Face Model Fitting based on Machine Learning from Multi-band Images of Facial Components

Matthias Wimmer*
Perceptual Computing Lab
Waseda University
Tokyo, Japan

Christoph Mayer
Image Understanding Group
Technische Universität München
Garching, Germany

Freek Stulp
Cognitive Neuroinformatics Group
University of Bremen
Bremen, Germany

Bernd Radig
Image Understanding Group
Technische Universität München
Garching, Germany

Abstract

Geometric models allow to determine semantic information about real-world objects. Model fitting algorithms need to find the best match between a parameterized model and a given image. This task inherently requires an objective function to estimate the error between a model parameterization and an image. The accuracy of this function directly influences the accuracy of the entire process of model fitting. Unfortunately, building these functions is a non-trivial task.

Dedicated to the application of face model fitting, this paper proposes to consider a multi-band image representation that indicates the facial components, from which a large set of image features is computed. Since it is not possible to manually formulate an objective function that considers this large amount of features, we apply a Machine Learning framework to construct them. This automatic approach is capable of considering the large amount of features provided and yield highly accurate objective functions for face model fitting. Since the Machine Learning framework rejects non-relevant image features, we obtain high performance runtime characteristics as well.

1. Introduction

Model-based techniques have proven successful for extracting high-level information from images. A priori knowledge such as shape or texture allows to reduce the large amount of image data to a small number of model parameters. The model's parameter vector $\boldsymbol{p}$ represents its configuration, such as position, rotation, scaling, and deformation.

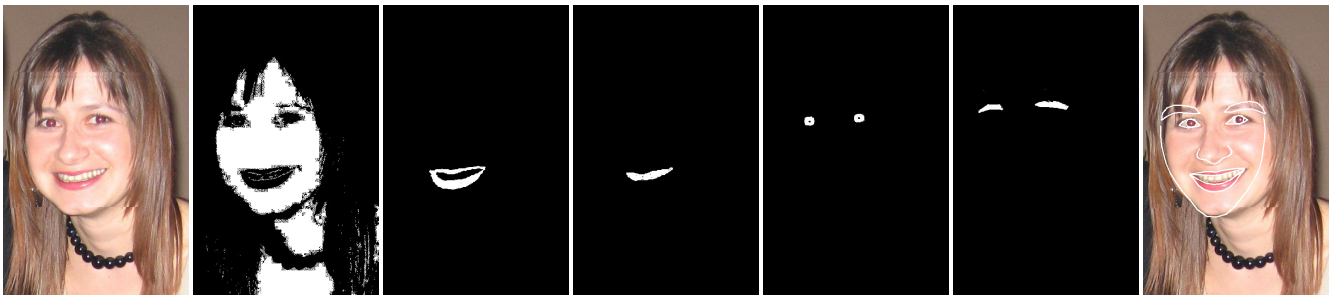
Model fitting is the computational challenge of finding the model parameters that best describe the given image and it usually consists of two components: the fitting algorithm and the objective function. The objective function $f(I, \boldsymbol{p})$ yields a comparable value that indicates how accurately a parameterized model $\boldsymbol{p}$ fits to an image I. In this paper, smaller values denote a better model fit. The fitting algorithm searches for the model parameters that minimize the objective function. Since the described methods are independent of the used fitting algorithm, this paper shall not elaborate on them but we refer to [5] for a recent overview and categorization.

Problem Statement: The accuracy of model fitting greatly depends on the quality of the objective function. This function is often designed manually using the designer's intuition and domain knowledge. He selects a small number of image features and formulates mathematical calculation rules [10, 3]. Afterwards, the function's appropriateness is subjectively determined by inspecting its result on example images and example model parameters.

To ensure well fit models, we calculate image features not only from the original image but also from a set of derived feature bands. Therefore, a vast amount of image features have to be evaluated for their benefit when the objective function is created. Humans are not able to consider this vast amount of data and therefore the traditional approach to design the objective function is not applicable.

Solution Idea: This paper proposes to automatically learn the objective function rather than designing it manually. Machine Learning allows to consider a large amount of image features, objectively evaluate their importance, pick

*This research is partly funded by a JSPS Postdoctoral Fellowship for North American and European Researchers (FY2007).



original skin lip teeth iris eye brows fitted model

Figure 1. Our multi-band image representation indicates the location of various facial components. An objective function that is able to consider this vast amount of information supports correctly fitting a face model.

the relevant ones and reject the rest. We provide the learning algorithm with image features from every band of a multi-band image representation that describes various facial components. Each pixel of a facial component band indicates the probability that it corresponds to a certain facial component, see Figure 1. It is robust to clutter and motion in the background and emphasizes the transitions between the facial components which are crucial for face model fitting.

