FUSION OF MULTIPLE VIEWPOINT INFORMATION TOWARDS 3D FACE ROBUST ORIENTATION DETECTION

C.Canton-Ferrer, J.R.Casas, M.Pardàs

Image Processing Group
Technical University of Catalonia

ABSTRACT

This paper presents a novel approach to the problem of determining head pose estimation and face 3D orientation of several people in low resolution sequences from multiple calibrated cameras. Spatial redundancy is exploited and the heads of people in the scene are detected and geometrically approximated by an ellipsoid using a voxel reconstruction and a moment analysis method. Skin patches from each detected head are located in each camera view. Data fusion is performed by back-projecting skin patches from single images onto the estimated 3D head model, thus providing a synthetic reconstruction of the head appearance. Finally, these data are processed in a pattern analysis framework thus giving a reliable and robust estimation of face orientation. Tracking over time is performed by Kalman filtering. Results are provided showing the effectiveness of the proposed algorithm in a SmartRoom scenario.

1. INTRODUCTION

The current paper addresses the problem of detecting and tracking the head of people present in a SmartRoom and estimating the orientation of their faces in the framework of multiple view geometry. Multi camera systems are widely used for image and video analysis tasks in SmartRooms, surveillance, body analysis or computer graphics. From a mathematical viewpoint, multiple view geometry has been addressed in [1] , but there is still work to do for the efficient fusion of information from redundant camera views and its combination with image analysis techniques for object detection, tracking or higher semantic level analysis such as attitudes and behaviors of individuals.

A number of methods for head pose estimation has been proposed in the literature [2]. The general approach involves estimating the position of specific facial features in the image (typically eyes, nostrils and mouth) and then fitting these data to a head model. The accuracy and reliability of the feature extraction process plays an important role in the head pose estimation results. In practice, many of these methods still require manually selecting feature points, as well as assuming that near-frontal views and high-quality images are available. For the applications addressed in our work, such conditions are usually difficult to satisfy. Specific facial features are typically not clearly visible due to lighting conditions and wide angle camera views. They may also be entirely unavailable when faces are not oriented towards the cameras. Methods which rely on a detailed feature analysis followed by head model fitting would fail under these circumstances. Furthermore, most of the existing approaches are based on monocular analysis of images but few have addressed the multiocular case for face or head analysis [3].

This material is based upon work partially supported by the IST programme of the EU through the IP IST-2004-506909 CHIL and by TEC2004-01914 project of the Spanish Government.

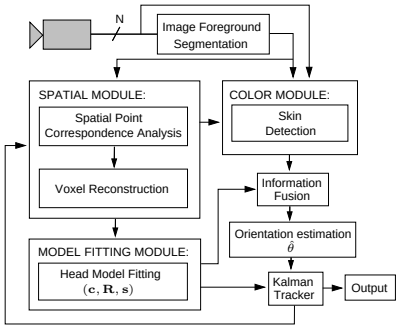


Fig. 1. System flowchart: acquisition, spatial and color analysis, ellipsoid model fitting, face orientation estimation and tracking.

We propose a method for 3D face orientation estimation which is both robust to enviromental conditions and computationally simple for real-time applications. Redundancy among camera views is exploited to obtain robust estimations of head 3D positions and to fit a model of the head. A fusion process of color and spatial information from all cameras is performed obtaining a synthetic reconstruction of face appearance in 3D. Finally, two analysis methods on these data are proposed in order to obtain the orientation of the face.

This method has been successfully applied to a multi-camera SmartRoom scenario in the framework of a scene understanding project involving tracking of attention in meetings. Other fields where our algorithm has potential applicability are vehicle driver attention tracking, disabled people interfaces and face recognition.

2. LOW LEVEL SIGNAL ANALYSIS MODULES

According to the flowchart depicted in Fig.1 the system comprises four low level image processing modules: image acquisition, spatial and color analysis and head model fitting. These modules provide data to the higher level analysis module that performs the information fusion required to estimate the orientation of heads and faces, and to the Kalman tracking module as well.

For a given frame in the video sequence, a set of N images are obtained from the N cameras. Each camera is modeled using a pinhole camera model based on perspective projection. Accurate calibration information is available. Foreground regions from

input images are obtained using a segmentation algorithm based on Stauffer-Grimson's background learning and substraction technique [4]. It is assumed that the moving objects are human people. Original and segmented images are the input information for the rest of image analysis modules described here.

