Head Pose Detection based on Fusion of Multiple Viewpoint Information

C.Canton-Ferrer, J.R. Casas, and M.Pardàs

Technical University of Catalonia, Barcelona, Spain,
{ccanton,josep,montse}@gps.tsc.upc.es

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 head in the scene is detected and geometrically approximated by an ellipsoid. 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 an estimation of face orientation. Tracking over time is performed by Kalman filtering. Results of the proposed algorithm are provided in the SmartRoom scenario of the CLEAR Evaluation.

1 Introduction

The current paper addresses the problem of estimating the head orientation of people present in a SmartRoom 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 [4] , 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 detection of attitudes and behaviors of individuals.

A number of methods for head pose estimation has been proposed in the literature [1]. 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, some 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 this work, such conditions are usually difficult to satisfy. Specific facial features are typically not clearly visible due to far-field conditions, inadequate lighting and wide angle camera views. They may also be entirely unavailable when faces are not oriented towards the cameras.

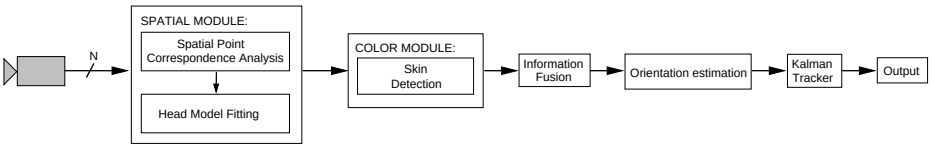

Fig. 1. System flowchart: acquisition, spatial and color analysis, head model fitting, face orientation estimation and tracking.

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].

We propose a method for 3D face orientation estimation which produce a fair estimation of the angle and is computationally simple for real-time applications. Redundancy among camera views is exploited to define a fusion process of color and spatial information in order to obtain a synthetic reconstruction of face appearance in 3D. Finally, an analysis method on these data is proposed in order to obtain the orientation of the head.

This method has been applied to a multi-camera SmartRoom scenario in the framework of a scene understanding project. 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 two low level image processing modules: spatial and color analysis. These modules provide data to the higher level analysis module that performs the information fusion required to estimate the orientation of head, 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. Bounding boxes describing the head of a person in multiple views are used to segment the interest area where the colour module will be applied. Center and size of the bounding box allow defining an ellipsoid model $\mathcal{H} = \{\mathbf{c}, \mathbf{R}, \mathbf{s}\}$ where $\mathbf{c}$ is the center, $\mathbf{R}$ the rotation along each axis centered on $\mathbf{c}$ and $\mathbf{s}$ the length of each axis. Colour information is processeced as described in subsection 2.1.

Information obtained by these two modules is combined in order to generate a 3D representation of the head and perform an estimation of its orientation.

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. 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 Color Module

Interest regions provided as a bounding box around the head 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 [9] 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.

Let us denote with $\mathcal{S}_n$ all skin pixels in the n-th view. It should be recalled that there could be empty sets $\mathcal{S}_n$ due to occlusions or under-performance of the skin detection technique. However, tracking information and redundancy among views would allow to overcome this problem.

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 $\mathcal{H}$ of the head and a set of skin patches at each view belonging to this head $\{\mathcal{S}_n\}$, $0 \leq n < N$. The output of this technique is a fusion of color and space information set denoted as Ω. An analysis technique of the data contained in Ω is provided in Sec.4.

