Localization of Human Faces Fusing Color Segmentation and Depth from Stereo

Francesc Moreno, Juan Andrade-Cetto, and Alberto Sanfeliu
Institut de Robòtica i Informàtica Industrial, UPC-CSIC
Llorens i Artigas 4-6, Edifici U, 2a pl. Barcelona 08028, Spain

Abstract—**This paper describes a method to localize faces in color images based on the fusion of the information gathered from a stereo vision system and the analysis of color images. Our method generates a depth map of the scene and tries to fit a head model taking into account the shape of the model and skin color information. The method is tailored for its use in factory automation applications where the detection and localization of humans is necessary for the completion or interruption of a particular task, such as robot manipulator safety or the interaction of service robots with humans. Keywords: Face localization, human detection.**

I. INTRODUCTION

The ability to recognize a human face or a facial expression is of great importance for the interaction of computers and robots with humans. At the Institut de Robòtica i Informàtica Industrial, UPC-CSIC, we are interested in providing our mobile platform Marco [1] with the ability to recognize people. Some results from our group on the recognition of human faces with the aid of a computer vision system are reported in [2]. However, this technique does not address the problem of locating faces in the scene prior to recognition. For this reason, we present now a method for the localization of faces that complements our recognition module.

The method is tailored for its use in factory automation applications where the detection and localization of humans is necessary for the completion or interruption of a particular task. This is particularly useful in robotic workcells that require automatic safety precautions such as speed reduction or sound warnings when a human operator approaches its workspace, or immediate motion interruption if such operator interferes with robot motion. These systems are tailored to multirobot workcells where motion sensors cannot accurately estimate human presence.

Another application field is that of human-machine interaction. It is desirable for a mobile service robot to be able to modify its behavior with respect to its interaction with people. Such is the case of surveillance systems or mobile delivery units that must modify their trajectory in the presence of humans. And ultimately, be able to recognize among different people and behave accordingly.

When no restrictions are imposed on the input images, human face localization can be a challenging task. Apart from scale variation and position uncertainty, there exist other artifacts that make this problem difficult including the a priori ignorance of the pose of the face in the image, i.e., frontal, sideways, or nodded; occlusions of the face by other objects; or the lighting conditions that may change the position of the skin color in the color space. Also, complex backgrounds could lead to inference of false head shapes.

Many approaches have been proposed for the detection and localization of human faces. A survey on face detection methods can be found in [3]. The reader should take into account the distinction in the literature between face localization and face detection. The former is aimed at finding the right position and orientation of a single face in an image, with the prior knowledge that the image does contain a face. This constraint is not necessarily true in the detection problem. Face detection techniques are divided in the following four groups:

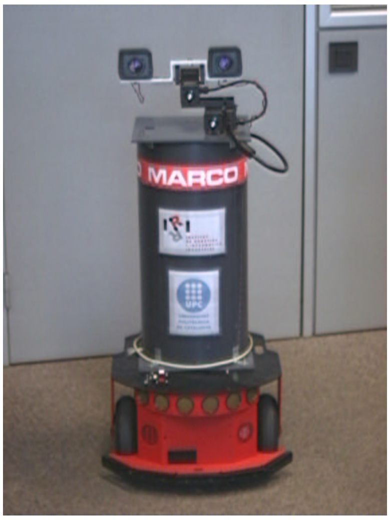

Fig. 1. MARCO mobile robot.

1. *Template matching*. This approach maximizes a correlation function of a human face pattern over the entire image. A sample image window with a face model is initially stored and sometimes normalized and scaled. Then, the saved model is searched on a query image, maximizing a localization hypothesis at the image point where the correlation value is the largest. An extension to this technique includes the use of deformable contours, due to the fact that not all face viewpoints have the same shape. The main drawback of this extension is the time response of deformable models, and its sensitivity to initial conditions and local minima.

