Proceedings of the IEEE Workshop on Omnidirectional Vision,
pages 169-174, Hilton Head Island, SC, June 2000.

Segmentation, Tracking and Interpretation Using Panoramic Video

Mircea Nicolescu Gérard Medioni
Integrated Media Systems Center
University of Southern California
Los Angeles, CA 90089-0273
{mnicoles, medioni}@iris.usc.edu

Mi-Suen Lee
Philips Research-USA
Briarcliff Manor, NY 10510-2099
Mi-Suen.Lee@Philips.com

Abstract

Choosing the appropriate type of video input is an important issue for any vision-based system and the right decision must take into account the specific requirements of the intended application. We present GlobeAll, a modular four-component prototype for a vision-based Intelligent Room: a panoramic video input component that uses an electronic pan-tilt-zoom camera array, a background learning and foreground extraction component, a tracking component and an interpretation component. In the context of Intelligent Room systems, we establish several qualitative criteria to evaluate their video input component. We use these principles to compare our panoramic video system with two other current solutions – mobile pan-tilt-zoom and wide-angle lens cameras – and we show that electronic pan-tilt-zoom systems best satisfy our criteria.

1. Introduction

Most of the existing attempts targeted at Intelligent Environments or Perceptual User Interfaces rely on (or are limited to) visual input. Several Computer Vision techniques are involved in processing this input, typically implemented as separate modules dedicated to different tasks: background learning and foreground extraction, tracking, 3D modeling, behavior interpretation.

While general design principles and requirements for Intelligent Environments have been frequently discussed in the literature [1, 2, 3], we focus here only on the problem of acquiring visual input for such systems. In this context we establish the following qualitative criteria that must be considered in order to choose the most appropriate type of video input:

- **Region of awareness** ("How much of the scene can I see?"). Essentially, it represents the portion of the environment being monitored. For a single-camera system or a compact array of cameras, the region of awareness is given by the overall field of view. It is desirable to acquire a region of awareness as large as possible.
- **Region of interest** ("What part of the scene am I looking at?"). Even if the system is visually aware of a larger part of the scene, it usually concentrates on a certain region, where some activity of interest is in progress (for example, a person walking). This region of interest is usually the one currently displayed. We characterize the region of interest through three criteria:

- *Quality*. This criterion is satisfied when the system is able to achieve a level of detail fine enough to allow further processing. It is given by the resolution of the extracted image for the region of interest.

- *Precision of location* ("Where exactly am I looking?"). It is desirable to know exactly where each region of interest is placed with respect to a common reference frame, or how different regions of interest are spatially related to each other.

- *Speed of redirection* ("How fast can I redirect my gaze?"). The system should be able to rapidly switch from a region of interest to another. For example, when an activity of interest (someone entering the room) occurs within the region of awareness, the region of interest must be appropriately positioned. Moreover, in order to track a moving person, the region of interest must be moved along fast enough to avoid losing the target.

- **Background model.** Ideally, a unique model of the background should be maintained for the entire region of awareness. Problems occur if the background model must be assembled from several (partially) overlapping parts (in the case of multiple or mobile cameras). The system must be able to maintain a correct and, if possible, unique background model for the whole scene.
- **Depth range.** A successful system should be able to deal with both close and distant objects.
- **Cost.** An obvious requirement is to use affordable camera systems and supporting hardware.

There is no camera that satisfies all these criteria. In this paper, we hope to give a more systematic view on visual input for Intelligent Rooms, through a twofold contribution: In the next section, we present GlobeAll, a modular prototype for a vision-based Intelligent Room, which uses an electronic pan-tilt-zoom camera array. Then, in Section 3 we compare our approach with other two types of camera systems - mobile pan-tilt-zoom platforms and wide-angle lens cameras - and show that electronic pan-tilt-zoom cameras offer the best trade-off solution for our requirements.

2. Description of our system

We developed GlobeAll, part of a wider research effort that is targeted at a modular framework for vision-based Intelligent Environments. GlobeAll is a modular four-component prototype based on panoramic video input through an electronic pan-tilt-zoom camera array. The physical camera setup is shown in Figure 1 and an overview of the system components is given in Figure 2.

