FACIAL POSE ESTIMATION USING ACTIVE APPEARANCE MODELS AND A GENERIC FACE MODEL

Thorsten Gernoth, Katerina Alonso Martínez, André Gooßen and Rolf-Rainer Grigat
Vision Systems E-2, Hamburg University of Technology, Harburger Schloßstr. 20, 21079 Hamburg, Germany

Keywords: Pose estimation, Active appearance model, Infrared imaging, Face recognition.

Abstract: The complexity in face recognition emerges from the variability of the appearance of human faces. While the identity is preserved, the appearance of a face may change due to factors such as illumination, facial pose or facial expression. Reliable biometric identification relies on an appropriate response to these factors. In this paper we address the estimation of the facial pose as a first step to deal with pose changes. We present a method for pose estimation from two-dimensional images captured under active infrared illumination using a statistical model of facial appearance. An active appearance model is fitted to the target image to find facial features. We formulate the fitting algorithm using a smooth warp function, namely thin plate splines. The presented algorithm requires only a coarse and generic three-dimensional model of the face to estimate the pose from the detected features locations. The desired field of application requires the algorithm to work with many different faces, including faces of subjects not seen during the training stage. A special focus is therefore on the evaluation of the generalization performance of the algorithm which is one weakness of the classic active appearance model algorithm.

1 INTRODUCTION

In the modern society there is a high demand to automatically and reliably determine or verify the identity of a person. For example, to control entry to restricted access areas. Using biometric data to identify a target person has some well known conceptual advantages, such as the identification procedure is immutable bound to the person which should be identified. Using facial images as a biometric characteristic has gained much attention and commercially available face recognition systems exist (Zhao et al., 2003, Phillips et al., 2007). However unconstrained environments with variable ambient illumination and changes of head pose are still challenging for many face recognition systems.

The appearance of a face can vary drastically if the intensity or the direction of the light source changes. This problem can be overcome by employing active imaging techniques to capture face images under invariant illumination conditions. In this work we use active near-infrared (NIR) illumination (Gernoth and Grigat, 2010). Possible surrounding light in the visible spectrum is filtered out.

Another benefit of active near-infrared illumination is the *bright pupil effect* which can be employed to assist eye detection. Pupils appear as unnaturally bright spots when an active near-infrared radiation source is mounted close to the camera axis (Figure 1). We use image processing to detect these bright spots in the images and thus can reliable detect the eyes (Zhao and Grigat, 2006).

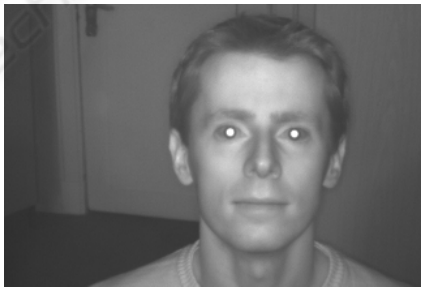


Figure 1: The bright pupil effect perceivable under active near-infrared illumination.

Challenging for face recognition systems are also changes of head pose. Appearance-based face recognition systems use the texture of faces in the form of two-dimensional frontal images to identify a target person. But faces are three-dimensional objects and due to head pose changes, their appearance in images can change significantly.

There are three different main strategies to over-

Gernoth T., Alonso Martínez K., Gooßen A. and Grigat R. (2010).
FACIAL POSE ESTIMATION USING ACTIVE APPEARANCE MODELS AND A GENERIC FACE MODEL.
In *Proceedings of the International Conference on Computer Vision Theory and Applications*, pages 499-506
DOI: 10.5220/0002835604990506

come this problem in appearance-based approaches for face recognition. The first is to use features which are invariant to these deformations, e. g. invariant to changes of the facial pose relative to the camera. Another strategy is to use or synthetically generate a large and representative training set. A third approach is to separate the factors which code the identity of a person from other sources of variation, such as pose changes. This is addressed in this paper. The posture of the head in-front of the camera is estimated from monocular images. The additional pose information may be utilized to register the facial images very precise and thereby make it possible to perform face recognition using a pose normalized representation of faces.

A survey of head pose estimation in computer vision was recently published by Murphy-Chutorian and Trivedi, 2009. We use active appearance models (AAM) to detect facial features in the images. Subsequently, the head pose is determined from a subset of the localized facial features using an analytical algorithm (DeMenthon and Davis, 1995, Martins and Batista, 2008). The algorithm can estimate the pose from a single image using four or more non-coplanar facial features positions and their known relative geometry. Using three-dimensional model points from the generic Candide-3 face model (Ahlberg, 2001, Dornaika and Ahlberg, 2006) and their image correspondences estimated using the active appearance model, the posture of the head in-front of the camera can be estimated.

