Learning Displacement Experts from Multi-band Images for Face Model Fitting

Christoph Mayer
Intelligent Autonomous Systems Group
Technische Universität München
München, Germany
mayerc@in.tum.de

Bernd Radig
Intelligent Automomous Systems Group
Technische Universität München
München, Germany
radig@in.tum.de

***Abstract*—Models are often used to gain information about real-world objects. Their parameters describe various properties of the modeled object, such as position or deformation. In order to fit the model to a given image, displacement experts serve as an update function on the model parameterization. However, building robust displacement experts is a non-trivial task, especially in real-world environments. We propose a novel approach that learns displacement experts from a multi-band image representation which is specifically tuned towards the task of face model fitting. We provide the fitting algorithm not only the original image but an image representation that reflects the location of several facial components within the face. To demonstrate its capability to work robustly not only in constrained conditions, we integrate the** *Labeled Faces In The Wild* **database, which consists of images that have been taken outside lab or office environments. Our evaluation demonstrates, that the information provided by this image representation significantly increases the accuracy of the model parameter estimation.**

Keywords*- *face model fitting; computer vision; human-maschine-interaction

I. Introduction

The interpretation of human face images is a traditional topic in computer vision research. The analysis of human faces provides information about person identity, facial expression or head pose and is the interest of several research groups. Model-based techniques have proven to be a successful method for extracting such high-level information from single images and image sequences. Face models reduce the large amount of image data to a small number of model parameters $\mathbf{p}$. These model parameters describe properties of interest of a single face, such as its position, shape or the visual appearance of its texture. However, in order to allow the extraction of high-level information to work robustly, model parameters that match the content of a single image have to be calculated to ensure that the model corresponds to the image content.

Model fitting is the computational challenge of finding these model configurations. Although a large number of face models and fitting strategies have been proposed, it is still a challenging task in uncontrolled environments when no restrictions towards face pose or lighting and only weak restrictions towards face size or image quality exist. Usually, approaches are evaluated on standard image databases that restrict the image content with respect to background, lighting, head pose or facial expression [19], [20]. Most image data is captured in controlled lab or office environments [12], [13], [14]. Some databases, such as the BioId Database [14] or the VALID database pose less constraints on background and lighting to create realistic image data. However, both databases do not include turned heads and half-profile views.

In contrast, we propose to evaluate face model fitting algorithms in an unconstrained environment. The contributions of this paper are two-fold: Firstly, we integrate the database *Labeled Faces In The Wild* database [18], which contains image material taken in real-world conditions. The images depict persons of public life and include both male and female, spanning a large variety of ethnic backgrounds, age, facial expressions, lighting and image backgrounds. See Figure 1 for some example images on which a face model has been fit automatically. Second, in order to tackle this challenging task, we provide the fitting algorithm not only with the original image data but with multi-band images which highlight facial components like eye brows or lips. These image-bands can be thought of as additional image channels but we restrain from this nomenclature in order to avoid confusions.

This paper continues as follows: Section II provides an overview about related approaches and the scientific background of this paper. Section III introduces our face model and displacement experts as well as our learning approach. Section IV demonstrates results of the experimental evaluation of our approach.

II. Related Work and Background

The interpretation of image data is often arranged in multiple steps with every step relying on the computation results of preprocessing steps. This section discusses related approaches in both important steps of this work: creating

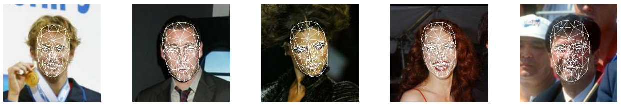

Figure 1. The face model is automatically fitted to every database image.

facial component related multi-band images and model fitting.

A. Computing the Facial Component Feature Bands

Skin color represents an important source of information to various computer vision applications. For a recent survey we refer to Phung et al. [3]. Most approaches identify a subspace of the color space to represent skin color, either by inspecting every single element of the color space or by defining clustering rules [4], [3]. Usually, this mapping is obtained by inspection of a (in some cases very large) number of training images. In contrast, we adapt our model to the specific image content to obtain a more robust estimation of the skin color distribution.

