Face Tracking in Meeting Room Scenarios Using Omnidirectional Views

Frank Wallhoff, Martin Zobl, and Gerhard Rigoll
Munich University of Technology
Institute for Human-Machine-Communication
Arcisstraße 21, 80290 München, Germany
{wallhoff,zobl,rigoll}@mmk.ei.tum.de

Igor Potucek
Brno University of Technology
Faculty of Information Technology
Bozetechova 2, 612 66 Brno, Czech Republic
potucek@fit.vutbr.cz

Abstract

The robust localization and tracking of faces in video streams is a fundamental concern for many subsequent multi-modal recognition approaches. Especially in meeting scenarios several independent processing queues often exist that use the position and gaze of faces, such as group action- and face recognizers.

The costs for multiple camera recordings of meeting scenarios are obviously higher compared to those of a single omnidirectional camera setup. Therefore it would be desirable to use these easier to acquire omnidirectional recordings.

The present work presents an implementation of a robust particle filter based face-tracker using omnidirectional views. It is shown how omnidirectional images have to be unwarped before they can be processed by localization and tracking systems being invented for undistorted material. The performance of the system is evaluated on a part of the PETS-ICVS 2003 Smart Meeting Room dataset.

1. Introduction

Research on smart environments has become focus for many activities in the field of human-machine interaction. One basic important step is the consideration about the input sensors and how to capture the activities in a defined environment. With a view to image based scene interpretation this means, how many cameras are needed and where have they to be installed. Depending on the scenario, several reasonable approaches exist for this task. However, the amount of collected data has to be small and the setup should be easily installable and configurable.

Especially for smart meeting room scenarios, which run under typical situation-given constraints (people sit around a table), the idea to capture the scene with just one single omnidirectional camera is desirable. In this case there will be no synchronization problem, which would take place for a multiple camera setup. A camera capturing 360° is simply located in the middle of the table facing all participants. The disadvantage for this easier setup is based on the distortion of the captured images. These view-point dependant deformations can be mostly reconstructed using sophisticated image transformations, but unfortunately not lossless.

This work copes with the problem of finding and tracking faces in omnidirectional image sequences. The output can be the basis for further detection cues, such as action or face recognition in meetings [2]. The system performance of the presented integrated approach is tested on the scenario A1, of the PETS-ICVS 2003 Smart Meeting Room data [1].

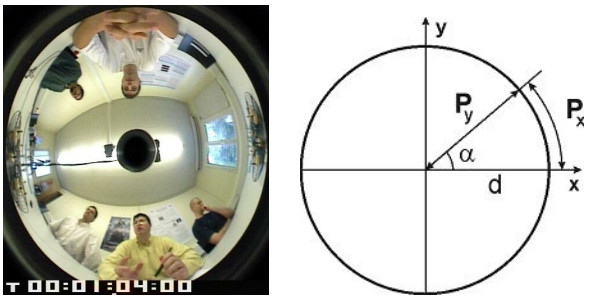
The structure of this paper is as follows. After the transformation from an omnidirectional image to an unwarped one, we briefly introduce two implementations for skin color segmentation and the computation of a face-likelihood for face detection in still images. The output of these systems is then be merged to a particle filter based face tracking system for image sequences. The paper concludes with achieved results.

2. Tranformation of Omnidirectional View

Our system is based on images captured by a standard video camera equipped with a hyperbolic mirror, which allows the capturing of a large portion of the space angle, here 360°. The obeyed image sequences from the PETS-ICVS database were acquired with the mirror under the camera and contain artificial meeting scenarios, with up to six people sitting around a table, see Figure 1 (a).

Before the face tracking, each image has to be transformed to a standard perspective view. Therefore we first apply a simple transformation that presumes a linear pixel distribution along the radius direction to get an panoramic view, see Figure 1 (b).

