Facial Model Fitting Based on Perturbation Learning and It's Evaluation on Challenging Real-World Diversities Images

Koichi Kinoshita[1,2], Yoshinori Konishi[1], Masato Kawade[1], and Hiroshi Murase[2]

[1] Technology and Intellectual Property HQ, Omron Corporation, Japan
[2] Graduate School of Information Science, Nagoya University, Japan

Abstract. We present a robust and efficient framework for facial shape model fitting. Traditional model fitting approaches are sensitive to noise resulting from scene variations due to lighting, facial expressions, poses, etc., and tend to spend substantial computational effort due to heuristic searching algorithms. Our work distinguishes itself from conventional approaches by employing (a) non-uniform sampling features unified by the shape model that affords robustness, and (b) regression analysis between observed features and underlying shape parameters that allow for efficient model update. We demonstrate the effectiveness of our framework by evaluating its performance on several new and existing datasets including challenging real-world diversities. Significantly higher localization accuracy and speedup factors of 15 have been observed comparing with the traditional approach.

1 Introduction

Facial model fitting methods are expected to fare well under variety of scene conditions resulting, for example, due to lighting changes (e.g., shadows), facial expressions, occlusions, etc. In addition to robustness, practical algorithms also require real time performance for acceptable use. In this work we propose a feature extraction and shape model estimation approach that is robust and computationally efficient.

Previous studies (e.g., [1]) have argued that appropriate combination of local information around each feature point (bottom up) and global information about their layout (top down) is important accurate facial feature localization.

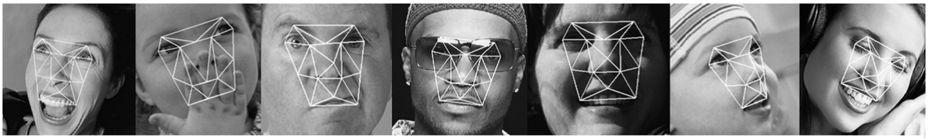


Fig. 1. Robust facial model fitting under variable lighting, complex expressions, and occlusion conditions

A. Fusiello et al. (Eds.): ECCV 2012 Ws/Demos, Part I, LNCS 7583, pp. 153–162, 2012.

Two most significant approaches that realize this concept are namely the Active Shape Model (ASM) [2] and Active Appearance model (AAM) [3], wherein, the local facial features are integrated by a global "Shape Model". These approaches recognized that the facial features lie on a low dimensional linear subspace within the high dimensional feature space.

ASM and AAM approaches have inspired several other facial feature point localization and representation methods. For instance, Li and Ito [4] employ AdaBoosted histogram classifiers that exploit local appearances by means of texture features. STASM [5] improves the model fitting accuracy of ASM by exploiting brightness gradient orthogonal to edge direction as local features, over multiple scales. Building further on the AAM framework, Blanz and Vetter [6, 7] employed 3D shape model and improved fitting performance to non-frontal face images. Matthews and Baker [8, 9] reduced computational time by utilizing the inverse compositional image alignment.

However, in each of these methods the fitting accuracy tends to decrease dramatically when the images contain unexpected variations such as shadow or facial expressions. Additionally, the previously proposed fitting techniques require a large number of iterations, making the real time robust model fitting a difficult problem to solve. In order to address these problems, our work proposes a novel shape model fitting algorithm which has the following main contributions: (a) non-uniform sampling features unified by the shape model that affords robustness, and (b) regression analysis between observed features and underlying shape parameters that allow for efficient model update.

Although [10][11] attempted to control shape parameters by regression, our proposed fitting method correlates the shape model to feature set sampled in a structured layout around each node of the shape model. Therefore, we call our approach as Active Structure Appearance Model (ASAM).

