming and computationally expensive to create since the goal is to best replicate a real world environment. But what about time critical situtaions such as Search and Rescue operations (SAR) where an environmental map must be created in real-time so that lives can be saved? There is a trade-off between representations of high quality and those that need to be constructed quickly. Our goal is to strike a balance between the two conflicting approaches: slowly constructed, high quality images, versus low resultion ones.

We wish to automatically construct usable three-dimensional representations of real world environments through the use of autonomous mobile robots in real time. We accomplish this by collecting sonar data for the environmental dimensions, video images for filling the interior of the environment, and combining the two sets of data via the Virtual Reality Modeling Language (VRML) for real-time viewing.

We have divided the process of automated scene construction into three sub-problems:

a. Environmental Map Construction,
b. Panorama Image Construction and
c. Map / Panorama Fusion.

The next sections of the paper will describe our approach for each of the sub-problems, followed by some preliminary experimental results in simple environments.

Problem Breakdown

Environmental Map Construction

"Environmental mapping" is the process of taking robot collected data from real world environments in order to construct a two dimensional layout, or map, of the area. The greater the required detail the map must contain, the more computationally expensive the environmental mapping procedure becomes, which, in turn, creates a lengthy production time. Therefore, we reduce the detail of the environmental map to nothing more than the environmental dimensions.

The process of mapping a real world environment is difficult for the following reasons. Since the environmental data are collected via an autonomous mobile robot, problems such as robot navigation, robot planning, obstacle avoidance and path following must be considered and overcome. Physical problems pertaining to the robot, such as wheel slippage, sensor limitations, and sensor failure must also be taken into account. These problems contribute to the lack of "dead reckoning," or inability for the robot to locate its exact position in an environment. If a robot is not equipped with a global positioning system (GPS), which performs satellite triangulation for very accurate location detection, a robot's exact position cannot be ascertained.

We propose an approach for data collection and map construction that limits robot exposure time in the environment while still collecting enough data to construct a usable environmental map. We simplify the problems of robot navigation, planning, obstacle avoidance and path following by implementing a process grounded in the concepts of a "perceptual search" (Murphy, Sprouse, & Hawkins 1995). A perceptual

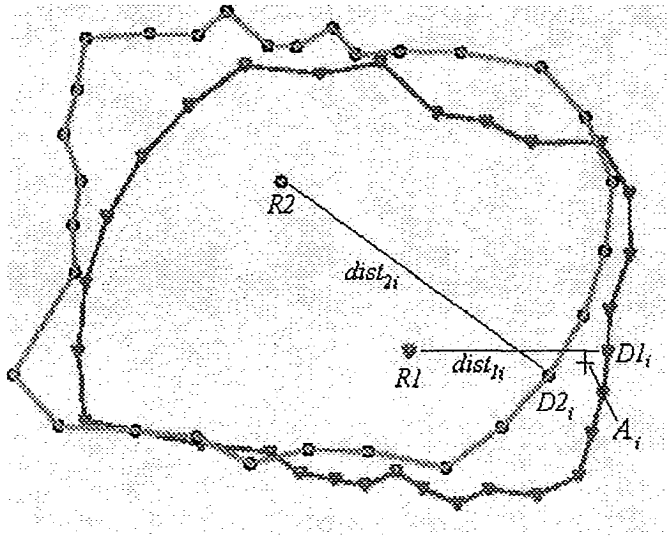

Figure 1: 2D average data point calculation at the ith location from a weighted average function.

search is a process of moving through an environment and collecting data, until, at a perceptual level, enough data has been collected to cover the environment. After a set of data has been collected, at a given location in the environment calculations on the collected data will facilitate the robot's navigation to unexplored areas in the environment. Once the robot has completed exploration, i.e. probabilistically covered the entire environment, the robot will exit. This ensures that the robot will only visit a minimal number of locations, which we call "reference locations," in order to collect enough data for the environmental mapping. Techniques, such as perceptual search, must be used for the identification of reference locations in the environment and for planning a route that the robot can take, to ensure the real-time aspect of this approach to scene construction.

