IEEE Conference on Computer
Vision and Pattern Recognition,
Santa Barbara, California,
June 1998

Elliptical Head Tracking Using Intensity Gradients and Color Histograms

Stan Birchfield
Computer Science Department
Stanford University
Stanford, CA 94305
birchfield@cs.stanford.edu

Abstract

An algorithm for tracking a person's head is presented. The head's projection onto the image plane is modeled as an ellipse whose position and size are continually updated by a local search combining the output of a module concentrating on the intensity gradient around the ellipse's perimeter with that of another module focusing on the color histogram of the ellipse's interior. Since these two modules have roughly orthogonal failure modes, they serve to complement one another. The result is a robust, real-time system that is able to track a person's head with enough accuracy to automatically control the camera's pan, tilt, and zoom in order to keep the person centered in the field of view at a desired size. Extensive experimentation shows the algorithm's robustness with respect to full 360-degree out-of-plane rotation, up to 90-degree tilting, severe but brief occlusion, arbitrary camera movement, and multiple moving people in the background.

1 Introduction

Robust, reliable visual tracking of an object in a complex environment will require the integration of several different visual modules, each using a different criterion and each employing different assumptions about the incoming images. The modules must be selected so that their assumptions are, as much as possible, orthogonal to each other so that when one module fails the other one can come to its aid.

According to elementary set theory, every closed set in the plane can be decomposed into two disjoint sets: the boundary and the interior [8]. Since these two sets are complementary (in the true, mathematical sense), it stands to reason that the failure modes of a tracking module focusing on the object's boundary will be orthogonal to those of a module focusing on the object's interior.

In this paper, we present a method for object tracking that combines the output of two different modules: one that matches the intensity gradients along the object's boundary and one that matches the color histogram of the object's interior. The present work applies the method to tracking a person's head, primarily because of the number of applications that could benefit from such a system, such as video conferencing, distance learning, automatic video analysis, and surveillance. Moreover, the head is well approximated by a simple two-dimensional model, namely an ellipse, thus simplifying the present investigation.

Despite their complementarity, the gradient and color modules operate in a symmetric fashion, thus making the combination step trivial and obviating the need for complicated sensor fusion techniques. The result is a robust tracker that is accurate enough to actively control the camera's pan, tilt, and zoom for long periods of time in order to keep the subject centered in the field of view at a desired size. The algorithm is insensitive to out-of-plane rotation, tilting, severe but brief occlusion, arbitrary camera movement, and multiple moving people in the background.

2 Searching for the Head

Assume that we have an estimate of the position (x, y) and size σ of the head in an image. The head is modeled as a vertical ellipse with a fixed aspect ratio of 1.2, so that (x, y) is the center of the ellipse and σ is the length of the minor axis. We will use the notation $\mathbf{s} = (x, y, \sigma)$ for the head's state or location.[1] The tracking task is to update the state by finding the location whose image values best match the values in the model. This is accomplished via a hypothesize-and-test procedure [4, 7] in which the goodness of the match is dependent upon the intensity gradients around the object's boundary and the color histogram of the object's interior:

$$\mathbf{s}^* = \arg\max_{\mathbf{s}_i \in S} \{\bar{\phi}_g(\mathbf{s}_i) + \bar{\phi}_c(\mathbf{s}_i)\}, \tag{1}$$

*This research was sponsored by Autodesk.

[1] Throughout this paper, the term *position* refers to (x, y), while *location* or *state* refers to (x, y, σ).

where $\bar{\phi}_g(\mathbf{s}_i)$ and $\bar{\phi}_c(\mathbf{s}_i)$ are the matching scores based on intensity gradients and color histograms, respectively.

The search space S is the set of all states within some range of the predicted location, using velocity prediction [2]. Somewhat surprisingly, this simple prediction scheme greatly improves the behavior of the tracker because it removes any restriction on the maximum lateral velocity of the subject — only the amount of acceleration is limited.

We now examine the gradient and color modules, in turn.