The visual input is acquired by the electronic pan-tilt-zoom component, which generates a planar mosaic and a synthesized view (Virtual Camera) corresponding to the desired region of interest. A background learning and foreground extraction component maintains an adaptive background model and segments moving objects as sprites. Among them, a target is selected and followed by the tracking component. Based on user-defined generic descriptions, the interpretation module analyzes the models generated by previous components (sprites, trajectories) and augments them with semantic labels. It can also issue commands for external actions (such as voice synthesis).

Depending on the activity in the observed scene, our system has two distinct modes of operation. When idle (no one is in the room), GlobeAll is learning the background and, in future developments, will recover 3D models, recognizing and labeling familiar objects. When someone enters the room, the system switches to the active mode, where it focuses on extracting sprites, recovering 3D positions, modeling persons, tracking and behavior interpretation.

In the following sections we give a description of each of these modules. Special emphasis is put on how they were designed in order to benefit from our panoramic video input system.

2.1. Electronic pan-tilt-zoom camera system

In our implementation, we use an array of five fixed CMOS cameras, mounted on a spherical setup and oriented radially, so that they acquire a field of view large enough (approximately 130° both vertically and horizontally), while also maintaining some overlapping required for calibration. Essentially, we create a two-dimensional mosaic of the observed scene with geometric and photometric correction, then we generate any arbitrary intermediate view, with the ability of performing electronic pan, tilt and zoom operations.

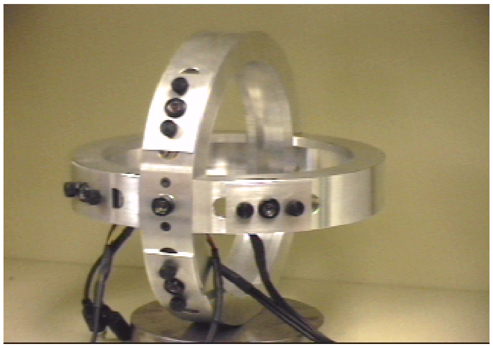


Figure 1. GlobeAll camera array.

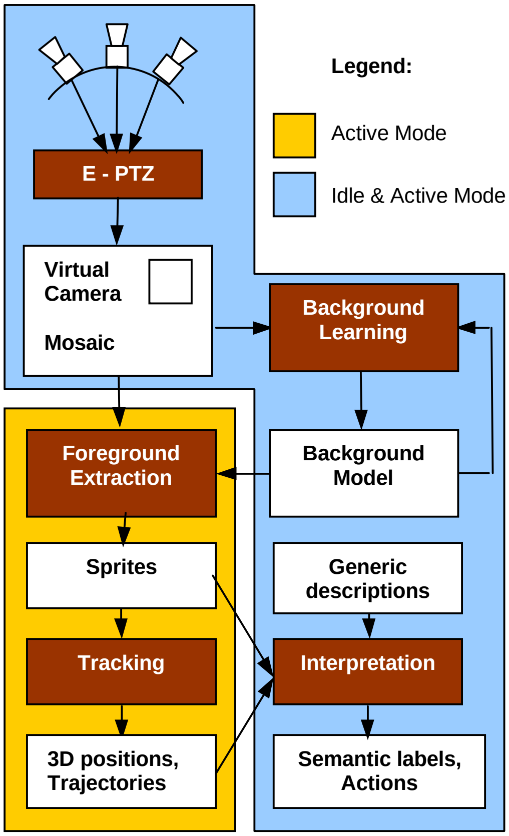
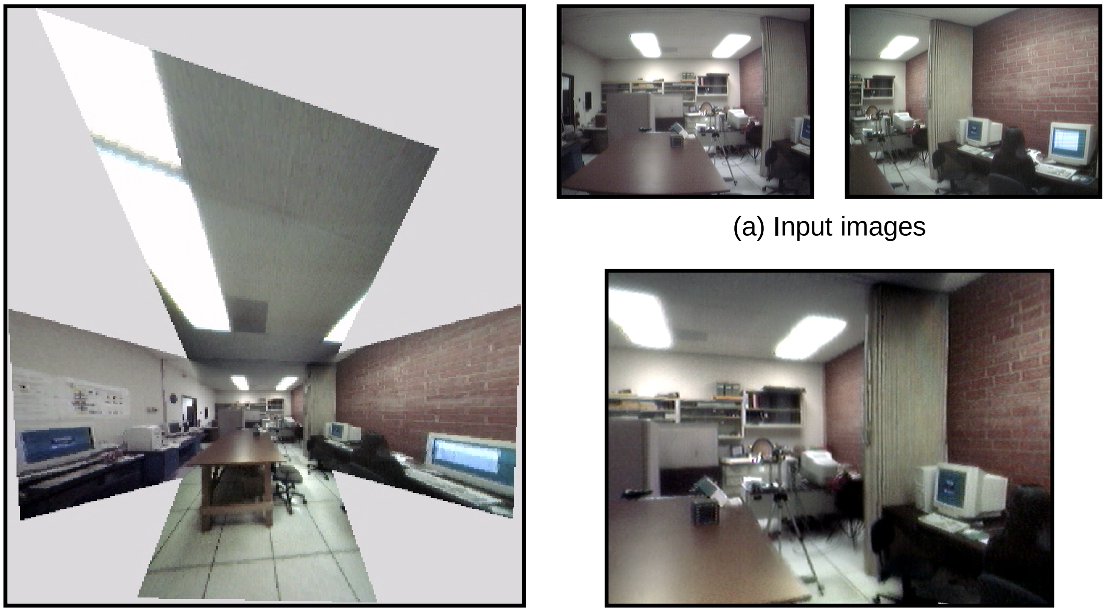