The number of reference locations the robot must visit varies per environment. For small areas, only a single location may be required. The actual number of locations visited per environment is inversely proportional to the limitations of the sensors. In other words, more powerful sensors mean greater range and accuracy thereby decreasing the number of locations needed to obtain enough data to cover the environment. By requiring only enough data to gain a sense of the layout and not a detailed mapping of the entire area as required by current algorithms, the amount of time the robot spends planning and navigating in the environment is minimized.

At each reference location, sonar data is collected using a method we call the "data collection spin", i.e. the robot will spin in a 360 degree circle, collecting sonar data all the while. Since sonar sensors calculate the time difference between a emitted sound wave and its return, this allows for distance calculations between the sonar and objects in, or on the boundaries of, the environment.

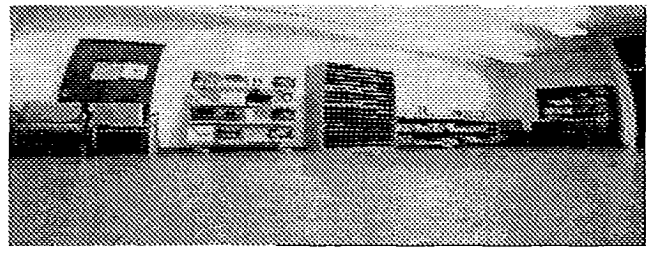

Figure 2: Example of an image afflicted with barrel distortion.

The result of each data collection spin at a reference location will be a set of sonar values corresponding distances at each of a series of angular rotational values. Once the environment has been covered, the resulting collection of sonar data sets will be averaged to form a single data set from which the dimensions of the environment can be calculated. This average calculation is a weighted average function that assigns a higher weight to sonar readings that are closer to the reference location, and a lower weight to readings that are closer to the maximum range of the sonar sensors. This type of weighted function is implemented in attempt to minimize the range error of the sonar sensors, i.e. a close reading is more accurate than a reading farther from the sensor. As readings becomes more distant, the less accurate they become, and, therefore, less useful in the calculation of the environmental dimensions. Figure 1 shows the average data point, denoted by A, is closer to D1 than it is to D2. This is due to the fact that dist1 is less than dist2, therefore D1 gets a higher weight than D2. The resulting averaged data set will then be interpreted and the dimensions of the environment will be sketched. This approach allows for the development of a environmental mapping from a small group of reference locations in real-time.

Panorama Image Construction

In order to construct a believable representation of an environment, video images must be taken from within the environment. Since we do not have a camera that will allow us to take a picture of the entirety of an environment in a single shot, we must pursue the process of "panoramic video imaging" (Szeliski & Johnson 1995). Panoramic video imaging is the process of combining a sequential series of video images obtained from a real world environment to produce a single seamless image. For example, the Mars Pathfinder collected a series of video images from the Martian surface which were processed via panorama image construction to produce panoramic images. The panoramic images were constructed by "stitching" the series of serial images together. Our method is based on video images being collected from the environment at each reference location as the robot performs its data collection spin.

The process of building panoramic images is difficult

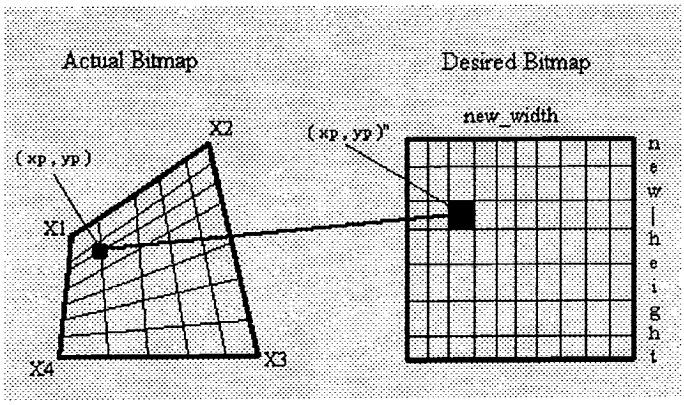

Figure 3: Normalization process.

for several reasons. The size of each video image is limited by the size of the camera's lens and the position of the camera in the environment. Since cameras typically have a wide angle lens, each image suffers from "barrel distortion". Barrel distortion is an affliction which warps the image into looking like a barrel, i.e. stretched around the middle and thinner on the top and bottom as can be seen in Figure 2.