2. *Knowledge-based model techniques*. A more general approach consists on describing a model by features that we derive from our knowledge of human faces, and their relation with each other. These features and their relations are typically expressed as sets of rules, and the search for a face on an image consists on the formulation of hypotheses and the verification of these hypotheses with the aid of a decision tree. An example application of this technique can be found in [4], where general rules that describe what a face looks like, and specific rules about the details of facial features are combined in a multiresolution system. One such rule used to find face candidates at a

low resolution level could be *the center part of a face has a region with a basically uniform gray level*. The main drawback of knowledge-based modeling is the necessity of an expert to come up with efficient rules for discerning.
3. *Feature-invariant approach*. This method for face detection also searches for sets of facial features. The difference between this method and the previous one resides on the technique used for feature verification. In the feature-invariant case, face detection candidates are obtained maximizing one or more search criteria, instead of decision rules. In this approach, the kind of features used for face detection are expected to be invariant regardless of the face pose or viewpoint. These features can be either geometric, such as the edges of the frontal view of a face or the curvature of the shape of the face; or based on appearance, i.e., face texture, and most importantly skin color. Recent contributions combine the extraction of various geometric and appearance-based facial features to improve their robustness (most of them use skin color and shape). For example [5] begins the detection stage with the search for skin-like regions, and after a clustering stage, facial candidates are considered in regard to the elliptical or oval shape of a connected region. The problem with this technique is that most of these features even when invariant to size and orientation are still sensitive to lighting conditions, occlusions and noise.
4. *Appearance-based methods*. This set of techniques characterize human faces as topological structures in a multidimensional feature space, namely, the image space. Several training images of a small window containing the same face with small viewpoint variations will usually map to different points in the attribute space, forming a manifold parameterized by pose. Scaled query windows of the candidate image to be analyzed must be projected to the attribute space, and the closer the projection of this window is to the manifold, the greater the probability that the candidate window will correspond to the trained human face. When the feature space is reduced with the aid of principal components, we call this method *eigenfaces* [6]. In [7] on the other hand, a neural network trained to output the presence or absence of a face is directly applied to portions of the input image. Another example is shown in [8], where a probabilistic model for 3D face detection is described with separate detectors tailored to specific face orientations.

We present an approach for face localization using a mixture of two techniques: the segmentation of skin regions through color histograms, and the localization of a head shape model with a known size in the depth map acquired with a stereo vision system. Our approach belongs to the set of feature-invariant techniques labeled above in the sense that we also search for the combination of facial features maximizing several criteria. The novelty of the approach resides on the partitioning of the head shape search space in terms of depth. The size of the model is accurately scaled at various depth slices in such a way that the model searched in a further region will appear smaller than a model searched in a region closer to the viewpoint. Moreover, when a face is detected, we know its exact three-dimensional position with respect to the camera.

Our system operates in three stages. First, two parallel low-level vision modules extract information from the scene. A stereo vision system builds a depth map of the scene; while at the same time, skin-like regions are extracted from the original image with a color histogram segmentation technique. Secondly, the system refines the depth map eliminating those regions that do not correspond to our previously stored skin model. Finally, a correlation based search for a head shape model is performed in the refined depth map. One last verification step analyzes the percentage of skin color pixels on the hypothesized face location according to our skin color model.

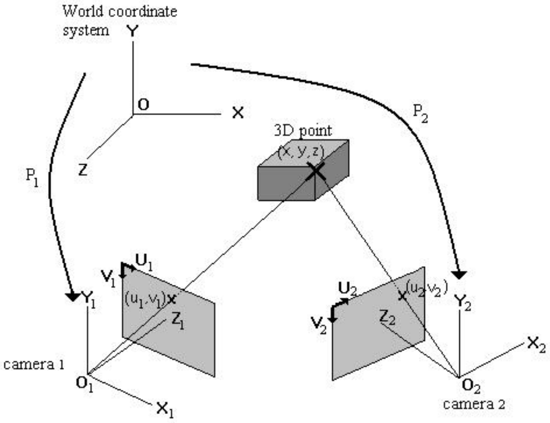

Fig. 2. Stereo geometry for a pair of pinhole cameras.

Detailed descriptions of the stereo vision and color segmentation modules are given in Section 2. In Section 3, the proposed fusion model is derived. Some performance issues and conclusions are presented in Section 4.