The contributions of this publication are threefold. 1) We propose a Machine Learning framework to create objective functions for face model fitting. 2) Great accuracy is obtained because the machine learning process is able to objectively evaluate the benefit of each image feature from a large set. 3) Since it rejects non-relevant image features and only considers the relevant ones, the obtained objective function shows high runtime performance as well.

This paper continues as follows. Section 2 provides an overview about related approaches and the scientific background of this paper. Section 3 introduces our approach that learns robust objective functions from a multi-channel image representation. Section 4 discusses our approach and shows advantages and disadvantages. Section 5 demonstrates results of the experimental evaluation of our approach. Section 6 summarized our approach and points out future work.

2. Related Work and Background

This section reviews the scientific background for fitting deformable shape models including related approaches that consider multi-band image representations.

2.1. Fitting Active Shape Models

Active Shape Models (ASM) represent the variation in shape of deformable objects as a linear combination of a small number of modes of variation. A small number of parameters is capable of expressing large variations in shape. The shape $\boldsymbol{x}$ of an ASM is a linearized vector of N shape points generated by Equation 1, where $\bar{\boldsymbol{x}}$ is the mean shape, P is the matrix of orthogonal modes of shape variation, and $\boldsymbol{b}$ is the vector of deformation parameters.

$$\boldsymbol{x} = \bar{\boldsymbol{x}} + P\boldsymbol{b} \tag{1}$$

Both $\bar{\boldsymbol{x}}$ and P are obtained from a set of registered training shapes by computing the mean and by applying PCA, respectively. P is restricted such that it contains a small number of the most important variations only. Applying affine transformation T yields the global shape $\hat{\boldsymbol{x}}$ that can be located anywhere within the image.

$$\hat{\boldsymbol{x}} = T(t_x, t_y, s, \alpha, \boldsymbol{x}) \tag{2}$$

As described by Cootes et al. [1], the fitting strategy for ASMs is to individually consider each shape point $\hat{\boldsymbol{x}}_n$ and search for a better hypothesis $\hat{\boldsymbol{x}}_n^H$ by minimizing an objective function $f_n(I, \boldsymbol{u})$. Equation 3 describes this search, where the value of f_n indicates how accurately the location $\boldsymbol{u}$ within the image I describes the n^{th} shape point.

$$\hat{\boldsymbol{x}}_n^H = \underset{\boldsymbol{u}}{\operatorname{argmin}}\, f_n(I, \boldsymbol{u}) \tag{3}$$

In order to reduce the computational cost of this local search, $\boldsymbol{u}$ is usually constrained to lie on the perpendicular to the shape line within a certain search radius.

The so determined hypothesis of the shape $\hat{\boldsymbol{x}}^H$ usually does not obey the model restrictions, i.e. model parameters that would yield the shape $\hat{\boldsymbol{x}}^H$ do not exist. Therefore, the model parameters $\boldsymbol{p}$ are approximated in a second step by minimizing the squared distance between the hypothesis $\hat{\boldsymbol{x}}^H$ and the shape $\hat{\boldsymbol{x}}$ described by $\boldsymbol{p}$.

The advantage of this two-step fitting approach is that the search along the perpendicular is only one-dimensional. This allows to conduct exhaustive search, which does not suffer from getting stuck in local minima. However, its drawbacks are that this second approximation step might decrease the previously achieved accuracy of $\hat{\boldsymbol{x}}^H$ and yield the final result $\hat{\boldsymbol{x}}$.

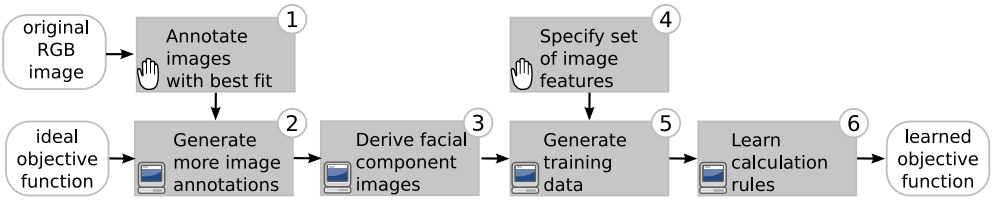


Figure 2. This figure illustrates how to learn an objective function from multi-band images. The image features are extracted not only from the original image data but also from a set of facial component feature bands.

