3D Real-Time Head Tracking Fusing Color Histograms and Stereovision

Francesc Moreno, Adrià Tarrida, Juan Andrade-Cetto and Alberto Sanfeliu

Institut de Robòtica i Informàtica Industrial, UPC-CSIC
Llorens i Artigas 4-6, 08028 Barcelona, Spain
fmoreno,atarrida,cetto,sanfeliu@iri.upc.es

Abstract

A system that performs the tracking of a human head in 3D in real time is presented. The head shape is modeled by an ellipse with a trained color histogram of skin and hair samples. The color histogram is dynamically updated based on incoming image data in order to accommodate for varying illumination conditions. On the other hand, the size of the searched ellipse projected on the image is scaled depending on the depth information gathered from stereo vision. The strength of our method resides on the use of a predictive filter to fuse color and depth information, iteratively refining the location of the head in 3D and the parameters of the head color histogram.

Keywords*: data fusion, depth from stereo, color histograms.*

1. Introduction

The ability to detect and track human heads on an image sequence is useful in a great number of applications, such as human-machine interaction, or face recognition and gesture tracking for surveillance systems. When no restrictions are imposed on the input data, the problem becomes a challenging task. Some typical difficulties we have to deal with, are varying illumination conditions that make face appearance to change over time, the a priori ignorance of the head scale and pose, occlusions, and complex or unknown backgrounds that could lead to false head shapes.

Many approaches have been proposed for detecting and tracking humans. Often, only a single technique is used to locate human features and to extract them from the rest of the image. Yang and Waibel [10] model the skin-color distribution as a multivariate normal distribution. To handle variations in illumination conditions they propose to update this distribution over time using an EM algorithm. Beymer and Konolige [1] only use stereo information and template matching techniques, to separate human shapes from the background, and a Kalman filter to track people. Improved and more robust results can be obtained by using multiple processing modalities. Birchfield [2] uses intensity gradients and color histograms to update the position of the head over time. Darrel, *et al.* [3] combine stereo and color via a pattern classification method. In [6] color an stereo are used independently, i.e., motion and color are used to track in a known scenario with stereo triangulation only used to estimate the 3D location afterwards. A Kalman filter is used to refine the results.

We present an approach to 3D human head tracking that achieves a robust and real time performance operating in relatively complex scenarios by combining two techniques: the search of skin-like color regions through color histograms, and the extraction of the scale and 3D location of the head using depth from stereo. The main difference between our method and the previously cited approaches is that we completely characterize the fusion of multiple sensors through the Kalman filter, combining the 3D location and color parameters in the state estimate.

An overview of our system is given in Section 2. In Section 3 we describe the color and depth modules, and in Section 4, the proposed fusion model is derived. Results and conclusions are presented in Sections 5 and 6, respectively.

2. System overview

The system flowchart shown in Fig. 1 comprises three main modules: color, stereo and Kalman filter. The process starts by capturing a pair of synchronized stereo color images. The left image is fed into the color module, and using the information of the previous state about the position on the image and scale of the head (modeled as an ellipse), it computes the position of the head in the new image. This search is done by maximizing an intersection function between the color histogram of the new head candidate and a model histogram. The later is updated by taking into account the color histogram of the *best* candidate. This is the key operation that makes our algorithm robust under varying illumination conditions.

The pair of images, the previous computed disparity and the position and scale of the ellipse of the best head candidate are fed to the *fast* stereo module. This module performs an area-based correlation on a gray-scale version of the im-

16th IAPR International Conference on Pattern Recognition
Quebec, August 2002, Vol I, pp 368-371.