II. LOW LEVEL MODULES

A. Depth Estimation

Stereo vision allows us to reconstruct the three-dimensional structure of a scene from its projection in two images.

Given a pair of perspective projection matrices $\mathbf{P}_1$ and $\mathbf{P}_2$ for the stereo vision model depicted in Fig. 2, the projection of the homogeneous coordinates of a point $\tilde{\mathbf{x}} = [x, y, z, 1]^\top$ from the origin of the i-th camera centered frame onto the i-th image plane is given by

$$\tilde{\mathbf{m}}_i = \mathbf{P}_i \tilde{\mathbf{x}} \qquad (1)$$

where $\tilde{\mathbf{m}}_i = [U_i, V_i, s_i]^\top$ are the homogeneous coordinates of that projection line. The intersection of $\tilde{\mathbf{m}}_i$ and the i-th image plane is given by $\mathbf{m}_i = [u_i, v_i, 1]^\top$.

Reconstruction consists on solving for $\tilde{\mathbf{x}}$ given a pair of image correspondences $\mathbf{m}_i$ and the known camera projection matrices $\mathbf{P}_i$. The technique is applicable not only to a pair of views, but for as many cameras as desired, provided the point correspondences have been found. When two or more cameras are used, the problem is overconstrained, and we choose the solution that minimizes the sum of the squared distances from $\tilde{\mathbf{x}}$ to the corresponding lines $\tilde{\mathbf{m}}_i$ in $\mathbb{P}^2$.

$$\tilde{\mathbf{x}} = (\mathbf{A}^\top \mathbf{A})^{-1} \mathbf{A}^\top \mathbf{b} \qquad (2)$$

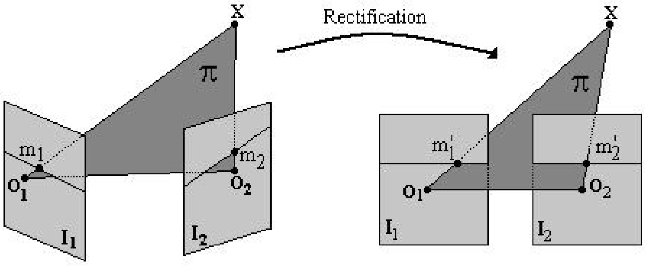

Fig. 3. Image rectification brings epipolar lines to collinearity, therefore reducing the dimensionality of the point correspondence search space.

where for the stereo case

$$\mathbf{A} = \begin{bmatrix} u_1 \mathbf{p}^\top_{<3>1} - \mathbf{p}^\top_{<1>1} \\ v_1 \mathbf{p}^\top_{<3>1} - \mathbf{p}^\top_{<2>1} \\ u_2 \mathbf{p}^\top_{<3>2} - \mathbf{p}^\top_{<1>2} \\ v_2 \mathbf{p}^\top_{<3>2} - \mathbf{p}^\top_{<2>2} \end{bmatrix} \tag{3}$$

$$\mathbf{b} = \begin{bmatrix} p_{<14>1} - u_1 p_{<34>1} \\ p_{<14>1} - v_1 p_{<34>1} \\ p_{<14>2} - u_2 p_{<34>2} \\ p_{<14>2} - v_2 p_{<34>2} \end{bmatrix} \tag{4}$$

and the perspective projection matrix associated with the i-th camera is written in the following form

$$\mathbf{P}_i = \begin{bmatrix} \mathbf{p}^\top_{<1>i} & p_{<14>i} \\ \mathbf{p}^\top_{<2>i} & p_{<24>i} \\ \mathbf{p}^\top_{<3>i} & p_{<34>i} \end{bmatrix} \tag{5}$$

The steps that we have used to implement our stereo vision system, are the following:

1. Calibration of the stereo head. In this stage the camera matrices $\mathbf{P}_i$ are determined using the calibration algorithm detailed in [11].