The approach of Hsu et al. is similar to ours such that it expects facial components to be located within an ellipsoid around the skin color area [4]. It constructs eye and mouth maps in order to verify the location of the entire face. A similar approach has recently been published by Beigzahed et al. who manually define rules to construct the eye and mouth maps for determining mouth and eye candidates [5]. We take the reverse approach, since we use an estimation of the face position to determine the location of facial features. Lips classifiers are able to provide useful information for speech recognition, speaker authentication and lip tracking [6], [7].

B. Related Approaches Towards Face Model Fitting

Most fitting strategies fall in one of two categories, utilizing either *objective functions* or *displacement experts*. Objective functions $f(I, \mathbf{p})$ yield a comparable value that determines how accurately a parameterized model $\mathbf{p}$ fits to an image I and are minimized to determine the optimal parameterization $\mathbf{p}_I$. This strategy is either applied on the raw image data with help of a photo-realistic appearance model [8], [10], [11] or rely on the extraction of image features, such as edges and color representation, Haar-like features, or color distribution [9], [21]. The draw-back of these approaches is that the minimization is usually computationally expensive and prone to local noise.

In contrast, displacement experts calculate a parameter update $\Delta\mathbf{p}$ on the initial parameters $\mathbf{p}_0$ to obtain the optimal model parameters $\mathbf{p}_I = \mathbf{p}_0 + \Delta\mathbf{p}$. The drawback of this approach is, that displacement experts can only be accurately learned when the displacement is small. This is only the case if a good initial pose estimate is available. Traditionally, these algorithms are evaluated on a set of well-known databases [12], [13], [14]. The images in these databases are manually annotated with model parameters or single points that reflect the position of eyes, mouth, nose tip or other facial features. The fitting algorithm is presented a subset of these images for training and is evaluated on another subset. Therefore, the fitting algorithm is able to consider only variances that are represented in the images. If, for instance, only frontal view images are included in the data, the algorithm is not able to work on half-profile views.

Cootes et al. [15] propose to utilize images with two feature bands for creating and fitting face appearance models. These feature bands reflect edge directions in two dimensions, where the magnitude indicates the degree of reliance in the orientation estimation. This approach is similar to our approach because not only the raw image data but an image representation with various additional feature bands is considered. Similarly, Stegmann et al. propose to utilize a multi-band image representation [16]. Edges and color bands obtained from converting the image to different color spaces are considered. Kahmaran et al. also take this approach but rely on a different color conversions [17]. In contrast to these approaches, our representation adapts to image conditions and the characteristics of the visible person.

III. Learning Displacement Experts from Multi-band Images

To fit the model and calculate the model parameters $\mathbf{p}_I$ for a single image I, we calculate a parameter update $\Delta\mathbf{p}$ to the current model parameters $\mathbf{p}$ as shown in Equations 1 and 2.

$$\mathbf{p}_I = \Delta\mathbf{p} + \mathbf{p}. \tag{1}$$

$$\Delta\mathbf{p} = \mathbf{g}(I, \mathbf{p}) \tag{2}$$

Since $\mathbf{p}$ is known, the challenge in this approach is the computation of $\Delta\mathbf{p}$. Because this computation has to be performed from the image data I and current model parameters $\mathbf{p}$ only, this requires to obtain robust calculation rules in $\mathbf{g}(I, \mathbf{p})$. To obtain this robustness, our approach provides the displacement expert not only with the original image I but with a multi-band image representation $\mathbf{I}$. The idea is, that since model parameters represent relative positions of facial components in the face, the fitting algorithm benefits from an image representation that specifically highlights these facial components. Although the displacement expert could

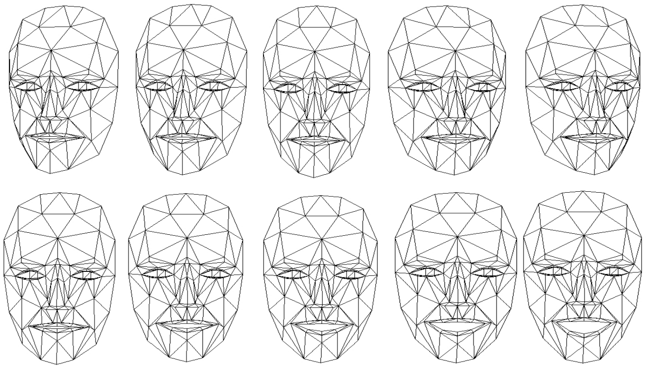

Figure 2. Model parameters change point positions to reflect face pose or shape change