Active appearance models are a common approach to build parametric statistical models of facial appearance (Cootes et al., 2001, Stegmann et al., 2000). The desired field of application requires the algorithm to work with many different faces, including faces not seen during the training stage (Gross et al., 2005). We use simultaneous optimization of pose and texture parameters and formulate the fitting algorithm using a smooth warping function (Bookstein, 1989). The thin plate spline warping function is parametrized efficiently to achieve some computational advantages. A special focus is on the evaluation of the generalization performance of the model fitting algorithm.

In Section 2 we introduce statistical models of facial appearance. Section 3 describes the smooth warping function. The pose estimation algorithm is explained in Section 4. With experimental results and discussion in Section 5, we conclude in Section 6.

2 STATISTICAL MODELS OF FACIAL APPEARANCE

We parametrize a dense representation of facial appearance using separate linear models for shape and texture (Matthews and Baker, 2004). The shape and texture parameters of the models are statistically learned from a training set.

2.1 Facial Model

Shape information is represented by an ordered set of l landmarks $\boldsymbol{x}_i, i = 1 \dots l$. These landmarks describe the planar facial shape of an individual in a digital image. The landmarks are generally placed on the boundary of prominent face components (Figure 2a). The two-dimensional landmark coordinates are arranged in a shape matrix (Matthews et al., 2007)

$$\boldsymbol{s} = \begin{pmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_l \end{pmatrix}^\top, \quad \boldsymbol{s} \in \mathbb{R}^{l \times 2}. \tag{1}$$

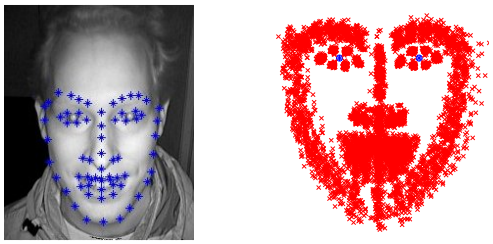
Active appearance models express an instance $\boldsymbol{s}_p$ of a particular shape as mean shape $\boldsymbol{s}_0$ and a linear combination of n eigenshapes $\boldsymbol{s}_i$, i.e.

$$\boldsymbol{s}_p = \boldsymbol{s}_0 + \sum_{i=1}^{n} p_i \boldsymbol{s}_i. \tag{2}$$

The coefficients p_i constitute the shape parameter vector $\boldsymbol{p} = \begin{pmatrix} p_i, \dots, p_n \end{pmatrix}^\top$. The mean shape $\boldsymbol{s}_0$ and shape variations $\boldsymbol{s}_i$ are statistically learned using a training set of annotated images (Figure 2a). Since reliable pupil positions are available (Zhao and Grigat, 2006), the training images can be aligned with respect to the pupils in a common coordinate system $\mathcal{I} \subset \mathbb{R}^2$ using a rigid transformation. The images are rotated, scaled and translated using a two-dimensional similarity transform such that all the pupils fall in the same position (Figure 2b). The mean shape $\boldsymbol{s}_0$ and basis of shape variations $\boldsymbol{s}_i$ are obtained by applying principal component analysis (PCA) on the shapes of the aligned training images (Cootes et al., 2001).

The texture part of the appearance is also modeled using an affine linear model of variation. Texture is defined as the intensities of a face at a discrete set $\mathcal{A}_0$ of positions $\boldsymbol{x}$ in a shape-normalized space $\mathcal{A} \subset \mathbb{R}^2$. The texture of a face is vectorized by raster-scanning it into a vector. Similar to the shape, $\boldsymbol{\lambda} = \begin{pmatrix} \lambda_i, \dots, \lambda_m \end{pmatrix}^\top$ denotes a vector of texture parameters describing a texture instance

$$\boldsymbol{a}_\lambda = \boldsymbol{a}_0 + \sum_{i=1}^{m} \lambda_i \boldsymbol{a}_i. \tag{3}$$



(a) Annotated image (b) Aligned shapes

Figure 2: Shapes of annotated images aligned with respect to pupil positions.

(a) $\boldsymbol{a}_0$ (b) $\boldsymbol{a}_1$ (c) $\boldsymbol{a}_2$

Figure 3: The mean texture $\boldsymbol{a}_0$ and the first two basis of texture variations $\boldsymbol{a}_i$.