2. Image rectification. Once the perspective projection matrices are computed, we transform the images in such a way that epipolar lines become collinear in each pair of images, see Fig. 3. By performing this transformation corresponding image points can be searched for in the same scanline, therefore reducing the dimensionality of the search space from two dimensions to one. To rectify our images, we have implemented the algorithm presented in [12].

3. Solution of the correspondence problem. The most critical step in any stereo vision system is the solution of the correspondence problem, i.e., identifying the projection of the same 3D point in the two images. To cope with it, we have used a correlation method that takes the grey level of a neighborhood around an interest pixel in one image, and searches for the pixel location with similar grey distribution in the other image. For rectified images, the search is performed along the same scanline, i.e., over collinear epipolar lines. In this application we have resorted to the *sum of absolute differences* operator to compute the similarity between the area around possible matching pixels in the grayscale version of the left and right images. The location of a matching pixel from image i to image j is given by

$$u_j = u_i + \arg\min_d \sum_{u,v \in W_i} |I_i(u,v) - I_j(u+d,v)| \tag{6}$$

where d is the disparity displacement along the direction of u on each rectified image, W_i is the search area window, I_i is the entire reference image, and I_j is the search image.

4. Depth map refinement. The last step in the implementation of our stereo vision module is the refinement of the correspondence computation. The following three refinement operations take place:

(a) Since the sum of absolute differences does not provide accurate disparity computation for homogeneous regions, we eliminate from the depth map building process those pixel locations whose gradient is lower than a given threshold t by convolving the original images with a set of interest operators W_j.

$$\min_j (I_i * W_j) \leq t \tag{7}$$

with

$$W_1 = \begin{bmatrix} \mathbf{1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & -\mathbf{1} \end{bmatrix} \quad W_2 = \begin{bmatrix} \mathbf{0} & \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & -\mathbf{1} & \mathbf{0} \end{bmatrix}$$

$$W_3 = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{1} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\mathbf{1} & \mathbf{0} & \mathbf{0} \end{bmatrix} \quad W_4 = \begin{bmatrix} \mathbf{0} & \mathbf{0} & \mathbf{0} \\ -\mathbf{1} & \mathbf{0} & \mathbf{1} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} \end{bmatrix}$$

and $\mathbf{1}$ and $\mathbf{0}$ are the $k \times k$ matrices of ones and zeros respectively, and k is a power of 3. When $k = 1$, it is called Moravec's interest operator. The size of k and the value of t will depend on the quantity of detail and the quality of the images for a particular setting.

(b) Left to right image comparison. To overcome the difficulties that arise on any stereo application due to occlusions or illumination variations, we exclude from the analysis those correspondences that have distinct disparity values computed in both directions with Eq. 6.

(c) Interpolation. The two refinement techniques used above might produce holes on the disparity map. The recovery of depth values for these artificial holes is computed interpolating over the average depth value on the neighboring points of each hole.

B. Depth Segmentation

Once we have determined the depth map of an image, we perform a depth segmentation step in order to divide the scene into sections according to their distance to the camera. Depth segmentation allows for the separation of objects over a desired depth range, even when they posses similar color and texture properties, provided enough detail is extracted to accurately estimate their distance to the camera. This technique is particularly useful for the extraction of objects of interest from a complex background. In our application to face localization, we will use this technique to extract the human head in an image.

Moreover, when depth segmentation is present, the size of each extracted region can be estimated from its projection in both images, resulting in an accurate location of the head in 3D space.

Fig. 4 shows sample result images of the depth extraction process where brighter points correspond to points with smaller depth values and further apart points appear dark. The figure also shows some slices of the image at distinct depth values. Note how in some of these images the human head is extracted from the background.

