https://doi.org/10.1016/j.patcog.2016.03.002

Shape-Appearance-Correlated Active Appearance Model

Huiling Zhou [a,b], Kin-Man Lam [a,b] and Xiangjian He [b]
[a] Centre for Signal Processing, Department of Electronic and Information Engineering,
The Hong Kong Polytechnic University, Kowloon, Hong Kong
[b] School of Computing and Communications,
University of Technology, Sydney, Australia
E-mail: hl.zhou@connect.polyu.hk, enkmlam@polyu.edu.hk,
Xiangjian.He@uts.edu.au

Abstract

Among the challenges faced by current active shape or appearance models, facial-feature localization in the wild, with occlusion in a novel face image, i.e. in a generic environment, is regarded as one of the most difficult computer-vision tasks. In this paper, we propose an Active Appearance Model (AAM) to tackle the problem of generic environment. Firstly, a fast face-model initialization scheme is proposed, based on the idea that the local appearance of feature points can be accurately approximated with locality constraints. Nearest neighbors, which have similar poses and textures to a test face, are retrieved from a training set for constructing the initial face model. To further improve the fitting of the initial model to the test face, an orthogonal CCA (oCCA) is employed to increase the correlation between shape features and appearance features represented by Principal Component Analysis (PCA). With these two contributions, we propose a novel AAM, namely the shape-appearance-correlated AAM (SAC-AAM), and the optimization is solved by using the recently proposed fast simultaneous inverse compositional (Fast-SIC) algorithm. Experiment results demonstrate a 5-10% improvement on controlled and semi-controlled datasets, and with around 10% improvement on wild face datasets in terms of fitting accuracy compared to other state-of-the-art AAM models.

Keywords– Facial-feature localization, Generic Active Appearance Model, Canonical correlation analysis, Orthogonal CCA

1. Introduction

Facial-feature detection and localization is a crucial process for various applications such as facial-expression recognition, face animation, 3D face reconstruction, etc. Among all competitive techniques, model-based algorithms have been proven to be most effective in automatic facial-information learning. The earliest work of such algorithms includes the deformable template [1] and the active contour model [2]. These approaches aim to extract facial features and locate face boundaries by studying feature points individually, and hence have limited robustness and accuracy. Most recently, more efficient methods, including the Active Shape Model (ASM) [3] and the Active Appearance Model (AAM) [4], have been proposed. ASM considers the facial-shape information (based on manually annotated facial-feature points) from a holistic perspective, while AAM also includes texture information (usually in terms of the pixel intensities within a face region). Due to these models' efficiency and accuracy, many variant ASM and AAM methods have been proposed in the past few decades, and they improve the localization performance. However, both ASM and AAM have problems in three different aspects, namely, insufficient robustness to variations, sensitivity to face-model initialization, and poor performance in generic situations. In the following, the challenges in these three aspects and those existing methods, which address these challenges, are discussed.

Insufficient robustness to variations. Since both ASM and AAM rely on global parametric models, they can work well for faces available in a training set with small variations in illumination, pose and expression. However, when these variations become greater, their performances usually degrade dramatically. One way to solve this problem is to integrate ASM and AAM [5] [6]. In [5], a texture-constrained shape model was used to prevent the local-minima problem, and it can achieve a robust performance under illumination variations. In [6], the profile-search step in ASM is changed into a gradient-based optimization problem to more accurately localize feature points. Recently, improved ASM models using 2-D profiles were proposed to achieve pose-adaptive localization [7] [8] [9]. It has been proven that the 2-D profiles can capture more information around each landmark than the original 1-D profiles. By properly setting the initial face model and using an optimization method, these methods can achieve accurate results, and thus have become popular model-based localization methods.

Sensitivity to initialization. In the process of refining the feature-point locations, both ASM and AAM usually perform gradient-descent optimization over a whole face, so their performances are sensitive to the initial face model. This issue has drawn much attention, and can be improved in two major steps, namely, constructing a more representative initial face model and using a robust feature-point refining scheme. For the first step, several frameworks [10] [11] [12] reformulate the original AAM as a sparse representation problem [13] and approximate the local appearance of feature points with locality constraints. After the shape and appearance priors are learned, the *K* nearest neighbors with similar patterns to the test face in terms of pose, expression, etc. are searched from a training set, and are used to model the face in a locally linear sub-space. It has been shown that this pre-processing step helps to reach faster convergence and to obtain better fitting results. Similarly, [9] [14] pre-define the number of face clusters and classify the test face into one of the clusters based on a statistical analysis. For the second step, in order to refine the face model, a stacking strategy is usually employed to search, in series, for a better location for each feature point in the face model iteratively [7] [15] [16].