2 Features and Shape Model

2.1 Shape Model

Facial feature point layout can be compactly represented by lower dimensional linear subspace. Let $[x_m, y_m]^T$ represent the coordinates of m-th feature point node in the face image. Taken together, the feature point coordinates form a $2 \times M$-dimensional feature point set for the n-th image, $\hat{\mathbf{x}} = \left[[x_1, y_1]^T, \ldots, [x_M, y_M]^T\right] \in \mathcal{R}^{2\times M}$, where M is the total number of feature points in the image.

Let $\mathbf{x}$ denote the the normalized coordinates of the feature point set, which are related to corresponding $\hat{\mathbf{x}}$ by "pose parameters" $\mathbf{p} = [t_x, t_y, t_\theta, t_s]^T$ corresponding to rotation $\mathbf{R}$, translation $\mathbf{T}$, and scaling t_s,

$$\hat{\mathbf{x}} = \begin{bmatrix} \cos t_\theta & -\sin t_\theta \\ \sin t_\theta & \cos t_\theta \end{bmatrix} \mathbf{x} t_s + \begin{bmatrix} t_x \\ t_y \end{bmatrix} = \mathbf{R}\mathbf{x}t_s + \mathbf{T} \tag{1}$$

The normalized feature point set $\mathbf{x}$ can be compactly represented in a linear subspace of the high $2 \times M$-dimensional feature space. Given $\mathbf{x}$ for several training images, we employ PCA to obtain the reduced orthonormal basis set retaining top k basis vectors corresponding to k largest eigenvalues of the subspace spanned by feature point sets. Let the reduced set of basis vectors be denoted by $\tilde{\mathbf{\Phi}}$. Normalized feature $\mathbf{x}$ can be reconstructed from its projection $\mathbf{b}$ as $\mathbf{x} \approx \bar{\mathbf{x}} + \tilde{\mathbf{\Phi}}\mathbf{b}$, implying

$$\hat{\mathbf{x}} \approx \mathbf{R}(\bar{\mathbf{x}} + \tilde{\mathbf{\Phi}}\mathbf{b})t_s + \mathbf{T} \tag{2}$$

where, $\bar{\mathbf{x}}$ is the average normalized face model and $\mathbf{b}$ is the "shape parameter". The pose and shape parameters together represent the model parameter $\mathbf{\Theta} = [\mathbf{p}, \mathbf{b}]$.

2.2 Feature Sampling

Several feature representations have been evaluated in context of faces [5, 8, 12–15]. Face images have been exhaustively scanned by Gabor [13] and Harr-like [14] features for feature extraction. Most of ASM based methods represent features as one dimensional sampling along the edge normal at the feature node [5]. Such a representation is relatively low dimensional and does not robustly capture the underlying feature. Instead, features are easily affected by noise caused by shadow, occlusions, facial expression, etc., and therefore cannot result in reliable shape model fitting.

On the other hand, AAM based methods generally define a homogeneous sampling grid on the average shape model, which is transformed to obtain sampling coordinates for other face images [8]. In these cases, the feature vector tends to be high dimensional, thus requiring high computational cost for shape transformation. Furthermore, such a representation captures unessential information in areas which do contribute to model deformation (e.g. cheek, forehead). In fact, this superfluous information tends to be harmful for model fitting under noisy conditions.

To address these problems, we employ feature sampling method called "Retinotopic Sampling" [12], in which sampling points radiate out from each node of the shape model. In contrast to [12], where sampling was done independently at each feature node, our work associates the non-uniformly sampled features together by means of the shape model. Since the sampling distribution is associated with the shape model structure, we call this sampling method as Structural Retinotopic Sampling. In this work we manually select a particular sampling pattern as shown in fig. 2. A sampling operation given the model parameters $\mathbf{\Theta}$ will be indicated as:

$$\mathbf{f} = \mathcal{S}(\mathbf{p}) \tag{3}$$

here $\mathbf{f}$ is a sampled feature vector.

3 Feature Perturbation Analysis

In this section we show the relative shift between the ground truth position can be inferred based on a feature subspace learned during off-line training.

