Three Dimensional Fingertip Tracking in Stereovision

S. Conseil[1], S. Bourennane[1], and L. Martin[2]

[1] Univ. Paul Cézanne, Institut Fresnel (CNRS UMR 6133),
Dom. Univ. de Saint Jérôme, F-13013 Marseille cedex 20, France
{simon.conseil, salah.bourennane}@fresnel.fr
[2] ST Microelectronics, ZI Rousset BP 2, F-13106 Rousset, France
lionel.martin@st.com

Abstract. This paper presents a real time estimation method of the three dimensional trajectory of a fingertip. Pointing with the finger is indeed a natural gesture for Human Computer Interaction. Our approach is based on stereoscopic vision, with two standard webcams. The hand is segmented with skin color detection, and the fingertip is detected by the analysis of the curvature of finger boundary. The fingertip tracking is carried out by a three dimensional Kalman filter, in order to improve the detection with a local research, centered on the prediction of the 3-D position, and to filter the trajectory to reduce the estimation error.

1 Introduction

Hand gestures are a natural and instinctive mean for humans to interact with their environment. They can be used to emphasize speech, to point or to manipulate objects in augmented environment, or to communicate with sign language. Over the past few years, there has been a growing interest in hand gestures recognition, thanks to the use of computer vision techniques [1].

Finger pointing is a simple gesture, well-fitted to replace the mouse. The finger is indeed a natural and very practical pointing device for Human Computer Interaction. Various assumptions have been used to ease fingertip detection and different configuration have been studied. Hence it is possible to determine the 3-D finger trajectory with a geometric model of the body, using a single camera and the detection of head and shoulders [2] or with a stereovision system and eye-to-fingertip pointing mode [3].

Other systems recognize 2-D trajectory with a single camera above the work plane. In the Digital Desk system from Crowley *et al.* [4] tracking is carried out on a single view by correlation with a fingertip template, but it is not robust to orientation and scaling changes. The resulting plane trajectory have been used for handwriting recognition with HMM [5]. In the EnhancedDesk system [6], several fingers are tracked in one view, with a two-dimensional Kalman filter for each finger. The detection of the fingertips is carried out thanks to an infrared camera and a normalized correlation. Finally two-dimensional trajectories (square, circle, triangle) are recognized using HMM.

J. Blanc-Talon et al. (Eds.): ACIVS 2005, LNCS 3708, pp. 9–16, 2005.

Segen and Kumar [7] use two cameras to determine the 3-D direction pointed by the finger. Strong curvature points are detected on the boundary of the hand region and used to classify three hand gestures. They apply this system to a computer game interface and a navigation software. The use of the disparity obtained from a stereoscopic pair has also been studied by [8], but the computation of a disparity map is computationally expensive.

In this paper, we focus on fingertip tracking for pointing gestures. Two calibrated cameras are used to compute 3-D fingertip trajectory, but errors in the fingertip localization and the absence of synchronization result in unprecise trajectory estimation. The originality of our approach is the use of a three dimensional Kalman filter to predict 3-D position and to smooth the fingertip trajectory. The predicted 3-D position is projected in the two images, and a local search is performed for the fingertip detection.

2 Finger Detection

2.1 Hand Segmentation

Hand segmentation is the first important step. Fast and robust segmentation is needed, without assumptions on the background color. In previous experiments, the classical background subtraction method has proved to be too much sensitive to shadows and illumination variations, even in a controlled environment.

Skin color detection is now commonly used in both hand and face segmentation. Many approaches have proved their efficiency, with different color spaces or learning techniques [9]. We have chosen to detect skin color pixels in the YCbCr color space, with the fast and simple approach presented in [10]. Figure 1.(b) shows that the result is convincing, even if some skin pixels are not detected on the right part of the hand. In order to reduce the noise from the binary silhouette, a median filter is applied, then a morphological filtering, and finally a connected components labeling to remove the non-hand regions and to fill holes in the hand blob.

(a) (b) (c)

Fig. 1. Hand segmentation: (a) original image, (b) silhouette obtained from CbCr thresholding and (c) final silhouette after filtering and connected components labeling

2.2 Fingertip Detection

When the finger enters in the cameras field of view, it is necessary to detect accurately the fingertip position in order to initialize the tracking. As we consider the case of pointing gesture, we assume that only one finger is pointing and so the fingertip is the point located at the extremity of the hand region.