The coordinates of this panoramic view are P_x and P_y, which can be transformed to the coordinates of the omnidirectional image. It is assumed that real world elements are projected onto a cylinder with radius d. The axis of the



(a) (b)

Figure 1. Original omnidirectional image (a) and Transformation (b)

cylinder is identical to the mirror- and camera-axis [8, 5, 4]. The horizontal size of the panoramic view is a perimeter of a cylinder $WIDTH = 2\pi d$.

$$X_M = (d - P_Y) \cdot cos(\alpha) + CenterX \quad (1)$$

$$Y_M = (d - P_Y) \cdot sin(\alpha) + CenterY \quad (2)$$

$$\alpha = \frac{P_x}{d} \quad (3)$$

The computed pixels in the camera image do not correspond ”one to one” to the pixels in the projected image. Therefore sub pixel anti-aliasing methods have to be used. In our case it is sufficient to use weighted averages of neighbord pixels, since the size of a output pixel is comparable to that of the input pixel. It is further suitable to crop a part of image that contains the center of the omnidirectional view, which usually only displays a part of the camera itself. The result of a part of the scene is given in Figure 2.

Figure 2. Part of ”unrolled” panoramic image together with alignment for equalization

After this first transformation, the resulting images are still deformed. The problem arises from different distances between the mirror and the observed objects. Geometrical corrections can be applied based on knowledge about the geometrical setup of the room and information about the positions of the participants. Therefore the user has to define two setup describing curves by marking three points in the image for each one (red lines in Figure 2). The curves are approximated by circles, each characterized by the center-coordinates and a radius. These circles are used to make pixel interpolations to solve deformations in the vertical direction. A second deformation arises by the cylindrical projection of the image. The impact is that depending on the angle, the vertical width of a pixel has to be different. This deformation can be transformed by using a perspective projection of the cylindrical image into the plane, see Fig. 3.

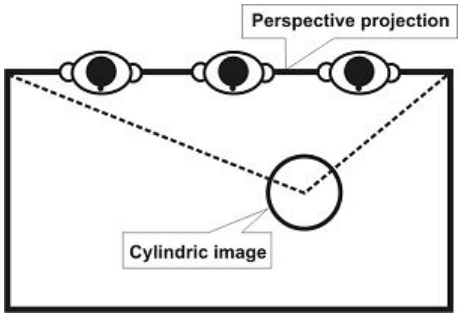


Figure 3. Perspective correction

An image can be transformed with equation 4, where x_{OMNI} is the position in the omnidirectional image, $Width_{OMNI}$ is width of the omnidirectional image, x is the position in the perspective image and d_M is the distance between the center of the cylinder and the projection plane.

$$x_{OMNI} = \arctan\left(\frac{x}{d_M}\right) \cdot \frac{Width_{OMNI}}{2\pi} \quad (4)$$

For smoothing the image and removing non-uniformity in the horizontal and vertical direction we again use a simple weighted interpolation method. Figure 4 depicts the result after applying all transformations.

Figure 4. Equalized image.

After all images within a sequence are preprocessed the way described above, they can be processed by the subsequent face tracking system, where skin color localization techniques as well as a neural network are used. The fundamental functionality is briefly described below.

3. Skin Detection and Segmentation

Color is a key feature for the detection of hands and heads in images. It is probably one of the most used methods for the detection of human body parts, which may be rested on its low computational cost. The disadvantage is

the low reliability, caused by the change of skin-tone color appearance under different lightning conditions.

3.1. Gaussian Mixture Skin Color Model

One approach to recognize skin color under varying illumination and brightness conditions is to transform the RGB-color intensities into the normalized rg-Chroma space.

The $r = \frac{R}{R+G+B}$ and $g = \frac{G}{R+G+B}$ components create a 2D color space. Skin colored pixels can be modeled with a normal probability distribution, respectively a Gaussian mixture model (GMM). To find the right GMM-parameters, various face color pixels are picked manually to estimate the distribution of the color class Ω_k. A color class Ω_k is determined by its mean vector μ_k and the covariance matrix K_k. The probability of an unknown pixel being skin colored can be computed by the following equation:

$$p(c|\Omega_k) = \frac{1}{2\pi\sqrt{||K_k||}} e^{-\frac{1}{2}(c-\mu_k)^T K_k^{-1}(c-\mu_k)} \quad (5)$$

A suitable parameter constellation, which fits on typical in-door conditions and for the current scenario is given by:

$$\mu_k = \begin{pmatrix} 44.548 \\ 28.935 \end{pmatrix} \quad , \quad K_k = \begin{bmatrix} 4.0916 & -0.3925 \\ -0.3925 & 1.53269 \end{bmatrix}$$

3.2. Global Skin Color Model

Because of the restriction, that the parameters of the GMM above are specialized to a certain environment, a second more robust approach for unconstrained environments desirable. The basic assumption is that a skin colored pixel lies within a certain area in the rg-Chroma plane, the so-called skin locus [7].