3 Gradient Module

Perhaps the most natural way to measure the goodness of match around the object's boundary is to compute the normalized sum of the gradient magnitude around the perimeter of the ellipse:

$$\phi_g(\mathbf{s}) = \frac{1}{N_\sigma} \sum_{i=1}^{N_\sigma} |\mathbf{g}_\mathbf{s}(i)|, \qquad (2)$$

where $\mathbf{g}_\mathbf{s}(i)$ is the intensity gradient at perimeter pixel i of the ellipse at location **s**, and N_σ is the number of pixels on the perimeter of an ellipse with size σ.

Except for the fixed shape of the object's perimeter, the above formulation is nearly identical to that employed by most contour trackers [1, 3]. One minor difference is that the gradient is summed around the entire perimeter rather than just at select points. A more significant difference is that the current hypothesize-and-test paradigm [4, 7] allows all of the data to be examined before a decision is made, in contrast to the typical contour tracker in which each control point independently decides how to move based on purely local information.

A more sophisticated measure than the one in (2) is the one proposed by Nishihara [13]. Rather than just desiring large gradient magnitudes around the perimeter, it also desires the gradient direction to be perpendicular to the perimeter:

$$\phi_g(\mathbf{s}) = \frac{1}{N_\sigma} \sum_{i=1}^{N_\sigma} |\mathbf{n}_\sigma(i) \cdot \mathbf{g}_\mathbf{s}(i)|, \qquad (3)$$

where $\mathbf{n}_\sigma(i)$ is the unit vector normal to the ellipse at pixel i and $(\cdot)$ denotes the dot product. The gradient magnitude still plays a part here since the gradient is unnormalized (Our experiments have indicated that normalization greatly increases sensitivity to image noise).

In practice, the performance of the gradient magnitude module is inferior to that of the gradient dot product module. Therefore, the term *gradient module* will hereafter refer to the latter, unless specifically stated otherwise.

To facilitate adding the gradient score to the color score, the former is converted to a percentage by subtracting the minimum and dividing by the range:

$$\bar{\phi}_g(\mathbf{s}) = \frac{\phi_g(\mathbf{s}) - \min_{\mathbf{s}_i \in S} \phi_g(\mathbf{s}_i)}{\max_{\mathbf{s}_i \in S} \phi_g(\mathbf{s}_i) - \min_{\mathbf{s}_i \in S} \phi_g(\mathbf{s}_i)}.$$

4 Color Module

Many researchers have exploited the relative uniqueness of skin color to track faces [4, 5, 9, 15, 16]. A weakness of these systems is their heavy reliance upon skin color that forbids skin-colored objects in the background and, more importantly, forbids the subject from turning around so that the back of his head, rather than his face, is visible. The color of human heads is complex, however, being at the very least bimodal due to the skin and hair, and any system attempting to handle out-of-plane rotation must address this issue.

The color histogram [17] is well suited to this task because of its ability to implicitly capture complex, multimodal patterns of color. Moreover, because it disregards all geometric information, it remains relatively invariant to many complicated, non-rigid motions.

The procedure is as follows. Off-line, the subject presents a three-quarters view to the camera in order to capture both face and hair, and a model histogram is constructed by counting the pixels inside the ellipse (the ellipse can be either manually placed or automatically placed via the gradient module, either of which takes about one to thirty seconds of user time). Then, at run time, the histogram intersection [17] is computed between the model histogram M and the image histogram I at each hypothesized location:[2]

$$\phi_c(\mathbf{s}) = \frac{\sum_{i=1}^{N} \min(I_\mathbf{s}(i), M(i))}{\sum_{i=1}^{N} I_\mathbf{s}(i)},$$

where $I_\mathbf{s}(i)$ and $M(i)$ are the numbers of pixels in the ith bin of the histograms, and N is the number of bins.

The power of histogram intersection results from the $\min()$ function, which matches no more image pixels of a certain color than are present in the model histogram. Thus, for example, the measure is more satisfied with a region containing both facial and hair color than a region containing all facial color.