With the hand silhouettes obtained from the previous stage, we describe the boundary of the hand region by a succession of points $P(i) = (x(i), y(i))$. The fingertip is the point of the boundary that maximizes the distance from the center of gravity of the hand region. The center of gravity is obtained with the computation of the geometrical moments.

However this measure is not very precise, depending on the hand orientation. Hence we refine the fingertip detection with the curvature of the boundary with the method presented in [7]. The k-curvature is given by the angle between the vectors $[P(i-k)P(i)]$ and $[P(i)P(i+K)]$. The fingertip is the point with the stronger curvature.

3 Three Dimensional Tracking

With the position of the fingertip in each of the two images, one can compute its 3-D position. However the 3-D positions are not precise for several reasons: unprecise detection of the fingertip due to a bad segmentation, discretization of the images (one pixel error on the fingertip localization can represent several millimeters in 3-D), temporal shift between the acquisition of the two images (the two cameras are not synchronized).

Furthermore it is not necessary to treat the whole image whereas we know the finger position. Thus the research of the fingertip can be reduced to a small window, thanks to the tracking of the finger and the prediction of its position with the preceding pair of images. The goal of the temporal tracking is thus to facilitate the localization of the finger and to smooth the trajectories.

3.1 Kalman Filter

Our approach is based on a Kalman filter [11] in three dimensions, with the fingertip's location and velocity. We assume that the movement is uniform and the velocity is constant, the frame interval ΔT being short. The state vector $\mathbf{x}_t$ is defined as:

$$\mathbf{x}_k = (x(k), y(k), z(k), v_x(k), v_y(k), v_z(k))^T$$

where $(x(k), y(k), z(k))$ is the position and $(v_x(k), v_y(k), v_z(k))$ the velocity of the fingertip in frame k. The state vector $\mathbf{x}_k$ and the observation vector $\mathbf{z}_k$ are related by the following equations:

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathrm{A}\,\mathbf{x}_k + \mathbf{w}_k \\ \mathbf{z}_k &= \mathrm{H}\,\mathbf{x}_k + \mathbf{v}_k \end{aligned} \tag{1}$$

with $\mathbf{w}_k$ and $\mathbf{v}_k$ the process and measurement noises, assumed to be independent white gaussian noises, A the state transition matrix and H the observation matrix:

$$\mathrm{A} = \begin{pmatrix} 1\,0\,0 & \Delta T & 0 & 0 \\ 0\,1\,0 & 0 & \Delta T & 0 \\ 0\,0\,1 & 0 & 0 & \Delta T \\ 0\,0\,0 & 1 & 0 & 0 \\ 0\,0\,0 & 0 & 1 & 0 \\ 0\,0\,0 & 0 & 0 & 1 \end{pmatrix}$$

$$\mathrm{H} = \begin{pmatrix} 1\,0\,0\,0\,0\,0 \\ 0\,1\,0\,0\,0\,0 \\ 0\,0\,1\,0\,0\,0 \end{pmatrix}$$

Writing $\mathbf{x}_k$ and $\mathbf{x}_k^-$ the *a posteriori* and *a priori* state estimates, P_k and P_k^- the *a posteriori* and *a priori* estimate error covariances, Q the process noise covariance, R the measurement noise covariance, and K_k the Kalman gain, one obtains the following equations:

Prediction equations:

$$\begin{aligned} \mathbf{x}_k^- &= A\mathbf{x}_{k-1} \\ P_k^- &= AP_{k-1}A^T + Q \end{aligned} \tag{2}$$

Update equations:

$$\begin{aligned} K_k &= P_k^- H^T (HP_k^- H^T + R)^{-1} \\ \mathbf{x_k} &= \mathbf{x_k}^- + K_k(\mathbf{z_k} - H\mathbf{x_k^-}) \\ P_k &= (I_6 - K_k H)P_k^- \end{aligned} \tag{3}$$

Parameters Setting. The three components are supposed to be independent, thus the covariance matrices are diagonal. As we assume constant velocity in our model, which may not be always true, the process noise covariance is supposed to be important on the velocity component whereas it is weak in the position one.

The measurement noise covariance is calculated with a sequence of images where the finger remains fixed. We obtain $Var(X, Y, Z) = (0.31, 2.39, 15.06)$, which shows that the measurement error is more significant on component Z than on X and Y.

