31st Asilomar Conference on
Signals, Systems, and
Computers, November 1997

An Elliptical Head Tracker *

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

Abstract

A simple algorithm for tracking a person's head is presented. A two-dimensional model, namely an ellipse, is used to approximate the head's contour. When a new image becomes available, a local search determines the position and size of the best ellipse by maximizing the normalized sum of the image gradient magnitude around the perimeter of the ellipse. The local search begins from a predicted position, using the head's velocity, which eliminates the tracker's dependence upon maximum velocity. The tracker operates at 30 Hz and actively controls camera pan and tilt in order to track a person moving in a real environment. The algorithm tolerates full 360-degree rotation of the body as well as moderate amounts of occlusion, and it performs reacquisition of the subject.

1 Introduction

Automatic visual tracking of a person in an unmodified environment is a promising goal for computer vision research, both for its usefulness and its feasibility. As to the former, such a system could easily find its way into commercial applications, such as video conferencing, automatic surveillance, and distance learning. As to the latter, recent research has shown great progress in developing systems that are capable of performing this task [1, 4, 6, 7, 9, 10, 11, 13, 14, 15, 16, 17].

Despite the recent progress, however, at least two difficulties continue to impede the development of a robust, reliable tracker. First, many trackers assume that the subject does not perform any out-of-plane rotation (rotation about an axis parallel to the image plane) and fail when this assumption does not hold. In particular, template-based motion trackers [7] are notorious for sliding off an object undergoing this type of rotation because parts of the object disappear while other parts reappear (The problem of sliding can be alleviated by using additional information, such as stereo depth [13]). Similarly, trackers relying primarily on facial color [4, 9, 14] fail when the subject turns so that the back of his head, rather than his face, is visible. To solve this latter problem, the bimodality of head color (that is, face and hair) must be addressed.

A second difficulty is that the common and effective technique of background differencing [6, 10, 17] does not allow for the flexibility of a moving camera,[1] which is necessary for many applications. Although related techniques [11, 16] allow the camera to move, they restrict its motion to rotation about the focal point and, in addition, require either background texture [16] or accurate, synchronized position feedback [11].

In this paper, we present a real-time algorithm to track a person's head using a rigid, two-dimensional model (namely, an ellipse) to approximate the projection of the head's contour in the image. Despite its simplicity, this algorithm is able to track a person in some unmodified environments with enough accuracy to actively control the camera's pan and tilt in order to keep the subject centered in the field of view. Although the algorithm performs a local search, the subject's velocity is not restricted, due to a simple prediction scheme. More importantly, because the computation does not use any features inside the face, rotation about any axis causes no problem, even full 360-degree turning of the body. In addition, the algorithm is insensitive to small amounts of occlusion, performs reacquisition of the subject, and requires no training for new subjects.

2 Motivation

A person's head has some interesting properties that make it an intriguing focus for research on tracking. First, it is the most distinguishing feature of the body, making it

*This research was sponsored by Autodesk and the Advanced Research Projects Agency's Real-Time Planning and Control program under Contract DACA76-93-C-0017.

[1] In [10], the camera can move occasionally but not continuously.

©*1997 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.*

the sole concern of applications such as face recognition and making its appearance in the image almost mandatory (people usually complain when their heads are cropped out of pictures) [5]. Secondly, the head has limited motion with respect to the torso, so that knowledge of the head's location is often sufficient for framing the torso and arms in an image and is also a good starting point for more complex algorithms attempting to track other body parts. Finally, the shape of the head in an image is well approximated by a rigid, two-dimensional model as the result of two convenient geometrical properties: (1) the head is nearly rigid, and (2) the head is roughly symmetrical about an axis parallel to the image plane.