Our color space consists of scaled versions of the three axes $B-G$, $G-R$, and $B+G+R$. The first two contain the chrominance information and are sampled into eight bins

[2] This equation is identical to the one in [17]. At first glance the denominator may look different, but this is because our goal is to match a single model to the best image patch, rather than to match a single image patch to the best model.

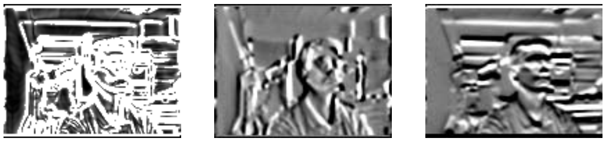


Figure 1: A cluttered background. (a) Gradient magnitude. (b) Horizontal and (c) vertical components of gradient.

each, while the last one contains the luminance information and is sampled more coarsely into four bins [17]. Some researchers have ignored luminance information completely [4, 9], but this is dangerous with out-of-plane rotation because, based on chrominance alone, dark brown hair looks similar to a white wall.

As in the case of the gradient scores, the color scores are converted to percentages:

$$\bar{\phi}_c(\mathbf{s}) = \frac{\phi_c(\mathbf{s}) - \min_{\mathbf{s}_i \in S} \phi_c(\mathbf{s}_i)}{\max_{\mathbf{s}_i \in S} \phi_c(\mathbf{s}_i) - \min_{\mathbf{s}_i \in S} \phi_c(\mathbf{s}_i)}.$$

5 Experimental Results

In this section we examine the performance of the individual modules, the ways in which they complement each other, and the robust behavior achieved with the complete system.

5.1 Gradient module alone

It is somewhat surprising that the simple gradient module is sufficient to control the camera's pan and tilt in order to track a person walking around an untextured, unmodified room [2]. Even in the rather cluttered environment shown in Figure 1, the gradient module was able to consistently track the subject's slowly-moving head for about fifty pixels or so of image motion before becoming distracted by the background (The gradient magnitude performed slightly worse than the gradient dot product).

However, in cluttered environments the gradient module fails too often for reliable tracking. Moreover, even in untextured environments the module is unable to control the camera's zoom because the ellipse tends to become attracted to gradients inside the face. Finally, the gradient score function has a small basin of attraction that prevents large accelerations (see below).

5.2 Color module alone

In nearly every respect, the performance of the color module is superior to that of the gradient module, which should not be too surprising since it looks at more pixels. Alone, this module is capable of controlling the camera's pan, tilt, and zoom in order to track a person in an unmodified environment, even when there are skin-colored objects in the background. For example, the subject was able to move in front of the board shown in Figure 2 without causing the tracker to become lost. Although the ellipse's location became unstable while the subject was in front of the board, yet it remained on the subject because of the non-skin pixels such as the hair, eyes, and mouth. When the subject subsequently moved away from the board, the ellipse's location quickly stabilized onto the head.

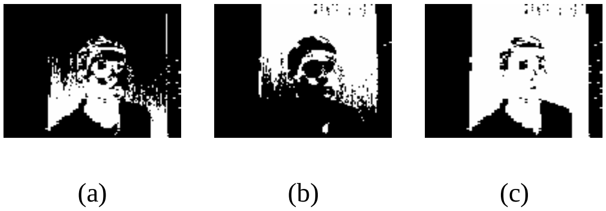


Figure 2: The subject in front of a skin-colored board. All the white pixels in (a) have the same quantized color, and similarly for (b). The logical OR is shown in (c).