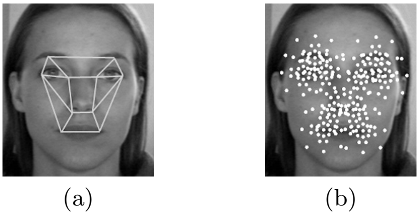


Fig. 2. (a)Shape model and (b) corresponding sampling points

This reduces the required number of search iterations and enables accurate feature localization.

Fig.3(a) shows a test image and selected feature sampling layout. In this experiment the inner corner of the right eye is selected as the basepoint. For training we use 200 images and we extract 5 perturbation samples from each image, resulting in a total of 1000 training examples. Features are perturbed in the following range: shift = within 20% of eye width, rotation = within +/ − 30 deg, and scale = 0.5 - 1.5 of the original. For this experiment, we employ pixel brightness at sample locations as feature descriptor. Reduced feature subspace is obtained by applying PCA to this perturbation feature set.

Fig.3(b) shows first two dimensions of the feature subspace, where half the perturbation features are centered to the left of the eye corner and the other half centered on the other side. Samples are obtained from 100 facial images which are different from training examples. Feature perturbation is limited to a distance of 20% of eye width. Similarly, fig.3(c) shows feature subspace spanned by 2nd and 3rd principal components. In this case, half the perturbation feature are sampled at +30 deg orientation with respect to the base feature, and the remaining are sampled at −30 deg orientation.

These examples clearly demonstrate that the perturbation feature subspace is discriminative in terms of the induced perturbations. These results, further reinforces the idea that we can estimate the induced perturbation by utilizing the compact subspace of perturbation features.

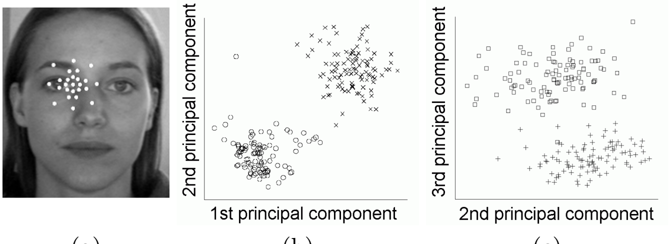


Fig. 3. (a) Sampling points. (b), (c) Feature vector plots on the feature subspace, circle: inside, x: outside, rectangle: −30 deg, +:+30 deg.

4 Shape Model Correlation with Perturbation Features

This section extends the concept and describes an algorithm to estimate the perturbation of the shape model parameters by learning a relationship between perturbed features and perturbation condition of the shape model during training. We employ Canonical Correlation Analysis (CCA) to learn the correlation between sampled features and the shape model parameters.

4.1 Canonical Correlation Analysis

Let $\mathbf{f} = [f_1, \cdots, f_k]^T$ and $\mathbf{p} = [p_1, \cdots, p_l]^T$ denote the k dimensional feature vector and l dimensional parameter vector, respectively. For some vectors, $\mathbf{a}$ and $\mathbf{b}$, $\mathbf{u} = \mathbf{a}^T\mathbf{f}$ and $\mathbf{v} = \mathbf{b}^T\mathbf{p}$ represent arbitrary linear transformations of $\mathbf{f}$ and $\mathbf{p}$, respectively. In order to find optimal values of $\mathbf{a}, \mathbf{b}$ that maximize the correlation between $\mathbf{u}$ and $\mathbf{v}$, the following covariance should be maximized:

$$Cov(\mathbf{u}, \mathbf{v}) = \mathbf{a}^T \mathbf{\Sigma} \mathbf{b} \tag{4}$$

where, $\mathbf{\Sigma}$ is the cross covariance matrix between $\mathbf{f}$ and $\mathbf{p}$. This problem can be solved as a standard eigenvalue problem by using Lagrange multipliers after normalizing variance of both $\mathbf{u}, \mathbf{v}$ to 1.