2.2. Multi-band Image Representation of Facial Component

Our multi-band image representation describes the positions of different facial components, such as skin, lips, eyes, and brows. The pixel values of the various feature bands represent the probability to belong to a certain facial component. In order to quickly determine this value, simple classifiers would consider single pixel color information only. Unfortunately, high intra-class and small inter-class variations prevent these classifiers from yielding robust results, whereas more elaborate classifiers do not achieve real-time performance.

Recently, we proposed a two-step approach to solve this task [14]. In the first step, we determine context information about the given image and the visible person. For this reason, the bounding box around the face is obtained from face locators, such as the approach of Viola and Jones [13]. The bounding box is utilized to estimate the parameters of the skin color distribution, which is assumed to be Gaussian. This color distribution together with the geometric coordinates of the facial bounding box represents the characteristics of the entire image.

In the second step, we learn a quick classifier that is able to determine the facial component of a pixel. This approach is superior to simple classifiers, because it considers pixel information as well as the image characteristics at the same time. Therefore, the calculation is both, accurate and very fast at the same time.

2.3. Multi-band Model Fitting

Cootes et al. [2] 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. Therefore, the appearance model is not rendered as intensity values but as edge directions. 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. [11] propose to utilize a multi-band image representation. Edges and color bands obtained from converting the image into various color spaces are considered. They experienced a significant gain in accuracy. Kahmaran et al. [7] also take this approach but rely on a different image representation.

In contrast to these approaches, our representation adapts to image conditions and the characteristics of the visible person. It does not consider simple local features such as edges but sophisticated global image features that depend on global image properties.

3. Fitting Contour Models with Objective Functions

The contour points $\hat{\boldsymbol{x}}$ that are partially connected with lines are obtained from the model's parameter vector $\boldsymbol{p}$. The common fitting strategy for contour models [1] is to search for the best hypothesis $\hat{\boldsymbol{x}}_n^H$ of each contour point $\hat{\boldsymbol{x}}_n$ individually by minimizing a local objective function $f_n(\mathcal{I}, \boldsymbol{u})$ with $\boldsymbol{u}$ being a position in the image. Therefore, the success of ASM fitting relies on the quality of the objective function.

Traditionally, objective functions are specified manually by first selecting a small number of image features, such as edges or corners, and then formulating calculation rules that compute the function value from the feature values [1]. Our approach provides the objective function not only with the original image but with a multi-band image representation $\mathcal{I} = \{I^{gray}, I^{skin}, I^{teeth}, I^{lip}, I^{iris}, I^{brows}\}$.

3.1. Properties of Ideal Objective Functions

Functions that perfectly support the process of model fitting are called *ideal* objective functions. Two properties ensure that minimizing the objective function always results in a perfect model fit. First, the global minimum has to correspond to the correct position of the shape point. This property is important because the fitting algorithm will estimate the model parameterization from this minimum. Second, the objective function must not have any further local minima. This property is important because fitting algorithms can get easily stuck in local minima. Equation 4 depicts an ideal objective function $f_n^\star$ for the n^{th} shape point. It computes the Euclidean distance between a given location $\boldsymbol{u}$ on the image plane and the correct location of the shape point $\hat{\boldsymbol{x}}_n^\star$. Note that the vector of correct shape points $\hat{\boldsymbol{x}}^\star$

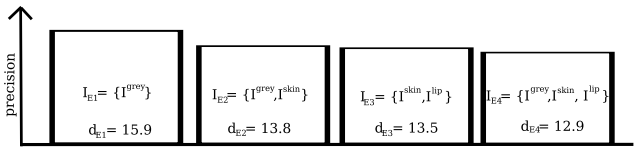


Figure 3. The selection of the feature bands that are considered by the objective function influences the precision of the objective function. The feature bands representing skin color and lip color provide more information than grey values.

must be specified manually.

$$f_n^\star(\mathcal{I}, \boldsymbol{u}) = |\boldsymbol{u} - \hat{\boldsymbol{x}}_n^\star| \tag{4}$$

Fitting Active Appearance Models (AAMs), such as the ones used by Stegmann et al. [11], requires a better initial parameter estimate than fitting ASMs. The reason is, that common AAM objective functions do not fulfill the second property. Therefore, the multi-band approach proposed here will outperform a multi-band implementation of AAMs.