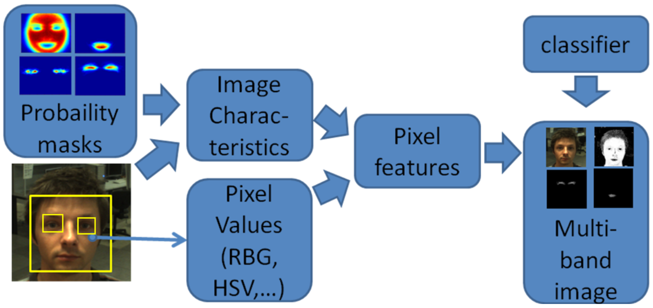

Figure 3. Pixel-based classifiers are trained on adjusted pixel features, that have been adapted to the image content via a set of image characteristics.

be applied to any model parameterization $\mathbf{p}$, we usually apply it to a fixed, initial parameterization $\mathbf{p}_0$:

$$\mathbf{p}_I = \mathbf{g}(\mathbf{I}, \mathbf{p}_0) + \mathbf{p}_0 \tag{3}$$

We apply this fitting approach to the Candide-3 face model, which is a wire-frame model consisting of $K = 116$ anatomical landmarks [2]. The $3K$ dimensional vector $\mathbf{v}$ contains the vertex x-, y- and z-coordinates. The model is controlled by applying the shape deformation $\mathbf{v}_{shape} = Ss + Aa$, a rotation matrix R, a scaling factor c and a translation $\mathbf{t}$ to the basic model structure $\mathbf{v}_{basic}$. The difference between the parameters in s and a is that A contains motion that may appear due to facial expressions whereas S contains vertex motions to adapt the general face structure to the face structure of a specific person. In total, the model vertex coordinates are computes according to Equation 4.

$$\mathbf{v} = cR(\mathbf{v}_{basic} + \mathbf{v}_{shape}) + t. \tag{4}$$

We denote the single vector elements, i.e. single parameters by p_i. The parameter vector is assembled according to Equation 5. Figure 2 depicts some example parameterizations.

$$\mathbf{p} = [c, t^T, r_x, r_y, r_z, a^T, s^T] = [p_1, ..., p_n, ..., p_N] \tag{5}$$

The remainder of this section will show learning displacement experts in two steps: First, the calculation of the multi-band image representation is discussed, and then the procedure to create a displacement expert from this image representation is presented.

A. Computation of Multi-Band Images

We rely on the approach presented by Mayer et al. and refer to their work for a detailed description and evaluation [1]. Still, for the sake of completeness, we will provide a short overview which is visualized in Figure 3. Firstly, characteristics of the current image are calculated that describe its content. Secondly, every pixel is inspected individually and a large amount of pixel features are calculated with help of the image characteristics. Since their calculation is based on the image characteristics they are adjusted to the image content and they are called ”adjusted pixel features”. Finally, pixel-based classifiers are trained on these pixel features that decide for a single pixel whether this pixel depicts a specific facial component or not. Figure 4 depicts examples for four different feature bands. We will inspect every single step in further detail.

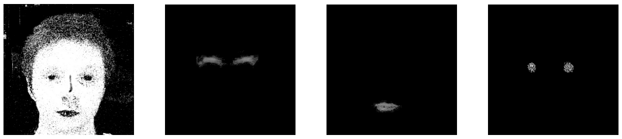

Figure 4. Image bands represent the probability of single pixels to depict certain facial components like skin, eye brows, lips and retinas for a specific image.