3 Finding the head

The head is modelled by an ellipse with a fixed vertical orientation and a fixed aspect ratio of 1.2.[2] Each time a new image becomes available, the ellipse's state $\mathbf{s} = (x, y, \sigma)$, where (x, y) is the position and σ is the size (length of the minor axis), is maintained by performing a local search to maximize the normalized sum of the gradient magnitude around the perimeter of the ellipse:

$$\mathbf{s}^* = \arg\max_{\mathbf{s} \in S} \left\{ \frac{1}{N_\sigma} \sum_{i=1}^{N_\sigma} |\mathbf{g}_i| \right\}, \qquad (1)$$

where $\mathbf{g}_i$ is the gradient at perimeter pixel i (the dependence of $\mathbf{g}_i$ on $\mathbf{s}$ is implicit), and N_σ is the number of pixels on the perimeter.

The search space S is the set of all states within some range of the predicted location:

$$S = \{\mathbf{s} : |x - x^p| \leq x_r, |y - y^p| \leq y_r, |\sigma - \sigma^p| \leq \sigma_r\},$$

where in our implementation the search range is $x_r = y_r = 8$ pixels and $\sigma_r = 1$ pixel. The predicted position (x^p, y^p) in frame t arises from the assumption of constant velocity [2], using the positions found in the previous two frames:[3]

$$\begin{aligned} x_t^p &= 2x_{t-1}^* - x_{t-2}^* \\ y_t^p &= 2y_{t-1}^* - y_{t-2}^* \\ \sigma_t^p &= \sigma_{t-1}^*. \end{aligned}$$

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. A Kalman filter [1, 3] might improve results even more but has not been tried.

[2] Since the gradients around the chin tend to be small, the particular choice of aspect ratio seems unimportant. Others [5] have advocated using the golden ratio (approximately 1.6).

[3] We have not found it necessary to predict the size.

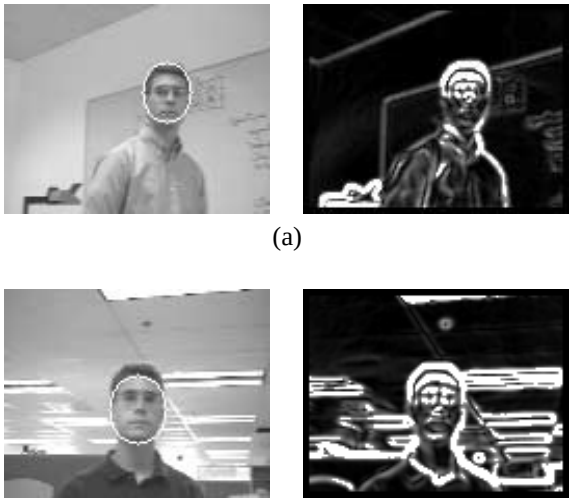


Figure 1. The two environments: (a) the "untextured" room, and (b) the "textured" room.

To make the computation less attracted to strong background gradients, all values of the gradient magnitude that are above a certain threshold are mapped to the maximum value. As an alternative to using the gradient magnitude, Nishihara [12] has suggested using the square of the dot product of the gradient vector and ellipse normal, which would reduce the attraction to background gradients and eliminate the need for a threshold.

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, 8] 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.

4 Experimental results

To demonstrate the performance of the tracker, a person was tracked for about thirty seconds in each of two different environments, shown in Figure 1. The first sequence was taken in the "untextured" room, which consisted mostly of a whiteboard that, although filled with writing, caused only weak gradients in the image. In contrast, the second sequence was taken in the "textured" room, in which the ceiling lights caused strong gradients. We now examine some excerpts from these sequences, shown in Figure 3, in

detail:

(a) *Occlusion.* The tracker handled occlusion, as long as neither the occluding object nor the background looked like an ellipse. In this example, the tracker maintained its fixation on the subject's head while the subject waved his arm in front, even though at one point the arm nearly completely covered the head.

(b) *Rotation.* The tracker was not confused by full 360-degree out-of-plane rotation.