Figure 2. System overview.

During an off-line calibration process, the images from each camera are registered by computing a full perspective transformation that aligns them to a common reference frame.

At run-time, as the input images are continuously captured by their camera array, they are first corrected for radial lens distortion. We pre-compute pixel relocation maps, so that, in real time, we just use them as



(a) Input images

(b) Planar mosaic

(c) Virtual Camera synthesized view

Figure 3. Results from the electronic pan-tilt-zoom system.

lookup tables to remap pixels and thus compensate for lens distortion [4].

In order to generate intermediate views, the images acquired by the cameras must be first registered and merged into a panorama (environment map) of the entire scene observed. Because our application requires less than 180° pan and tilt, we chose a planar environment map (planar mosaic), which is the least expensive in processing time compared to other types of environment maps, such as cubes, cylinders or spheres. This planar mosaic represents an imaginary plane placed at an arbitrary distance in front of one of the cameras, chosen as reference (typically the one in the middle). Each frame is warped into this mosaic by using the perspective transformation determined in the calibration process [5].

At this point, the individual frames acquired by the cameras have been registered, so that their geometric alignment into the mosaic should be almost perfect. However, due to different light conditions, drastic exposure differences and even physical differences in manufacture between cameras, the same object may appear with different intensity from a frame to another. As a result, there are visible boundaries between frames in the mosaic. In order to deal with this problem we developed an intensity blending algorithm, based on the weighted average of pixel values over the transition regions between images. In this approach, the intensity of the resulting pixel in the overlapping region is defined as a weighted sum of the intensities of corresponding pixels in the frames that overlap.

In order for our system to perform all the functions of a standard pan-tilt-zoom camera, we need to have the ability to synthesize novel views for any given intermediate pan-tilt angles or zoom factor. What we get is a *Virtual Camera*, which is functionally equivalent to a regular pan-tilt-zoom platform. For this purpose, the appropriate portion of the planar mosaic is determined according to the desired pan-tilt angles and zoom factor, and then unwarped back to the novel view in the Virtual Camera. A synthesized view, together with the planar mosaic and two input images are shown in Figure 3.