Each image also contains information that is duplicated in other images. The overlap is corrected by removing duplicate parts of each of the images. However, there is a downside to this approach: the overlapping regions usually create a seam in the panoramic image.

Each image has a slightly different perspective view, due to the orientation of the camera in the environment, causing the process of vertical alignment to become a problem. Since a discrepancy exists between the perspective views of each individual image and since the information in each image has a slightly different orientation, the images must be "normalized" before the construction of the panoramic image. Normalization is the process of re-sizing the information contained within an image in order to change its perspective view, thus creating a standard perspective for the set of images. A diagram of the normalization process can be seen in Figure 3.

Occlusion, either partial or full, is the blockage of certain areas of an environment due to the presence of objects located in the interior of the environment. Our method circumvents the problem of occlusion through the use of multiple reference locations where environmental data are collected. If an area of the environment is occluded at one reference location, the environmental data from another reference location can be utilized for the blocked regions. Our initial experiments were performed in simplified environments free of objects. In order to handle objects in the interior of the environment, we must be able to consider the environment not as an independent set of reference locations, but consider how those reference locations interact. Objects located between the reference locations must also be located and referenced for their accurate representation and placement in the environment.

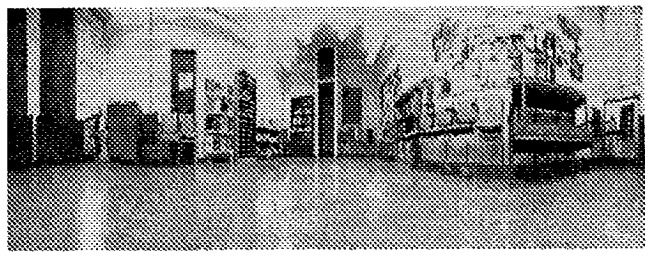

Figure 4: Panoramic Image.

Once each image has been corrected through normalization and all have the same perspective view, the process of combining these images to form a single image is simplified, thus resulting in a more seamless view of the environment. But even with image normalization, vertical inconsistencies can still exist. More intelligent processes must then be defined to automatically overcome all existent inconsistencies. An example panoramic image can be seen in Figure 4.

Map / Panorama Fusion

At this point we have both a horizontal 2D mapping, constructed from sonar data, and a vertical 2D panoramic image, constructed from video data. From these two representations, the layout of an environment can be realized by combining these two mappings by hand, but this goes against our goal of autonomy. A better solution to the problem is an autonomous process that takes the two representations and merges them into a single, seamless 3D representation.

The mapping representation is a 2D map with x-y dimensionality, while the video representation is a 2D panoramic image with x-z dimensionality. By fusing a 2D x-y map and a 2D x-z panoramic a 3D x-y-z panoramic map of the environment is created, thus fulfilling our desired result.

"Data fusion" is the process of combining data modalities into a single, data representation (Murphy, Sprouse, & Hawkins 1995). The process of fusing sonar data and video images for the production of a 3D representation is a prime example. We have broken the process of fusing sonar and video data into three steps:

a. Extract the "wall regions" from a panoramic image through edge detection and wall extraction.

b. Calculate the dimensions of the 3D representation from sonar data through min/max average functions.

c. Create the 3D representation by applying both the extracted wall regions and the sonar dimensionality to a "framework".

Data modalities can vary in terms of dimensionality, i.e. sonar spanning x-y and video spanning x-z. Therefore data fusion requires a multidimensional framework to act as a canvas. The data modalities are applied to this framework in order to construct the 3D representation.