Poor performance in generic situations. In the survey work of [17], statistical evaluation has shown that person-specific active models (i.e. images of a query also exist in the training set) are both easier to build and more robust to fitting than generic ones (i.e. no images of a query in the training set). To solve the generalization problem, frameworks [18] [19] based on AAM were proposed to learn a discriminative fitting function and establish a mapping between the facial appearance and the face shape in order to improve the alignment accuracy. Unlike AAMs − which model a whole facial region − the family of Constrained Local Models (CLMs) [20] [21] [22] extracts templates around each landmark and matches them to new instances of an object using a shape-constrained search and iterative template generation. This process always relies on the response surfaces generated by fitting the current feature templates using normalized correlation at each point. Recently, an approach which can handle unseen faces and variations was proposed, and is known as the Active Orientation Model (AOM) [23]. It establishes a generative deformable appearance model based on the principal components of images' gradient orientations, and it uses the project-out inverse compositional algorithm to optimize the results. An improved AAM model [24] using more efficient optimization algorithms was also proposed for generic situations.

As discussed in some survey papers [25] [26] [27], AAMs take advantage of all grey-level information across faces to build a convincing model with a relatively small number of landmarks, while ASM is just a special case of AAM. Therefore, in this paper, we focus on establishing a shape-appearance-correlated AAM (SAC-AAM) framework to tackle the above-mentioned three challenges at the same time, especially under a generic localization environment.

The contributions of this paper are given as follows. In order to fulfill the goals, we first propose a fast initialization scheme, which retrieves the most similar faces to a test face in terms of both poses and textures. Based on the idea of locality constraint, these nearest neighbors form a locally linear subspace. Then, the shape and appearance of the selected images are analyzed, and their correlation is maximized by applying Canonical Correlation Analysis (CCA) [28] (actually, the orthogonal CCA (oCCA) [29] is employed in our framework due to its superior data reconstruction property). We will show that our approach can increase the correlation between the principal components learned for face appearances and shapes, as well as the respective projection coefficients. This can improve the convergence speed and the fitting accuracy, while almost no additional computational cost will be added. By conducting experiments on different face datasets and comparing our proposed framework with state-of-the-art model-based methods, experimental results show that our framework can achieve a great improvement in terms of fitting accuracy, especially for faces under large pose, expression, and occlusion variations, as well as for unseen faces.

The remainder of the paper is organized as follows. In Section 2, we briefly introduce the well-known AAM model and some of its latest improved models. The Canonical Correlation model and its orthogonal variant are also discussed there. In Section 3, our shape-appearance-correlated AAM (SAC-AAM) framework is presented, and the details of generating initial face models and obtaining more correlated principal components are described. Experimental results and analysis are given in Section 4. The conclusion is outlined in Section 5.

2. Related work

In this section, we will give a brief overview of the Active Appearance Model (AAM) and its latest variants. We will also introduce the concept of Canonical Correlation Analysis (CCA) and its extension to orthogonal CCA, together with its efficiency for various applications.

2.1 Active Appearance Model

As mentioned in the previous section, unlike ASM − which only deals with shape information – AAM also takes texture information into consideration. The shape vector is usually presented by concatenating the position coordinates of labeled landmarks, while texture is modeled in terms of the demeaned pixel intensities or colors within the convex hull of a facial shape. When given a training set of face images with corresponding labeled landmarks, the shape model is established from 2*N* fiducial points denoted as $\mathbf{s} = (x_1, y_1, x_2, y_2, ..., x_N, y_N)^T$. The shapes are normalized by using the Procrustes analysis [30], which is a commonly used method to align shapes to a common coordinate system (usually, the mean shape of the training objects). Then, the principal component analysis (PCA) is applied to project the normalized and aligned shapes onto the shape subspace. Thus, the shape instance ***s*** can be presented as a linear combination of principal shapes as follows:

$\hat{\mathbf{s}} = \bar{\boldsymbol{s}} + \mathbf{P}_s \cdot \boldsymbol{\alpha}$, and (1)

$$\boldsymbol{\alpha} = \mathbf{P}_s^T(\boldsymbol{s} - \bar{\boldsymbol{s}}), \tag{2}$$

where $\bar{\boldsymbol{s}}$ is the mean shape, $\mathbf{P}_s$ is the matrix whose columns form a set of orthonormal base vectors, and the weight vector $\boldsymbol{\alpha}$ (also known as projection parameters) is used to control the shape variations.