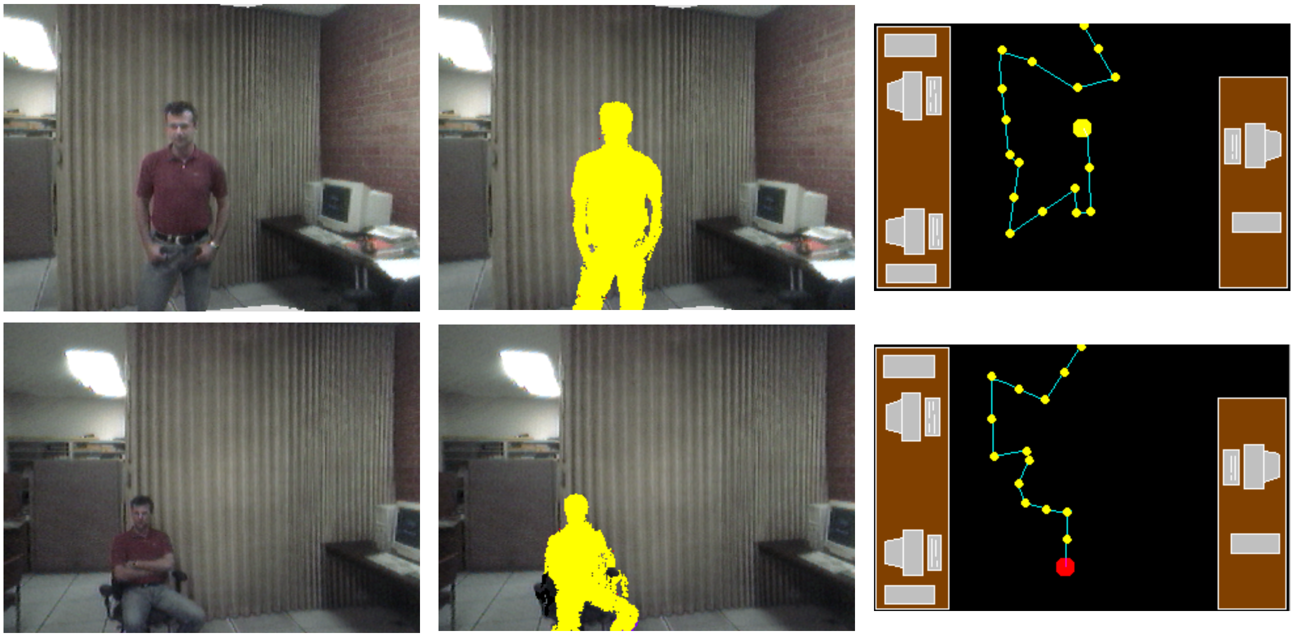
2.2. Background learning / foreground extraction

The second component in our system is responsible for maintaining an adaptive background model for the entire region of awareness, and for segmenting the moving objects that appear in foreground. Our approach involves learning a statistical color model of the background, which is used for detecting changes produced by occluding elements. Each background pixel value is modeled as a multi-dimensional Gaussian distribution in RGB space, characterized by its mean value μ and standard deviation σ.

When a new frame is processed, the observed pixel values are compared to the current distribution in order to segment foreground elements. A new pixel observation x is marked as foreground if:

$$(x-\mu)^2 > (2\sigma)^2$$

The pixels detected as foreground are then grouped in connected components, so that each moving object is represented as a sprite (silhouette and texture) [6, 7].



(a) Input frames (b) Detected object (c) Overhead trajectory and attitude

Figure 4. Foreground detection and 3D trajectories. In the diagrams at right, a yellow (light) spot represents "standing" attitude and a red (dark) spot represents "sitting" attitude.

Foreground regions whose size is below a certain threshold are rejected as false positives.

After the foreground has been detected, pixel distribution values are updated in order to consider the case of slowly changing background. The Gaussian distribution is updated for each color component as follows:

$$\mu \leftarrow \alpha x + (1-\alpha)\mu$$

$$\sigma^2 \leftarrow \max\left(\sigma^2_{\min}, \alpha(x-\mu)^2 + (1-\alpha)\sigma^2\right)$$

where α is the learning rate and $\sigma_{\min}$ is introduced to prevent the standard deviation from decreasing below a threshold value if the background remains constant over a long period of time.

The background model is kept in mosaic space, allowing the system to detect new objects appearing anywhere in the scene, even if it happens outside the region of interest (the view of the Virtual Camera). This is a clear advantage of using an electronic pan-tilt-zoom system instead of a standard mobile camera platform.

2.3. Tracking

The tracking component has two roles: to select the target from the objects detected in the previous step, and to follow the target motion, keeping it permanently within the region of interest. At this stage, the selection process is a choice based on the size, height or color of the detected sprites.

As a target is selected, a pan/tilt command is issued for the Virtual Camera, so that it redirects the region of interest. The visual effect on screen is similar to that of a mobile camera tracking the moving object, although no mechanical movement is involved. By recording previous target positions, we adjust the speed of Virtual Camera according to the target speed, to ensure smooth camera movement. A main advantage of the electronic pan-tilt-zoom camera becomes apparent because the region of interest is instantly redirected, with no danger of losing the target.