The texture at position $\boldsymbol{x} \in \mathcal{A}_0$ of $\boldsymbol{a}_\lambda$ is a function of the domain $\mathcal{A}$, with

$$a_\lambda : \mathcal{A} \to \mathbb{R}; \quad \boldsymbol{x} \mapsto a_\lambda(\boldsymbol{x}). \tag{4}$$

To create a texture model, all the aligned training images are warped into the shape-normalized space. The shape-normalized space is given by the mean shape $\boldsymbol{s}_0$ of the shape model. A smooth warping function that maps one image to another by relating two sets of landmarks is used as described in Section 3. $\mathcal{A}_0$ contains positions that lie inside the mean shape $\boldsymbol{s}_0$. PCA is applied on the training textures to obtain the mean texture $\boldsymbol{a}_0$ and basis of texture variations $\boldsymbol{a}_i$.

Photometric variations of the texture $\boldsymbol{a}_\lambda$ are modeled separately by a global texture transformation $T_u(a_\lambda(\boldsymbol{x})) = (u_i + 1)a_\lambda(\boldsymbol{x}) + u_2$ (Baker et al., 2003). The intensities of the texture vector $\boldsymbol{a}_\lambda$ are scaled by a global gain factor $(u_i + 1)$ and biased by u_2.

To simplify the notation, the parameters describing shape, texture and photometric variations are combined into the single parameter vector

$$\boldsymbol{q} = \begin{pmatrix} \boldsymbol{p}^\top & \boldsymbol{u}^\top & \boldsymbol{\lambda}^\top \end{pmatrix}^\top. \tag{5}$$

2.2 Model Fitting

The parameters of the generative model described in Section 2.1 need to be estimated to fit the model to a target image. The target image can be aligned to the common coordinate system with respect to the pupils (Section 2.1). The target image is regarded as a continuous function of the domain $\mathcal{I}$:

$$I : \mathcal{I} \to \mathbb{R}; \quad \boldsymbol{x}' \mapsto I(\boldsymbol{x}'). \tag{6}$$

Fitting the model to an image is generally done by minimizing some error measure between the modeled texture and the target image. The error at the position $\boldsymbol{x} \in \mathcal{A}_0$ between the generated texture and the target image is

$$e(\boldsymbol{x}, \boldsymbol{q}) = a_\lambda(\boldsymbol{x}) - T_u(I(\boldsymbol{W}(\boldsymbol{x}, \boldsymbol{p}))). \tag{7}$$

$\boldsymbol{W}(\boldsymbol{x}, \boldsymbol{p})$ is a non-linear warping function that maps positions $\boldsymbol{x} \in \mathcal{A}$ of the model to positions $\boldsymbol{x}' \in \mathcal{I}$ of the target image. The warping function is parametrized by the shape parameters $\boldsymbol{p}$ as described in Section 3.

Typically the sum-of-squared error of all positions $\boldsymbol{x}$ is minimized to find the parameters $\boldsymbol{q}$, such that

$$\underset{\boldsymbol{q}}{\operatorname{argmin}} \frac{1}{2} \sum_{\boldsymbol{x} \in \mathcal{A}_0} [e(\boldsymbol{x}, \boldsymbol{q})]^2. \tag{8}$$

This is a non-linear optimization problem. General optimization techniques can be used to find a solution. Commonly used is an iterative Gauß-Newton style algorithm (Matthews and Baker, 2004). We assume a current estimate of $\boldsymbol{q}$ and solve for incremental updates $\Delta\boldsymbol{q}$ in each step. The update can be combined with the previous estimate in several ways (Matthews and Baker, 2004). The simplest update is the linear additive increment $\boldsymbol{q} \leftarrow \boldsymbol{q} + \Delta\boldsymbol{q}$. The following expression is minimized with respect to $\Delta\boldsymbol{q}$:

$$\underset{\Delta\boldsymbol{q}}{\operatorname{argmin}} \frac{1}{2} \sum_{\boldsymbol{x} \in \mathcal{A}_0} [e(\boldsymbol{x}, \boldsymbol{q} + \Delta\boldsymbol{q})]^2. \tag{9}$$

Performing a first-order Taylor expansion of the residual $e(\boldsymbol{x}, \boldsymbol{q} + \Delta\boldsymbol{q})$ around $\boldsymbol{q}$ yields:

$$e(\boldsymbol{x}, \boldsymbol{q} + \Delta\boldsymbol{q}) \approx e(\boldsymbol{x}, \boldsymbol{q}) + \left(\frac{\partial e(\boldsymbol{x}, \boldsymbol{q})}{\partial \boldsymbol{q}}\right)^\top \Delta\boldsymbol{q} \tag{10}$$

$$\approx e(\boldsymbol{x}, \boldsymbol{q}) + \left(\frac{\partial e(\boldsymbol{x}, \boldsymbol{q})}{\partial \boldsymbol{p}}\right)^\top \Delta\boldsymbol{p} + \left(\frac{\partial e(\boldsymbol{x}, \boldsymbol{q})}{\partial \boldsymbol{u}}\right)^\top \Delta\boldsymbol{u} + \left(\frac{\partial e(\boldsymbol{x}, \boldsymbol{q})}{\partial \boldsymbol{\lambda}}\right)^\top \Delta\boldsymbol{\lambda}. \tag{11}$$