(c) *Tilting.* Although the orientation of the ellipse was fixed as vertical, the tracker remained fixated on the subject's head when it was tilted sideways.

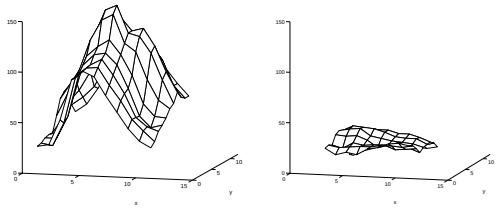
(d) *Reacquisition.* The tracker exhibited an uncanny ability to reacquire the subject when he returned to the camera's field of view. By frame 933, the ellipse had been stuck on the background for an extended period of time, and when the subject reappeared the ellipse slid down to lock onto his head. This behavior is remarkable when one considers that, at the time of the reacquisition, the center of the subject's head was still 30 pixels away from the ellipse's original center (Recall that the search range was only ± 8 pixels).

(e) *Scaling.* As the subject walked closer to the camera, the size of the ellipse grew. (The subject's head in the last frame was cropped because the camera's tilt limit had been reached.) However, it was difficult to repeat this behavior because the ellipse tended to be attracted to gradients within the head, and therefore our system was unable to control zoom reliably.

(f) *Textured background.* The tracker was also successful in the textured room.

In general, we have found the tracker to be successful in untextured environments with subjects whose hair color is different from the background. Surprisingly, the tracker worked fine on a woman who had long black hair, in which case the ellipse enlarged itself to trace the outline of her hair rather than her face. However, the tracker was less successful with balding subjects of light complexion, whose head outline did not contain strong gradients. Even less promising were the results in highly textured environments, such as a room full of stacked cardboard boxes in which the tracker failed because the corners of the boxes yielded strong gradients that were roughly elliptical.

The algorithm was implemented on a Hewlett-Packard personal computer equipped with a 133 MHz Pentium microprocessor. A Canon VC-C1 camera supplied the algorithm with 128×96 images every 33 milliseconds and received pan and tilt velocity commands at the same rate.



(a) (b)

Figure 2. The normalized sum of Equation (1) versus x and y for a particular choice of σ, in two different situations. (a) The ellipse is locked onto the subject's head in an unambiguous environment. (b) The tracker is in the process of drifting from the head to some gradients in the background.

5 Measuring confidence

When the head tracker determines the best state, it also computes a measure of confidence in that decision by examining the likelihood values of the states encountered during the search. Such a measure would be important if the head tracking module were embedded in a larger system because it would help to indicate when the module had failed and needed to be overridden by another module [6, 15].

Recall from Equation (1) that each state $\mathbf{s}$ in the search region has an associated likelihood value. To compute confidence, the curvature of each of the three likelihood surfaces (one at each scale) is approximated by counting the number of values that are at least a certain distance (in terms of value) away from the global maximum. For example, if the peak is sharp then many of the values will be far from the maximum, yielding a high curvature (Figure 2a). On the other hand, a broad peak will yield a low curvature because many values will be near the maximum (Figure 2b). The curvature of the three surfaces are averaged to produce a single confidence measure.

The time history of confidences for the two sequences of the previous section are shown in Figure 4. The subject was successfully tracked throughout the second sequence — notice that the confidences remained high, around 92 percent. On the other hand, in the first sequence the tracker lost the subject around frame 650 (due to the subject's large acceleration and the camera's poor mechanical response) and reacquired the subject around frame 945. Although the confidence correctly dipped when the subject was lost and returned to 92 percent after reacquisition, interpreting the confidence is not trivial. For example, the confidence also dipped around frame 560 even though the subject was still being tracked, because the subject's head had become

(a) Occlusion (0.3 sec) — Sequence I — frames 284 through 292

(b) Rotation (1.6 sec) — Sequence I — frames 360 through 408

(c) Tilting (1.1 sec) — Sequence I — frames 444 through 476

(d) Reacquisition (0.4 sec) — Sequence I — frames 933 through 945

(e) Scaling (1.5 sec) — Sequence I — frames 948 through 992

(f) Textured background (4.0 sec) — Sequence II — frames 110 through 230

Figure 3. Demonstration of the head tracker's performance under various conditions. The number in parentheses indicates the amount of elapsed time between the first and last frames of the row. These image sequences can be obtained from the World Wide Web at http://vision.stanford.edu/~birch.