In this approach a skin color candidate has to be between two circles g_{up} and g_{down} in the rg-plane, where $g_{up} = a_{up}r^2 + b_{up}r + c_{up}$ and $g_{down} = a_{down}r^2 + b_{down}r + c_{down}$ ($a_{up} = -1.8423$, $b_{up} = 1.5294$, $c_{up} = 0.0422$, $a_{down} = -0.7279$, $b_{down} = 0.6066$ and $c_{down} = 0.1766$,). Furthermore whitish and gray pixels which lie in a circle with the radius 0.02 around the color white ($r = g = 0.333$) are discarded. Together with the neighbored pixels, a skin color probability $p(c|\Omega_k)$ can also be introduced.

Figure 5. Skinmask with potential skin color candidates

4. Neural Network based Face Detection

In addition to the search of faces using skin color, a second technique is involved to calculate face-likelihoods and to reject non-face regions. For this purpose we use an implementation of an artificial neural network based face detector, similar to [6]. This technique has established itself as being highly robust but computationally expensive. However, a combination of both cues will lead to a fast and robust system.

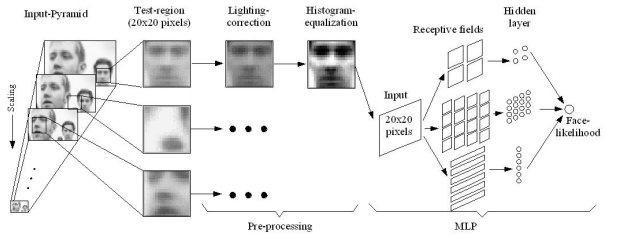


Figure 6. MLP based face-likelihood

With the help of a MLP structure as shown in Figure 6 (being trained with frontal upright, tilted faces and faces rotated in depth) we can compute the likelihood of a given image to be a face or not by a certain threshold. In the initialization phase, a sliding window samples all possible regions in the image, and is then enlarged until the window fits the image dimensions. For each sample a face-likelihood can be measured an possible face locations are merged.

5. Particle Filter based Tracking

Assuming a Markov-State-Space model with hidden states $\{\mathbf{x}_t\}$ describing position, size and dynamics of a face, the prior described observations of skin color and face-likelihood $\{\mathbf{z}_t\}$ are used to estimate the state of the system through the filtering distribution $p(\mathbf{x}_t|\mathbf{z}_{1:t})$. In most cases, this probability cannot be derived directly but calculated recursively by

$$p(\mathbf{x}_t|\mathbf{z}_{1:t}) \propto p(\mathbf{z}_t|\mathbf{x}_t) \int p(\mathbf{x}_t|\mathbf{x}_{t-1}) p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1}) d\mathbf{x}_{t-1}$$

. The prior distribution $p(\mathbf{x}_{t-1}|\mathbf{z}_{1:t-1})$ describing the system state in the last time step is predicted with dynamics $p(\mathbf{x}_t|\mathbf{x}_{t-1})$. Then the observation $p(\mathbf{z}_t|\mathbf{x}_t)$ updates the predicted distribution according to the measurement of the image to generate the current distribution.

The filtering distribution is approximated with a set of weighted samples, called particles. These are containing information about the system state, such as position, size, and dynamics. This way the distribution becomes $\widehat{p}_N(\mathbf{x}_t|\mathbf{z}_{1:t}) = \sum_{i=1}^{N} \pi_t^{(i)} \delta(\mathbf{x}_t - \mathbf{x}_t^{(i)})$. This method is known as condensation algorithm (particle filter, sequential monte carlo) [3].