The Gauß-Newton algorithm uses the update

$$\Delta\boldsymbol{q} = -\boldsymbol{H}^{-1} \sum_{\boldsymbol{x} \in \mathcal{A}_0} e(\boldsymbol{x}, \boldsymbol{q}) \frac{\partial e(\boldsymbol{x}, \boldsymbol{q})}{\partial \boldsymbol{q}} \tag{12}$$

with

$$\boldsymbol{H} = \sum_{\boldsymbol{x} \in \mathcal{A}_0} \frac{\partial e(\boldsymbol{x}, \boldsymbol{q})}{\partial \boldsymbol{q}} \left(\frac{\partial e(\boldsymbol{x}, \boldsymbol{q})}{\partial \boldsymbol{q}}\right)^\top. \tag{13}$$

Solving for $\Delta \boldsymbol{q}$ and using a Gauß-Newton step to optimize Eq. (9) involves computing an approximation of the Hessian matrix (Eq. (13)) and its inverse in each iteration. Assuming a constant Hessian matrix results in significant computational savings (Amberg et al., 2009, Cootes et al., 2001).

Matthews and Baker, 2004 showed that the linear additive increment is not the only parameter update strategy. They introduced compositional update strategies which permit overall cheaper algorithms for person-specific active appearance models. Their *simultaneous inverse compositional algorithm* for person-independent active appearance models (Gross et al., 2005) is however not as computationally efficient.

3 THIN PLATE SPLINE WARP

A warp function maps positions of one image to positions of another image by relating two sets of landmarks. The most common warp functions are the piecewise affine warp (Glasbey and Mardia, 1998) and the thin plate spline (TPS) warp (Bookstein, 1989). The affine warp function has the advantage of being simple and linear in a local region. But although it gives a continuous deformation, it is not smooth. Thin plate spline warping as an alternative produces a smoothly warped image. However, it is more expensive to calculate and non-linear due to the interpolating function used. In this paper we focus on the thin plate spline warp.

In the case of our active appearance model, the warp maps the positions $\boldsymbol{x}$ from the shape-normalized space $\mathcal{A}$ to positions $\boldsymbol{x}' \in \mathcal{I}$ of the target image. The transformation is such that the landmarks $\boldsymbol{x}_i, i = 1 \dots l$ are mapped to corresponding landmarks $\boldsymbol{x}'_i, i = 1 \dots l$ of a shape instance $\boldsymbol{s}_p$ in the target image. Since the landmark positions in the target image depend on the shape parameters $\boldsymbol{p}$, we parametrize the warp function by the shape parameter vector $\boldsymbol{p}$.

The thin plate spline warp function $\boldsymbol{W} : \mathcal{A} \rightarrow \mathcal{I}$ is vector valued and defined as (Bookstein, 1989, Cootes and Taylor, 2004):

$$\boldsymbol{W}(\boldsymbol{x}, \boldsymbol{p}) = \left(\sum_{i=1}^{l} \boldsymbol{w}_i U\left(\|\boldsymbol{x} - \boldsymbol{x}_i\|\right) \right) + \boldsymbol{c} + \boldsymbol{C}\boldsymbol{x} \tag{14}$$

$$= \underbrace{\boldsymbol{W}(\boldsymbol{p})}_{2\times(l+3)} \cdot \underbrace{\boldsymbol{k}(\boldsymbol{x})}_{(l+3)\times 1}, \tag{15}$$

with

$$\boldsymbol{W}(\boldsymbol{p}) = \begin{bmatrix} \boldsymbol{w}_1 & \dots & \boldsymbol{w}_l & \boldsymbol{c} & \boldsymbol{C} \end{bmatrix}, \tag{16}$$

$$\boldsymbol{k}(\boldsymbol{x}) = \begin{bmatrix} U(r_1(\boldsymbol{x})) & \dots & U(r_l(\boldsymbol{x})) & 1 & \boldsymbol{x}^\top \end{bmatrix}^\top, \tag{17}$$

where $r_i(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{x}_i\|$ is the Euclidean distance between a position $\boldsymbol{x}$ and landmark $\boldsymbol{x}_i$ of the mean shape $\boldsymbol{s}_0$. $U(r)$ is the TPS interpolating function (e.g. $U(r) = r^2 log r^2$ with $U(0) = 0$) that makes the warp function non-linear.

