STEREO-BASED ELLIPTICAL HEAD TRACKING

Karthik Narayanan, Raghunandan Kumaran and John Gowdy

Electrical and Computer Engineering Department, Clemson University
29634, Clemson, USA
{ knaraya, ksampat, jgowdy }@clemson.edu

ABSTRACT

A novel algorithm for head-tracking based on stereo is presented in this paper. Employing stereo vision makes the tracker robust to many factors such as clutter, color and lighting variations, that affect contemporary head trackers. The head is modeled as an ellipse and tracking is performed on the foreground obtained by modeling depth. This is facilitated by using stereo vision. Further, we present a simple head size estimation technique that is critical for modeling considerable motion by the subject. The results obtained demonstrate the robustness of the head tracker and the accuracy of the head-size estimation technique. This would find extensive use in speaker tracking in smart rooms.

1. INTRODUCTION

Head tracking has been an area of active research over the years. This involves addressing different factors such as rotation, pose, sensitivity to background clutter, effect of lighting variations and erratic movements by the subject, besides localizing the human head . Several techniques based on intensity edges, skin color information and pattern classifiers have been proposed and successfully implemented addressing the above issues.

Stereo vision can be used in negating the shortcomings of contemporary head trackers. Background clutter and color that affect the performance of intensity edge-based trackers [2] and skin color based trackers [1, 3] do not affect stereo. Stereo based trackers also perform well in varying light conditions [4]. Besides, availability of commercial stereo-cameras has encouraged the use of stereo vision for tracking. Stereo also enables localization in 3-D space that would find extensive use in *smart-room* based applications. We intend using the proposed algorithm in consonance with acoustic localization techniques for speaker tracking. Their utility would be enhanced in challenging acoustically environments, where localization based on audio alone is less efficient. Stereo has been used as one of the modes in a multimodal head tracker in [5]. Very little work has been done in using stereo alone for head tracking. The stereo based head tracker proposed in [6] fails when the subject does not face the camera.

In this paper, we present a robust head tracker based on stereo alone. A correlation based stereo algorithm is used to yield disparity images, which are used for further tracking. A background subtraction scheme, modeling depth as proposed in [6], is used to yield the foreground that contains the region of interest. We then present a tracking algorithm that models the head as a 2-D ellipse and assumes constant velocity of motion of the subject as in [2]. In addition, a simple head size estimation technique incorporating the disparity values is presented and used for tracking. This would find use in human-computer interfaces and biometric person authentication.

The paper is organized as follows: The stereo algorithm and the background subtraction scheme used are presented in Section 2. Section 3 comprises of the head tracking algorithm, with the head size estimation explained in detail. The experimental results and a discussion on them are presented in Section 4. A conclusion is presented in Section 5.

2. STEREO AND SEGMENTATION

Correlation based correspondence algorithms involve the correspondence of intensity values between a stereo pair of images. The task involves obtaining the best match for each pixel of an image on the other corresponding image. As individual intensity values have many potential matches, blocks of data or windows are used. The disparity images thus obtained have the relative depth information embedded in them. These images become highly unreliable for those areas in the image devoid of texture. This issue needs to be addressed, as the disparity images are subject to segmentation before tracking. Other demerits include the over-reliance on window length and its inability to model depth discontinuities. These though, do not have a direct impact on the performance of the tracker, as indicated in Section 4. Several real time systems have been developed using this algorithm [10], and this motivates our choice of it over other techniques despite its drawbacks. The disparity images obtained are segmented as explained below.

The foreground segmentation technique used is one of modeling depth as proposed in [6]. Every pixel in the background disparity image is modeled as a univariate Gaussian with mean and standard deviation . The mean $_{i,j}$ and standard deviation $_{i,j}$ for each pixel in the disparity image $D_{i,j}$ are obtained using equations (1) and (2).

$$_{i,j} = \frac{1}{N} \sum_{k=1}^{N} D_{i,j}(k) \tag{1}$$

$$_{i,j} = \sqrt{\frac{1}{N-1} \sum_{k=1}^{N} (D_{i,j}(k) - _{i,j})^2} \tag{2}$$

N is the number of background images used to obtain the model. For well-textured regions in the background images, the disparity values are reliable. Thus, ideally, the mean would be the be the pixel's disparity value and the standard deviation would be zero. Even otherwise, the standard deviation would be a small value and the mean close to the images' disparities. Thus, the foreground can be defined as those pixels whose disparity exceeds their mean by at least a standard deviation. A pixel Di, j having a disparity greater than