2.4. Interpretation

From the previous modules we obtain a description of the scene in terms of planar layers (geometry and texture). These models are then augmented with semantic information by the interpretation module. Currently, this component is able to determine the 3D trajectories of the selected target in the room and to detect simple events such as a person standing up or sitting down on a chair.

The position of the person in the room is computed as follows: by knowing the floor position and detecting the person's head we determine the person's height and thus the head position in 3D. The trajectory is then built by tracking the head, assuming that it will never be occluded. As we retrieve the position in the room based on the person's height, a simple event such as sitting on

Table 1. Performance of camera systems

	Region of Awareness	Region of Interest			Background Model	Depth Range	Cost
		Quality	Precision	Speed			
Mobile PTZ Camera	Small	High	Low	Slow	Bad	Good	High
Wide Angle Lens Camera	Large	Low	High	Fast	Bad	Good	High
Electronic PTZ Camera	Large	High	High	Fast	Good	Limited	Low

a chair is recognized by a sudden change in location. The results are shown in Figure 4.

We plan to extend the interpretation module to handle more complex demonstration scenarios [8, 9].

3. Discussion

Using the criteria defined above we compare our approach with two other current solutions for acquiring video input in an Intelligent Room system. The discussion is summarized in Table 1.

Mobile pan-tilt-zoom cameras. They represent a *mechanical solution*, where the desired region of interest is observed on demand only, by mechanical movement of the camera. Potentially, such a system could observe a wide area overall, but at a given moment (and pan-tilt orientation) its visual awareness is confined to a much smaller area. Therefore, for a mobile pan-tilt-zoom camera system the region of awareness is the same as the region of interest, which is not wide enough to satisfy our criteria.

Since the entire camera resolution is dedicated for observing the region of interest, its quality is high. Not as good are the precision of location and speed of redirection. When receiving a pan-tilt-zoom command, the response has a certain delay until the actual movement of the camera is performed. In addition, it is never possible to know exactly how the camera is oriented, or to accurately position the camera at a precise given orientation.

This imprecision in location is a major problem when trying to maintain a background model for the entire scene. When a new frame is used to update the background, the registration errors lead to large differences between current and previous pixels, hence false positives in the detected foreground.

There is no inherent limitation in the range of depths for this type of camera system. Nevertheless, because of their mechanical components, they are quite expensive and not very robust, being subject to wear with time.

An extension to this approach is to use a set of mobile cameras (placed in different corners of the room) [10]. However, such a configuration involves complex protocols to coordinate the camera movements, in order to follow a moving target.

Wide-angle lens cameras. This is an *optical solution*, where a special lens – also known as “fish-eye lens” – is used to capture an extremely large field of view (almost 180°), thus having a good region of awareness [11].

Since this is possible only by distorting the image around the periphery of the field of view, the spatial resolution varies around the optical axis. The way to use such a system is to extract a smaller region of interest and unwarp it, but this leads to a region of interest with poor quality, due to the significant loss in spatial resolution. However, the precision of location and speed of redirection are better than those for mobile pan-tilt-zoom cameras, because no mechanical movement is involved.

Although a full background model can be maintained for the whole region of awareness, the variations in spatial resolution have a major impact on its quality.

Such cameras offer a good depth range, but as in the case of mobile pan-tilt-zoom systems, the wide-angle lenses are an expensive solution.

Electronic pan-tilt-zoom cameras. Such a system, described in Section 2, acquires a wide overall field of view, its region of awareness being maintained as a two-dimensional mosaic [12, 13]. The system is functionally equivalent to a mobile camera platform, but it performs its pan-tilt-zoom operations electronically rather than mechanically, so it can be considered a *digital solution*. The region of interest has good quality, its resolution being similar to the resolution of any one camera in the array. Because the region of interest is digitally extracted, it can be redirected rapidly and precisely.

We showed that the background model can be successfully maintained over the whole region of awareness. Once the cameras have been calibrated, the position of each frame inside the mosaic never changes and is known precisely, so there are no false foreground positives due to misalignments. There is also enough resolution for maintaining a good background model.

In terms of depth range, the electronic pan-tilt-zoom system is slightly inferior to the other two approaches. The cameras can never be arranged so that their optical axes converge, so 3D effects are bound to occur. Typically, there is a minimum working distance of about one or two meters. Outside this range, images appear seamlessly aligned, while objects closer than the working distance appear blurred. However, when used for an Intelligent Room system, the camera system is usually placed farther that this (for example, in a room corner), so this problem is less important.