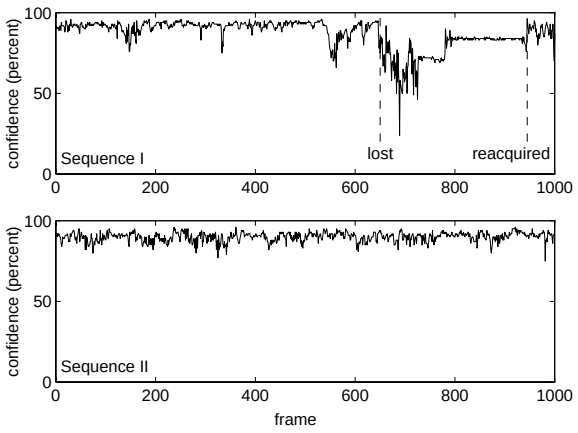


Figure 4. The confidence of the head tracker. TOP**: In Sequence I, the subject is lost around frame 650 and reacquired around frame 945.** BOTTOM**: The subject is successfully tracked throughout Sequence II.**

smaller than the smallest allowable ellipse, leading to an unstable position. In addition, even though the subject was still lost between frames 800 and 945, the confidence remained a stable 84 percent because the tracker had settled onto a good location of the background. This latter example demonstrates that, although a low confidence can signify that the subject is being lost, a high confidence means little unless it is known that the subject has not already been lost.

6 Conclusion

This paper has presented a simple head tracker that is able to follow a person in some unmodified environments. The tracker overcomes several common problems, including full body rotation, occlusion, and reacquisition. However, the tracker is heavily dependent on the single assumption that intensity gradients outline the subject's head (Note that other trackers exhibit a similar dependence [1, 3]). Therefore, it is too fragile to be used alone and must be augmented by other techniques in order to provide a robust, general tracking system. Its low overhead and ability to measure confidence make it a viable candidate as one module in such a system.

Acknowledgements

Thanks to all the people at Autodesk who made this work possible, especially Dan Perrin who helped me learn a new computer environment, and Rick Marks who wrote some of the low level code. Also thanks to Carlo Tomasi for his useful comments on the writing of this paper. 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] M. J. Black. Recursive non-linear estimation of discontinuous flow fields. In *Proceedings of the 3rd European Conference on Computer Vision*, pages 138–145, 1994.

[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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–27, 1997.

[5] V. Govindaraju, S. N. Srihari, and D. Sher. A computational model for face location based on cognitive principles. In *Proceedings of the 10th National Conference on Artificial Intelligence(AAAI-92)*, volume 1, pages 350–355, 1992.

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

[7] G. D. Hager and P. N. Belhumeur. Real-time tracking of image regions with changes in geometry and illumination. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 403–410, 1996.

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

[9] M. Hunke and A. Waibel. Face locating and tracking for human-computer interaction. In *Proceedings of the 28th Asilomar Conference 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 *Proceedings of the 4th International Conference on Computer Vision*, pages 93–101, 1993.

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

[12] H. K. Nishihara. Personal communication, 1996.

[13] H. K. Nishihara, H. J. Thomas, and E. Huber. Real-time tracking of people using stereo and motion. In *Machine Vision Applications in Industrial Inspection II:Proceedings of the SPIE*, volume 2183, pages 266–273, 1994.

[14] 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.

[15] K. Toyama and G. D. Hager. Incremental focus of attention for robust visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 189–195, 1996.

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

[17] 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.