Let $(\mathbf{ea}_1, \ldots, \mathbf{ea}_k)$ and $(\mathbf{eb}_1, \ldots, \mathbf{eb}_l)$ denote the k and l eigenvectors, respectively, obtained as a solution to this problem. Assuming, $k > l$, $\mathbf{u}, \mathbf{v}$ can be written as $\mathbf{u} = [\mathbf{ea}_1, \ldots, \mathbf{ea}_l]^T\mathbf{f} = \mathbf{A}^T\mathbf{f}$ and $\mathbf{v} = [\mathbf{eb}_1, \ldots, \mathbf{eb}_l]^T\mathbf{p} = \mathbf{B}^T\mathbf{p}$.

If $\lambda_1, \cdots, \lambda_l$ denote the corresponding eigenvalues, the liner regression from $\mathbf{u}$ to $\mathbf{v}$ can be written as:

$$\mathbf{v} = diag[\lambda_1, \cdots, \lambda_l]\mathbf{u} = \mathbf{\Lambda}\mathbf{u} \tag{5}$$

Finally, the mapping $\mathbf{f} \Rightarrow \mathbf{p}$ can be obtained as:

$$\mathbf{p} = \mathbf{G}\mathbf{f}, \quad \text{such that} \quad \mathbf{G} = (\mathbf{B}^T)^{-1}\mathbf{\Lambda}\mathbf{A}^T. \tag{6}$$

4.2 Training Procedure and Model Fitting

The training procedure for learning regression model is outline in Algorithm 1. Ground truth feature point locations are assumed to be available for the training images. Briefly, for each training image model parameters, random perturbations are generated and corresponding sampled features are obtained. The relationship between known perturbed model parameters and sampled features is learnt using CCA (Sect. 4.1).

As an example to demonstrate the efficacy of the learning algorithm, fig. 4 shows scatter plots of ground truth values (horizontal axis) versus the estimation result (vertical axis) of model parameters predicted by learned regression model. Correlation coefficient, r, is shown under each plot. All parameters have a positive correlation value greater than 0.5 indicating that transformation matrix, G captures the relationship between features and model parameters effectively.

```
**Input**: $N$ Training images with annotated feature point coordinates $\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_N$
**Output**: Regression matrix, $\mathbf{G}$
for $n \leftarrow 1$ to $N$ do
    Obtain model parameters $\mathbf{\Theta}_n$ from $\hat{\mathbf{x}}_n$ as described in Sect. 2.1;
    for $r \leftarrow 1$ to $R$ do
        Generate, $\mathbf{\Theta}_{\text{err}} = \mathbf{\Theta}_n + \mathbf{\Delta\Theta}_r$, where $\mathbf{\Delta\Theta}_r$ is random a perturbation;
        Sample feature $\mathbf{f}_r$ corresponding to $\mathbf{\Theta}_{\text{err}}$ according to eqn. 3;
    end
end
Apply CCA to $\{\mathbf{\Delta\Theta}\}$ and $\{\mathbf{f}\}$ to obtain $\mathbf{G}$, such that $\mathbf{\Delta\Theta} = \mathbf{Gf}$;
return $\mathbf{G}$
```

Algorithm 1. Training procedure to learn regression function $\mathbf{G}$

At run time the model fitting procedure described in Algorithm 2 is employed. As an input to model fitting, we assume that rough face parameters (location, rotation, and size) are available from the underlying detection method, e.g., [14]. The fitting algorithm starts from the initial model parameters to obtain corresponding sampled features, f. The learned transformation $\mathbf{G}$ (eqn. 6) is used to determine the parameter perturbation $\mathbf{\Delta\Theta}$. A correction is applied to update the model parameters and the process is repeated until convergence. Several different convergence criterions are conceivable, e.g., maximum number of iterations, $||\mathbf{\Delta p}_i|| < \varepsilon$, or use of a trained classifier to evaluate the feature score, etc.