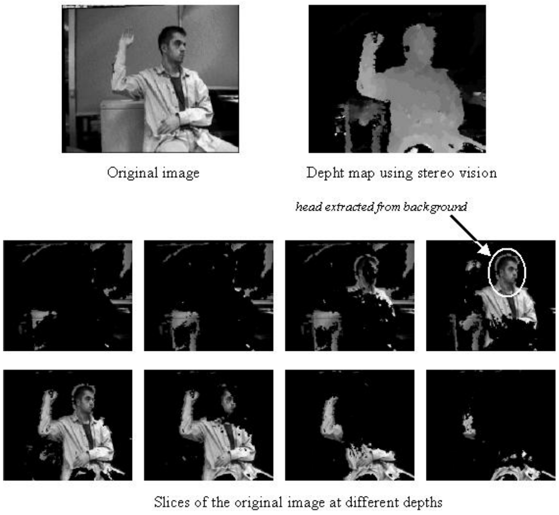
Original image
Depht map using stereo vision
head extracted from background
Slices of the original image at different depths

Fig. 4. Depth map and scene segmentation at various depth intervals.

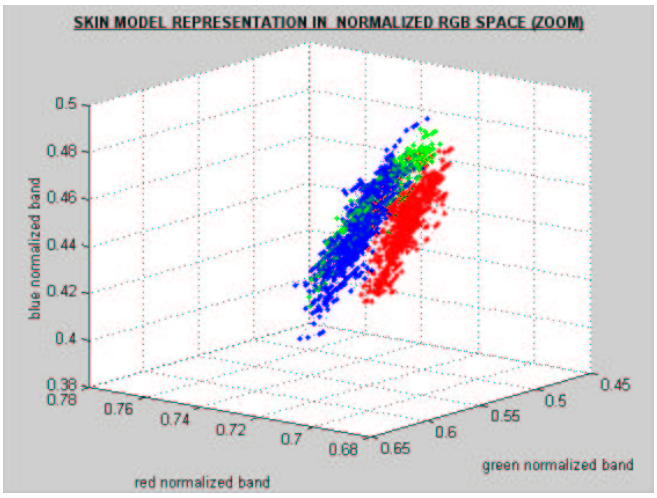

Fig. 5. Skin model distributions in the normalized RGB space for three different training sets.

C. Color Segmentation

While the depth map of the scene is computed, an independent process takes place over the same input images for the extraction of skin color regions. We have developed an algorithm for skin-like supervised region classification based on the computation of color histograms, and the lookup for a match on a hash-table made up of a subsampled version of this histogram.

During an off-line training session, the system is provided with user-selected regions of skin texture extracted from various images of human faces under different illumination conditions. To reduce the sensitivity of the system to the illumination source, the RGB color values for each pixel on these sample windows is normalized with

$$\begin{bmatrix} \bar{R} \\ \bar{G} \\ \bar{B} \end{bmatrix} = \frac{1}{\sqrt{R^2+G^2+B^2}} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{8}$$

and stored in fixed size buckets on a normalized RGB histogram. The discretization of the histogram space depends on the task at hand. If for example, we wish to differentiate among different classes of skin texture or among various illumination settings, a fine discretization is required. If on the other hand, we only wish to recognize skin-like texture from images, the bucket size can be larger. After several empirical tests, we found that for our particular application, a suitable number of buckets that achieved robust skin-color characterization on the normalized histogram space was 25^3. Fig. 5 depicts a three-dimensional distribution of non-empty buckets on the histogram space for a set of various training samples of skin-color taken from different people under varying illumination conditions.

Only after training the system with a skin color model, we can start the online process of image segmentation of skin-like regions. For each pixel on an image we must calculate its normalized RGB value with Eq. 8. Skin classification is achieved if this pixel color value falls on a non-empty bucket on the trained hash-table [13].

One last morphological operation is necessary in order to fill for those pixels on the sample image that, even when they fall inside the face area, their pixel color values fall onto an empty bucket in the color histogram. This can be achieved with just one dilation over the resulting image. Fig. 6 shows several skin segmentation results.

III. FUSION MODEL

In this section we will describe how a human head is characterized in terms of its geometry and its color properties. Also, the searching of this model on input images is explained, paying special attention to the data fusion aspects of the different low-level image processing modules.

A. Model Definition

Our face model consists on two distinctive features. One of them is purely geometric, i.e., the shape of the head, which size is modified under perspective projection as the search is performed at different depth ranges on the scene. The head and neck silohuete follows the model depicted in Fig. 7. In our implementation, we have considered human heads with parameters $a \approx 200mm$ and $b \approx 300mm$; and the size of the window containing the head shape plus a section of the neck is of size $1.2a \times 1.2b$.