The **image characteristics** are calculated by applying a set of object locaters to determine a region of interest (ROI) around the face and the position of the eyes in the image, represented by yellow boxes in Figure 3. The approach of Viola et al. is integrated for this task [22].We generate a set of masks from training images that determine the probability for each pixel within the face bounding box to depict a specific facial component, see Figure 5 for a visualization of some example masks. By aligning the masks to the face bounding box, we map each pixel within the face bounding box to an entry in the probability masks. We denote the pixel by $\mathbf{x}$, its position within the image by $\mathbf{x}_s$ and its color values by $\mathbf{x}_c$. Furthermore, $\mathbf{x}$ is mapped to the mask entry $w^f_{\mathbf{x}}$ corresponding to the facial component f. We calculate the spatial mean of each facial component according to equation 6 and an estimation of average facial component colors according to equation 7.

$$\mu_f = \sum_{\mathbf{x} \in ROI} w^f_{\mathbf{x}} \mathbf{x}_s \tag{6}$$

$$\nu_f = \sum_{\mathbf{x} \in ROI} w^f_{\mathbf{x}} \mathbf{x}_c \tag{7}$$

The **adjusted pixel features** are computed for each pixel separately. They contain the pixel coordinates within the face bounding box and relative to the eye positions. Furthermore, distances $\mathbf{x}_s - \mu_f$ for different facial components are computed in different distance measurement, such as Euclidean distance and Mahalanobis distance. Additionally, the pixel's color relative to the estimated color means are calculated by $\mathbf{x}_c - \nu_f$ with respect to skin color distribution, brow color distribution and lip color distribution.

Training classifiers requires to manually annotate a set of training images with the facial component each pixel depicts. A face locater is applied to determine the face bounding box and the probability matrices are applied to calculate the image characteristics. Although, all pixels in all images' face

ISBN: 978-1-61208-117-5

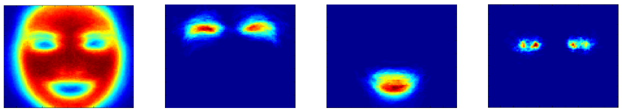

Figure 5. Probability masks are created beforehand to represent the spatial distribution of facial components in the face bounding box. Left to right: skin, eye brows, lips and retinas.

bounding boxes could be used as training data, due to the large amount of data only a subset of them is utilized. For every facial component, a classifier is trained from the extracted data to determine whether a single pixel depicts that specific facial component or not. In this paper we assemble a multi-band image $\mathbf{I} = \{I^{grey}, I^{skin}, I^{brow}, I^{lip}, I^{retina}\}$. Please note again, that this is only a rough overview of this approach. To learn more on how the spatial and color distributions are calculated and which distance measures are applied in detail, we refer to the work of Mayer et al [1].

B. Learning Displacement Experts

Ideally, a displacement expert should always determine the correct parameter update to obtain a perfect model fit $\mathbf{p}_I^\star$. Equation 8 depicts such an ideal displacement expert which simply computes the difference between the ideal model parameterization $\mathbf{p}_I^\star$ and the current model parameterization $\mathbf{p}$. Note, that the ideal model parameterization is usually not known, unless it is specified manually, which prevents this function from being practically applied. However, it will be applied to generate training data to train a further displacement expert $\mathbf{g}^l(\mathbf{I}, \mathbf{p})$ by combining Equation 3 and Equation 8. This displacement expert will be independent of $\mathbf{p}_I^\star$ and instead utilize the image data I as presented in Equation 9.

$$\mathbf{g}^\star(\mathbf{p}_I^\star, \mathbf{p}) = \mathbf{p} - \mathbf{p}_I^\star \tag{8}$$

$$\mathbf{g}^\mathbf{l}(\mathbf{I}, \mathbf{p}_I^\star + \Delta\mathbf{p}) = \Delta\mathbf{p} = \mathbf{g}^\mathbf{l}(\mathbf{I}, \mathbf{p}) \tag{9}$$

The first step is to annotate a set of images with the correct model parameters $\mathbf{p}_I^\star$. Obviously, it is still desirable to have $\mathbf{g}^l(\mathbf{I}, \mathbf{p}_I^\star) = 0$, since no parameter update is required in this case. However, we required training examples that reflect cases, when a parameter update is required. This training data is acquired automatically, by inducing random model parameter variations $\Delta\mathbf{p}$ to obtain new model parameterization $\mathbf{p} = \mathbf{p}_I^\star + \Delta\mathbf{p}$. In these cases the displacement expert is expected to determine the induced parameter variation according to Equation 9. Please see Figure 6 for a visualization of the feature extraction for two example annotations.