3.2 Developed Algorithm

Figure 2 summarizes the different stages of the treatment: starting from the computation of the 3-D position with a pair of images, one can predict the 3-D position corresponding to the following pair of images. The predicted 3-D position is projected in the two images to obtain a 2-D prediction of the fingertip position. Then the research of the fingertip in each image can be reduced to a

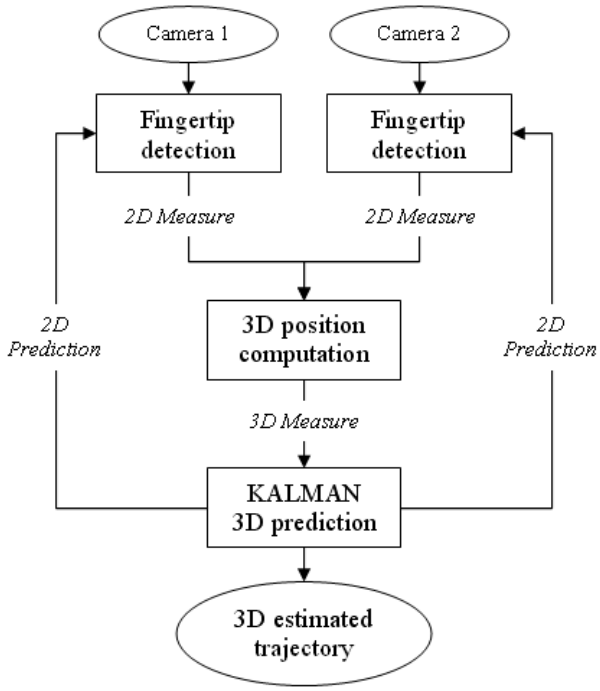


Fig. 2. Diagram summarizing the different stages

neighborhood of the predicted fingertip position. The size used for the research window is 80×80 pixels.

The detection of the finger is then carried out with the method described in Sec. 2. Finally, the epipolar constraint is checked to ensure the good detection of the fingertip in the two images. However, because of the non-synchronisation, a little error is admitted in the epipolar constraint checking.

4 Results

We use two common webcams, with 352×288 image resolution. Images are transmitted by USB connection, with a MJPEG compression which introduces noise on the images. Moreover, the two cameras are not synchronized, which can induce a small difference in position between two images, and can result in an oscillation in the finger trajectory: during the time interval between the two frame grabbings, the finger can have moved, depending on the velocity of movement. Consequently, the triangulation is skewed, mainly on the depth dimension (corresponding to the optical axis of the cameras).

4.1 Example: Case of a Circle

In order to be able to measure computation errors of the 3-D position, it is necessary to know the ground truth, which is often difficult in stereovision. In

our configuration the reconstruction error is found mainly on the component Z, corresponding to the depth (direction of the optical axes). Thus we are interested in a plane trajectory, a circle realized on the desk, which corresponds to the plane $z = 0$ (Fig. 3). As the trajectory is plane, standard deviation of component Z can easily be computed to compare the reconstruction errors.

We can see on Fig. 4.(a) the estimated 3-D trajectory of the circle, as well as measurements in dotted lines. The reconstruction error is more important on the Z component, corresponding to the depth. Figure 4.(b) reveals that the Kalman filter smoothes the component Z, in this case the standard deviation on the depth is reduced from 9.77 to 5.46.

4.2 Velocity Influence

The velocity of the movement influences the reconstruction error. Indeed, the faster is the movement, the more the finger can have moved between the acquisition of the two images, which results in a more significant error. Table 1 illustrates this with the study of two plane trajectories (circle and square), treated in real time (30 Hz).

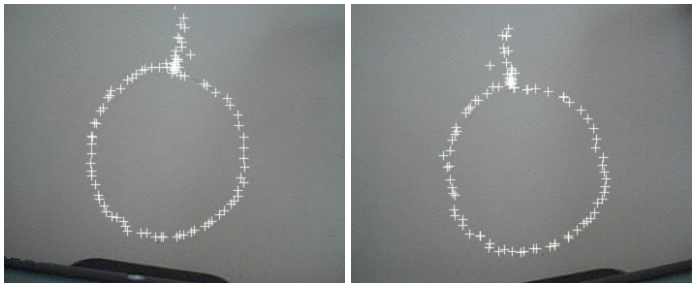
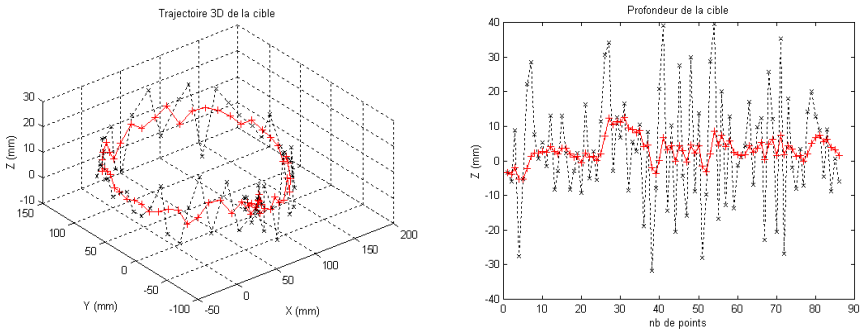