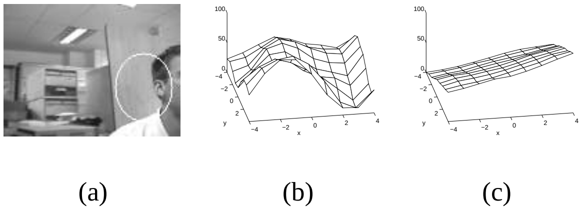


Figure 3: (a) A situation in which the ellipse was far from the true solution. (b,c) The matching scores as a function of x and y at a particular scale for the gradient and color modules, respectively. Looking at the maximum, we see the former incorrectly pulling the ellipse left but the latter correctly pulling to the right.

Not only can the color module control zoom and handle background clutter, it has the added advantage of a large basin of attraction. For example, Figure 3 shows a situation in which the ellipse was barely hanging on to the head because of the subject's quick acceleration and the camera's slow dynamics (notice the skin-colored board behind the subject). Although the gradient module was distracted by the background and tried to pull the ellipse to the left, the color module correctly pulled to the right, even though a large percentage of the ellipse's interior contained the potentially distracting skin-colored board.

5.3 Module complementarity

It is therefore clear that the color module greatly helps the gradient module by ignoring background clutter, correctly

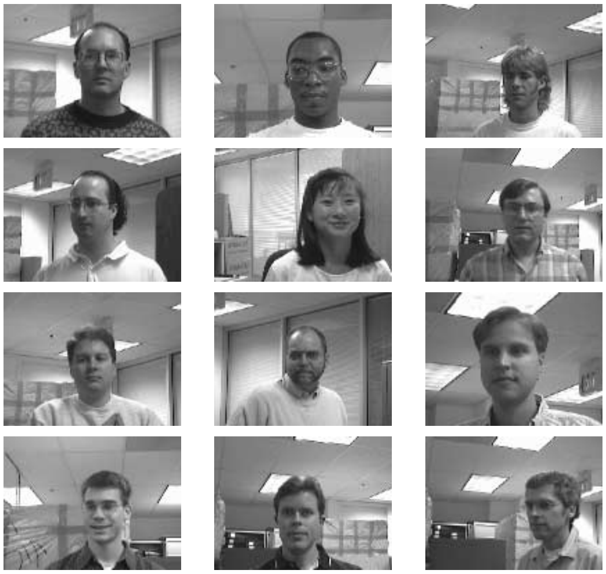


Figure 4: People in the experiments.

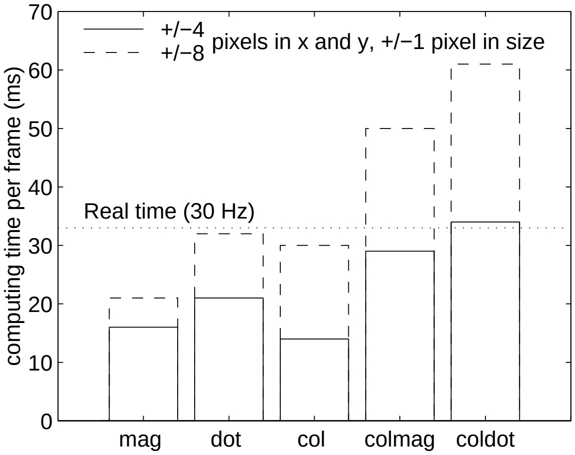


Figure 5: From left to right, the computing time for the gradient magnitude, gradient dot product, color, color and gradient magnitude, and color and gradient dot product modules, for two different search ranges.

handling changes in scale, and providing a larger basin of attraction. In a similar manner, sometimes the gradient module helps the color module.

One situation occurred when the subject turned his back toward the camera and moved farther or closer. Left to itself, the color module had difficulty finding the correct scale when the face was thus invisible; but with the gradient module, the tracker was able to succeed. In another scenario, the subject moved away from the camera while the skin-colored board remained behind him. Although the color module did not lose the position of the subject due to his hair, it was unable to properly scale the head since the board looked like skin. By adding the gradient module, the ellipse correctly scaled to the subject's head and the camera zoomed in. Finally, the color module was found to occasionally slip down to the subject's neck, a problem that was solved by adding the gradient module, which tended to get a strong response from the outline of the top of the head.