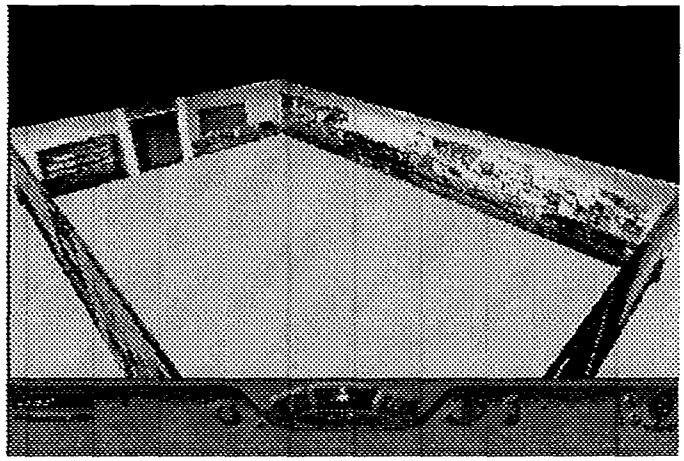

Figure 5: Top view of 3D VRML representation built from the fusion of sonar and video information collected from the environment

In current scene construction algorithms, this framework is represented as a "3D mesh" (Szeliski & Johnson 1995). A 3D mesh is a contiguous set of polygonal regions, or faces, which map the layout of an environment. Each face of the mesh corresponds to a set of video data points in the panoramic image. The video data points might be mapped onto the polygonal regions through "texture mapping", a computationally expensive process of point to point application. For texture mapping onto a 3D mesh, the number of faces in the mesh is directly proportional to the mapping time, i.e. more faces take more time (Szeliski & Johnson 1995).

However, the use of a 3D mesh and texture mapping is too computationally expensive for our use. Instead, our framework is constructed in the Virtual Reality Modeling Language (VRML), a 3D web-based, graphical scripting language. VRML is simple to construct from multiple data modalities, i.e. sonar dimensions and video images. The dimensions of the VRML framework are calculated from the sonar data and each wall of the VRML framework holds a video image that has been extracted from a panoramic image. VRML was chosen because of its simplicity, both for its creation ease and because of its wide-range use for real-time graphics on the World Wide Web. Figure 5 shows the final result of the process of automatic scene construction by scripting the VRML file based on the processed data.

Conclusion and Future Work

Environmental map construction, as with many of the current scene construction approaches, utilizes an autonomous mobile robot for data collection in an environment. The process of environmental map construction builds a 2D horizontal representation of the environmental dimensions from collected sonar data via a weighted average function. Currently there exist no algorithms to accomplish the environmental data collection in real-time, so further research is needed in the area of robot navigation and data collection in real-world environments. All other computations of environmental map construction are real-time compliant, so once the algorithms of robot navigation have been defined, environmental map construction can also be accomplish in real-time for simplified, object-free environments.

The process of panoramic image construction utilizes video images to build a 2D vertical panoramic image of an area. The process of panoramic image construction has been implemented and automated in real-time, given a set of perfect serial images. Further research in the area of panoramic image construction is necessary for the automatic production of more seamless panoramic images of real-world environments. The current research has been shown to produce usable representations of real-world environments in real-time.

The final process of map / panorama fusion takes the 2D horizontal representation and combines it with the 2D vertical representation. This results in a 3D representation of a real-world environment. Currently, there are no algorithms defined for the location and extraction of wall sections from a panoramic image, therefore this process is completed manually. Further research in the area of wall section identification and extraction must be defined for the process of map /panorama fusion to be automated in real-time. We have outlined the process for map / panorama fusion which results in the production of a useful 3D representation of a real-world environment.

We have identified a new niche in the automatic, real-time construction of quick and dirty 3D representations of real-world environments. We have also demonstrated a generalized solution for the problem. By following the outline we have constructed in this paper, we believe the automatic production of 3D representations for real-world environments is possible.

References

Debevec , P.; Taylor, C.; and Milik, J. 1996. Modeling and rendering architecture from photographs: A hybrid geometry-and-image-based approach. In *SIGGRAPH-96*. ACM.

Murphy, R. R.; Gomes, K.; and Hershberger, D. 1996. Ultrasonic data fusion as a function of robot velocity. *SPIE Sensor Fusion and Distributed Robotic Agents* 114–126.

Murphy, R. R.; Sprouse, J. J.; and Hawkins, D. 1995. An action-oriented perception approach to perceptual search. *Autonomous Robotics*.

Szeliski, R. Weiss, R., and Johnson, A. 1995. Scene sensing (aka videocopying of 3-d scenes). Technical report, Cambridge Research Laboratory, Digital Equipment Corporation.