OMNIDIRECTIONAL TRACKING AND RECOGNITION OF PERSONS IN PLANAR VIEWS

Sascha Schreiber, Andre Störmer, Gerhard Rigoll

Institute for Human-Machine Communication
Technische Universität München
Arcisstr. 21, 80333 München

ABSTRACT

In this paper a view-independent head tracking system applying an Active Shape Model based particle filter is used to find precise image sections. DCTmod2 feature sequences are extracted from these sections and given as input to Cyclic Pseudo two-dimensional Hidden Markov Model based classifiers. These classifiers are trained to recognize the identity of the shown persons. The video material is recorded in an office environment with changing lighting conditions and thus results in a challenging task for both tracking and recognition. The overall performance of the system is evaluated depending on the various views of persons rotating on a swivel chair.

Index Terms— Monte Carlo Methods, Face Recognition, Hidden Markov Models, Surveillance

1. INTRODUCTION

The human desire for security steadily increases due to certain events in the last years, thus public video surveillance and automatic analysis of recorded data is clamored. For most of these applications numerous steps in a long chain of processing tasks like localization, tracking, feature extraction and classification have to be executed to derive the desired result. In the last two decades there has been a lot of progress in each of these steps in research and diverse algorithms have been developed, especially in the field of face and head tracking as well as face recognition. The probably most popular cue for tracking tasks might be the usage of color information extracted from skin or other person specific regions [1]. Other methods apply techniques like template matching, facial features, contour analysis, optical flow or exploit a combination of these features. A survey on these techniques can be found in [2]. Also face recognition is a widely researched topic. The first popular approach has been a facial description using linear subspaces called Eigenfaces. Similarly the Elastic Bunch Graph Matching has risen much attention. Later Active Appearance Models were introduced and it was suggested to solve different classification tasks using their parameters. There is a huge variety of different approaches in face recognition, a survey on these can be found in [3]. Excellent results in face recognition have also been achieved by Hidden Markov Model (HMM) based algorithms. A comparison between different statistical classifiers like Gaussian Mixture Model, 1D-HMM and Pseudo 2D-HMM (P2D-HMM) on the face recognition task can be found in [4]. In this paper a system for fully automatic tracking and recognition of a person's identity near real-time is presented. In Section 3 the fundamental procedure to localize and track humans is explained. Based on the tracked position a Cyclic Pseudo 2D-HMM based approach is then used to classify the extracted DCTmod2 features of the image patch which is described in Section 4. Finally promising results are presented (Section 5).

2. DATABASE

One of the motivations while creating a database for the identity recognition task of heads rotated in depth is to use standard technology in every days office life. A database containing 28 video sequences showing 14 persons rotating on a swivel chair in front of a low budget and noisy VGA webcam has been created. In order to simulate the challenges that occur in real data, the recordings have been captured in a typical office environment with a complex background and changing lighting conditions due to various times of day. A single frame of the data has a resolution of 640×480 pixels, the head in these images have typically a size of about 150×200 pixels. The dataset has been strictly split into a training set, which is only used for building the object model and to train the recognition system, and a test set, which is exclusively utilized for the evaluation of the proposed approach. Each set consists of roughly 2000 frames showing each of the persons in different views.

3. FEATURE EXTRACTION

3.1. Active Shape Tracking

To separate background information from the human head, which typically shows a very non-rigid projection in 2D when e.g. turning from profile to frontal view, an active shape model (ASM) [5] is used to parameterize the head. For this rea-

This work has partially been funded by the European Union within FP7 project PROMETHEUS, www.prometheus-fp7.eu

978-1-4244-1764-3/08/$25.00 ©2008 IEEE

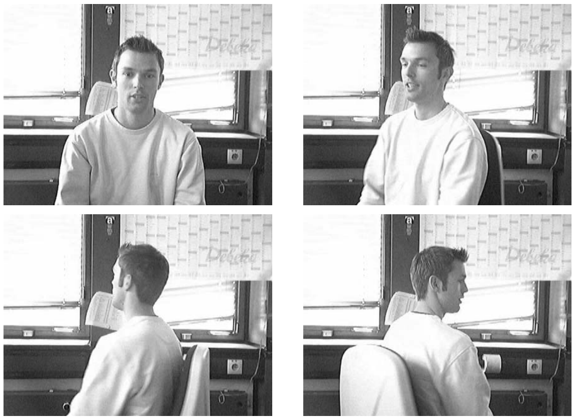

Fig. 1. Typical examples of a video sequence originating of the proposed database

son a head is modeled by $L = 20$ landmark points, which have been manually labeled for all training sequences, and an ASM is created by applying principal component analysis to the aligned labeled data. In this way a head at position $c = (x, y)$ with a size s can be fully described by a shape $S = (c, s, p)$, where p is a vector representing the silhouette. Contrary to the standard ASM approach where the gray values of the pixels are observed along the normal of the contour to detect learned histogram characteristics, in this approach a modified technique is applied directly on the gradient image and thus benefits from not only the fact, that there is an edge at a certain pixel position within the image, but also the direction of this edge. Regarding this modification, the adaptation instruction for any shape to the image data is as follows:

- Compute gradient image by applying a Sobel filter in x- and y-direction.
- Calculate the normal vectors at each landmark of the shape.
- Compute the angle between the normal vector and the gradient vector at each of the p pixel coordinates along the normal. Choose the pixel with the smallest corresponding angle as the new position for each landmark.
- Update the model parameters (position, scale, rotation) by least squares fitting to fit the new landmarks best.
- Repeat until the shape does not change significantly any more.

Finally a quality score $\theta = \sum_{i=1}^{L} \vec{n}_i \circ \vec{g}_i$ can be obtained by summing the maximum dot product between the normal vector $\vec{n}_i$ and the gradient direction $\vec{g}_i$ of each landmark. The described method is embedded in a stochastic particle filtering framework called ICondensation [6] to stabilize the tracked shape. Thus basically several different hypotheses are generated, where each represents diverse shape appearances. In this way the probability distribution $p(\vec{w}_t)$ for heads at each time step t is modeled. The aim is to track the position of the persons throughout the posterior probability $p(\vec{w}_t|\vec{z}_{1:t})$, based on the observations $\vec{z}_t$, representing the image features. Since there is no functional representation for this conditional probability available, it has to be approximated iteratively by:

$$p(\vec{w}_t|\vec{z}_{1:t}) \propto p(\vec{z}_t|\vec{w}_t) \int p(\vec{w}_t|\vec{w}_{t-1}) p(\vec{w}_{t-1}|\vec{z}_{1:t-1}) d\vec{w}_{t-1} \qquad (1)$$

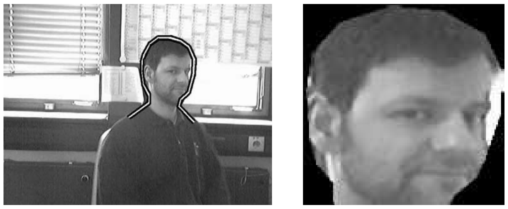

Fig. 2. Output of the shape tracker (left), cropped and masked image section (right)

Updating the posterior distribution $p(\vec{w}_{t-1}|\vec{z}_{1:t-1})$ from the previous time step by prediction with dynamics $p(\vec{w}_t|\vec{w}_{t-1})$ leads to the effective prior $p(\vec{w}_t|\vec{z}_{1:t-1})$ for the actual time step. Finally the current state density $p(\vec{w}_t|\vec{z}_{1:t})$ results from multiplying the prior distribution with the actual measurement $p(\vec{z}_t|\vec{w}_t)$ derived from the quality score θ of the ASM. The hypotheses of the particle filter are initialized by a simple skin color detector. The major advantage of this principle is that a decrease in the computation time can be achieved, because of lower precision requirements of the ASM detector. Due to the ability of shape structure improvements during the measurement, where each shape hypothesis is locally adapted to the image data, a significant higher tracking performance than using only a plain ASM structure is achieved.

3.2. Feature preparation

As mentioned in Section 2 the sequences have been partially exposed to changes in lighting and thus the cropped images obtained by the tracking system (cf. Fig. 2) have to be subsequently corrected. For this reason a technique called Contrast-limited adaptive histogram equalization (CLAHE) [7] is applied. The idea is to divide the image into non-overlapping $N \times N$ blocks and then limit the maximum slope of the cumulative distribution function within these blocks, i.e. to transform the given histogram to the desired equal distribution. To avoid hard boundaries at the edge of each block, bilinear interpolation is applied followed by a median filter to smooth the cropped image. In this way overexposed regions in an image region as shown in Fig. 3 on the left side can be reliably corrected, while images, which are not influenced by any lighting keep nearly unchanged. The resultant CLAHE corrected image section is then used for the computation of the features.

3.3. DCTmod2 features

It has been shown that DCTmod2 features are robust to out-of-plane rotation and changing lighting conditions. A short overview of the feature extraction scheme is given. Please note that an elaborate description of the feature extraction method can be found in [8]. In this approach the features are computed from 8×8 pixel sized blocks. Each block is

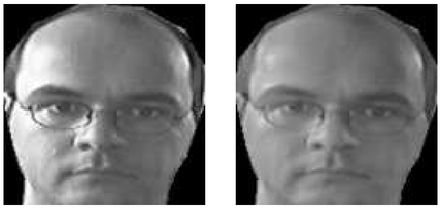

Fig. 3. Cropped image (left), that is obtained by the tracking system, and the CLAHE corrected image section (right)

overlapping the neighboring blocks by half the block size to derive an approximately smooth feature sequence. A standard 2D-DCT is applied to each block. The first three DCT coefficients (constant component and the lowest horizontal and vertical cosine frequency) are replaced by their horizontal and vertical deltas to the neighboring blocks. This leads to a 18 dimensional feature vector (6 delta + 12 DCT coefficients) for each 8×8 pixel sized block. The complete feature sequence used to classify is derived by concatenating the feature vectors column-wise and inserting an unique marker feature at the beginning of each column.