5.4 Demonstrations of robust performance

To demonstrate the tracker's robust behavior in various situations, it was tested on twelve people with a wide variety of facial complexion, hair color, amount of hair, head shape, and type and color of shirt, as shown in Figure 4. Several of the people wore spectacles, and one of them had a beard. Due to hardware restrictions, all the testing was performed in the same environment.

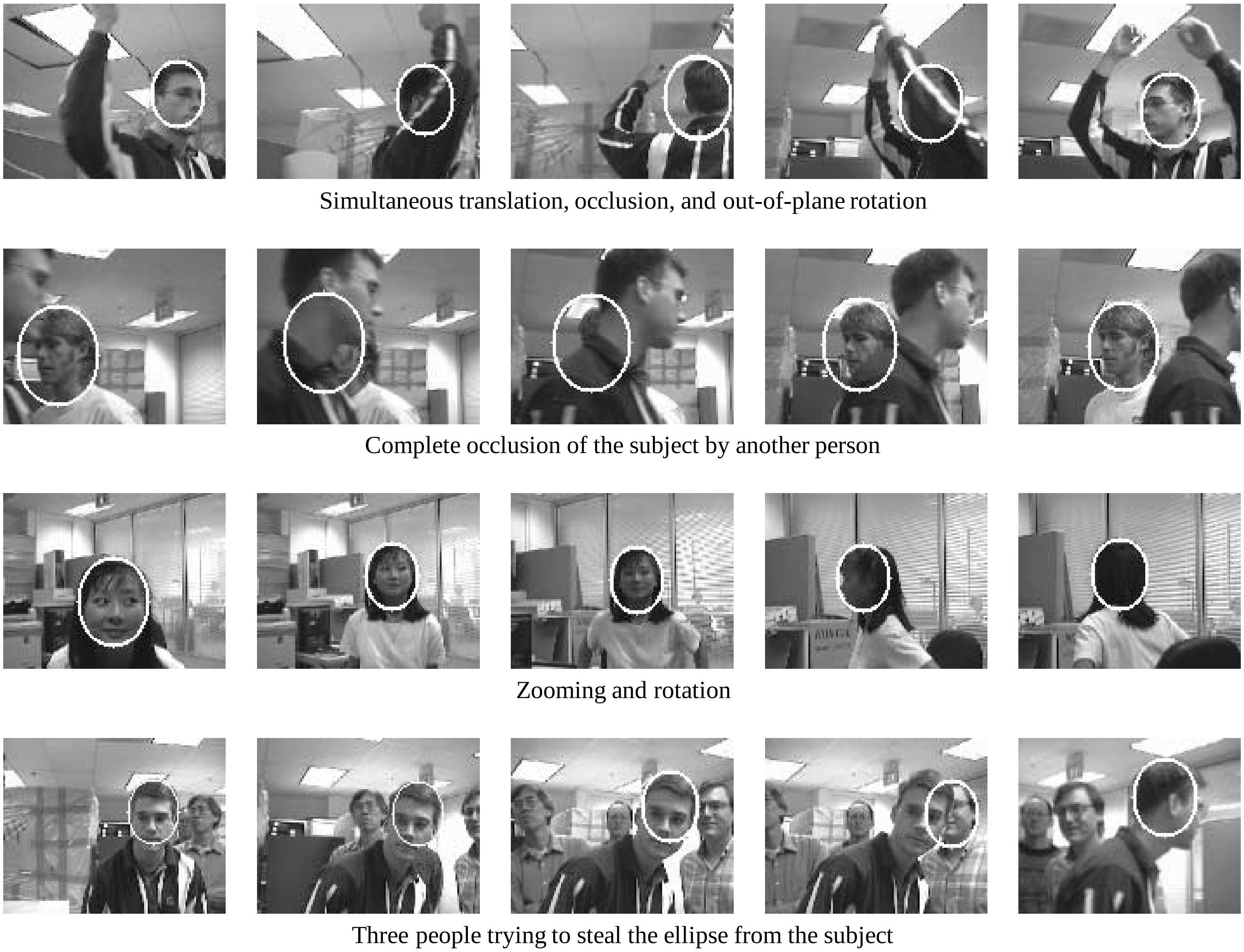
During the tests, the output of the tracker was used to automatically control the camera's pan, tilt, and zoom to keep the subject centered in the field of view at a desired size. A few snapshots from the various video clips are shown in Figure 6. The tracker was able to handle full 360-degree rotation and up to 90-degree tilting of the head, arbitrary camera motion, and severe occlusion (at least for short periods of time). With multiple people in the scene the tracker usually succeeded but was occasionally distracted when the two faces occupied adjacent regions in the image. For example, in the fourth row of the figure the ellipse temporarily preferred one of the background people but quickly returned to the subject when his continued motion caused the other person to be occluded. Had the subject changed direction at that point, the tracker probably would have lost him.