The final low level signal analysis module employed in our system is a standard Kalman tracker with a constant velocity model. With respect to our model of parameters evolution, it computes the predictions and adds the information coming from the measurements in an optimal way to produce a posteriori estimations of the parameters. Moreover, the tracking loop helps rejecting false detections and dealing with oclusions. The tracked parameters are the geometric parameters defining the head and the estimated face orientation angle. For the initialization of this filter, hand marked sequences were analyzed in order to estimate the noise correlation matrices.

2.1. Spatial Analysis Module

Prior to any further image analysis, the analyzed scene must be characterized in terms of space disposition and configuration of the foreground volumes, i.e. people candidates, in order to select those potential 3D regions where the head of a person could be present. Images obtained from a multiple view camera system allow exploiting spatial redundancies in order to detect these 3D regions of interest. This task is carried out by the spatial analysis module.

Once foreground regions are extracted from the set of N original images at time t, a set of M 3D points $\mathbf{x}^k$, $0 \leq k < M$, corresponding to the top of each 3D detected volume in the room is obtained by applying a robust Bayesian correspondence algorithm described in [5]. Information coming from the tracking loop speeds up the process narrowing the search space of these correspondences on time $t+1$ and allows rejecting false head detections.

The information given by the established correspondences allows defining a bounding box $\mathcal{B}^k$, centered on each 3D top $\mathbf{x}^k$ with an average size adequate to contain the human head candidate. Afterwards, a voxel reconstruction [6] is computed on each bounding box $\mathcal{B}^k$, thus obtaining a set of voxels $\mathcal{V}^k$ defining the k-th 3D foreground volume candidate as a head. In order to refine and verify whether the set $\mathcal{V}^k$ indeed belongs to an ellipsoidal geometric shape, a template matching evaluation [6] is performed. Results for this module are shown in Fig.2(a) and 2(b).

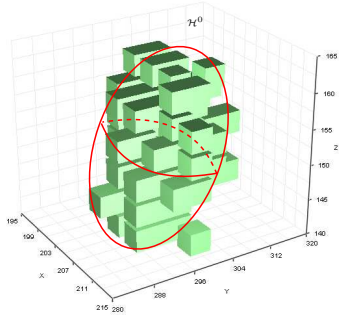
2.2. Color Module

Foreground regions detected by the segmentation algorithm provide 2D masks within the original images where skin color pixels are sought. The masked original images are processed in the CbCr color space since different skin types mostly differ in the luminance component and not with regard to the hue value. Afterwards, a probabilistic classification is computed on the CbCr information [7] where the color distribution of skin is estimated from offline hand selected samples of skin pixels in the same light conditions of the online experiments and approximated by a Gaussian function.

Finally, color information is combined with spatial information obtained from the former module. For each pixel classified as skin, $\mathrm{p}^n_{\mathrm{skin}}$, in the view n, $0 \leq n < N$, we check whether

$$\mathrm{p}^n_{\mathrm{skin}} \in P_n\left(\mathcal{V}^k\right), \qquad 0 \leq k < M, \tag{1}$$

(a)



(b)

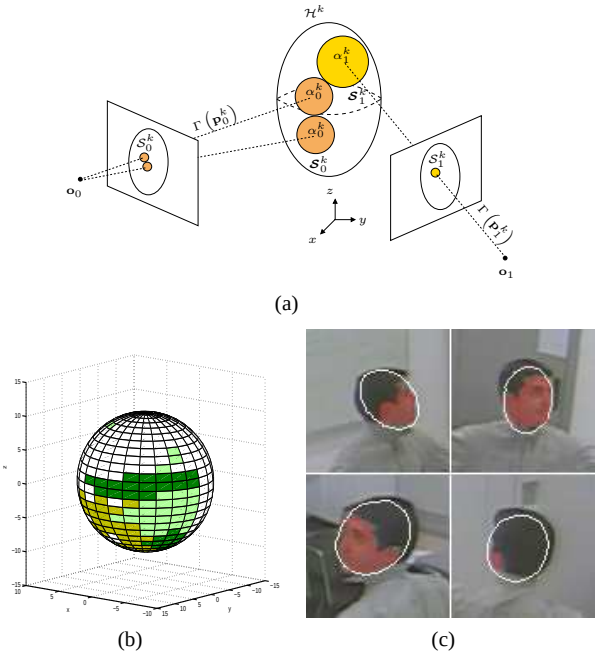
Fig. 2. Example of the outputs from the spatial analysis and model fitting modules. In (a), multiview correspondences among heads are correctly established. The projection of the bounding box $\mathcal{B}^0$ containing the head is depicted in white. In (b), voxel reconstruction is applied to $\mathcal{B}^0$ thus obtaining the voxels belonging to the head (green cubes). Model fitting module result is depicted in red.