5 Experiments

We conducted exhaustive experimental evaluation of the proposed algorithm on various complex face databases and compared the performance with a state-of-the-art algorithm. The four datasets employed for testing include two public datasets, BioID [16] and the extended Yale face database B [17], and two new datasets INC and Snap. These datasets contain large variations in lighting conditions, facial expressions, occlusions, etc. A summary of dataset composition can be found in table 1. A total of 10,000 images across all datasets were used for training.

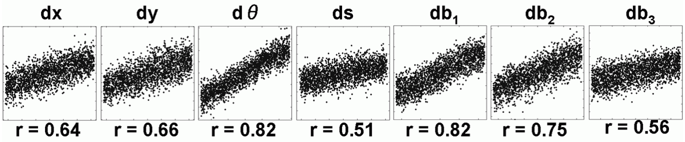


Fig. 4. Scatter plots of model parameters: Ground truth (horizontal axis) vs. estimated values (vertical axis) are shown for pose parameters $(t_x, t_y, t_\theta, t_s)$ and shape parameters (b_1, b_2, b_3). r denotes correlation value in each case.

```
**Input**: Face image with rough estimate of pose parameters $\mathbf{p}_0 = [\hat{t}_x, \hat{t}_y, \hat{t}_\theta, \hat{t}_s]$
**Output**: Optimal model parameters, $\mathbf{\Theta}_{opt}$
initialize $i \leftarrow 1$, $\mathbf{\Theta}_i = [\mathbf{p}_0, \mathbf{0}^T]$;
**repeat**
    Sample feature $\mathbf{f}_i$ corresponding to $\mathbf{\Theta}_i$ according to eqn. 3;
    Obtain parameter perturbation $\mathbf{\Delta\Theta} = \mathbf{G}\mathbf{f}_i$;
    Update parameter $\mathbf{\Theta}_{i+1} = \mathbf{\Theta}_i + \eta\mathbf{\Delta\Theta}$        /*$\eta$ is learning rate */
    $i \leftarrow i + 1$;
**until** *convergence*;
$\mathbf{\Theta}_{opt} \leftarrow \mathbf{\Theta}_{i-1}$;
**return** $\mathbf{\Theta}_{opt}$
```

Algorithm 2. Model fitting procedure estimates optimal model parameters given rough initialization.

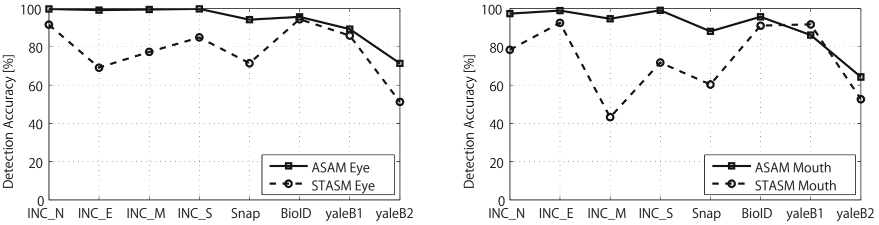
Table 1. Evaluation Datasets

	Condition	Facial expression	Number of images
BioID	indoor, homogeneous lighting	including open, closed eyes and mouth	1,521
YaleB	indoor, homogeneous and directional lighting	neutral (frontal pose only)	601
Snap	various lighting condition including indoor room lighting and outdoor natural lighting	various expressions including smile	2,325
INC	indoor, homogeneous lighting	neutral(N), close eyes(E), open mouth(M), smile(S).	300 for each expression

For experiments, Haar-like features [14] with different shapes and orientations were extracted at sampling locations. A total of 6 features at 235 sampling locations resulted in feature dimension of $6 \times 235 = 1410$. The proposed algorithm is compared with the state-of-the-art ASM based STASM [5] approach as baseline. Unlike AAM approaches which estimate appearance, a comparison with ASM based approach used for feature localization is more in line with the proposed framework.

All algorithms are implemented in C/C++ and executed on a Pentium D 3.2GHz PC. Shape model parameters for test images are initialized by face detection algorithm of [18]. Model fitting relies on 6 feature point locations, namely, corners of eyes and mouth. To accommodate for error in ground truth, feature localization results that fall within 10% of eye-to-eye distance are considered positive detections.