$\boldsymbol{W}(\boldsymbol{p})$ contains the warp weights. The weights $\boldsymbol{c}, \boldsymbol{C}$ represent the affine part of the mapping in Eq. (14). The warp weights are defined by the sets of source and destination landmarks to satisfy the constraints $\boldsymbol{W}(\boldsymbol{x}_i, \boldsymbol{p}) = \boldsymbol{x}'_i \; \forall \, i \in \{1 \dots l\}$ and to minimize the bending energy. Combining all constraints yields in a linear system (Bookstein, 1989):

$$\begin{bmatrix} \boldsymbol{K} & \boldsymbol{P} \\ \boldsymbol{P}^\top & \boldsymbol{O} \end{bmatrix} \boldsymbol{W}(\boldsymbol{p})^\top = \boldsymbol{L}\, \boldsymbol{W}(\boldsymbol{p})^\top = \begin{bmatrix} \boldsymbol{s}_p \\ \boldsymbol{o} \end{bmatrix} \tag{18}$$

where $\boldsymbol{K}$ is a $l \times l$ matrix and $K_{ij} = U\left(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|\right)$, the i'th row of the $l \times 3$ matrix $\boldsymbol{P}$ is $\begin{pmatrix} 1 & \boldsymbol{x}_i^\top \end{pmatrix}$, $\boldsymbol{O}$ is a 3×3 matrix of zeros and $\boldsymbol{o}$ is a 3×2 matrix of zeros. If $\boldsymbol{L}$ is non-singular the warp weights are given by

$$\boldsymbol{W}(\boldsymbol{p}) = \left(\boldsymbol{L}^{-1} \boldsymbol{B} \boldsymbol{s}_p \right)^\top \tag{19}$$

with

$$\boldsymbol{B} = \underbrace{\begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \dots & 1 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \end{bmatrix}}_{(l+3)\times l}. \tag{20}$$

During model fitting (Section 2.1) we estimate iteratively the parameters $\boldsymbol{p}$ of the shape $\boldsymbol{s}_p$ that corresponds to the shape of the target image. Since the residual in Eq. (7) is defined within the shape-normalized space $\mathcal{A}$ given by $\boldsymbol{s}_0$, the matrix $\boldsymbol{L}$ depends only on the landmarks $\boldsymbol{x}_i$ of the mean shape $\boldsymbol{s}_0$. The mean shape does not change for a give training set. $\boldsymbol{L}$ and its inverse can therefore be precomputed.

Parametrization of the warp function with respect to $\boldsymbol{p}$ yields another computational advantage. The texture is defined at a discrete set of positions $\mathcal{A}_0$ (Section 2.1). The residual in Eq. (7) only need to be evaluated at these positions. The positions depend on $\boldsymbol{s}_0$ and do not change for a give training set. $\boldsymbol{k}(\boldsymbol{x})$ can also be precomputed for all $\boldsymbol{x} \in \mathcal{A}_0$.

Using Gauß-Newton to minimize Eq. (9) requires the Jacobian matrix $\frac{\partial \boldsymbol{W}(\boldsymbol{x}, \boldsymbol{p})}{\partial \boldsymbol{p}}$ of the warp function with respect to the shape parameters (Eq. (11)):

$$\left(\frac{\partial e(\boldsymbol{x},\boldsymbol{q})}{\partial \boldsymbol{p}}\right)^{\top} = -\left(\frac{\partial T_u(I(\boldsymbol{W}(\boldsymbol{x},\boldsymbol{p}))}{\partial \boldsymbol{p}}\right)^{\top} \tag{21}$$

$$= -\left(\frac{\partial T_u(I(\boldsymbol{W}(\boldsymbol{x},\boldsymbol{p}))}{\partial \boldsymbol{W}(\boldsymbol{x},\boldsymbol{p})}\right)^{\top} \frac{\boldsymbol{W}(\boldsymbol{x},\boldsymbol{p})}{\partial \boldsymbol{p}} \tag{22}$$

Using the parametrization of the warp function with respect to $\boldsymbol{p}$ gives the components of the Jacobian matrix as follows:

$$\frac{\partial \boldsymbol{W}(\boldsymbol{x},\boldsymbol{p})}{\partial p_i} = s_i\left(\boldsymbol{L}^{-1}\boldsymbol{B}\right)^{\top} \boldsymbol{k}(\boldsymbol{x}). \tag{23}$$

The Jacobian does not depend on the value of the evaluation point $\boldsymbol{p}$ and can be precomputed for all $\boldsymbol{x} \in \mathcal{A}_0$.