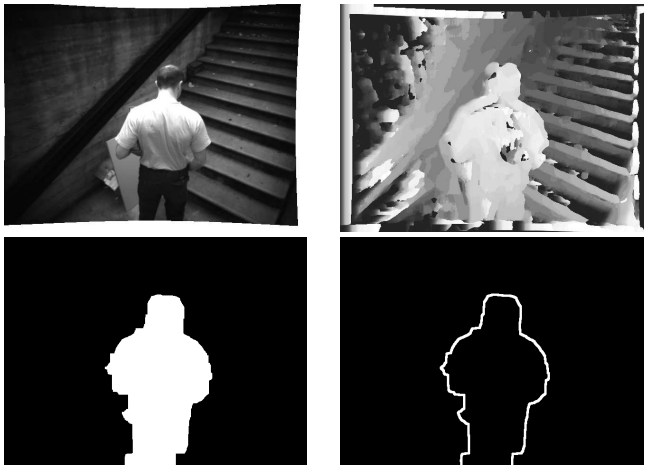

Figure 1: Top left: Test image used for segmentation. Top right: Corresponding disparity image (histogram equalized for better viewing). Bottom right: Binary image after connected components. Bottom left: Image obtained after edge detection (image modified for better viewing)

$_{i,j} +$ $_{i,j}$ is classified as foreground. Considering the inverse relationship between depth and disparity, regions closer are classified as foreground, which is physically intuitive. For regions in the background that lack texture the disparity is unreliable. An unreliable pixel would have a high standard deviation since the disparity values in the background images used for training the model would be highly deviant from the mean disparity. Thus, all pixels with a standard deviation $_{i,j}$ greater than a pre-determined value are deemed unreliable. This is due to many close matches for every pixel in the corresponding stereo pair. Any pixel labeled unreliable $_{i,j} >$ is considered foreground. The choice of the threshold is important, though not critical ($= 2$ in our case). Once the foreground segmentation is achieved, a binary connected components algorithm is executed, retaining the biggest blob. We assume that our region of interest (human) is a continuous blob of pixels. An edge detector is then employed to obtain the edges of the region of the subject's body as shown in figure 1.

3. HEAD TRACKING

The projection of a head on a 2-D plane can be closely modeled by an ellipse. Our tracking algorithm is based on the one used in [2]. The equation of an ellipse with center (x_c, y_c) and minor and major axis lengths a and b, respectively, is given by equation (3). The ratio of major and minor axis lengths is called the aspect ratio. The ellipse equation in terms of the aspect ratio is given in equation (4).

$$\frac{(x-x_c)^2}{a^2} + \frac{(y-y_c)^2}{b^2} = 1 \tag{3}$$

$$k^2(x-x_c)^2 + (y-y_c)^2 = b^2 \text{ where } k = \frac{b}{a} \tag{4}$$

It is observed that for a fixed aspect ratio k, the position of the ellipse (x_c, y_c) and the scale (major axis length b) describe the state of the ellipse. Thus for each frame the most likely position of the ellipse (x_c, y_c) is determined as one with the maximum normalized sum about the circumference of the ellipse. In other words, for a given scale the co-ordinates that yield the maximum likelihood, as defined in equation (5), is chosen as the center of the ellipse.

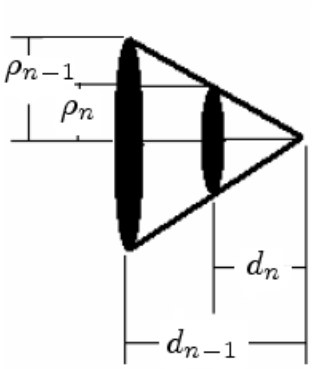

Figure 2: Head size estimation of person moving away from camera (Side view)

$$(x^*, y^*) = \arg\max_{\{|x-x^p| \le x_r, |y-y^p| \le y_r\}} \{\frac{1}{N} \sum_{k=1}^{N} p_k\} \tag{5}$$

The new position of the ellipse would be the one with maximum likelihood over a search range of (x_r, y_r)[1] pixels about the predicted position (x^p, y^p). N denotes the number of pixels on the ellipse's circumference. The position in frame n is then predicted assuming constant velocity of motion of the subject and is obtained using (6). Though other state tracking methods employing complex models exist, they would be complicating the task as tracking is performed on a sequence of binary images delineating the subject alone.

$$(x_n^p, y_n^p) = 2 * (x_{n-1}^*, y_{n-1}^*) - (x_{n-2}^*, y_{n-2}^*) \tag{6}$$

A local search about the predicted position over the search range would yield the most likely position of the ellipse. For the first two frames, the search space has to span a greater area, and in our case it was confined to the subject's body. A global search would prove computationally expensive, thereby justifying the employment of constant velocity prediction. As for the scale, the other parameter describing the state of the ellipse, we use our head size estimation technique described in Section 3.1.