3.2. Generating Image Annotations

Although $f_n^\star$ already provides ideal characteristics, it is not able to be used for previously unseen images, because it requires to manually specify the correct shape points $\hat{\boldsymbol{x}}_n^\star$ to compute its value. However, we propose to use $f_n^\star$ to generate training data for learning a further objective function f_n^ℓ that does not require knowledge of $\hat{\boldsymbol{x}}_n^\star$.

This six-step procedure is depicted in Figure 2. The key idea behind our approach is that $f_n^\star$ has the properties for idealness and generates ideal training data. f_n^ℓ is learned from this training data and will therefore approximately have these properties, too.

A set of manually annotated images with the correct shape points $\hat{\boldsymbol{x}}_n^\star$ forms the basis of this approach. For each $\hat{\boldsymbol{x}}_n^\star$, the ideal objective function returns the minimum $f_n^\star(\mathcal{I}, \hat{\boldsymbol{x}}_n^\star) = 0$ by definition. These correspondences between pixel coordinates and the function's result values do not yet represent sufficient training data to learn f_n^ℓ. In the second step, further correspondences between image coordinates and function values are automatically acquired by considering locations in the neighborhood of $\hat{\boldsymbol{x}}_n^\star$. Evaluating the ideal objective function at these locations returns a value greater than 0. This training data is sufficient to learn f_n^ℓ. For a more detailed inspection of this approach we refer to [15].

3.3. Learning Objective Functions from Multi-band Image Data

For more robustness, we do not learn f_n^ℓ from the plain image content, but we compute the multi-band image representation $\mathcal{I} = \{I^{\text{gray}}, I^{\text{skin}}, I^{\text{lip}}, I^{\text{brow}}, I^{\text{iris}}, I^{\text{tooth}}\}$ containing a set of facial components including I^{gray} that denotes the original image as described in Section 2.2. From each band, we extract a large set of Haar-like features [13] of different sizes and different styles. Now, we have a mapping from a set of feature values to the result value of the ideal objective function.

The last step learns $f_n^\ell(\mathcal{I}, \boldsymbol{p})$ from this large amount of training data. We choose tree-based linear regression [9] as Machine Learning algorithm, because this approach objectively investigates the benefit of each feature, picks the relevant ones, and rejects all other features. Therefore, the values of only a small number of image features need to be computed during runtime.

Performing model fitting for previously unseen images is now a two-step approach. In the first step, the multi-band image representation is calculated from the original image as shown in Section 2.2. Their feature bands are adapted to the current image because their calculation is based on image-specific properties. In the second step, the learned objective function is optimized in order to estimate the correct model parameters. This function calculates its value from the multi-band image representation.

4. Discussion

This procedure holds several advantages. First, the remaining manual step (annotating the images) is less error-prone than designing the entire function: It is intuitive to specify the correct model parameters for a moderate number of example images. According to its definition, the objective function returns zero for these examples. In contrast, explicitly formulating rules that yield the correct result for a variety of model parameters is difficult. Second, the procedure is mostly automated and therefore, it gets rid of the labor-intensive task of designing the objective function. Third, by using an ideal objective function to generate training data, the learned objective function is also approximately ideal. Fourth, this approach does not rely on expert knowledge and therefore, it is generally applicable and not domain-dependent.

Finally, extracting the image features from the original image data as well as from facial component feature bands holds further advantages. Since these image representations provide information about the location of various facial components, they decrease the influence of the image content of these components to the fitting process. The bottom line is that our approach yields more accurate objective functions, which greatly facilitate the model fitting task.

However, the quality of the objective function is limited by the quality of the training data. For instance, if the training set does not contain bearded men, the learned objective function will not be able to fit the model to images with bearded men correctly.