5.5 Computing times

Figure 5 shows the raw computing times using a 200 MHz Pentium Pro microprocessor. Notice that increasing the number of search locations by 257% caused only about an 80% increase in computing time, which indicates that the fixed time needed to convolve the image, transform the color space, and set up the search is much greater than the time needed to actually conduct the search.

6 Comparison with Previous Work

Compared with previous work, this tracker is the only one of which we are aware that can handle significant out-of-plane rotation, arbitrary camera motion, textured foregrounds and backgrounds, and multiple moving people in



Simultaneous translation, occlusion, and out-of-plane rotation

Complete occlusion of the subject by another person

Zooming and rotation

Three people trying to steal the ellipse from the subject

Figure 6: Demonstration of the tracker's performance in various situations. These and other MPEG sequences are available from `http://vision.stanford.edu/˜birch.`

the background, all simultaneously.

Template- and neural network-based trackers [6, 9, 11, 18], as well as trackers based on facial color [4, 5, 9, 15, 16, 18], cannot handle severe out-of-plane rotation because such a rotation causes the face to disappear. The color-based techniques also tend to have difficulty with skin-colored objects or other people in the background.

Trackers utilizing some form of background differencing [5, 10, 11, 12, 18, 19, 20] either require a static camera or restrict the camera's motion to rotation about its focal point.[3] Moreover, many of these techniques perform motion-based figure-ground segmentation, which tends to fail when the camera zooms or when multiple objects move in the scene.

Reliable tracking was reported by combining a template-based tracker with stereo depth [14]. However, besides the additional hardware, it is not clear whether this system would be able to handle multiple people at a similar depth as the subject.

Also, promising results have been achieved using a shape-based contour tracker [1] that is more sophisticated than ours because it allows the shape to deform over time. However, in its present implementation the tracking criterion is the gradient magnitude alone, which will probably fail with quick movements in cluttered scenes.

Finally, it must be mentioned that some of the systems cited above contain multiple modules. However, it is of-

[3] In [10], the camera may move occasionally but not continuously.

ten the case that one of the modules utilizes background-differencing and another uses facial color. Although the former can handle out-of-plane rotation and the latter can handle a dynamic camera, the system resulting from combining the two cannot handle both situations simultaneously.

7 Conclusion

In this paper we have presented a method for robustly tracking a person's head undergoing complex motions such as 360-degree out-of-plane rotation, severe occlusion, and scale changes in front of a dynamic, unstructured background. Robustness is achieved using two orthogonal modules, one based on the intensity gradient around the head's perimeter and another based on the color histogram of the head's interior.

One limitation of the current work is that the color histogram of the model is not adaptive. Therefore, changing lighting conditions or automatic gain adjustments by the camera will cause the color module to become confused. To solve this problem, the histogram must be able to quickly update itself because conditions can change quickly (imagine the subject walking in front of a window through which sunlight shines), but the ellipse does not always provide a good segmentation of the head at each frame, thus preventing the update routine from completely trusting the pixels inside the ellipse. An additional module is needed.