3.1 Head-size estimation

Here we present a technique to predict the size of the head using the disparity values already obtained. The motion of the subject in 3-D space can be considered as one from a disparity plane to another. This calls for the interpretation of 3-D space as a set of discrete disparity planes, parallel to each other and the plane of view. The plane closest to the camera has the highest disparity and the farthest has zero disparity. The size of the head is assumed to appear diminishing when the subject is moving away from the camera and increasing when moving towards the camera. This is driven

[1] We used $x_r = y_r = 20$ pixels

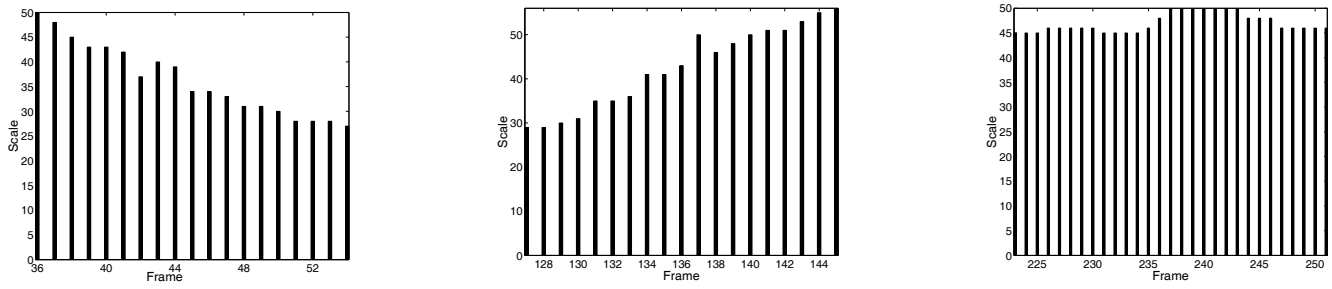

Figure 3: Scales corresponding to figures 4,5 and 6

Figure 4: Subject walking away from camera: Selected frames between frames 36-54

Figure 5: Subject walking towards camera: Selected frames between Frames 127-145

Figure 6: Subject walking towards camera: Selected frames between Frames 223-251

by intuition and is particularly useful when the motion of the subject seems pronounced between frames. From figure (2), using similar triangles, the predicted scale $\overset{p}{n}$ for a frame is given by equation (7). Here d_n represents the disparity in the current frame and $_{n-1}$ the scale in the previous frame. This provides us with a scale that is approximate, since we assume the vertex of the cone generated to lie on the zero disparity plane. It is critical to note the irrelevance of the position of the head in a frame. It is to be noted that the size of the head is assumed to appear the same when the motion of the subject is confined to the same disparity plane.

$$\overset{p}{n} = \frac{\overset{*}{n-1} * d_n}{d_{n-1}} \tag{7}$$

$$\overset{*}{z} = \arg\max_{\{|\ z-\ \overset{p}{z}|\leq\ r\}} \{\frac{1}{N} \overset{N}{k=1} p_k\} \tag{8}$$

This assumption is consistent with equation (7). A local search about the predicted scale can be conducted to obtain the closest scale, but was found to be redundant in our case. As our algorithms were run offline, the scale in the first frame was used recursively to yield the scales in the subsequent frames. The approximate scale as in (7) worked remarkably well, given the simplicity of the expression. Equation (8) can be used for accurate modeling of the head, for frame z. [2]

Here we make the assumption that the subject has a single disparity. In most cases, due to other factors such as occlusion, the pixels corresponding to the human would have no unique disparity. In that case, the statistical mode of the corresponding disparity values is taken as the required disparity. Once the position of the head in the image plane is known, the position in 3-D space can be obtained with knowledge of the camera parameters.

4. RESULTS AND DISCUSSION

A stereo database, comprising of sequences of rectified images with considerable motion of the subject is used [8]. The stereo algorithm used a window 15 pixels in length and a 64 pixel search for correspondence, assuming the epipolar constraint [9]. The depth model was developed using 30 images of the background. For tracking, the head was modeled as an ellipse with aspect ratio 1.1 and the search space was 20 pix-

[2] $_r = 0$ in our case

els ($x_r = y_r = 20$). The results obtained are shown in figures 4,5 and 6. Figures 4 and 5 illustrate the performance of the tracker for a person moving towards and away from the camera, respectively. Figure 6 shows the results for the subject changing the direction of motion.