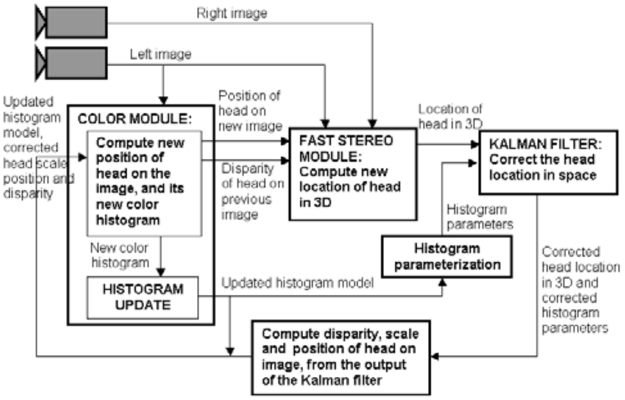
Right image
Left image
Updated histogram model, corrected head scale position and disparity
COLOR MODULE:
Compute new position of head on the image, and its new color histogram
New color histogram
HISTOGRAM UPDATE
Position of head on new image
Disparity of head on previous image
FAST STEREO MODULE: Compute new location of head in 3D
Location of head in 3D
KALMAN FILTER: Correct the head location in space
Histogram parameters
Histogram parameterization
Updated histogram model
Corrected head location in 3D and corrected histogram parameters
Compute disparity, scale and position of head on image, from the output of the Kalman filter

Figure 1. Flowchart for the head tracking system.

ages to solve the correspondence problem for the interior of the ellipse, computing an estimate of the position in 3D of the head, focusing the search of the head around the prediction done on the previous state.

Both the 3D head position and a parameterization of the updated model histogram are fused via a Kalman filter to compute a new estimate of the head position and the parameters modeling the histogram. This estimate is used to compute the new disparity, scale and position of the head on the image, which in turn will be the input parameters to the next state estimate.

3. Low level modules

3.1. Color module

Our color module is highly inspired in Birchfield's real-time head tracking system [2], where the projection of a head in the image plane is modeled by an ellipse. We initialize the process by detecting a human head on an image, using the technique described in our previous work [5]. This method, executed off-line, gives an initial position and scale of the subject head on the image, and lets us construct a model color histogram by filling the buckets of a discretized color space (B-G,G-R,B+G+R), with the pixels inside the ellipse. As proposed by Birchfield, to cope with situations where the subject turns around, we use a bimodal histogram containing skin and hair data samples.

At run time, when a new image is presented, a head candidate is searched on a local region around the previous position trying to maximize the intersection between the model histogram *M* and the candidate histogram C. The size of the candidate ellipse is given by the previous iteration of the stereo module. $C(i)$ and $M(i)$ represent the number of pixels inside the i-th bucket of the candidate and model histograms respectively, with N the total number of buckets. Swain and Ballard [8] propose the following expression as a measure of histograms intersection:

$$\phi(C) = \frac{\sum_{i=1}^{N} \min\left(C(i), M(i)\right)}{\sum_{i=1}^{N} C(i)} \tag{1}$$

In the search for an ellipse that maximizes the intersection function in an interest region, there will be an overlapping between adjacent ellipses, meaning that adjacent ellipses have common pixels. This redundancy can be exploited since for each new candidate ellipse, its color histogram can be computed from the adjacent ellipse by only subtracting those common pixels of the histogram and adding the new ones. We have adopted this strategy from Birchfield's work, in order to fulfill real-time results.

To obtain a model histogram robust to varying illumination, we have updated it over time with the equation [7]:

$$M_k(i) = (1-a)M_{k-1}(i) + (a)C_k(i) \qquad i = 1..N \tag{2}$$

In order to avoid updating with false head candidates, Eq. 2 is applied only when the measure of the model and candidate intersection is above an empirically determined threshold.

3.2. Depth module

The stereo module receives a pair of images taken with a calibrated stereo rig. The other input parameters are the elliptical interest zone on the left image (estimation of the position of the head candidate computed by the color module), and the disparity of the previous state (coming from the corrected 3D head location).

Restricting the stereo algorithm described in [5] to only those pixels on the left image that are inside the given ellipse, and to those pixels on the right image that have a range of disparity centered on the previous disparity value, we can considerably speed up the search for a stereo match, up to real-time values (over 50 Hz). With this restriction, we make the assumption that the velocity of displacement in depth of the tracked head will be lower than a certain value $Vmax_z$.