Therefore, training data consists of pairs of images and example parameterizations $< \mathbf{I}, \mathbf{p} >$, that are labeled with the induced parameter error $\Delta\mathbf{p}$. The displacement expert will be trained to determine $< \mathbf{I}, \mathbf{p} > \rightarrow \Delta\mathbf{p}$, which allows to fit the model according to Equation 1. In order to simplify this learning problem, we train single displacement experts for each single parameter separately. Equation 10 formalizes this step.

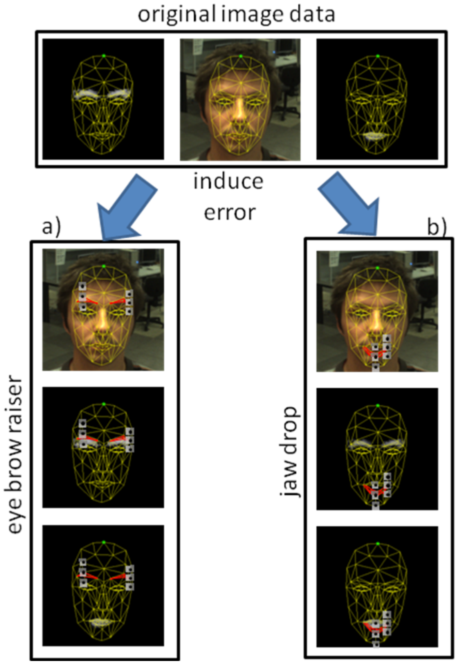

Figure 6. Features are extracted at points influenced by single parameter changes. Additional features are extracted along the direction of point motion from all image bands Due to space limitations, only selected image bands are presented here. Top: image bands of the original image with manually fit model. a) error induced in the eye brow raiser parameter. b) error induced in the jaw drop parameter.

$$\Delta p_i = g_i^l(\mathbf{I}, \mathbf{p}_I^\star + \Delta p_i) \tag{10}$$

It is usually beneficial in computer vision applications to apply a data reduction to the image data by extracting descriptive image features. In this paper, we extract Haar-like features in different styles and sizes. These features are calculated by summing up pixel intensity values in two image regions and then subtracting one of these sums from the other. These image regions are visualized in light grey and dark grey in Figure 6.

As mentioned, model parameters change the relative positions of model points. However, most model parameter influence only a small subset of model points. For instance, the parameter that represents rising eye brows has no influence on model points at the chin. Therefore, we extract image features only in the neighborhood of influenced model points. Features are extracted at the model point positions and at positions along the model point motion defined by the model parameter. To integrate all image-bands, the image features are calculated from every image-band, see Figure 6. Afterwards, they are assembled in an image feature vector.

We integrate model trees to create a mapping of these feature values to the model point's displacement. During execution time, we calculate the multi-band image representation from the probe image as demonstrated in Figure 3. We extract feature values from all image bands as shown in Figure 6. The trained model tree infers the model parameter update from these features and the model parameters are adapted, thus fitting the face model onto the image.

IV. Experimental Evaluation

This section conducts a twofold evaluation of our approach. First, the computation of the image-bands is inspected. Second, the accuracy of learned displacement expert is measured in unconstrained environments.

A. Classification of Facial Components

This experiment inspects the pixel classification accuracy. We collect a large number of images from the internet to cover a wide variety of people. The reason why we decided against using images of a standard database is, that we require the image content to vary in a large extend to train robust classifiers. We decided against using the "Labeled Faces in the Wild" database, because the images in this database are difficult to annotate due to their small size. The facial components are manually annotated in the images and the complete data set is split to images for training and testing. Adjusted pixel features are calculated from the training images and the classifiers are trained on them. Table I illustrates the accuracy of each classifier. The first row presents the number of correctly classified pixels (true positives and true negatives) divided by the total number of pixels within the face bounding box. The second row presents the number of pixels correctly classified as depicting a facial component is divided by the ground truth number of pixels depicting that facial component.