Another problem with the current system is that the head's acceleration is fairly limited when the background is confusing, but this is really a limit of the 30 Hz NTSC video signal and the speed of the computer than it is a limitation of the algorithm. A higher temporal sampling rate, coupled with a slightly faster machine, should cause a noticeable improvement in this area.

Acknowledgments

This work was largely inspired by several discussions I had with J. Brian Burns. Many thanks to him and to the subjects who willingly gave their time. S.D.G.

References

[1] A. M. Baumberg and D. C. Hogg. An efficient method for contour tracking using active shape models. In *Proceedings of the IEEE Workshop on Motion of Non-Rigid and Articulated Objects*, pages 194–199, 1994.

[2] S. Birchfield. An elliptical head tracker. In *Proc. of the 31st Asilomar Conf. on Signals, Systems and Computers*, 1997.

[3] A. Blake, R. Curwen, and A. Zisserman. A framework for spatiotemporal control in the tracking of visual contours. *Intl. Journal of Computer Vision*, 11(2):127–145, 1993.

[4] P. Fieguth and D. Terzopoulos. Color-based tracking of heads and other mobile objects at video frame rates. In *Proc. of the IEEE CVPR*, pages 21–27, 1997.

[5] H. P. Graf, E. Cosatto, D. Gibbon, M. Kocheisen, and E. Petajan. Multi-modal system for locating heads and faces. In *Proc. of the Second Intl. Conference on Automatic Face and Gesture Recognition*, pages 88–93, 1996.

[6] G. D. Hager and P. N. Belhumeur. Real-time tracking of image regions with changes in geometry and illumination. In *Proc. of the IEEE CVPR*, pages 403–410, 1996.

[7] R. M. Haralick and L. G. Shapiro. *Computer and Robot Vision*, volume 2. Reading, Mass.: Addison-Wesley, 1993.

[8] F. Hausdorff. *Set Theory*. New York: Chelsea Publishing Company, third edition, 1978.

[9] M. Hunke and A. Waibel. Face locating and tracking for human-computer interaction. In *Proc. of the 28th Asilomar Conf. on Signals, Systems and Computers*, pages 1277–1281, 1994.

[10] D. P. Huttenlocher, J. J. Noh, and W. J. Rucklidge. Tracking non-rigid objects in complex scenes. In *Proc. of the 4th Intl. Conference on Computer Vision*, pages 93–101, 1993.

[11] S. McKenna and S. Gong. Tracking faces. In *Proc. of the Second International Conference on Automatic Face and Gesture Recognition*, pages 271–276, 1996.

[12] D. Murray and A. Basu. Motion tracking with an active camera. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(5):449–459, May 1994.

[13] H. K. Nishihara. Personal Communication, 1996.

[14] H. K. Nishihara, H. J. Thomas, and E. Huber. Real-time tracking of people using stereo and motion. In *SPIE Proceedings*, volume 2183, pages 266–273, 1994.

[15] Y. Raja, S. J. McKenna, and S. Gong. Segmentation and tracking using colour mixture models. In *Proceedings of the 3rd Asian Conference on Computer Vision*, volume I, pages 607–614, 1998.

[16] K. Sobottka and I. Pitas. Segmentation and tracking of faces in color images. In *Proc. of the Second Intl. Conf. on Automatic Face and Gesture Recognition*, pages 236–241, 1996.

[17] M. Swain and D. Ballard. Color indexing. *International Journal of Computer Vision*, 7(1):11–32, 1991.

[18] K. Toyama and G. D. Hager. Incremental focus of attention for robust visual tracking. In *CVPR*, pages 189–195, 1996.

[19] J. I. Woodfill. *Motion Vision and Tracking for Robots in Dynamic, Unstructured Environments*. PhD thesis, Stanford University, 1992.

[20] C. R. Wren, A. Azarbayejani, T. Darrell, and A. P. Pentland. Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(7):780–785, 1997.