4 POSE ESTIMATION

The pose of a face in-front of the camera is defined as its position and orientation relative to a three-dimensional camera coordinate system. We denote by $\hat{\boldsymbol{x}}_i$ a three-dimensional feature point of a face model in a coordinate system attached to the model. The perspective projection of a feature point onto the image plane of the camera is $\boldsymbol{x}_i$. The pose of a face with respect to a camera can be defined with a translation vector $\boldsymbol{t} \in \mathbb{R}^3$ and a rotation matrix $\boldsymbol{R} \in \mathrm{SO}(3)$:

- The translation vector is the vector from the origin of the coordinate system attached to the camera to the origin of the face model: $\boldsymbol{t} = \begin{pmatrix} t_x & t_y & t_z \end{pmatrix}^{\top}$.
- The rotation matrix is the matrix whose rows are the unit vectors of the camera coordinate system expressed in the coordinate system of the face model: $\boldsymbol{R} = \begin{bmatrix} \boldsymbol{i} & \boldsymbol{j} & \boldsymbol{k} \end{bmatrix}^{\top}$.

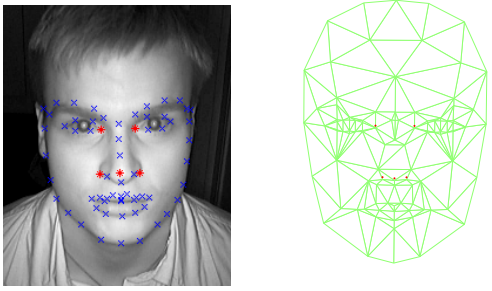
Under perspective projection with a camera having focal length f, $\boldsymbol{x}_i$ is related to the corresponding feature point of the model $\hat{\boldsymbol{x}}_i$ as follows:

$$\boldsymbol{x}_i = \frac{f}{\boldsymbol{k}^{\top}\hat{\boldsymbol{x}}_i + t_z}\left(\begin{bmatrix} \boldsymbol{i}^{\top} \\ \boldsymbol{j}^{\top} \end{bmatrix}\hat{\boldsymbol{x}}_i + \begin{bmatrix} t_x \\ t_y \end{bmatrix}\right). \tag{24}$$

The equation may be written as in DeMenthon and Davis, 1995,

$$\boldsymbol{x}_i(1+\epsilon_i) - \begin{bmatrix} t'_x \\ t'_y \end{bmatrix} = \begin{bmatrix} \boldsymbol{i}'^{\top} \\ \boldsymbol{j}'^{\top} \end{bmatrix}\hat{\boldsymbol{x}}_i, \tag{25}$$

with $\boldsymbol{i}' = \frac{f}{t_z}\boldsymbol{i}$, $\boldsymbol{j}' = \frac{f}{t_z}\boldsymbol{j}$, $t'_x = \frac{f}{t_z}t_x$, $t'_y = \frac{f}{t_z}t_y$ and $\epsilon_i = \frac{\boldsymbol{k}^{\top}\hat{\boldsymbol{x}}_i}{t_z}$. By setting $\epsilon_i = 0$, $\boldsymbol{x}_i$ equals the scaled orthographic projection of the face model point $\hat{\boldsymbol{x}}_i$. Scaled orthographic projection is similar to perspective projection if the depth of the face is small compared to its distance to the camera. For the case of fixed ϵ_i and assuming known projection of the model origin onto the image plane, the unknown $\boldsymbol{i}$, $\boldsymbol{j}$ and t_z can be computed from Eq. (25) from at least 4 non-coplanar feature points (DeMenthon and Davis, 1995, Martins and Batista, 2008). The third row of the rotation matrix $\boldsymbol{R}$ can be obtained from the cross product $\boldsymbol{k} = \boldsymbol{i} \times \boldsymbol{j}$. Using an iterative scheme and estimating ϵ_i from $\boldsymbol{k}$ and t_z of the previous iteration, an approximation of the pose can be computed.



(a) AAM Landmarks (b) Candide Model

Figure 4: (a) The selected landmarks for pose estimation and (b) the corresponding Candide-3 model.

4.1 Generic Face Model

A three-dimensional model of the face shape is required to estimate the pose. The Candide-3 model (Ahlberg, 2001) is used. This general model is used in its neutral state but adapted in scale for facial pose estimation of all subjects (Figure 4b).

4.2 Feature Point Selection

As stated above, the algorithm requires at least 4 non-coplanar feature points to estimate the facial pose. These points are chosen from the estimated landmark positioning provided by the AAM. A correspondence was established between the Candide-3 model points and the landmarks of the active appearance model.

Since only one generic face model is used, the landmarks which do not vary much between faces of different subjects are chosen as feature points for pose estimation. The variability of each landmark was studied in order to pick the most stable ones. As it can be seen in Figure 4a, two landmarks from the eye contour and landmarks on the nose are chosen. The nose tip is used as the origin of the coordinate system of the face model.