Fig. 5 shows the average localization accuracy results for eye and mouth corners for different datasets. For a better perspective, the results for Yale database are broken down as yaleB1, consisting of lighting angles less than 20 deg, and yaleB2, denoting other lighting conditions. Similarly, the INC database is



(a) (b)

Fig. 5. Detection accuracy for (a) Eye and (b) Mouth of the proposed approach (ASAM) compared to the baseline (STASM) on several evaluation datasets

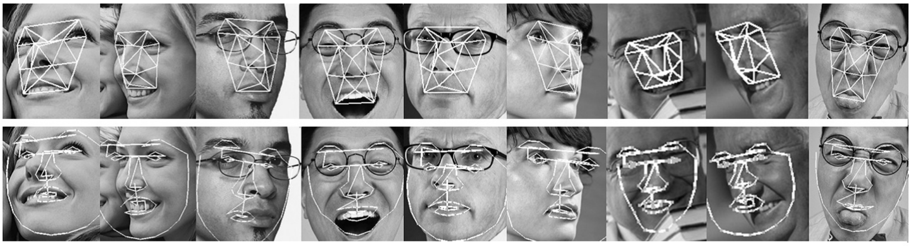


Fig. 6. Qualitative model fitting results comparing the outputs of proposed ASAM (1^{st} row) and baseline STASM (2^{nd} row) algorithms

sub-categorized based on facial expressions. Detection accuracy is defined as the ratio of number images with successful feature localization to the total number of evaluation images.

Detection accuracy for both of ASAM and STASM are similar for datasets which contain only neutral expression and homogeneous lighting (INC_N, BioID, yaleB1). However, ASAM demonstrates significantly superior performance compared to STASM for datasets with complex facial expression (INC_E, INC_M, INC_S). Although both methods show lower performances for extreme directional lighting condition in yaleB2, ASAM still has a better performance than STASM. Lastly, ASAM again outperforms STASM with significant margins on the Snap dataset which includes various facial expressions and lighting conditions.

The average frame processing time for ASAM is 0.017 sec compared to 0.264 sec for STASM. This corresponds to a speedup of 15.

5.1 Discussion

The superior performance and efficiency of the proposed method can be attributed mainly to the novel (1) structural retinotopic sampling integrated by the shape model and (2) perturbation estimation using CCA.

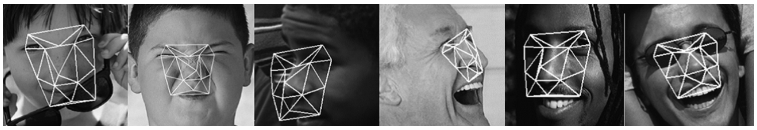


Fig. 7. Model fitting failure results

Structural Retinotopic Sampling: For facial model fitting, areas such as cheek and forehead carry less discriminative information compared to other features such eye, nose, and mouth corners. Homogenous feature sampling approaches get distracted by noise due to shadows, facial expressions (wrinkles), etc., in these relatively unimportant facial regions and tend to result in poor model fitting accuracy. On the other hand, if model fitting approaches only focus on the distinctive features such as eye, nose, and mouth corners, then they are susceptible to minor misalignments and tend to getting stuck in local minima, thus again resulting in poor fitting.

To address the limitations of both global and local feature sampling approaches, our proposed non-uniform feature sampling strategy offers a robust solution. The discriminative corner features are densely sampled to give higher weight to local information, while at the same time modeling semi-global appearance through sparse sampling to allow for smooth search space for model parameters.

Fig. 1 shows fitting results obtained by our proposed approach on several difficult examples including various facial expressions, occlusions, deformations, poses, and lighting variations. Additionally, fig. 6 provides further qualitative assessment by showing examples of model fitting results of ASAM compared with those obtained by STASM. As shown, our method can successfully recover the face model under various challenging, real-world, diversities.