4. THE RECOGNITION SYSTEM

The recognition system uses DCTmod2 features computed from the image section found by the tracking system. A CP2D-HMM classifier uses these feature sequences as input to identify the persons shown in the tracked sections.

4.1. Statistical model based classifier

In this approach Cyclic Pseudo 2-dimensional Hidden Markov Models (CP2D-HMMs) are used to classify person identities based on the described feature sequences. A CP2D-HMM is computed for every person. In the recognition stage the production probabilities for each model on a feature sequence are computed and the one with the highest probability is selected. Since HMMs are well known state of the art [9] only a short description of the basic principle is given. A major focus is given to the training of the cyclic version of the HMM, since the alignment to the different rotational views of the persons is fundamental for the recognition.

4.1.1. Pseudo 2-D-Hidden Markov Model

P2D-HMMs are nested 1D-HMMs, every column of a two-dimensional field is modeled as a one-dimensional HMM. Additionally so called markerstates are put at the beginning of each 1D-HMM. Transitions from the last state of each 1D-HMM back to its marker state (column-wise selftransition) and to the markerstate of the 1D-HMM of the next column (column-wise forward transition) are inserted (see Fig. 4). This leads to a warping ability in two directions separately. The well-known Baum-Welch algorithm [10] is applied to train the P2D-HMM, exactly as it is used for the 1D-HMM. The mentioned markerstates contain a maximum output probability for the unique marker features which were inserted in the feature sequence. During the computation of the most likely state sequence of the overall P2D-HMM, this leads to an enforced alignment at the beginning of every column.

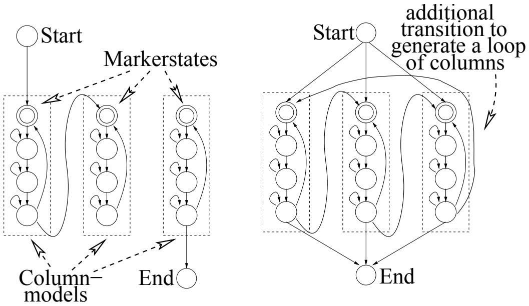

Fig. 4. Structure of a 3×3-state left-right P2D-HMM (left), which is extended to Cyclic P2D-HMM (right)

4.1.2. Cyclic P2D-HMMs

For this recognition task a structure to model heads in arbitrary views has been developed. Since it is unknown how the head texture is rotated in depth, the P2D-HMM has been adopted (Fig. 4) to model a loop of column-models, that are 1D-HMMs referring to the columns of the image, able to start and end at each position of the loop. This structure of column-models can be supposed as a model of column-wise views projected on the surface of a cylinder (see Fig. 5). Since we allow the model to start and end at each column-model, 2D projections of 3D objects rotated by an arbitrary angle around the y-axis can be recognized. Since the model is able to start and end at each position, it has to be ensured during training, that a full rotation of the person is mapped to the model. This is achieved by using initialization images which are generated by concatenating the kth column of each frame over time. In this way a set of complete omnidirectional views of the head is produced for each training video (see Fig. 6). These pictures provide a basis to initialize a global head model by training it with the Baum-Welch method. In most cases a piece of background as well as the head mask (see Fig. 2) is seen at

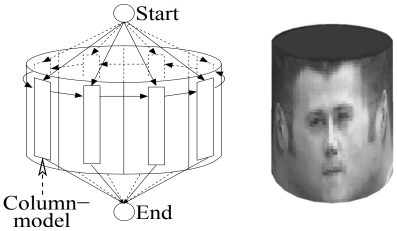

Fig. 5. Interpretation of a Cyclic P2D-HMM, which is a loop of column-models, as a cylinder

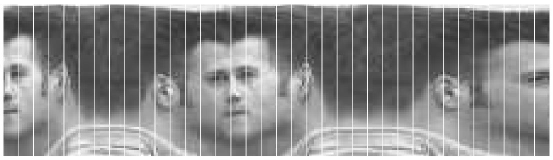

Fig. 6. Image for the initialization of the model during training, at least a full rotation is trained on the loop of column-models. Those images are generated by concatenating image columns over time

the borders of the analyzed image section. For this reason the global head model is nested between garbage models before the second stage of training is entered. After that person specific models are retrained applying an embedded training version of the Baum-Welch method on this nested global head model with the manually labeled image sections of the training videos.