5 EXPERIMENTS

We use the TUNIR database (Zhao et al., 2007) for all experiments. The database consists of recordings

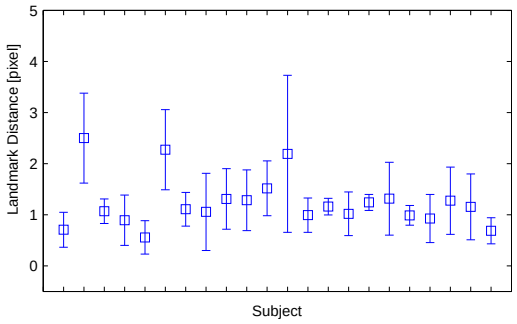


Figure 5: Mean and standard deviation of the distance between estimated and hand-labeled landmarks used for pose estimation for each subject (Person-specific AAM).

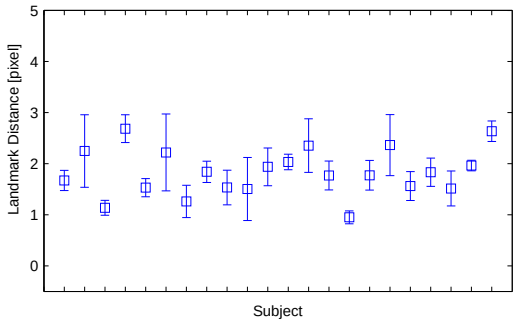


Figure 6: Mean and standard deviation of the distance between estimated and hand-labeled landmarks used for pose estimation for each subject (Generic AAM).

of 74 people in a typical access control scenario under active near-infrared illumination. The subjects move in-front of the camera and were asked to speak to recreate a realistic scenario.

5.1 Model Fitting

To evaluate the performance of the AAM fitting algorithm with smooth warp function, 5 images of 22 subjects from the TUNIR database were labeled in a semi-automatic way with 67 landmarks. Before fitting, the shapes were prepositioned according to the pupil positions, as described in Section 2.1. A hierarchical approach with two levels was chosen. To evaluate the fitting ability of the algorithm and for comparison, an experiment with known subject but novel image was conducted. Person-specific active appearance models were trained for each subject and evaluated in a leave-one-out cross-validation manner. 90 % of shape variance and 95 % of appearance variance of each training set were retained in the model. This corresponds to the optimal settings for the experiment with generic active appearance models described below. In Figure 5 the mean and standard deviation of the distance between the estimated and the hand-labeled landmarks are shown for each subject. Only the landmarks which are used for pose estimation contribute to the evaluation of the landmark distance.

Of more interest for the desired field of application is the performance with subjects not seen during training. In a second experiment generic active appearance models were trained from all but one subject. All images of the remaining identity were used to evaluate the performance of the fitting algorithm. Of interest is again the distance between the estimated and the hand-labeled landmarks used for pose estimation. This is shown in Figure 6.

As expected, the mean distance between estimated and hand-labeled landmarks is lower for person specific active appearance models. Nevertheless, the total mean distance is also just 1.8 pixels for generic active appearance models. In many cases the estimated landmarks fit well the face components but do not match the hand-labeled landmarks. This is because the hand-labeled positioning is not necessary the optimum one. Data refitting (Gross et al., 2005) could improve the performance. After a visual inspection of the estimated landmarks, an error up to 3 pixels was considered as good performance for the application.

5.2 Pose Estimation

To test the accuracy of the pose estimation algorithm quantitatively, we used the three-dimensional Candide-3 model. Of interest was the quality of the pose estimation from landmarks perturbed in the range of what can be expected from the AAM fitting algorithm. The Candide-3 model was situated in 729 different positions. We obtained simulated landmarks by projecting the three-dimensional model points perspectively to an image plane corresponding to the application scenario. We uniformly perturbed these landmarks from 1 pixels to 10 pixels and estimated the pose using the algorithm described in Section 4.

In Figure 7, the mean distance of the true model points to the model points of the Candide-3 model with estimated pose for different ranges of landmark perturbation is shown. Figure 7 shows that for a perturbation between 3 and 4 pixels, the mean model point distance is just around 7 pixels.

A typical result of pose estimation is shown in Figure 8. Since the posture of the subjects in-front of the camera is not known for the test images of the TUNIR database, the performance could only be evaluated qualitatively for this database.

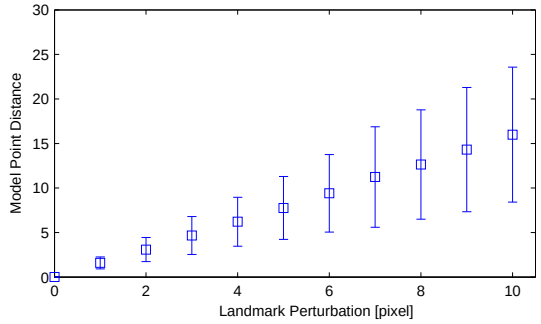


Figure 7: Mean and standard deviation of the distance between true model points and model points with estimated pose for different ranges of landmark perturbation.