where $P_n(\cdot)$ is the perspective projection operator from 3D to 2D coordinates on the view n [1]. In this way $\mathrm{p}^n_{\mathrm{skin}}$ can be identified as being a projection of a voxel of the set $\mathcal{V}^k$ and therefore correctly handled when establishing orientation of multiple heads and faces in later modules. Let us denote with $\mathcal{S}^k_n$ all skin pixels in the n-th view classified as belonging to the k-th voxel set. It should be recalled that there could be empty sets $\mathcal{S}^k_n$ due to occlusions or under-performance of the skin detection technique. However, tracking information and redundancy among views would allow to overcome this problem.

2.3. Head Model fitting

In order to achieve a good fitting performance, a geometrical 3D configuration of human head must be considered. For our research work, an ellipsoid model of human head shape has been adopted. In spite of this fairly simple approximation compared to more complex geometries of head shape [2], head fitting still achieves enough accuracy for our purposes (see Fig.2(b) for an example).

Let $\mathcal{H}^k = \{\mathbf{c}^k, \mathbf{R}^k, \mathbf{s}^k\}$ be the set of parameters that define the ellipsoid modelling the k-th detected human head candidate where $\mathbf{c}^k$ is the center, $\mathbf{R}^k$ the rotation along each axis centered on $\mathbf{c}^k$ and $\mathbf{s}^k$ the length of each axis. After obtaining the set of voxels $\mathcal{V}^k$ belonging to k-th candidate head $\mathcal{H}^k$, the ellipsoid shell modelling is fit to these voxels. Statistic moment analysis is em-



(a) (b) (c)

Fig. 3. In (a), color and spatial information fusion process scheme. Pixels in the set $\mathcal{S}_n^k$ are back-projected onto the surface of the ellipsoid defined by $\mathcal{H}^k$, generating the set $\boldsymbol{\mathcal{S}}_n^k$ with its weightening term α_n^k. In (b), result of information fusion obtaining a synthetic reconstruction of face appearance from images in (c) where the skin patches are plot in red and the ellipsoid fitting in white.

ployed to estimate the parameters of the ellipsoid from the centers of the marked voxels thus obtaining a 3D spatial mean $\bar{\mathcal{V}}^k$ and a covariance matrix $\mathbf{C}_{\mathcal{V}^k}$. The covariance can be diagonalized via an eigenvalue decomposition into $\mathbf{C}_{\mathcal{V}^k} = \mathbf{\Phi}\mathbf{\Delta}\mathbf{\Phi}^\top$, where $\mathbf{\Phi}$ is orthonormal and $\mathbf{\Delta}$ is diagonal. Identification of the defining parameters of the estimated ellipsoid $\mathcal{H}^k$ with moment analysis parameters is then straighforward:

$$\mathbf{c}^k = \bar{\mathcal{V}}^k, \qquad \mathbf{R}^k = \mathbf{\Phi}, \qquad \mathbf{s}^k = \mathrm{diag}(\mathbf{\Delta}). \tag{2}$$

3. MULTIPLE VIEW COLOR AND SPATIAL INFORMATION FUSION

Fusion of both color and space information is required in order to perform a high semantic level classification and estimation of face orientation. Our information fusion procedure takes as input the information generated from the low level image analysis for each person: an ellipsoid estimation $\mathcal{H}^k$ of the head and a set of skin patches at each view belonging to this head $\{\mathcal{S}_n^k\}$, $0 \leq n < N$. The output of this technique is a fusion of color and space information set denoted as Ω^k. Analysis techniques of the data contained in Ω^k are provided in Sec.4.

The procedure of information fusion we define is based on the assumption that all skin patches $\{\mathcal{S}_n^k\}$ are projections of a region of the surface of the estimated ellipsoid defining the head of a person. Hence, color and space information can be combined to produce a synthetic reconstruction of the head and face appearance in 3D. This fusion process is performed for each head separately starting by back-projecting the skin pixels of $\mathcal{S}_n^k$ from all N views onto the k-th 3D ellipsoid model. Formally, for each pixel $\mathrm{p}_n^k \in \mathcal{S}_n^k$, we compute

$$\Gamma\left(\mathrm{p}_n^k\right) \equiv P_n^{-1}\left(\mathrm{p}_n^k\right) = \mathbf{o}_n + \lambda\mathbf{v}, \qquad \lambda \in \mathbb{R}^+, \tag{3}$$