Fig. 3. Left and right images with a circle trajectory and detected fingertip positions



(a) (b)

Fig. 4. (a) 3-D trajectory with a circle gesture and (b) component Z, corresponding to the depth dimension (measures in dotted lines, estimations in full lines)

These trajectories have been realized at three different velocities, thus a faster trajectory is made up of a lower number of points. The standard deviation on component Z is then computed to compare the reconstruction errors.

In both cases we see that the standard deviation increases with speed, and the standard deviation is weaker for the trajectory estimated by the filter of Kalman than for measurements. We also see that the reconstruction error is smaller for the square, the linear movement being better adapted to the model than the circular one.

Table 1. Evolution of the standard deviation on the depth according to the speed of realization of the movement

Trajectory	Velocity	Number of points	Std dev Mesures	Std dev Estimation
Circle	slow	306	9.7673	5.4587
	medium	189	11.3158	8.3916
	fast	108	14.7552	10.8265
Square	slow	290	10.4771	4.8718
	medium	185	11.1463	4.4337
	fast	106	12.2786	6.0401

5 Conclusion

We presented a three dimensional finger tracking system based on a Kalman filter, which performs robust detection thanks to the reduction of the fingertip research to a small window and the reduction of the estimation error with the smoothing of the 3-D trajectories. The system runs in real time on real data, on a 2.6 GHz PC. With adapted detection method, other applications are possible, like people or vehicles tracking. To improve the system, the computation of the research window width could be adapted to the movement velocity. We also plan to extend the tracking to multiple fingers and to deal with occlusion problems.

References

1. Pavlovic, V., Sharma, R., Huang, T.: Visual interpretation of hand gestures for human-computer interaction : A review. IEEE trans. on Pattern Analysis and Machine Intelligence **19** (1997)
2. Wu, A., Shah, M., da Vitoria Lobo, N.: A virtual 3d blackboard: 3d finger tracking using a single camera. In: Proc. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition. (2000)
3. Hung, Y., Yang, Y., Chen, Y., Hsieh, I., Fuh, C.: Free-hand pointer by use of an active stereo vision system. In: Proc. of the IEEE Int. Conf. on Pattern Recognition, Brisbane (1998) 1244–1246
4. Crowley, J., Berard, F., Coutaz, J.: Finger tracking as an input device for augmented reality. In: IEEE Int. Workshop on Automatic Face and Gesture Recognition, Zurich (1995) 195–200

5. Martin, J., Durand, J.B.: Automatic handwriting gestures recognition using hidden markov models. In: Proc. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition. (2000)
6. Oka, K., Sato, Y., Koike, H.: Real-time fingertip tracking and gesture recognition. IEEE trans. on Computer Graphics and Applications **22** (2002) 64–71
7. Segen, J., Kumar, S.: Human-computer interaction using gesture recognition and 3d hand tracking. In: Proc. of the IEEE Int. Conf. on Image Processing. (1998)
8. Jojic, N., Huang, T., Brumitt, B., Meyers, B., Harris, S.: Detection and estimation of pointing gestures in dense disparity maps. In: Proc. of the IEEE Int. Conf. on Automatic Face and Gesture Recognition. (2000)
9. Phung, S., Bouzerdoum, A., Chai, D.: Skin segmentation using color pixel classification: analysis and comparison. IEEE trans. on Pattern Analysis and Machine Intelligence **27** (2005)
10. Chai, D., Ngan, K.: Face segmentation using skin-color map in videophone applications. In: IEEE Trans. on Circuits and Systems for Video Technology. Volume 9. (1999)
11. Welch, G., Bishop, G.: An introduction to the kalman filter. Technical report (1995)