A MULTI-STEP ALIGNMENT SCHEME FOR FACE RECOGNITION IN RANGE IMAGES

Andre Störmer, Gerhard Rigoll

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

ABSTRACT

Face Recognition in range images is a challenging task, especially if the pose of the shown face is unknown. To solve this, an alignment procedure consisting of facial feature hypotheses extraction by invariant curvature features, PCA-based classification and Iterative Closest Point alignment will be introduced to create aligned and normalized patches. These patches will then be used in a recognition algorithm, a discrete Pseudo 2- Dimensional Hidden Markov Model approach based on vector quantized DCTmod2 features. The results of this processing chain are discussed and compared to previous works.

Index Terms— Face Recognition, Statistic modeling

1. INTRODUCTION

Recently more and more recognition tasks shift from the use of traditional 2D image data to the additional or exclusive use of depth information. A big advantage of depth based solutions is, that they are less sensitive to lighting and pose variations. Many face recognition approaches on range images have been reported, a survey can be found in [1].
For all of these recognition algorithms the alignment of the data is fundamental. It is essential to find the rigid transformation (rotation and translation), that transforms the given face to a centered, frontal looking view, or to align it to a model. A successful approach to solve this is the Iterative Closest Point Algorithm (ICP) [2]. Another well known technique is to identify some facial features by curvature, and compute the alignment based on them [3].
In this paper an efficient alignment scheme consisting of facial feature hypotheses detection by curvature, hypotheses classification by PCA and Trimmed Iterative Closest Point alignment is proposed. After the faces are aligned, a Pseudo-2-Dimensional Hidden Markov model (P2D-HMM) using vector quantized DCTmod2 features, is tested on the range images. This is an adaptation of the work of Eickeler in [4], who used DCT-Features of JPEG-compressed 2D-images.
For the following experiments data is taken from the 3D face images database GavabDB [5]. It contains 3D surface meshes from faces, offering different views as well as different facial expressions per individual. The texture has been omitted, the database contains only depth information. There are scans from 61 different persons provided. An example showing one person in the seven different datasets is shown in Fig. 1.

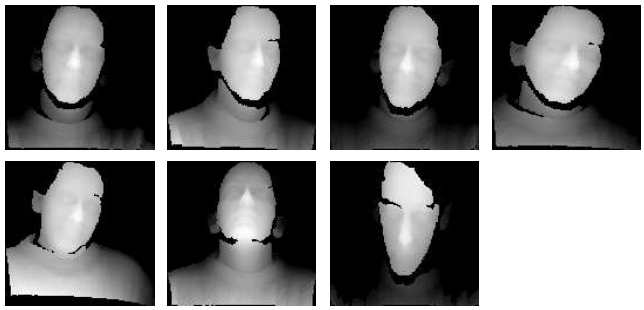

Fig. 1. The same person as range image (interpolated and smoothed) in the seven different datasets, from upper left to lower right: *neutral1*, *neutral2*, *smile*, *laugh*, *random*, *look up*, *look down*.

2. PREPROCESSING AND ALIGNMENT

2.1. Facial Features

The finding of facial features is the first step of the preprocessing. The input to this processing part is the data presented as depth image. The spatial discrete data is interpolated bilinearly and low-passed filtered to get a closed and smoothed surface. After that the mean curvature H and Gaussian curvature K is computed on the range image I based on the partial derivatives.

$$H = \frac{(1+I_x^2)I_{yy} - 2I_xI_yI_{xy} - (1+I_y^2)I_{xx}}{2(1+I_x^2+I_y^2)^{\frac{3}{2}}} \quad (1)$$

$$K = \frac{I_{xx}I_{yy} - I_{xy}^2}{(1+I_x^2+I_y^2)^2} \quad (2)$$

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

Fig. 2. Low-pass filtered range image (left), mean curvature (second) and Gaussian curvature (third), negative values are black, positives white. Combining mean and Gaussian curvature extracts convex (white) and concave (black) regions (right)

The curvature maps are divided into three discrete values (near zero, positive and negative curvature) for each mean and Gaussian curvature. Since it is known that the nose tip is an elliptic concave region and the inner eye corners are elliptic convex regions, several hypotheses of eye corners (concave regions, $H < 0$, $K > 0$) and nose tips (convex regions, $H > 0$, $K > 0$) can be generated (see Fig. 2).