The second feature is the skin-color information modeled as a distribution over a normalized RGB histogram. The collection of training samples of skin-color under various illumination conditions was explained in Section 2.3.

B. Model Search

Once both a depth map has been computed, and a color region segmentation has been obtained from a given scene, we need to search for our model. The original depth map is filtered with the trained color information. This reduced depth map is used for the search of our head model. The search is performed at various depth intervals starting from further regions and approaching the camera viewpoint.

Two special considerations are taken into account during this search. First, the interval to be analyzed from the color-reduced depth map is constrained to a distance of approximately 400mm

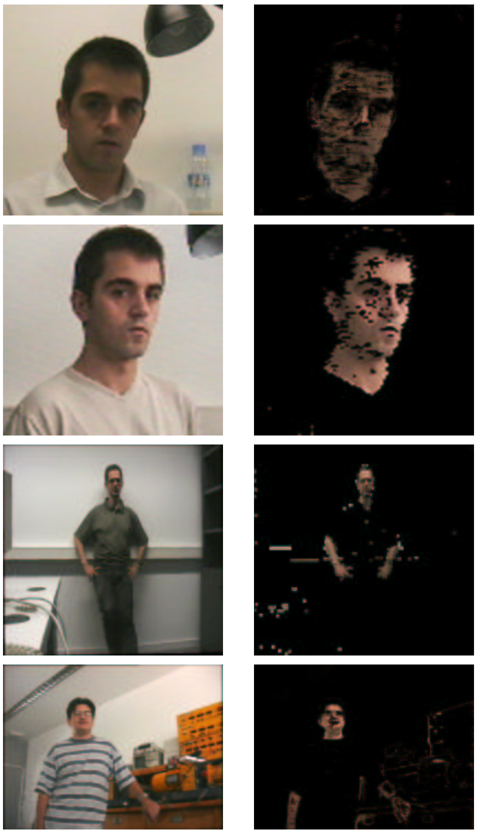

Fig. 6. Various skin segmentation samples.

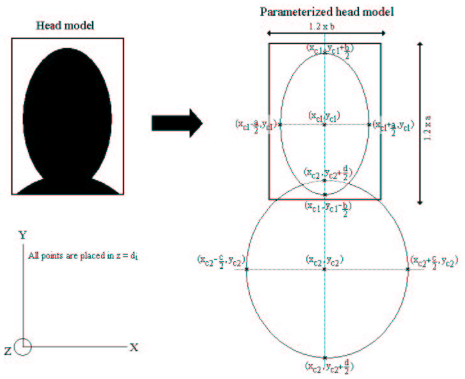
Head model
Parameterized head model
1.2 x b
1.2 x a
$(x_{c1}, y_{c1}+\frac{b}{2})$
(x_{c1}, y_{c1})
$(x_{c1}-\frac{a}{2}, y_{c1})$
$(x_{c1}+\frac{a}{2}, y_{c1})$
$(x_{c2}, y_{c2}+\frac{d}{2})$
$(x_{c1}, y_{c1}-\frac{b}{2})$
$(x_{c2}-\frac{c}{2}, y_{c2})$
(x_{c2}, y_{c2})
$(x_{c2}+\frac{c}{2}, y_{c2})$
$(x_{c2}, y_{c2}+\frac{d}{2})$
Y
All points are placed in $z = d_i$
Z
X

Fig. 7. Parameterized head model.

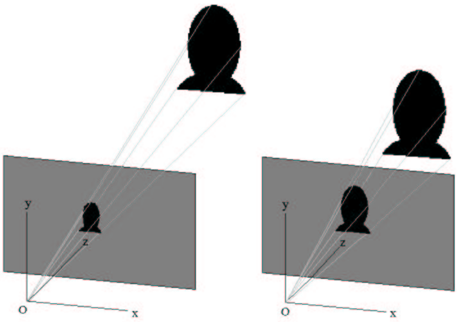

Fig. 8. The perspective projection of the parameterized head model creates different size correlation windows according to the depth slice to be processed.

at each depth interval. To do so, our color-reduced depth map is filtered using the appropriate projection matrix, providing as a result an even more constrained version of the depth map with not only the desired color range, but the desired depth slice as well.