The procedure of information fusion we define is based on the assumption that all skin patches $\{\mathcal{S}_n\}$ 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 $\mathcal{S}_n$ from all N views onto the 3D ellipsoid model. Formally, for each pixel $\mathrm{p}_n \in \mathcal{S}_n$, we compute

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

thus obtaining its back-projected ray in the world coordinate frame passing through p_n in the image plane with origin in the camera center $\mathbf{o}_n$ and director vector $\mathbf{v}$. In this equation, $P_n(\cdot)$ is the perspective projection operator from 3D to 2D coordinates on the view n [4]. In order to obtain the back-projection of p_n onto the surface of the ellipsoid modelling the head, Eq.1 is substituted into the equation of an ellipsoid defined by the set of parameters $\mathcal{H}$ [4]. It gives a quadratic in λ,

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

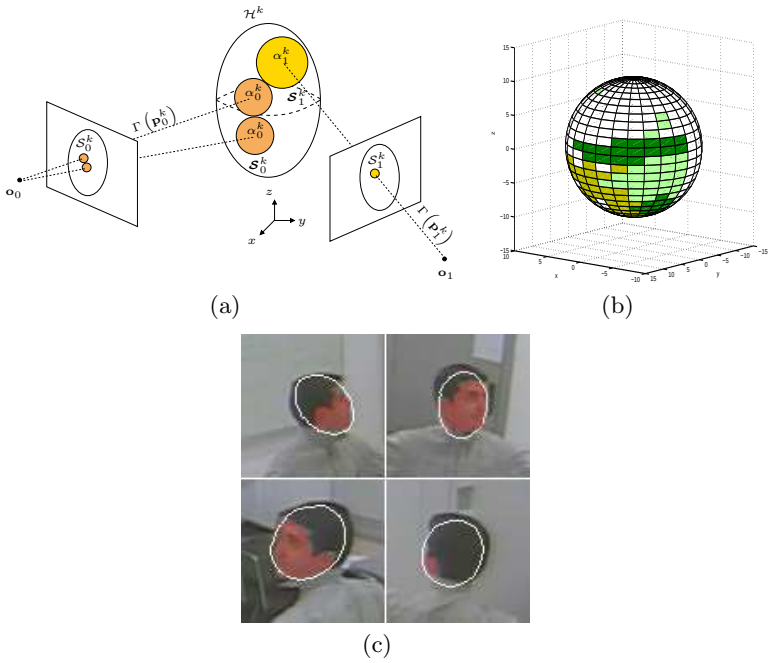

Fig. 2. In (a), color and spatial information fusion process scheme. Pixels in the set $\mathcal{S}_n$ are back-projected onto the surface of the ellipsoid defined by $\mathcal{H}$, generating the set $\boldsymbol{\mathcal{S}}_n$ with its weightening term α_n. In (b), result of information fusion obtaining a synthetic reconstruction of face appearance from images in (c) where the skin patches are plot in red and the ellipsoid fitting in white.

The case of interest will be when Eq.2 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.2(a).

This process is applied to all pixels of a given patch $\mathcal{S}_n$ obtaining a set $\boldsymbol{\mathcal{S}}_n$ 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 $\{\boldsymbol{\mathcal{S}}_n\}$, 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 can be estimated by projecting a sphere with radius $r = \max(\mathbf{s})$ 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 should be estimated for each element in $\boldsymbol{\mathcal{S}}_n$ but, since the *far-field* condition

$$\max(\mathbf{s}) \ll \|\mathbf{c} - \mathbf{o}_n\|_2, \qquad \forall n, \tag{3}$$

is fulfilled, α_n can be considered constant for all intersections in $\boldsymbol{\mathcal{S}}_n$. A schematic representation of the fusion procedure is depicted in Fig.2(a). Finally, after applying this process to all skin patches we obtain a fusion of color and spatial information set $\Omega = \{\boldsymbol{\mathcal{S}}_n, \alpha_n, \mathcal{H}\}$, $0 \leq n < N$, for the study head in the scene. A result of this fusion is shown in Fig.2(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]. We propose a method in order to estimate the value of the orientation angle $\hat{\theta}$.