After the hypotheses have been found, every possible tuple consisting of two eye corner hypotheses and one nose tip hypothesis is built and will be called inner face hypothesis. The number of tuples is reduced by applying a ruleset of a priori knowledge about distances and positioning of these facial features. The ruleset which is applied to the hypotheses looks like this:

- $2\times$ eye corners distance $>$ eye corner to nosetip distance
- the nosetip is below each eye corner
- each tuple should be unique (commutativity of eye corners)
- each curvature region has a minimum size

This is a sparse ruleset to reduce the overall number of inner face hypotheses which have to be classified. It can easily be extended by common knowledge but is intentionally kept small, to focus on the validation of the hypotheses. Patches of the remaining hypotheses are cropped out of the range image, based on the positions of the assumed eye corners and nose tips, resized to 96×64 pixels and put into a classifier, to decide if they are valid. Only the hypothesis with the best classification score survives. The classifier is based on an eigenspace computation and will be trained with 96×64 pixel sized example patches of inner face regions (see Fig. 3), which are selected from the training data. A PCA is made on the given

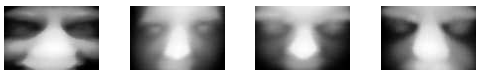

Fig. 3. Example patches of inner face regions from the training images. These images will be used to train the PCA-based classifier.

training set of inner face regions to find the main modes of variation within this region. If there are N training images x_i with $i = 1..N$, the first step is to compute the mean x_m.

$$x_m = \frac{\sum_{i=1}^{N} x_i}{N} \tag{3}$$

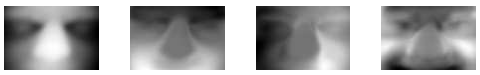

Fig. 4. The mean patch (left) and the three main modes (eigenvectors with largest eigenvalues) backprojected into the original space

Then the $N \times (nm)$ matrix A of zero-mean data is generated by

$$A = (a[1] = x_1 - x_m|..|a[N] = x_N - x_m) \tag{4}$$

The covariance matrix C is calculated by

$$C = AA^T \tag{5}$$

The eigenvector matrix $E = (e_1|..|e_k)$ of this covariance matrix C is computed and sorted by their eigenvalues v_i so that v_1 is the largest. The l eigenvectors with the l largest eigenvalues are chosen, that the proportion to the total variation achieves 98% (see Eq. 6).

$$l : \frac{\sum_{i=1}^{l} v_i}{\sum_{i=1}^{k} v_i} \geq 0.98 \tag{6}$$

During the experiments that lead to using $l = 50$ eigenvectors. They correspondend to the main modes of variation within the data and span a linear subspace which will be used to describe the original data. Each depthmap x can be projected into the linear subspace by computing the weights w.

$$w = E_l^T (x - x_m) \tag{7}$$

The matrix E now only holds the l eigenvectors $E = (e_1|..|e_l)$ with the l largest eigenvalues.

For each patch of inner face hypothesis x_h a resynthesis x_r of it is generated by projecting it onto the l eigenvectors to compute weight vector w(see Eq. 7), and then backproject it into the original range image space (see Eq. 8).

$$x_r = x_m + \sum_{i=1}^{l} w_i e_i \tag{8}$$

If the cross correlation coefficient d_{cc} of hypothesis x_h and resynthesis x_r is larger than a threshold, the hypothesis x_h is considered as a valid inner face region. An empirical determined threshold value of 0.98 turned out to work well.

$$d_{cc} = \frac{x_r x_h}{|x_r||x_h|} \tag{9}$$

If more than one hypothesis is considered to be true, the one with the highest cross correlation coefficient d_{cc} is chosen. Remember that the inner face regions are defined by the eye corners and the nose tip, so by using this scheme these facial features are extracted (see Fig.5). The first coarse alignment is then generated by rotating and scaling so, that the eye to eye connection line is horizontal in the viewplane, the nose is under the eyes and looks toward the viewpoint. Then a larger patch, namely the face, is cropped out of the range image, based on the extracted facial feature positions.

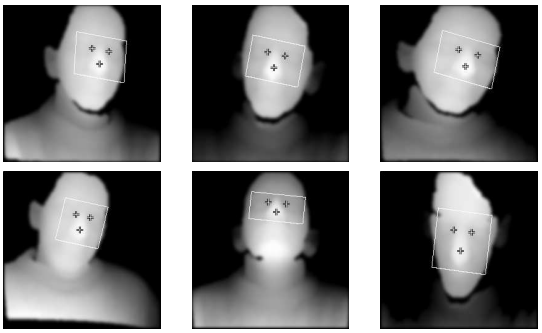

Fig. 5. Result of the feature detection, crosses denote eye corners and nose tip, the region of the best hypothesis found by the PCA-based classifier is boxed, from upper left to lower right: *neutral2*, *smile*, *laugh*, *random*, *look up*, *look down*.

