Multiple Person Tracking in Advanced Meeting Environments

Sascha Schreiber and Gerhard Rigoll

Institute for Human-Machine Communication, Technische Universität München,
Arcisstrasse 16, 80333 Munich, Germany
{schreiber, rigoll}@mmk.ei.tum.de

Abstract. This paper addresses the problem of tracking an unknown number of persons in meeting scenarios with a monocular camera. Although there might arise several problems caused by cluttered or noisy video data, it will be shown in this proposal that tracking of multiple persons in such scenarios can be realized by our novel idea. Our architecture basically utilizes a factored sampling technique - also known as ICondensation proposed by Isard and Blake [2] - to generate several hypotheses of a possible location for the tracked object. To enable a stable tracking we implemented an approach combining flexible properties of active shape models (introduced by Cootes et al. [1]) with the particle filter framework mentioned above and thus we could achieve robust results for tracking human heads in cluttered environments. Usually hypotheses will concentrate over time on the most likely location for the head in the image, thus a hierarchical structure has been invented which indicates both the number of persons and their position.

1 Basic System Overview

At the beginning of the tracking procedure N hypotheses $\{\mathbf{s}_t^{(i)}, i = 1, \ldots, N\}$ are initialized on skin colored regions in the image, which are detected after transforming the RGB-color intensities of every pixel into the normalized rg-chromatic color space. Using the properties of the resulting skin colored areas, hypotheses $\mathbf{s}_t^{(i)}$ can be initialized with the scale σ, the rotation ψ, the location τ, and the (mean) shape $\mathbf{c}$ of the tracked head. To model the head with all possible variations of the contour, we have chosen an active shape model as a flexible feature to describe the appearance of a head in the picture. For generating the model, we have manually labeled n points (landmarks) along the shape of the head in m training images, showing a large spectrum of variations in the contour. After removing all euclidic transformations from the training material and applying PCA to the landmark vectors, we finally obtain a matrix Φ, containing the eigenvectors of the training shapes. Using the matrix Φ and the mean shape $\overline{\mathbf{x}}$, which can be computed as the arithmetic mean of all landmark vectors, any possible shape can be generated by

$$\mathbf{x}' \approx \overline{\mathbf{x}} + \Phi\mathbf{c}, \tag{1}$$

From: MLMI 2005, Edinburgh, UK

where the vector $\mathbf{c}$ is used for weighting each eigenvector η_i in the matrix Φ to produce the variations of the shape. According to the principle of the factored sampling theory a score Θ is computed for each of the hypotheses $\mathbf{s}_t^{(i)}$ based on the true image data by the following procedure:
At first the normal vector $\{\kappa_i, i = 1, \ldots, n\}$ through each landmark of the shape, described by the corresponding hypothesis, is calculated. Along this straight line the dot product $\mu_{i,j}$ (cf. equation 2) of the unit vector normal and the gradient $\mathbf{g}(x, y)$ at each pixel position inside a certain distance ϵ to the respective landmark is computed.

$$\mu_{i,j} = \mathbf{n}_i \circ \mathbf{g}(\mathbf{p}_i + j \cdot \mathbf{n}_i), \forall i \in \{1, \ldots, k\}, j \in [-\epsilon, \epsilon] \tag{2}$$

For each of the n straight lines the pixel $(\hat{\mathbf{x}}_i, \hat{\mathbf{y}}_i)$ with the highest score $\mu_{i,j}$ is chosen for a new shape $\hat{\mathbf{x}}$, which represents the image data best. After that optimal parameters for the euclidic transformations σ, ψ and τ are computed to minimize the sum of squared distances between the model landmarks and the new shape $\hat{\mathbf{x}}$. Finally the weight $\mathbf{c}$ is obtained by applying model constraints on the new shape, and thus all properties of the hypotheses can be updated. The measurement score Θ can be computed as the squared sum of the dot product between the unit normal vector and the gradient at the landmark position of the final contour, normalized with the number of landmarks. In the last step of the algorithm N hypotheses for the next time step are sampled from the actual hypotheses set with their probability $\Theta^{(i)}$. Finally the chosen hypotheses are predicted by a stochastic linear dynamical model and the ASM-based measurement can be started for the next frame. In a second step this framework was extended by a so called master layer, which controls the allocation of hypotheses to the objects detected in the frame. For this layer skin colored blobs are extracted and N hypotheses are generated on each blob. After performing the measurement step described above, the master layer generates L hypotheses representing possible combinations of the blobs. For each hypothesis a score Γ is computed, which indicates the probability for the respective constellation of objects represented by the combination of the blobs. This score is composed of the relation between the area A_{skin} of the skin colored blob and the area A_{shape} covered by the mean shape per blob, weighted with the mean score $\overline{\Theta}$ per blob. In this way a set of hypotheses in the master layer is obtained, which will be running through the basic PF framework. Thus the true object constellation in the image will be shown by the most probable hypothesis.

References

1. T.F. Cootes, D. Cooper, C.J. Taylor and J. Graham, "Active Shape Models - Their Training and Application." *Computer Vision and Image Understanding*, Vol. 61, No. 1, pp. 38-59, January 1995.
2. M. Isard and A. Blake, "ICONDENSATION: Unifying Low-level and High-level Tracking in a Stochastic Framework," *Proc. of the Fifth European Conference on Computer Vision (ECCV '98)*, Vol I. pp. 893-908, Freiburg, Germany, June 1998.