B. Fitting Accuracy on Multi-band Images

This section conducts an evaluation of our fitting approach. We collect random images from the Internet and annotate them to both, generate pixel classifiers and train displacement experts. For evaluation, we make use of the "Labeled Faces in the Wild" database, which provides images of people of public live that have been publish in the media. Therefore, these images have not been taken with a computer vision application in mind and include many challenges that have to be faced in real-world conditions. Due to the size of the database, only images of persons starting with the letter "A" are considered as a representative subset. To measure the accuracy of learned displacement experts, we induce errors $\Delta\mathbf{p}_I^{error}$ in manually specified models $\mathbf{p}_I^\star$ to create erroneous model parameterizations $\mathbf{p}^{error} = \mathbf{p}_I^\star + \Delta\mathbf{p}_I^{error}$. We apply the fitting algorithm to them to compute a parameter update $\Delta\mathbf{p}$ that is intended to compensate the artificially induced error. The difference $\Delta\mathbf{p} - \Delta\mathbf{p}_I^{error}$ between the induced error $\Delta\mathbf{p}_I^{error}$ and the suggested parameter update $\Delta\mathbf{p}$ serves as a measurement of accuracy. The maximum error induced is 1.0.

skin	lip	brow
91.5%	96.2%	92.6%
90.1%	95.1%	89.8%

Table I
The extraction of the additional image bands is robust even on images that are taken in real-world environment. This supports the subsequent task of model-fitting.

image bands	original	skin	lip	brow	retina
error rate	100.0%	78.1%	64.2%	60.7%	60.2%

Table II
Providing additional image bands reduces the fitting error.

C. Image Bands

In Table II the impact of provided image bands on the fitting accuracy is inspected. As a baseline, a displacement expert is trained and applied on the original image data only. Then, additional image bands are provided and the accuracy of these displacement experts is compared to the baseline displacement expert. The first row of the table shows which image band has been added to train the displacement expert. Single displacement experts are represented by table columns. Please note, that for every displacement expert, additional image bands are provided and therefore, the number of image bands provided increases with the table column index. For instance, the displacement expert represented by the third column, has been trained on the original image data, the skin color image band and the lip color image band and its error is 64.1% of the baseline error. The error is computed as the average of all single parameter error values for all probe images. Since error values are reduced with increasing table column index, providing additional image bands increases the fitting accuracy.

D. Absolute Fitting Accuracy

To obtain an absolute error measurement also, we visualize the cumulative error distribution of Δp_i for all images and a selection of parameters after fitting in Figure 7. We present only a subset of the complete parameter vector to prevent results from being difficult to read. We inspect five shape parameter and four facial expression parameters. This measures the fraction of models that have at most a specific error in a specific parameter after fitting. For instance, 60% of all models are fitted with an error of 0.2 or less in the jaw drop parameter. The displacement expert evaluated here is provided the complete set of image bands. The parameters are selected from the order proposed by the face model and intuitively reflect the most important changes. As can be seen, the displacement expert significantly compensates the induced error. In some few cases, however, the initial error is even increased by the fitting algorithm, represented by the

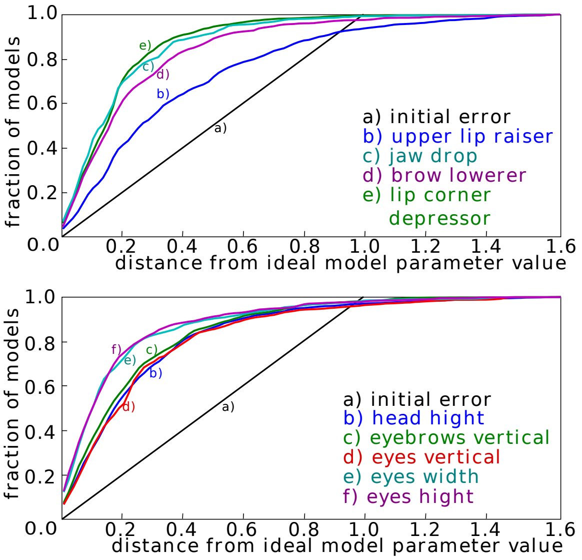

Figure 7. The fitting algorithm compensates the induced error to a large extend. For clarity of presentation, only selected parameters are evaluated.

models with errors larger than 1.0. This is due to occlusion by glasses, beards or long hair that covers the eye brows.