2.2. Fine Alignment and Normalization

The previous step aligned the data by using features based on curvature regions and not on precise feature points, thus the alignment is not very accurate. To resolve this, a fine alignment using the Trimmed Iterative Closest Point Algorithm (TrICP) [6] is applied. It uses a reference dataset $B = (b_1, .., b_N)^T$, b_i is a vector containing point coordinates, and tries to find the rigid transformation consisting of Rotation R and Translation t, so that a dataset $A = (a_1, .., a_M)^T$ will be aligned to B. Since the correspondence problem (which point in dataset A belongs to which point in dataset B) is unsolved, an iterative approach is taken. It approximates this correspondence in every iteration simply by the square distance nearest neighbor. Trimmed Iterative Closest Point Algorithm:

1. For each point b_i in B compute the nearest neighbor a_j in A.
2. Sort corresponding point pairs by distance, to derive the sorted datasets A_s and B_s containing only the n pointpairs $A_s = (a_{s,1}, .., a_{s,n})$, $B_s = (b_{s,1}, .., b_{s,n})$ with the shortest distances
3. Compute the Centers of Gravity C_A, C_B of A_s, B_s

$$C_A = \frac{\sum_{i=1}^{n} a_{s,i}}{n};\ C_B = \frac{\sum_{i=1}^{n} b_{s,i}}{n} \tag{10}$$

4. Compute the MSE estimation of rotationmatrix R , so that

$$B_s - C_B \approx R(A_s - C_A) \tag{11}$$

5. Apply Transformation on complete dataset A so that

$$A_{new} = RA - RC_A + C_B = RA + t \tag{12}$$

6. Update $A = A_{new}$, Repeat Steps 2.-3. until convergence is declared

Convergence is declared if the MSE is not reduced anymore. Since the data is already coarsely aligned, the number n of point pairs, used to estimate the transformation can be a high proportion of the overall number of points (e.g. 80%). To find an approximation for the rotation R (see Eq. 11), typically the quaternion solution is used. An overview about this and other techniques is found in [7]. TrICP converges to a local minimum, but because of the initial alignment from the previous step, good results are achieved in only very few iterations. After all faces are aligned, a rectangular patch in the (x, y)-plane including eyes and mouth is cropped (see Fig. 6), resized to 128×128 pixels and will be used for recognition.

Fig. 6. Examples of aligned and normalized facial patches

3. P2D-HMM ON DCTMOD2 FEATURES

Now that all data sets are aligned and normalized facial patches are given, a classificator based on Pseudo 2 Dimensional Hidden Markov Models (P2D-HMM) on DCTmod2 features extracted from the depth data is applied.

3.1. DCTmod2 features

The patches are divided into 8×8 pixel sized blocks, each overlapping the neighboring block by 4 pixels. For each block the DCTmod2 features are written into a features sequence in a column-wise order. These are 2d-DCT features where the mean and the lowest horizontal and vertical frequency coefficients are replaced by the first order delta coefficients to their horizontal and vertical neighbouring blocks. Here 18 coefficients (6 deltas + 12 DCT) per block were used. A detailed description how to compute DCTmod2 features can be found in [8]. Additional markers are inserted to the feature sequence at the beginning of every column, these are unique values that can be distinguished from every possible coefficient value.

3.2. Pseudo 2 Dimensional Hidden Markov Models

The feature streams will be processed by P2D-HMMs, which are a well known approach for classification of 2D patterns. They have been successfully used for face recognition based on texture information and will now be tested on features derived from range images. The purpose of a P2D-HMM is to model the columns of a 2D field with one-dimensional HMMs. At the beginning of each column-model a marker state is inserted (see Fig. 7), which forces alignment to markers in the feature sequence. A detailled description on P2DHMMs for face recognition can be found in [4]. For each individual in the trainset a P2D-HMM is trained, so that the states model the probability distribution of the features. Since as much training material as possible is needed, variations of the training data are generated by translation and rotation. To recognize, the probability that a pattern is produced by the trained P2D-HMMs is computed for each model. It is then classified to the model with the maximum probability.