For each pixel (u, v) contained in the ellipse ε, we compute its disparity with a matching algorithm based on the sum of the absolute differences:

$$disp_k(u,v) = \arg\min_{d} \sum_{(u,v)\in\varepsilon_k} |I_l(u,v) - I_r(u+d,v)| \tag{3}$$

$$\hat{d}_{k-1} - \delta^- \leq d \leq \hat{d}_{k-1} + \delta^+ \tag{4}$$

$$d_{k-1} = mean_{(u,v)\in\varepsilon_{k-1}}\left(disp_{k-1}(u,v)\right) \tag{5}$$

where d_{k-1} is the averaged disparity in ϵ_{k-1}, (the ellipse in the previous state), $\hat{d}_{k-1}$ is the value of d_{k-1} corrected by means of the Kalman filter and $\{\delta^-, \delta^+\}$ are constants that define the range of acceptable disparities around $\hat{d}_{k-1}$.

We should note that as the images have been locally rectified, we can exploit in Eq. 3 the use of the epipolar restriction, speeding even more the correspondence problem.

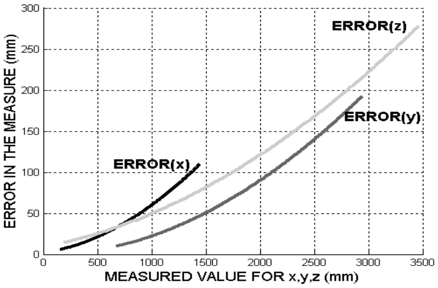

Figure 2. Estimated measurement errors introduced by the stereo module assuming a disparity error of 1 pixel.

Once the value $disp_k(u, v)$ has been computed for all the pixels in ϵ_k, we can extract the mean disparity d_k, and provided both camera calibration matrices, the estimation of the coordinate $\mathbf{p} = [x, y, z]^\top$ of the center of the ellipse is straightforward. These calibration matrices will also be used to compute the disparity, scale and head projection on the image, from the location of the head in 3D.

4. Fusion model and tracking

Our fusion model works with a Kalman filter [4, 9] receiving as inputs the estimation of the 3D head position and a parameterization of the updated histogram. Its goal is to combine the information from both modules to correct an estimate of the actual position of the tracked person minimizing the expected value of the error of this estimate.

4.1. System modelization

The process is governed by the following stochastic difference equation:

$$\mathbf{x}_k = \mathbf{x}_{k-1} + \mathbf{w} \tag{6}$$

The *state vector* $\mathbf{x}_k = [\mathbf{p}_k^\top, \boldsymbol{\mu}_k^\top, \phi_k(C_k)]^\top$ includes the head position $\mathbf{p}_k = [x_k, y_k, z_k]^\top$ in a world reference frame, the mean $\boldsymbol{\mu}_k = [\mu_{1k}, \mu_{2k}, \mu_{3k}]^\top$ of the updated model histogram M_k in the $(B - G, G - R, B + G + R)$ color space, and the value $\phi_k(C_k)$ corresponding to the intersection between the candidate histogram C_k and the model histogram before updating, M_{k-1}. The state vector is then represented by:

$$\mathbf{x}_k = [x_k, y_k, z_k, \mu_{1k}, \mu_{2k}, \mu_{3k}, \phi_k(C_k)]^\top \tag{7}$$

The random variable $\mathbf{w}$ in Eq. 6 is a *process noise vector* representing the uncertainty in the estimate of $\mathbf{x}$ over any time interval $k - 1$ and k. It is assumed that $p(\mathbf{w}) \sim N(0, \mathbf{Q})$. The *process noise covariance matrix* $\mathbf{Q}$ is typically set to a constant value, and we choose it to be a diagonal matrix, with each element $\mathbf{Q}(i, i)$, representing the uncertainty in the value of $\mathbf{x}_k(i)$ given its previous estimate $\mathbf{x}_{k-1}(i)$. The criteria for their adjustment are the following:

1. *Variance in head motion*. The movement of the head in each iteration is a random process, so the uncertainty in the values of $\mathbf{p}_k$ should represent the maximum displacement that we expect for the head in each axis.
2. *Variance in histogram mean*. The limit in the standard deviation of the histogram mean $\boldsymbol{\mu}_k$, depends on the level of discretization used to divide our color space to build the color histogram.
3. *Variance in Swin Ballard's distance*. If the value is normalized we can set the standard deviation of $\phi(C_k)$ to unity.