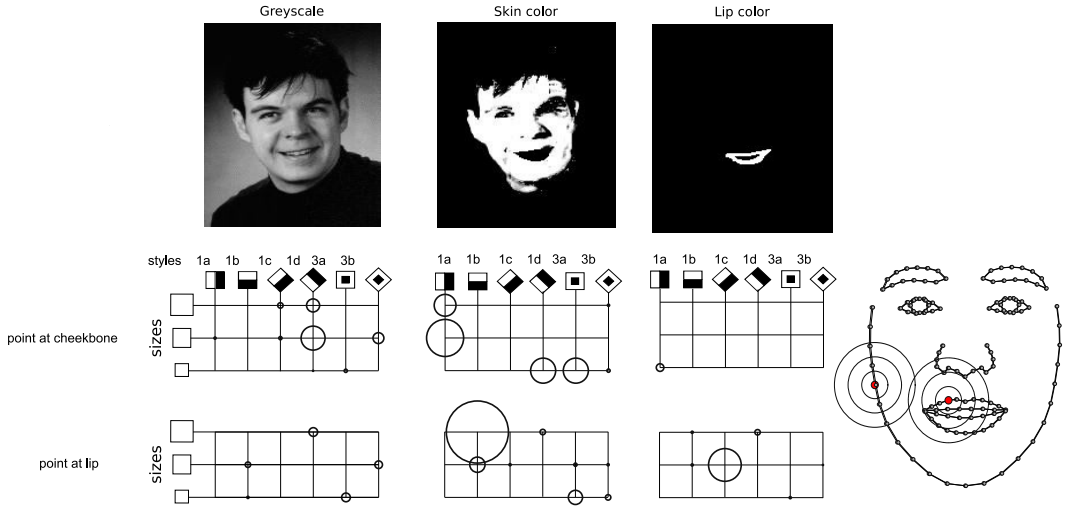


Figure 4. This chart illustrates the use of different Haar-like features within objective functions of two different shape points, the first one on the cheek line and the second one on the upper lip. Larger circles denote a more frequent use of a certain feature within the objective functions.

5. Experimental Evaluation

This section conducts a threefold evaluation of the learned objective functions. First, the accuracy of learning the objective functions is measured. Second, it inspects the image features selected for computing the function value. Finally, the runtime characteristics of learned objective functions are investigated.

The entire evaluation is conducted on a random collection of 500 images with faces acquired from the Internet. They show individual persons and they are taken without a computer vision application in mind. The images originate from different camera types and show different illumination conditions, head orientation, and facial expression. Their sizes vary between 0.1 and 1.0 M pixel, and the interocular distance varies between 30 and 230 pixels. We manually specified the correct model parameters ourselves. Therefore, evaluation on these images proves the general applicability[1].

We learned the objective functions on 67% of the images and the remaining 33% are considered for evaluation (167 images).

5.1. Comparing the Use of Different Facial Component Bands

This evaluation investigates how accurately an objective function is learned being provided with different facial component bands. We conduct four experiments E1 to E4 that learn objective functions from different image representations, providing the same number of features within each experiment to ensure comparability: $\mathcal{I}_{E1}=\{I^{\text{gray}}\}$, $\mathcal{I}_{E2}=\{I^{\text{gray}}, I^{\text{skin}}\}$, $\mathcal{I}_{E3}=\{I^{\text{skin}}, I^{\text{lip}}\}$, and $\mathcal{I}_{E4}=\{I^{\text{gray}}, I^{\text{skin}}, I^{\text{lip}}\}$. The objective functions for the individual shape points are evaluated on a separate test set showing different accuracy. Figure 3 demonstrates the accuracy by measuring the mean errors d_{E1} to d_{E4} between the value of the ideal objective function and the value of the learned objective functions. Comparing E1 to E2, Figure 3 clearly shows that providing additional information about skin decreases the error of the objective function because we mostly model transitions between skin colored regions and non-skin colored regions.

Since d_{E4} which considers the original image data, skin color and lip color is almost as accurate as d_{E3} which considers skin color and lip color only, providing the original image data does not greatly influence the fitting accuracy. However, providing a lip color feature band further improves the fitting accuracy.

5.2. Inspection of the Most Relevant Features

This section investigates which features are considered relevant in Experiment E4. We compare two shape points as visible in Figure 4. These shape points are chosen, because they are located at semantically different locations within the face: the cheek and the upper lip. As expected, they demonstrate the automatic selection of different facial component bands and different Haar-like features very well.

Figure 4 illustrates that Feature 1a in the skin color band is considered most important for the shape point located on the cheek. This feature is able to determine horizontal transitions. This is intuitive because the skin color band clearly separates the face from the background by a horizontal transition. In contrast to the gray value band, objects in the background do not affect the image content around

[1]The reader may verify this by uploading his own holiday photos to our Web-Service: http://www.someuniversity.com/author/FitFaceModel.php

this shape point.