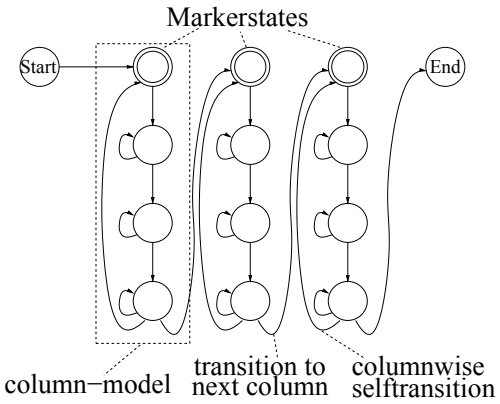

Fig. 7. A 3×3-state P2D-HMM

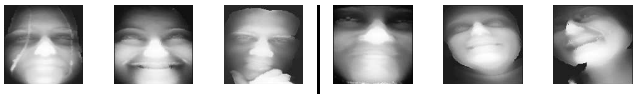

Fig. 8. Correctly aligned (left) and misaligned patches (right)

4. RESULTS

Now the results of the proposed approach are given and compared to previous works [9][10] (see Table 2). As trainset the *neutral1* series is used. Testings are performed by identifying images from all remaining datasets (see Fig. 1). The described alignment scheme automatically extracts the facial patches. The results after coarse alignment (curvature features) and fine alignment (TrICP) are reviewed manually (see Table 1). It is reviewed if the cropped patches contain forehead, eyes, nose and mouth, is straight, and contains no non-facial parts as the neck.

	curvature features + PCA	TrICP
neutral1	1.6	0.0
neutral2	0.0	0.0
random	4.9	1.6
laugh	8.1	1.6
smile	6.5	0.0
look up	14.7	13.1
look down	4.9	0.0

Table 1. Proportion of misalignments in the different datasets in % of total size of the dataset after the first and second alignment step.

All P2D-HMMs are trained using a 3×3-state structure. The feature vectors are clustered to 1024 codewords using k-means, leading to discrete probability distributions in the P2D-HMMs states. The proposed approach has the best recognition performance, if it is tested on *neutral2*. Since it is trained on a frontal dataset with neutral expression, it is obviously suited for this recognition task. The results show, that the warping ability of the P2D-HMM leads to acceptable recognition rates on the non-neutral facial expressions and different poses, too, but degrades if strong variations occur in the face. The suggested approach can compete with other state of the art algorithms (see Table 2). Note that the recognition results include misaligned datasets.

	P2D-HMM	salient wrinkles [9]	iso-geodesic stripes [10]
neutral 2	**95.1**	91	94.5
random	73.8	77	**80.5**
laugh	72.1	80	**81.1**
smile	**90.2**	83	84.4
look up	80.3	77	**92.8**
look down	88.5	80	**93.3**

Table 2. Recognition Rate in % on the different datasets

5. CONCLUSION

A fully automatic system able to detect and recognize faces in range images has been introduced. Coarse alignment based on curvature features and fine alignment applying TrICP leads to normalized facial patches which are suited for classification. It is shown, that recognition with the described P2D-HMM-approach leads to outstanding results for neutral facial expressions and to acceptable results over a wide range of different facial expressions and poses.

6. REFERENCES

[1] K. W. Bowyer, K. I. Chang, and P. J. Flynn, "A survey of approaches and challenges in 3D and multi-modal 3D + 2D face recognition," *CVIU*, vol. 101, no. 1, pp. 1–15, 2006.

[2] B. Ben Amor, K. Ouji, M. Ardebilian, and L. Chen, "3D face recognition by ICP-based shape matching," in *ICMI*, 2005.

[3] A.B. Moreno, A. Sanchez, J.F. Velez, and F.J. Diaz, "Face recognition using 3D local geometrical features: PCA vs. SVM," *IVC*, , no. 23, pp. 339–352, 2005.

[4] S. Eickeler, S. Mueller, and G. Rigoll, "Recognition of JPEG compressed face images based on statistical methods," *IVC*, vol. 18, no. 4, pp. 279–287, 2000.

[5] A. B. Moreno and A. Sanchez, "GavabDB, a 3D face database," *2nd COST Workshop on Biometrics on the Internet: Fundamentals, Advances and Applications*, pp. 77–82, 2004.

[6] D. Chetverikov, D. Svirko, D. Stepanov, and P. Krsek, "The trimmed iterative closest point algorithm," in *ICPR*, 2002.

[7] A. Lorusso, D. Eggert, and R. Fisher, "A comparison of four algorithms for estimating 3-d rigid transformations," in *BMVC*, 1995, pp. 237–246.

[8] C. Sanderson and K. Paliwal, "Fast features for face authentication under illumination direction changes," *Pattern Recogn. Lett.*, vol. 24, pp. 2409–2419, 2003.

[9] G. Antini, S. Beretti, A. Bimbo, and P. Pala, "3D face identification based on arrangement of salient wrinkles," in *ICME*, 2006.

[10] S. Berretti, A. Del Bimbo, and P. Pala, "Description and retrieval of 3D face models using iso-geodesic stripes," in *MIR*, 2006, pp. 13–22.