The use of a Kalman filter requires also a measurement model:

$$\mathbf{z}_k = \mathbf{x}_k + \mathbf{v}_k \tag{8}$$

The measurement vector $\mathbf{z}_k$, contains the value of the state vector $\mathbf{x}_k$ corrupted by a *noise vector* $\mathbf{v}_k$, representing any random error in the measurement, along with unmodeled nonlinearities on the observation model. It is assumed that $p(\mathbf{v}_k) \sim N(0, \mathbf{R}_k)$. Unlike the estimation of matrix $\mathbf{Q}$, we want the *measurement noise covariance matrix* $\mathbf{R}_k = E\{\mathbf{e}(\mathbf{z}_k)\mathbf{e}(\mathbf{z}_k)^\top\}$ to change in terms of $\mathbf{p}_k$ to express the decrement in accuracy of depth estimation that suffers any stereo system as the distance increases (see Fig. 2). We wish the variables x_k, y_k and z_k to have a growing error with respect to depth, and model this error with the 3rd-order polynomial:

$$\mathbf{e}(\mathbf{p}_k) = \begin{bmatrix} a_{00}x_k^3 + a_{01}x_k^2 + a_{02}x_k + a_{03} \\ a_{10}y_k^3 + a_{11}y_k^2 + a_{12}y_k + a_{13} \\ a_{20}z_k^3 + a_{21}z_k^2 + a_{22}z_k + a_{23} \end{bmatrix} \tag{9}$$

where each row fits one of the curves shown in Fig. 2. With this assumption we do not violate the basic restriction of the Kalman filter, that needs to be applied over *unimodal* Gaussian densities. We are just modifying the standard deviation of the population representing the measured value, but the probability distribution function remains unimodal and Gaussian.

The other two functions $\mathbf{e}(\boldsymbol{\mu}_k)$ and $\mathbf{e}(\phi_k(C))$ are set to constant values with off-line measurements of their deviation when computing the color histograms of a static head over time.

One may think that we could use the Kalman filter to update the whole histogram, i.e., to include Eq. 2 in the Kalman filter. But this, would imply high values in the dimension of vectors $\mathbf{x}$, $\mathbf{z}$ and matrices $\mathbf{R}$ and $\mathbf{Q}$, tampering with the possibility of a real-time solution.

4.2. Tracking algorithm

In this section we give the basic details of our iterative tracking algorithm. Assume an initial estimate of the *error covariance matrix* $\mathbf{P}_0$ and the initial state vector estimate $\hat{\mathbf{x}}_0$, at time state k we have computed the measurement vector $\tilde{\mathbf{z}}_k$ (approximation to the model given in Eq. 8), provided the previous state vector $\hat{\mathbf{x}}_{k-1}$. The steps to predict and correct the value of $\mathbf{x}_k$ are the following:

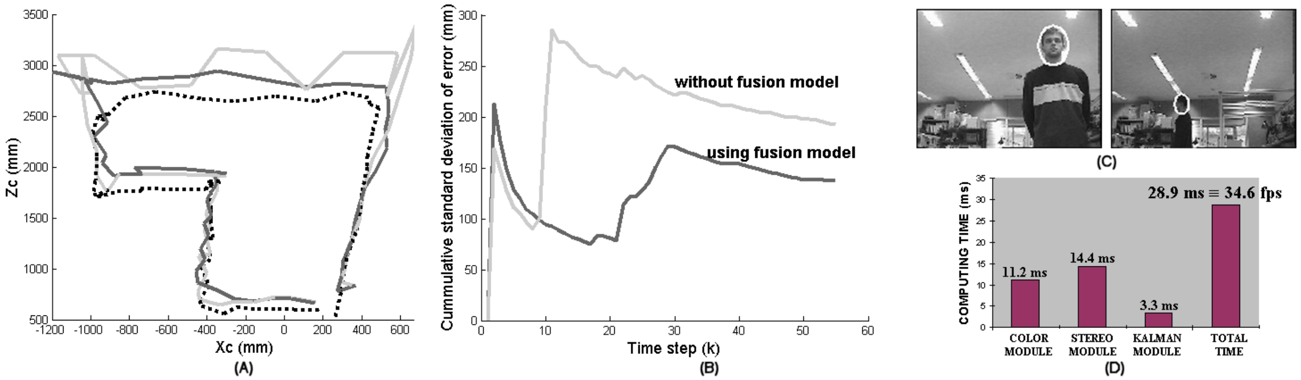

Figure 3. Results of our tracking system

1. Predict a priori estimates of the state vector, $\hat{\mathbf{x}}_k^- = \hat{\mathbf{x}}_{k-1}$ and error covariance matrix, $\mathbf{P}_k^- = \mathbf{P}_{k-1} + \mathbf{Q}$.
2. Update $\mathbf{R}_k$ using Eq. 9.
3. Compute the Kalman gain, $\mathbf{K}_k = \mathbf{P}_k^-(\mathbf{P}_k^- + \mathbf{R}_k)^{-1}$
4. Update the state estimate with measurement $\tilde{\mathbf{z}}_k$, i.e, $\hat{\mathbf{x}}_k = \hat{\mathbf{x}}_k^- + \mathbf{K}_k(\tilde{\mathbf{z}}_k - \hat{\mathbf{x}}_k^-)$
5. Update the error covariance, $\mathbf{P}_k = (\mathbf{I} - \mathbf{K}_k)\mathbf{P}_k^-$

For more details on the statistical theory behind the Kalman filter, the reader is referred to [4, 9].

5. Experiments

We have experimented the performance of our system on standard PC hardware (Pentium III at 1GHz). The images were 160×120 pixels, the search region used on the color module was 16×4 pixels and the range of disparity for the stereo computations was of 9 pixels ($\delta^- = \delta^+ = 4$ pixels).

Shown in the above figures is the response of the tracking algorithm in a scene that collects several complex factors: variation on head scale, illumination conditions do not remain constant, the subject turns his face, and objects with skin-like color appear on the background. We can see some of these disturbances on Fig. 3(c).

Fig. 3(a) shows the tracking results from a zenital view. The dotted line represents the real path followed by the subject computed a posteriori by manually specifying the center of the head in each pair of stereo images and reconstructing from stereo. The lighter solid line corresponds to the tracking result without fusing color and depth information. This result is clearly improved when we fuse color and depth with the Kalman filter as can be seen in the darker solid line. The improvements can be clearly observed at large distances, where small variations in disparity values induce high changes in the estimation of the position of the head in 3D. The reduction of the variance that the Kalman filter performs is shown on Fig. 3(b).

Fig. 3(d) shows the time specifications of our algorithm, certifying that it works in real time (over 30 Hz).

6. Conclusions

We have presented a method that robustly tracks a human head basing its strength on the fusion of information from color and depth using a Kalman filter. This lets the position of the head in 3D and a parameterization of the color histogram of the head to be iteratively updated in such a way that the covariance of the error is minimized. With this method we are able to undergo real time people tracking in complex situations such as in unstructured backgrounds, varying illumination conditions and with rotations of the person head.

Acknowledgments

This work was supported by CICYT projects DPI2000-1352-C02-01/02, DPI2001-2223, and a fellowship from the Spanish Ministry of Science and Technology.

References

[1] D. Beymer and K. Konolige. Real-time tracking of multiple people using continuous detection. In *ICCV*, 1999.

[2] S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *CVPR*, pages 232–237, 1998.

[3] T. Darrell, G. Gordon, M. Harville, and J. Woodfill. Integrated person tracking using stereo, color, and pattern detection. In *CVPR*, pages 601–609, 1998.

[4] P. S. Maybeck. *Stochastic Models, Estimation, and Control*, volume 1. Academic Press, New York, 1979.

[5] F. Moreno, J. Andrade-Cetto, and A. Sanfeliu. Localization of human faces fusing color segmentation and depth from stereo. In *ETFA*, pages 527–534, 2001.

[6] J. Rehg, M. Loughlin, and K. Waters. Vision for a smart kiosk. In *CVPR*, pages 690–696, 1997.

[7] L. Sigal, S. Sclaroff, and V. Athitsos. Estimation and prediction of evolving color distributions for skin segmentation under varying illumination. In *CVPR*, 2000.

[8] M. Swain and D. Ballard. Color indexing. *Int. J. Comp. Vision*, 7:11–32, 1991.

[9] G. Welch and G. Bishop. An introduction to the kalman filter (tutorial). In *SIGGRAPH*, 2001.

[10] J. Yang, W. Lu, and A. Waibel. Skin-color modeling and adaption. In *ACCV*, pages 687–694, 1997.