4.1 Weighted centroid

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

$$\mathbf{d} = \frac{1}{\sum_{n=0}^{N-1} |\boldsymbol{\mathcal{S}}_n|} \sum_{n=0}^{N-1} \alpha_n \sum_{\mathbf{p}_n \in \boldsymbol{\mathcal{S}}_n} (\mathbf{p}_n - \mathbf{c}) \,, \tag{4}$$

$$\hat{\theta} = \tan^{-1} (\mathbf{d}_y / \mathbf{d}_x) \,, \tag{5}$$

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

5 Results and Conclusions

Sequences of the seminar type have been evaluated and results are reported in Table 1. By analyzing the results in detail we reached the following conclusions. Orientation is highly depending on the detection of skin patches thus being sensitive to its performance. Typically, skin detection underperforms when the face is being illuminated by a coloured light, i.e. the beamer. In this cases, we estimate a wrong orientation angle and the Kalman filter looses track after a short while. On the other hand, our method is conditioned by the hair style, the presence of beard or baldness.

Future research towards solving the aforementioned weak points of our algorithms would involve employing more sophisticated skin detectiors robust to the bias introduced when the face is illuminated by coloured light [5]. Particle filtering tracking schemes would also be introduced to cope with fast changes in the head orientation.

Table 1. Results of the proposed method for the UKA Seminars. M1 states for Pan Mean Absolute Error, M3 for Pan Mean Absolute Error per Pose, M5 for Pan Correct Classification, M7 for Correct Pan Classification per Pan Pose Class and M6 for Pan Correct Classification within range of neighbouring pose classes.

	M1	M5	M6
UKA_20050427_B	79.45^o	17.39%	39.93%
UKA_20050622_A	96.01^o	08.14%	31.60%
UKA_20050511	61.91^o	16.96%	55.49%
UKA_20050601	79.52^o	16.21%	48.63%
UKA_20050615_A1	74.68^o	27.87%	48.72%
UKA_20050615_A2	67.65^o	28.76%	59.80%
UKA_20050504_B	84.03^o	09.31%	32.48%
UKA_20050525_C	53.20^o	27.91%	70.74%
Average	73.63^o	19.67%	48.83%

M3

Pose	0^o	45^o	90^o	135^o	180^o	225^o	270^o	315^o
Average Error	40.63^o	55.83^o	67.50^o	146.25^o	112.97^o	94.36^o	54.41^o	32.59^o

M7

Pose	0^o	45^o	90^o	135^o	180^o	225^o	270^o	315^o
Correct Classification	36.17%	14.81%	0%	0%	8.61%	2.42%	18.05%	40.52%

References

1. Brolly, X., Stratelos, C., Mulligan, J.: Model-based head pose estimation for air-traffic controllers. Proc. IEEE Int. Conf. on Image Processing, pp. 113–116, 2003.
2. Canton-Ferrer, C., Casas, J.R., Pardàs, M.: Towards a Bayesian Approach to Robust Finding Correspondences in Multiple View Geometry Environments. LNCS 3515:2, pp. 281–289, 2005.
3. Chen, M., Hauptmann, A.: Towards Robust Face Recognition from Multiple Views. Proc. IEEE Int. Conf. on Multimedia and Expo, 2004.
4. Hartley, R.I., Zisserman, A.: Multiple View Geometry in Computer Vision. 2nd Edition. Cambridge University Press. 2004.
5. Martinkauppim, B.: Face colour under varying illumination-Analysis and applications. PhD Thesis, University of Oulu, 2002.
6. Mikic, I.: Human Body Model Acquisition and Trackign using Multi-Camera Voxel Data. PhD Thesis, Univ. of California. 2002.
7. Stauffer, C., Grimson, W.: Adaptive background mixture models for real-time tracking. Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, pp. 252–259, 1999.
8. Stiefelhagen, R.: Tracking Focus of Attention in Meetings. Proc. IEEE Int. Conf. on Multimodal Interfaces, pp. 273–280, 2002.
9. Yang, J., Lu, W., Waible, A.: Skin-colour modeling and adaptation. Technical Report, Carnegie Mellon University, CMU-CS-97-146.