The second consideration consists on reprojecting the head model onto this slice, so that the shape of the head and the size of the search window are modified dynamically as we move from depth slice to depth slice, see Fig. 8. Both our model and the sliced depth map are binarized prior to the computation of a match score with the following normalized correlation equation

$$C(u,v) = \frac{\sum IM - \sum I \sum M}{\sqrt{\left(\sum I^2 - (\sum I)^2\right)\left(\sum M^2 - (\sum M)^2\right)}} \qquad (9)$$

where I represents the reduced and binarized depth map position $I(u+i, v+j)$, and M indicates the projected model position $M(i,j)$. All summations take place on the interval $(i,j) \in W$, with W being the dynamically modified search window.

The partitioning of the depth map may lead to false formation of head shapes. One last verification step is performed to overcome this issue. The correlation value $C(u,v)$ is augmented with the percentage $P(u,v)$ of pixel values in the search window W that fall within the trained skin color histogram model according to the linear form

$$s(u,v) = tC(u,v) + (1-t)P(u,v) \qquad (10)$$

By changing the value of the parameter t in the range $0 \leq t \leq 1$ the user can give more or less importance to either the geometry of the model or its color property. The resulting hypothesized head location (u,v) with the largest matching score s is considered as the most probable position for a human face on the scene. Fig. 10 shows the various face localization steps over a sample image.

IV. CONCLUSIONS

We have presented a method to localize faces in color images based on the fusion of the information gathered from a stereo vision system and the analysis of color images. Our method generates a depth map of the scene and tries to fit a head model taking into account the shape of the model and skin color information.

The fusion of different data acquisition modules simplifies considerably the search complexity of a model to scene match, therefore reducing the possibility of incurring on false matches. The shortcomings of the individual low-level processing modules are overcome in an integrated environment. However, the inherent variability of these low-level modules and the data formats and noise levels they produce make sensor fusion a challenging task.

One of the particularities that make our model robust is that on images containing human faces, the depth map covering the region close to the face is usually dense since human faces contain enough detail information. So even for those cases when the initial depth map is sparse or contains false matches due to the homogeneity of certain regions or occlusions, the region containing a human head will still be well defined on the depth

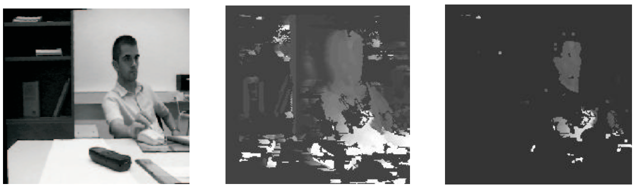

Fig. 9. The shortcomings of individual low-level processing modules are overcome in an integrated environment.

map. Fig. 9 shows an example illustration of such condition, where the initial depth map contains many false matches. However head localization is still possible.

On the other hand, large regions on an image with color similar to our skin model will also be partitioned thanks to the slicing of the depth map, thus reducing the possibility of false matches due to color similarity.

V. Acknowledgments

Work partially funded by CICYT project TAP98-0473, CONACYT and CETYS.

References

[1] A. Sanfeliu, J. Andrade-Cetto, R. Alquézar, J. Aranda, J. Climent, A. Grau, F. Serratosa, and J. Vergés-Llahí. MARCO: A mobile robot with learning capabilities to perceive and interact with its environment. In J. S. Sánchez and F. Pla, editors, *Proc. 9th Spanish Sym. Pattern Recog. Image Anal.*, volume 2, pages 219–224, Benicasim, May 2001.

[2] J. Vergés-Llahí, A. Sanfeliu, F. Serratossa, and R. Alquézar. Face recognition: Graph matching versus neural techniques. In M. I. Torres and A. Sanfeliu, editors, *Proc. 8th Spanish Sym. Pattern Recog. Image Anal.*, pages 259–266. Ediciones Geneve, 1999.