V. Conclusion and Future Work

We presented an algorithm for face model fitting in real-world environments, based on preprocessing the raw image data to generate multi-band images that highlight the position of various facial components. To demonstrate the robustness of the complete approach, images from the ”Labeled Faces in the Wild” database are taken for evaluation. Future work will consider applying the proposed techniques for face model tracking by calculating the multi-band images for subsequent images.

References

[1] C. Mayer and M. Wimmer, and B. Radig Adjusted Pixel Features for Facial Component Classification *Image and Vision Computing Journal, 2009*

[2] *J. Ahlberg. Candide-3 – an updated parameterized face. Technical Report LiTH-ISY-R-2326, Linköping University, Sweden, 2001.*

[3] *S.L. Phung, A. Bouzerdoum, and D.Chai. Skin segmentation using color pixel classification: analysis and comparison.* IEEE Transactions on Pattern Analysis and Machine Intelligence, *27(1), 2005.*

[4] *Rein-Lien Hsu, Mohamed Abdel-Mottaleb, and Anil K. Jain. Face detection in color images.* IEEE Transactions on Pattern Analysis and Machine Intelligence, *2002.*

[5] *M.Beigzahed and M.Vafadoost. Detection of face and facial features in digital images and video frames. In* Proceedings of the 4th IEEE Cairo International Biomedical Engineering Conference, *2008.*

[6] *M.T. Sadeghi, J.V. Kittler, and K. Messer. Modelling and segmentation of lip area in face images.* Vision, Image and Signal Processing, *2002.*

[7] *V. S. Sadeghi and K. Yaghmaie. Vowel recognition using neural networks.* International Journal of Computer Science and Network Security, *2006.*

[8] *D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In* 17th British Machine Vision Conference, *2006.*

[9] *D. Cristinacce and T. F. Cootes. Boosted regression active shape models. In* Proceedings of the British Machine Vision Conference, *2007.*

[10] *V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. In* Siggraph, Computer Graphics Proceedings, *1999*

[11] *T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In H. Burkhardt and Bernd Neumann, editors,* 5th European Conference on Computer Vision, *1998*

[12] *K. I. Chang, K. W. Bowyer, and P. J. Flynn. Face recognition using 2d and 3d facial data. In* ACM Workshop on Multimodal User Authentication, *2003.*

[13] *P. J. Phillips, H. Moon, S. A. Rizvi, and Patrick J. Rauss. The feret evaluation methodology for face-recognition algorithms.* IEEE Transactions on Pattern Analysis and Machine Intelligence, *2000.*

[14] *O. Jesorsky, K. J. Kirchberg, and R. Frischholz. Robust face detection using the hausdorff distance. In* Proceedings of the Third International Conference on Audio- and Video-Based Biometric Person Authentication, *2001*

[15] *T. F. Cootes and C. J. Taylor. On representing edge structure for model matching.* Computer Vision and Pattern Recognition, *2001.*

[16] *Mikkel Bille Stegmann and R. Larsen. Multi-band modelling of appearance.* Image and Vision Computing Journal, *2003.*

[17] *F. Kahmaran and M. Gokmen. Illumination invariant face alignment using multi-band active appearance models. In* Pattern Recognition and Machine Intelligence, *2005.*

[18] *Gary B. Huang, M. Ramesh, T. Berg, and E. Learned-miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments, 2008.*

[19] *H. Wu, X. Liu, and G. Doretto. Face alignment using boosted ranking models. In* Proceedings of the Conference on Computer Vision and Pattern Recognition, *2008.*

[20] *P. Tresadern, H. Bhaskar, S. Adeshina, C. Taylor, and T.F. Cootes. Combining local and global shape models for deformable object matching. In* Proceedings of the British Machine Vision Conference, *2009.*

[21] *S. Romdhani.* Face Image Analysis using a Multiple Feature Fitting Strategy. *PhD thesis, University of Basel, Computer Science Department, Basel, CH, January 2005.*

[22] *P. Viola and M. J. Jones. Robust real-time face detection.* International Journal of Computer Vision, *57(2):137–154, 2004.*