thus obtaining its back-projected ray in the world coordinate frame passing through p_n^k in the image plane with origin in the camera center $\mathbf{o}_n$ and director vector $\mathbf{v}$. In order to obtain the back-projection of p_n^k onto the surface of the ellipsoid modelling the k-th head, Eq.3 is substituted into the equation of an ellipsoid defined by the set of parameters $\mathcal{H}^k$ [1]. It gives a quadratic in λ,

$$a\lambda^2 + b\lambda + c = 0. \tag{4}$$

The case of interest will be when Eq.4 has two real roots. That means that the ray intersects the ellipsoid twice in which case the solution with the smaller value of λ will be chosen for reasons of visibility consistency. See a scheme of this process on Fig.3(a).

This process is applied to all pixels of a given patch $\mathcal{S}_n^k$ obtaining a set $\boldsymbol{\mathcal{S}}_n^k$ containing the 3D points being the intersections of the back-projected skin pixels in the view n with the ellipsoid surface. In order to perform a joint analysis of the sets $\{\boldsymbol{\mathcal{S}}_n^k\}$, each set must have an associated weighting factor that takes into account the real surface of the ellipsoid represented by a single pixel in that view n. That is, to quantize the effect of the different distances from the center of the object to each camera. This weighting factor α_n^k can be estimated by projecting a sphere with radius $r = \max(\mathbf{s}^k)$ on every camera plane, and computing the ratio between the appeareance area of the sphere and the number of projected pixels. To be precise, α_n^k should be estimated for each element in $\boldsymbol{\mathcal{S}}_n^k$ but, since the *far-field* condition

$$\max(\mathbf{s}^k) \ll \|\mathbf{c}^k - \mathbf{o}_n\|_2, \qquad \forall n, \tag{5}$$

is fulfilled, α_n^k can be considered constant for all intersections in $\boldsymbol{\mathcal{S}}_n^k$. A schematic representation of the fusion procedure is depicted in Fig.3(a). Finally, after applying this process to all skin patches we obtain a fusion of color and spatial information set $\Omega^k = \{\boldsymbol{\mathcal{S}}_n^k, \alpha_n^k, \mathcal{H}^k\}$, $0 \leq n < N$, for every head in the scene. A result of this fusion is shown in Fig.3(b).

4. HEAD AND FACE ORIENTATION

The final part of our system deals with the identification of head and face orientation using the output data of the previous fusion method. The angle of interest to be estimated for our purposes in a SmartRoom scenario has been chosen as a direction onto the xy plane. Since this angle gives information about where the people is looking at in the scene, it can be used for further analysis such as tracking of attention in meetings [8]. Two methods have been proposed and tested in this paper in order to estimate the value of the orientation angle $\hat{\theta}$. The performance of these two estimators is addressed in Sec.5.

4.1. Weighted centroid

A first estimation method of the orientation angle $\hat{\theta}$ would be the computation of the weigthed centroid of the fusion data Ω^k as

$$\mathbf{d}^k = \frac{1}{\sum_{n=0}^{N-1} |\boldsymbol{\mathcal{S}}_n^k|} \sum_{n=0}^{N-1} \alpha_n^k \sum_{\mathbf{p}_n^k \in \boldsymbol{\mathcal{S}}_n^k} \left(\mathbf{p}_n^k - \mathbf{c}^k\right), \tag{6}$$

$$\hat{\theta}^k = \tan^{-1}\left(\mathbf{d}_y^k / \mathbf{d}_x^k\right), \tag{7}$$

where $|\boldsymbol{\mathcal{S}}_n^k|$ denotes the number of elements (pixels) in the set.

4.2. Weighted histogram

A more robust estimation of head orientation is based on the computation of a weighted histogram. The sets $\boldsymbol{\mathcal{S}}_n^k$ containing the 3D Euclidean coordinates of the ray-ellipsoid intersections are transformed on the plane $\theta\phi$, in elliptical coordinates with origin at $\mathbf{c}^k$, describing the surface of $\mathcal{H}^k$. Then, the histogram on the axis θ over bins of width $2\pi/L$ is computed obtaining $\mathbf{H}_n^k(i)$, $0 \leq i < L$, for every $\boldsymbol{\mathcal{S}}_n^k$. A weighted histogram is computed as:

$$\hat{\mathbf{H}}^k(i) = \sum_{n=0}^{N-1} \alpha_n^k \mathbf{H}_n^k(i). \quad (8)$$

Finally, the estimation of the orientation is found by smoothing the histogram and computing the maximum over i.