In the first step the N particles are initialised with the output of the face detector (see Figure 7). Then

each particle is predicted by a linear regressive dynamical model with constant velocity. The parameters of this dynamical model are determined by training an adaptive linear network (ADALINE). For each particle the probability for containing a skin colored region out of the skin color mask is derived, and a face likelihood using the predescribed neural network is measured. These observations are linked together by multiplication $p(\mathbf{z}_t|\mathbf{x}_t^{(i)}) = p(\mathbf{z}_t^{skin}|\mathbf{x}_t^{(i)}) \cdot p(\mathbf{z}_t^{face}|\mathbf{x}_t^{(i)})$ and deliver the weight for each particle $\pi_t^{(i)} \propto \pi_{t-1}^{(i)} \cdot p(\mathbf{z}_t|\mathbf{x}_t^{(i)})$. A resampling step for the particle set, using the new weights keeps the particles in regions with high "face likeliness". To allow tracking of faces of people entering the scene, 10% of the particles are initialized by the face detection algortihm at each time step. For determination of the number and locations of faces, connected regions of particles with a specific size, minimum amount of particles and minimum probability are searched. For each so found location of a face a minimum number of particles is kept.

6. Results and Conclusion

Especially in critical tracking situations, where for a example a hand overlaps a part of the face, the combination of skin color detection with neural networks results in robust and reliable tracking performance, even on the transformed images.

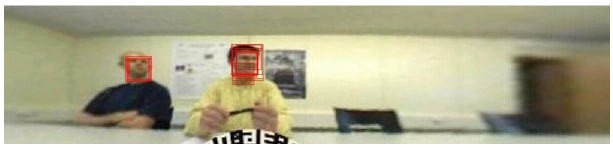


Figure 7. Initialisation of particle tracker with the face detection algorithm.

Figure 8. Border regions of the image are a problem, since the unwarping leads to rough distortion in these regions.

Keeping a minimum amount of particles at every face location prohibits the loss of partially bad tracked faces. From the view of the tracker there is no difference between the two methods for the extraction of skin colored regions. When the lighting conditions of the room are known the first model is more precise, but the second approach is preferred here because of its generality.

Figure 9. Faces of persons entering the room can be tracked through initialization with the face detector at every time step.

A comparison of the tracking results on the unwarped images from the omnidirectional views (camera 3) with real views (camera 1 and 2) showed no major differences in performance. Because the distortions of the omnidirectional camera can be mostly reconstructed, we think such a camera can be used as independent capture device for tracking task in meeting events in a room.

7. Acknowledgement

Parts of this work were funded by the EU IST Programme (project IST-2001-34485). It is part of CPA-2: the Cross Programme Action on Multimodal and Multisensorial Dialogue Modes, and is linked to the activity on Human Language Technologies. For further information see [2].

References

[1] Performance Evaluation on Tracking and Surveillance: Smart Meeting Rooms, In Conjunction with the IEEE-ICVS Conference 2003, Graz, Austria. http://petsicvs.visualsurveillance.org.

[2] The MultiModal Meeting Manager (M4) Project Homepage. http://www.dcs.shef.ac.uk/spandh/projects/m4/.

[3] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision 29(1)*, pages 5–28, 1998.

[4] J. Jones M., Rehg. Statistical Color Models with Application to Skin Detection. *Cambridge Research Laboratory, Computer Vision and Pattern Recognition (CVPR99), Ft.Collins, CO*, pages 274–280, June 1999.

[5] S. B. Kruppa H., Bauer M. Skin Patch Detection in Real-World Images. *Perceptual Computing and Computer Vision Group, ETH Zurich, Schwitzerland*, 2002.

[6] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. In *IEEE Transactions on PAMI*, pages 23–38, Jan. 1998.

[7] M. Soriano, S. Huovinen, B. Martinkauppi, and M. Laaksonen. Skin detection in video under changing illumination conditions. In *Proc. 15th International Conference on Pattern Recognition, Barcelona, Spain*, pages 839–842, 2000.

[8] T. Svoboda. Central Panoramic Cameras Design, Geometry, Egomotion. *PhD Thesis, Center for Machine Perception, Faculty of Electrical Engineering, Czech Technical University*, September 1999.