Nevertheless, the approach still has difficulties in obtaining a good model fit under extremely adverse conditions such as sudden contrast variations, substantial occlusions, extreme poses, and their combinations. Some examples of incorrect model estimation are shown in fig.7.

Perturbation Estimation from Feature: Conventional model fitting approaches, including the baseline STASM, refine the model parameters by iteratively searching for individual feature node positions, thus requiring long processing times. In contrast, our proposed framework achieves significant speed up because instead of tracking individual features, it can update the entire shape model quickly by relying on the learned correlation between features and model parameters in a single step matrix multiplication. The model parameters can be refined with fewer iterations.

Although we have found CCA to be very effective in learning the correlations between said features and model parameters, one aspect of our future work involves evaluating other frameworks such as support vector machines, relevance vector machines, etc., for regression modeling.

6 Conclusion

We presented a novel shape model fitting framework which is robust to noise due to structural retinotopic sampling features and is efficient in estimating model parameters by directly obtaining model perturbation based on feature observations. As future work, we will study optimal sampling patterns and investigate into other regression techniques to improve performance even for more extreme conditions.

Acknowledgment. The authors thank Dr. A.Tyagi from Omron STI Corp. for his valuable comments.

References

1. Cristinacce, D., Cootes, T.: Facial feature detection using adaboost with shape constraints. In: Brit. Mach. Vis. Conf. (2003)
2. Cootes, T.F., et al.: Active shape models – their training and application. Computer Vision and Image Understanding 6, 38–59 (1995)
3. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active Appearance Models. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407, pp. 484–498. Springer, Heidelberg (1998)
4. Li, Y., Ito, W.: Shape parameter optimization for adaboosted active shape model. In: Proc. Comp. Vis. and Pattern Rec., vol. 1, pp. 251–258 (2005)
5. Milborrow, S.: Active feature models. Master's thesis, University of Cape Town (2007)
6. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proc. of the SIGGRAPH 1999, pp. 187–194 (1999)
7. Blanz, V., Vetter, T.: Face recognition based on fitting a 3d morphable model. IEEE Trans. Patt. Analy. and Mach. Intell. 25, 1063–1074 (2003)
8. Matthews, I., Baker, S.: Active appearance models revisited. Int. J. of Comp. Vis. 60, 135–164 (2004)
9. Baker, S., Matthews, I.: Equivalence and efficiency of image alignment algorithms. In: Proc. Comp. Vis. and Pattern Rec. (2001)
10. Donner, R., et al.: Fast active appearance model search using canonical correlation analysis. IEEE Trans. Patt. Analy. and Mach. Intell. 28 (2006)
11. Langs, G.: et al.: Active feature models. In: Proc. Int. Conf. Pat. Rec., vol. 1, pp. 417–420 (2006)
12. Smeraldi, F., Bigun, J.: Retinal vision applied to facial features detection and face authentication. Patt. Recogn. Lett. 23, 463–475 (2002)
13. Wiskott, L., et al.: Face recognition by elastic bunch graph matching. IEEE Trans. Patt. Analy. and Mach. Intell. 19, 775–779 (1997)
14. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proc. Comp. Vis. and Pattern Rec. (2001)
15. Kinoshita, K., Konishi, Y., Lao, S., Kawade, M., Murase, H.: Perturbation feature and it's apprication to shape model fitting for facial images. IEICE Trans. D J94-D(4), 721–729 (2011) (in Japanese)
16. Jesorsky, O., et al.: Robust Face Detection using the Hausdorff Distance. Springer (2001), Audio and Video based Person Authentification - AVBPA
17. Lee, K.C., Ho, J., Kriegman, D.: Acquiring linear subspaces for face recognition under variable lighting. IEEE Trans. Patt. Analy. and Mach. Intell. (2005)
18. Yamashita, T.: et al.: A fast omni-directional face detection system. In: Proc. Int. Conf. Comp. Vis. (2005)