To analyze the performance of the tracker, we consider issues related to the environment and the algorithm. The former would address distractions from the ambience and issues concerning the subject's movement. Algorithm based issues includes addressing the palpable shortcomings of the stereo, segmentation and tracking algorithms. The tracker's performance for any orientation of the head, was observed to be satisfactory as seen in figures 4, 5 and 6 [3]. The tracker is expected to handle occlusions fairly well, as long as the occluding object is not elliptical in shape. Erratic movements of the subject, especially those involving change in direction of motion are not expected to be tracked successfully. However, the tracker performs well when the subject changes direction of motion as observed figure 6. This is because of the relatively larger search space employed, and it manifests the effectiveness of the constant velocity prediction technique used. The advantages of using stereo is clear in frames where the subject's head (hair) color blends with the background and when the head appears relatively bald. These are instances where color or intensity-edge based trackers would fail. The tracker is observed to be resistant to mild tilting of the head, as seen in figure 6. The effectiveness of the scale prediction technique used is evident in the results shown. The order of scale variation between frames 36-51 was as high as 40%, ranging from 50 in frame 36 to 30 pixels in frame 51, as shown in figure 3.

We now consider the algorithm related issues. The stereo algorithm used assumes all pixels in a window to have the same depth, which results in poor modeling of depth discontinuities. This would result in the edges obtained being offset by at most half the window length. As our window length is relatively small (15 pixels), this does not affect the results. The behavior of the tracker was unpredictable when the subject was along the boundaries. This stems from one of the major shortcomings of the correlation based stereo algorithm. Although the tracker required manual scale initialization, it was found to be tolerant to a wide range of values. The same was the case with the threshold ( ) chosen for segmentation. The value was not critical as 3.2% of the total pixels were deemed unreliable for = 2, while the unreliable pixels' percentage dropped marginally to 2.7% for = 3. The segmentation and tracking modules posed no problems related to the tracker's performance.

The tracker was thus found to be robust against out-of plane rotation, background distractions, tilting and considerable motion towards and away from the camera. Further, the tracker's performance is relatively unconstrained by the background except for the highly unlikely case where the background is completely devoid of texture.

5. CONCLUSION

A stereo based elliptical head tracker is presented in this paper. The approach is based on depth modeling facilitated by the disparity images obtained from stereo. Tracking is achieved by modeling the head as an ellipse and incorporating constant velocity prediction. A simple head size estimation technique is used and its usefulness is demonstrated.

The results illustrate the robustness of the adopted approach to various factors that affect contemporary head trackers. As long as the shape of the head above the shoulders is discernible, the tracker would work. Employing stereo would also enable spatial localization of the subject, which would find use in smart room based applications and Human computer interfaces. Future work will include integrating the tracker into a multimodal system with acoustic input realized by a real-time algorithm, for accurate speaker localization.

[3] View entire sequence at www.clemson.edu/ ˜ knaraya/stereo.html

REFERENCES

[1] P. Fieguth and D. Terzopoulos, "Color-based tracking of heads and other mobile objects at video frame rates," in *Proceedings of the Conference on Computer Vision and Pattern Recognition*, San Juan, Puerto Rico, 1997, pp. 21–27.

[2] Stan Birchfield, "An Elliptical Head Tracker," in *Proceedings of the 31st Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, California, November 1997, pp. 1710–1714.

[3] Stan Birchfield, "Elliptical Head Tracking Using Intensity Gradients and Color Histograms,"in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,*, Santa Barbara, California, June 1998, pp. 232–237.

[4] Louis-Philippe Morency, Ali Rahimi, Neal Checka, and Trevor Darrell, "Fast Stereo-Based Head Tracking for Interactive Environments," in *Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, Washington D.C. , May 20 - 21, 2002, pp. 390–395.

[5] T. Darrell, G. Gordon, M. Harville and J. Woodfill, "Integrated Person Tracking Using Stereo, Color, and Pattern Detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,*, Santa Barbara, California, June 23 - 25, 1998, pp. 601–609.

[6] Daniel B. Russakoff, Martin Herman, "Head tracking using stereo," *Machine Vision and Applications*,vol. 13, Issue 3, pp. 164–173, July 2002.

[7] Heiko Hirschmuller, "Improvements in Real-time Correlation-Based Stereo Vision ,"in *Proc. of IEEE Workshop on Stereo and Multi-Baseline Vision 2001*, Kauai, Hawaii, 9-10 December 2001, pp. 141–148.

[8] http://labvisione.deis.unibo.it/ smattoccia/stereo.htm

[9] Milan Sonka, Vaclav Hlavac and Roger Boyle, *Image Processing, Analysis and Machine Vision* 2000.

[10] K. Konolige, "Small vision systems: Hardware and implementation,"in *Eighth International Symposium on Robotics Research*, Hayama, Japan, October 1997, pp. 203212