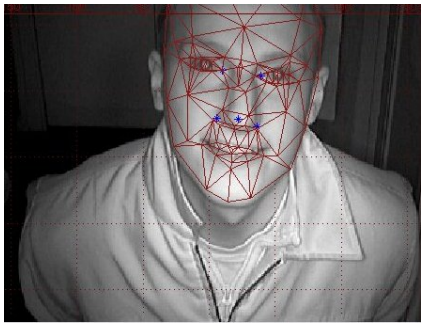


Figure 8: The estimated pose of a subject of the TUNIR database.

6 CONCLUSIONS

We presented an approach for facial pose estimation from two-dimensional images using active appearance models. Only a generic three-dimensional face model is required for pose estimation. We formulated the active appearance model fitting algorithm in an efficient manner with a smooth warp function. Our experiments show that the fitting accuracy of the algorithm is sufficient to estimate the pose from the detected landmark positions. Estimated poses from test images of the TUNIR database emphasize this result qualitatively.

ACKNOWLEDGEMENTS

This work is part of the project *KabTec – Modulares integriertes Sicherheitssystem* funded by the German Federal Ministry of Economics and Technology.

REFERENCES

Ahlberg, J. (2001). Candide-3 – an updated parameterized face. Technical Report LiTH-ISY-R-2326, Dept. of Electrical Engineering, Linköping University, Linköping, Sweden.

Amberg, B., Blake, A., and Vetter, T. (2009). On compositional image alignment, with an application to active appearance models. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009)*, pages 1714–1721.

Baker, S., Gross, R., and Matthews, I. (2003). Lucas-Kanade 20 years on: A unifying framework: Part 3. Technical Report CMU-RI-TR-03-35, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

Bookstein, F. (1989). Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585.

Cootes, T., Edwards, G., and Taylor, C. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685.

Cootes, T. and Taylor, C. (2004). Statistical models of appearance for computer vision. Technical report, University of Manchester, Manchester, U.K.

DeMenthon, D. F. and Davis, L. S. (1995). Model-based object pose in 25 lines of code. *International Journal of Computer Vision*, 15(1-2):123–141.

Dornaika, F. and Ahlberg, J. (2006). Fitting 3d face models for tracking and active appearance model training. *Image and Vision Computing*, 24(9):1010 – 1024.

Gernoth, T. and Grigat, R.-R. (2010). Camera characterization for face recognition under active near-infrared illumination. In *Proc. SPIE*, volume 7529, page 75290Z.

Glasbey, C. A. and Mardia, K. V. (1998). A review of image-warping methods. *Journal of Applied Statistics*, 25(2):155–171.

Gross, R., Matthews, I., and Baker, S. (2005). Generic vs. person specific active appearance models. *Image and Vision Computing*, 23(1):1080–1093.

Martins, P. and Batista, J. (2008). Monocular head pose estimation. In *Proc. 5th international conference on Image Analysis and Recognition (ICIAR '08)*, pages 357–368.

Matthews, I. and Baker, S. (2004). Active appearance models revisited. *International Journal of Computer Vision*, 60(2):135–164.

Matthews, I., Xiao, J., and Baker, S. (2007). 2d vs. 3d deformable face models: Representational power, construction, and real-time fitting. *International Journal of Computer Vision*, 75(1):93–113.

Murphy-Chutorian, E. and Trivedi, M. (2009). Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(4):607–626.

Phillips, P. J., Scruggs, W. T., O'Toole, A. J., Flynn, P. J., Bowyer, K. W., Schott, C. L., and Sharpe, M. (2007). FRVT 2006 and ICE 2006 Large-Scale Results. Technical Report NISTIR 7408, National Institute of Standards and Technology, Gaithersburg, MD.

Stegmann, M. B., Fisker, R., Ersbøll, B. K., Thodberg, H. H., and Hyldstrup, L. (2000). Active appearance models: Theory and cases. In *Proc. 9th Danish Conf. Pattern Recognition and Image Analysis*, pages 49–57.

Zhao, S. and Grigat, R.-R. (2006). Robust eye detection under active infrared illumination. In *Proc. 18th International Conference on Pattern Recognition (ICPR 2006)*, pages 481– 484.

Zhao, S., Kricke, R., and Grigat, R.-R. (2007). Tunir: A multi-modal database for person authentication under near infrared illumination. In *Proc. 6th WSEAS International Conference on Signal Processing, Robotics and Automation (ISPRA 2007)*, Corfu, Greece.

Zhao, W., Chellappa, R., Phillips, P. J., and Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458.