Finally, such a system is significantly cheaper, especially when it is implemented with off-the-shelf CMOS cameras.

After evaluating these three camera systems based on the criteria previously defined, we conclude that electronic pan-tilt-zoom cameras represent the most appropriate solution for acquiring the visual input in an Intelligent Room application.

4. Conclusions and future work

In the context of Intelligent Room systems, we have studied the requirements imposed on their video input. We established six qualitative criteria and analyzed three types of camera systems as potential solutions. The electronic pan-tilt-zoom camera array is shown to perform best with respect with these criteria. Its main advantages are:

- Captures an overall field of view as large as the one offered by the other systems, with enough resolution for focusing on a certain region of interest.
- Is functionally equivalent to a mobile pan-tilt-zoom camera platform, but performs its pan-tilt-zoom operations electronically.
- Has a better precision and response time in redirecting the region of interest.
- Is cheaper and more robust, because involves a digital solution, instead of using expensive and fragile mechanical or optical components.

To support our claim, we developed GlobeAll, a four-component prototype for a vision-based Intelligent Room system that uses panoramic video input through an electronic pan-tilt-zoom camera array.

We plan to extend the tracking module by developing a more complex strategy for target selection, ensuring that the same target is selected across the sequence of frames. A more appropriate criterion here would be the color or texture of clothes. Once all moving objects are reliably identified from a frame to the next one, then we can even allocate a separate Virtual Camera to independently follow each target, since there is no limitation on the number of Virtual Cameras we generate. Future developments also include adding a second camera system for stereo reconstruction of 3D models, using the HSV color space instead of RGB for the background model, and more sophisticated interpretation of complex scenarios.

Acknowledgements

This research has been funded in part by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, Cooperative Agreement No. EEC-9529152, with additional support from the Annenberg Center for Communication at the University of Southern California and the California Trade and Commerce Agency. The support of the Philips Multimedia Center is also gratefully acknowledged.

References

[1] M. Coen, “Design Principles for Intelligent Environments”, *Proc. AAAI-98*, Madison, WI, July 1998, pp. 547-554.

[2] M. Turk, “Moving from GUIs to PUIs”, *Proc. Symposium on Intelligent Information Media*, Tokyo, December 1998.

[3] R. Szeliski, “Video mosaics for virtual environments”, *IEEE Computer Graphics and Applications*, March 1996, 16(2), pp. 22-30.

[4] H. Sawhney, R. Kumar, “True Multi-Image Alignment and its Application to Mosaicing and Lens Distortion Correction”, *IEEE Trans. on PAMI*, 1997, vol.21: 3, pp. 235-243.

[5] Y. Xiong, K. Turkowski, “Registration, Calibration and Blending in Creating High Quality Panoramas”, *Proc. WACV-98*, 1998, pp. 69-74.

[6] A. François, G. Medioni, “Adaptive Color Background Modeling for Real-Time Segmentation of Video Streams”, *Proc. of the ICISST*, Las Vegas, NV, June 1999, 227-232.

[7] K. Toyama, J. Krumm, B. Brumitt, B. Meyers, “Wallflower: Principles and Practice of Background Maintenance”, *Proc. ICCV-99*, Corfu, Greece, September 1999, pp. 255-261.

[8] G. Medioni, R. Nevatia, I. Cohen, “Event Detection and Analysis from Video Streams”, *DARPA IUW-98*, Monterey, CA, November 1998, pp. 63-72.

[9] C. Wren, A. Azarbayejani, T. Darell, A. Pentland, “Pfinder: Real-Time Tracking of the Human Body”, *PAMI*, vol. 19, no. 7, pp. 780-785, July 1997.

[10] T. Matsuyama, “Cooperative Distributed Vision”, *DARPA-98*, pp. 365-384.

[11] Interactive Pictures Corporation, http://www.ipix.com.

[12] R. Swaminathan, S. Nayar, “Polycameras: Camera Clusters for Wide Angle Imaging”, *Technical Report CUCS-013-99*, Columbia University, April 1999.

[13] V. Nalwa, “A True Omnidirectional Viewer”, *Technical Report*, Bell Laboratories, February 1996.