[3] M-H. Yang, N. Ahuja, and D. Kriegman. A survey on face detection methods. Draft. Computer Vision Lab. University of Illinois Urbana-Champaign, Mar. 1999.

[4] G. Yang and T. S. Huang. Human face detection in a complex background. *Pattern Recogn.*, 27(1):53–63, 1994.

[5] K. Sobottka and I. Pitas. Face localization and facial feature extraction based on shape and color information. In *Proc.IEEE Int. Conf. Image Process.*, volume 3, pages 483–486, Lausanne, Sep. 1996.

[6] M. Turk and A. Pentland. Eigenfaces for recognition. *J. Cognitive Neurosci.*, 3(1):71–86, 1991.

[7] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE T. Pattern Anal.*, 20(1):23–38, 1998.

[8] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces and cars. In *Proc.IEEE Conf. Comput. Vision Pattern Recog.*, volume 1, pages 746–751, Head Island, Jun. 2000.

[9] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge, 2000.

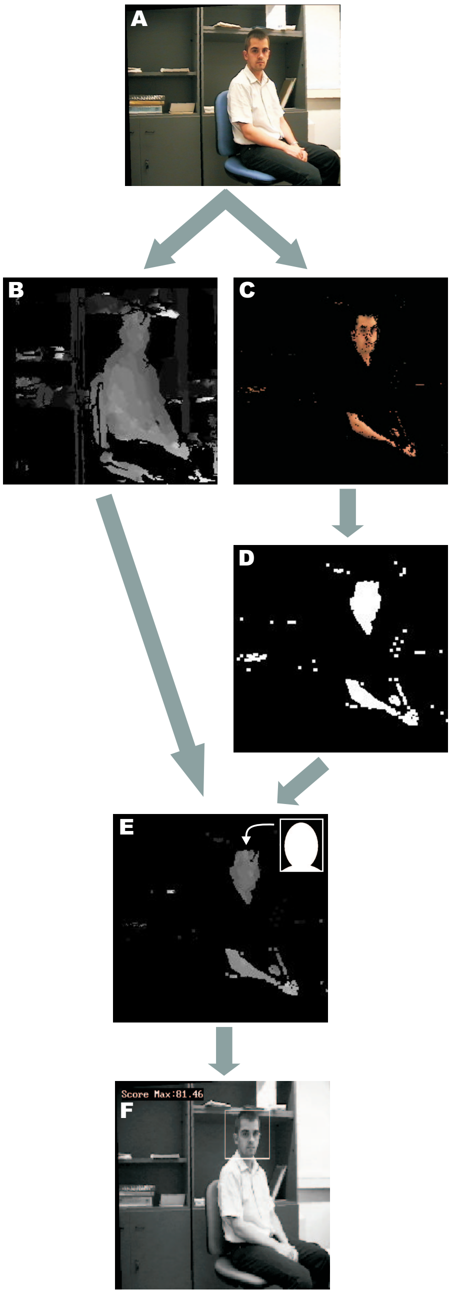

Fig. 10. A) Original Image. B) Depth map. C) Color segmentation of skin-like regions. D) Skin-like color regions after morphological processing. E) Fusion of depth and color information. F) Face localization.

[10] O. Faugeras. *Three-Dimensional Computer Vision. A Geometric Viewpoint*. The MIT Press, Cambridge, 1993.

[11] J. Andrade-Cetto. Camera calibration. Technical Report IRI DT 2001/2, IRI, UPC-CSIC, Jun. 2001.

[12] A. Fusiello, E. Trucco, and A. Verri. Rectification with unconstrained stereo geometry. In A. F. Clark, editor, *Proc.British Machine Vision Conf.*, pages 400–409, Colchester, Sep. 1997.

[13] J. Vergés-Llahí, J. Aranda, and A. Sanfeliu. Object tracking system using colour histograms. In J. S. Sánchez and F. Pla, editors, *Proc. 9th Spanish Sym. Pattern Recog. Image Anal.*, volume 2, pages 225–230, Benicasim, May 2001.