5. RESULTS

To evaluate the proposed approach the described DCTmod2 features are quantized by k-means clustering. With the resultant discrete features, 6×6 states CP2D-HMMs are trained on 14 videos each showing different individuals. For a more detailed discussion of the recognition performance all evaluations are analyzed with respect to the rotation angle.

In a first experiment a reclassification is performed to evaluate the principle applicability of the proposed system setup for recognizing persons from any planar view. As Fig. 7 (cf. dotted outer line) shows, a recognition rate of about 100% is achieved nearly independent of the angle, which confirms the idea of training all different views of an individual into a single person-specific cyclic HMM.

To analyze the recognition performance independent of the tracking system, the CP2D-HMM classifier is evaluated on manually cropped test data comprising 14 videos (≈ 2000 images), which are disjoint to the data used for training. The results are depicted also in Fig. 7 by the bright solid line. For frontal heads in the sector of $\pm 30°$ the classifier performs well with an average recognition rate of 66.1% keeping in mind that there occurred large variations in the lighting conditions. Obviously views in the frontal range are recognized more robust than views on the back of the head ($< 60\%$), which is expected commonsensible. Also the recognition rate of profile views decreases, since in these cases the idea of mapping a 2D texure to a cylinder does not fit too well. Altogether an average recognition rate of 47.7% is achieved.

In a last scenario the complete system comprising tracking, lighting correction and classification is run on the videos of the same test set in a bothfully automatic mode. As the dark solid curve in Fig. 7 displays there is nearly no difference in the performance of the classifier and the dependency of the rotation angle. Thus the average recognition rate only slightly decreases down to 41.4 %.

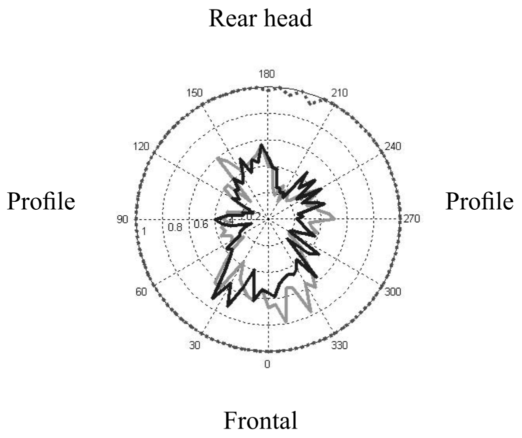

Fig. 7. Recognition results of reclassification (dotted outer curve) and classification once on the manually segmented test data (bright curve) and on the automatically tracked test data (dark curve)

6. CONCLUSION

A new approach for recognizing persons rotated by an arbitrary angle in depth was presented. First evaluations show the potential of the approach which can be further improved by both enhancing the tracking or the recognition system. Related to the difficulty of the presented task the results are impressive keeping in mind that only one view-independent model per person is used to classify real tracking outputs.

7. REFERENCES

[1] V. Vezhnevets, V. Sazonov, and A. Andreeva, "A survey on pixel-based skin color detection techniques," in *Proc. Graphicon*, 2003.

[2] Alper Yilmaz, Omar Javed, and Mubarak Shah, "Object tracking: A survey," *ACM Computing Surveys*, vol. 38, no. 4, 2006.

[3] W. Zhao, R. Chellappa, A. Rosenfeld, and P. Phillips, "Face recognition: A literature survey," Tech. Rep., 2000.

[4] F. Cardinaux, C. Sanderson, and S. Bengio, "Face verification using adapted generative models," in *Proc. FG*, 2004.

[5] T. Cootes and C. Taylor, "Statistical models of appearance for computer vision," Tech. Rep., 2004.

[6] M. Isard and A. Blake, "ICONDENSATION: Unifying low-level and high-level tracking in a stochastic framework," *Lecture Notes in Computer Science*, vol. 1406, pp. 893–908, 1998.

[7] S.M. Pizer, R.E. Johnston, J.P. Ericksen, B.C. Yankaskas, and K.E. Muller, "Contrast-limited adaptive histogram equalization: speed and effectiveness," *Proc. Visualization in Biomedical Computing*, pp. 337–345, 22-25 May 1990.

[8] C. Sanderson and K.K. Paliwal, "Fast features for face authentication under illumination direction changes," *Pattern Recognition Letters*, 2003.

[9] L.R. Rabiner, "A tutorial on hidden markov models and selected applications in speech recognition," in *Proc. IEEE*, 1989, pp. 77(2):257–286.

[10] S. Young, D. Kershaw, J. Odell, D. Ollason, V. Valtchev, and P. Woodland, *The HTK Book*, 2000.