Figure 4 illustrates that Feature 1b in the skin color band and Feature 1c in the lip color band are considered most important. This is intuitive because both features reflect the transition between the lips and the surrounding skin. In some cases, the lip color provides more information, e.g. beards cover the area around the mouth and make this transition difficult to detect in the skin color band. The lip color feature band in contrast still reflects this transition.

Both points consider the gray scale image only to a small extent. This proves that information obtained from the multi-band image representation is more important.

5.3. Runtime Characteristics

This section evaluates the timing characteristics of executing the learned objective functions of Experiment E4. Depending on the number of features provided, we inspect the number of features selected, which is similar to the number of arithmetic operations that have to be performed. As shown in Figure 5, increasing the number of features provided makes the tree-based regression algorithm reject more non-relevant features and therefore, the number of selected features converges to a fixed amount.

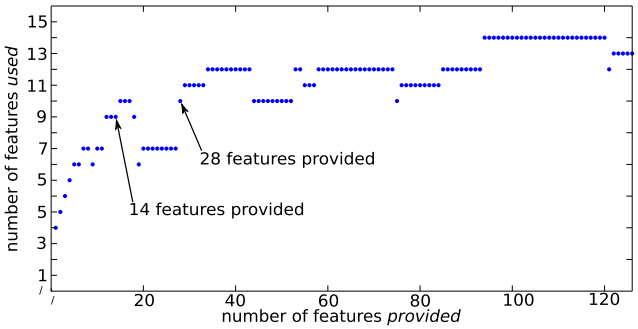


Figure 5. The number of features selected by tree-based regression is less than the number of features provided.

6. Conclusion and Outlook

In this paper, we propose a novel way to obtain robust objective functions for face model fitting. Their high accuracy is obtained, because they consider a large amount of image features, which are computed from a multi-band image representation that indicates the location of the facial components. Instead of conducting the tedious and erroneous process of manually formulating the objective function, we automatically learn this function from manually annotated images.

Our evaluation demonstrates that the Machine Learning algorithm rates the information gained from the facial component bands to be more relevant than the information of the original image. The execution time of the objective function converges to a fixed number of operations with a increasing number of features. In our ongoing research we will evaluate learning objective functions with multi-band for three-dimensional models.

References

[1] T. F. Cootes and C. J. Taylor. Active shape models – smart snakes. In *Proceedings of the 3rd British Machine Vision Conference*, pages 266 – 275. Springer Verlag, 1992.

[2] T. F. Cootes and C. J. Taylor. On representing edge structure for model matching. *CVPR*, 1:1114–1119, 2001.

[3] D. Cristinacce and T. F. Cootes. Facial feature detection and tracking with automatic template selection. In *International Conference on Automatic Face and Gesture Recognition*, 2006.

[4] B. Ginneken, A. Frangi, J. Staal, B. Haar, and R. Viergever. Active shape model segmentation with optimal features. *Transactions on Medical Imaging*, 21(8):924–933, 2002.

[5] R. Hanek. *Fitting Parametric Curve Models to Images Using Local Self-adapting Seperation Criteria*. PhD thesis, Department of Informatics, Technische Universität München, 2004.

[6] R.-L. Hsu, M. Abdel-Mottaleb, and A. K. Jain. Face detection in color images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(5):696–706, May 2002.

[7] F. Kahmaran and M. Gokmen. Illumination invariant face alignment using multi-band active appearance models. In *Pattern Recognition and Machine Intelligence*, pages 118–127, 2005.

[8] K. M. Lam and H. Yan. Locating and extracting the eye in human face images. *Pattern Recognition*, 29(5):771–779, 1996.

[9] R. Quinlan. Learning with continuous classes. In *Proceedings of the 5th Australian Joint Conference on Artificial Intelligence*, 1992.

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

[11] M. B. Stegmann and R. Larsen. Multi-band modelling of appearance. *IVC*, 21(1):61–67, 2003.

[12] Y.-L. Tian, T. Kanade, and J. F. Cohn. Recognizing action units for facial expression analysis. *IEEE PAMI*, 23(2):97–115, February 2001.

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

[14] M. Wimmer and B. Radig and M. Beetz. A Person and Context Specific Approach for Skin Color Classification. International Conference of Pattern Recognition 2006, pages 39-42, IEEE Computer Society.

[15] M. Wimmer and F. Stulp and S. Pietzsch and B. Radig. Learning Local Objective Functions for Robust Face Model Fitting, IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), 2008