5. RESULTS

In order to evaluate the performance of the proposed fusion method and the analysis technique, we employed it to determine the head and face orientation of a person performing a head movement spaning 180^o. The analysis sequences were recorded with 4 fully calibrated wide angle lense cameras in the SmartRoom at UPC with a resolution of 768x576 pixels at 25 fps (see a sample in Fig.2(a)).

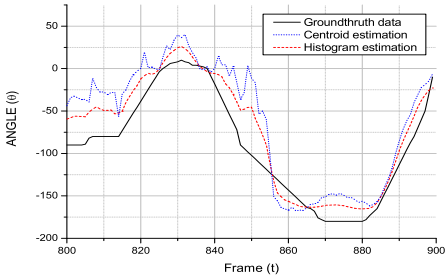
Fusion of the low resolution data obtained from the 4 cameras was fed into the two face orientation angle estimators thus producing the results depicted in Fig.4(a). Groundtruth data was labeled manually in order to compare the performance of the two estimators and both methods showed effective results. It can be seen that the weighted histogram estimation gives results closer to the groundtruth due to the dimension reduction and smoothing of the incoming data. On the other hand, the centroid estimation method achieves less accurate estimations but is 3:1 less computationally expensive than the histogram method. Results for multiple people face orientation algorithm (centroid) in a meeting environment are shown in Fig.4(b).

Within the scenario of low resolution video sequences where facial features can not be accurately detected our method proved to be efficient for our purposes. Finally, it must be pointed out that the performance of these methods is conditioned by the hair style and the presence of beard.

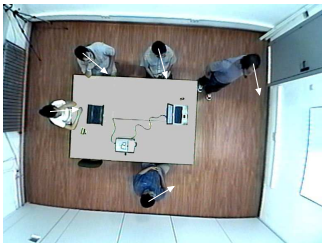
6. CONCLUSIONS AND FUTURE WORK

We presented an efficient technique for color and space information fusion in a multiple camera views environment. This technique allows integrating color information coming from different sources (views) onto an estimated 3D ellipsoid model of the head thus building up a synthetic reconstruction of the head appearance. This method has proven to produce reliable and robust data since this fusion exploits the redundancy among information sources even with the constrain of low resolution video data.

In this framework, we proposed two algorithms to estimate the orientation angle of faces in a SmartRoom environment based on the analysis of the data produced by this fusion method. Both algorithms have proven to generate reliable results as depicted in Fig.4(a) and 4(b). While we have yet to match the performance that existing methods obtain with high quality images, the results are nonetheless sufficiently accurate to be useful for automated behavioral analyses. Real-time performance is achieved at 2 fps in a 1.6Ghz Pentium.



(a)



(b)

Fig. 4. In (a), results of the two face orientation estimators (in degrees) described in this paper. In (b), multiple face orientation technique applied to a meeting scene towards attention tracking (zenital view).

Future research within this topic involve analysis of the fusion data towards tracking attention of people in meetings and understanding behaviors of individuals. Fusion of color, shape and audio information is currently under research towards multimodal analysis.

7. REFERENCES

[1] R.I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, Cambridge University Press, 2004.

[2] X. Brolly, C. Stratelos, and J. Mulligan, "Model-based head pose estimation for air-traffic controllers," in *Proc. IEEE Int. Conf. on Image Processing*, 2003, pp. 113–116.

[3] M. Chen and A. Hauptmann, "Towards robust face recognition from multiple views," in *Proc. IEEE Int. Conf. on Multimedia and Expo*, 2004.

[4] C. Stauffer and W. Grimson, "Adaptive background mixture models for real-time tracking," in *Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition*, 1999, pp. 252–259.

[5] C. Canton-Ferrer, J. R. Casas, and M. Pardàs, "Towards a Bayesian approach to robust finding correspondences in multiple view geometry environments," in *Proc. Work. on Computer Graphics and Geometric Modelling*, 2005, pp. 281–289.

[6] I. Mikic, *Human Body Model Acquisition and Trackign using Multi-Camera Voxel Data*, Ph.D. thesis, Univ. of California, San Diego, 2002.

[7] J. Yang, W. Lu, and A. Waible, "Skin-colour modeling and adaptation," Tech. Rep. CMU-CS-97-146, Carnegie Mellon University.

[8] R. Stiefelhagen, "Tracking focus of attention in meetings," *Proc. IEEE Int. Conf. on Multimodal Interfaces*, pp. 273–280, 2002.