The appearance model of a face image *I* is learned by first warping it into a "shape-free" model, usually the mean shape $\bar{\boldsymbol{s}}$. This is represented as a warping function $W(\boldsymbol{x};\boldsymbol{\alpha})$, where $\boldsymbol{x}$ denotes a set of pixels inside the mean shape $\bar{\boldsymbol{s}}$. Then, PCA is again applied to project the "shape-free" appearance of the image $I(W(\boldsymbol{x};\boldsymbol{\alpha}))$ on to the appearance subspace. The appearance instance $\boldsymbol{r}$ can be represented as a linear combination of principal appearances as follows:

$$\hat{\boldsymbol{r}} = \bar{\boldsymbol{r}} + \mathbf{P}_r \cdot \boldsymbol{\beta}, \text{ and} \tag{3}$$

$$\boldsymbol{\beta} = \mathbf{P}_r^T(\boldsymbol{r} - \bar{\boldsymbol{r}}), \tag{4}$$

where $\bar{\boldsymbol{r}}$ is the mean appearance, $\mathbf{P}_r$ is the matrix whose columns form a set of orthonormal base vectors, and the weight vector $\boldsymbol{\beta}$ is used to control the appearance variations. It should be noted that, in this paper, we focus on the AAM, which models the shape and appearance information independently, rather than combining shape and appearance with a single set of linear parameters as in [31].

With an appropriate initialized face, the fitting process for AAM aims to find the optimal shape and appearance parameters, which minimize the discrepancy between the synthesized image and the observed facial image. Various cost functions and optimization algorithms have been proposed to estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, among which the l_2-norm error minimization and the Inverse Compositional (IC-AAM) algorithm [31] are widely used, represented as follows:

$$\{\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0\} = \arg\min_{\{\boldsymbol{\alpha},\boldsymbol{\beta}\}} \left\| \boldsymbol{I}(W(\boldsymbol{x};\boldsymbol{\alpha})) - \bar{\boldsymbol{r}} - \mathbf{P}_r \cdot \boldsymbol{\beta} \right\|^2. \tag{5}$$

As discussed in a current work named Locality-constrained AAM (LC-AAM) [10], conventional AAMs assume a linear relationship across a whole data set, which is not always held, especially under a large variation of pose and expression. One efficient way to solve this problem is to explore the local linear subspace by modeling AAM as a sparsity-regularized problem. In [10], the original sparsity problem is approximated by adding locality constraints, as follows:

$$\{\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0\} \approx \arg\min_{\{\boldsymbol{\alpha},\boldsymbol{\beta}\}} \left\{ \sum_{\boldsymbol{x}\in\bar{\boldsymbol{s}}} \left[\boldsymbol{I}(W(\boldsymbol{x};\boldsymbol{\alpha})) - \sum_{i=1}^{K} \beta_i \cdot \mathbf{P}_{ri} \right]^2 + \lambda_1 \|\mathrm{d} \square\ \boldsymbol{\alpha}\|^2 + \lambda_2 \|\mathrm{d} \square\ \boldsymbol{\beta}\|^2 \right\}, \tag{6}$$

where the synthesized appearance image is represented as a linear combination of all the training faces, λ_1 and λ_2 are the regularization coefficients, and $\|\mathrm{d} \square\ \bullet\|^2$ denotes the distances between the input image and the respective appearance bases. In practice, Eq. (6) can be computed efficiently by directly selecting the *K* nearest neighbors of the input face image to form the shape and appearance bases, as shown in Eq. (7). With a smaller but similar training dataset, LC-AAM transforms the original non-linear problem into a locally linear one, and utilizes the popular project-out inverse compositional algorithm [31] to solve the optimization problem.

$$\{\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0\} \approx \arg\min_{\{\boldsymbol{\alpha}_K,\boldsymbol{\beta}_K\}} \{ \boldsymbol{I}(W(\boldsymbol{x};\boldsymbol{\alpha}_K)) - \bar{\boldsymbol{r}} - \mathbf{P}_r \cdot \boldsymbol{\beta}_K \}. \tag{7}$$

This approach can achieve a good performance on face images with pose and expression variations when images of the same subject are included in a training dataset (i.e. in a person-specific environment). However, if no images of the query face exist in the training set (i.e. in a generic environment) and the query face is partially occluded by facial hair, the performance of fitting the initial model to the query face deteriorates dramatically, as shown in Fig. 1(a). In our proposed framework, the sample faces, to be used to form the initial face, are selected by a weighted *K*-nearest neighbor (*K*-NN) searching scheme, which considers both pose and texture information. Compared to LC-AAM, the faces selected using our approach not only have similar poses, but also have similar facial textures (in particular, in the lower part of a face, e.g. the mouth and chin areas) to the ones in a query face. Giving a testing face, the corresponding top five similar faces selected by the method in LC-AAM and by using our proposed initialization scheme are shown in Fig. 1(b). Those faces selected by LC-AAM have similar poses to the query, but the appearance around the mouth area is different. Using our proposed scheme, the selected faces have greater similarity around the mouth areas. Hence, they can provide more useful information for learning the correlation between the face shape and the complex texture around mouth regions. Furthermore, some of the selected faces still have similar poses to the query. Consequently, our proposed scheme can improve the learning and the correlation of the principal components of the shape and appearance information. This can improve the fitting results by avoiding being trapped in local minima, as shown in Fig. 1(c).

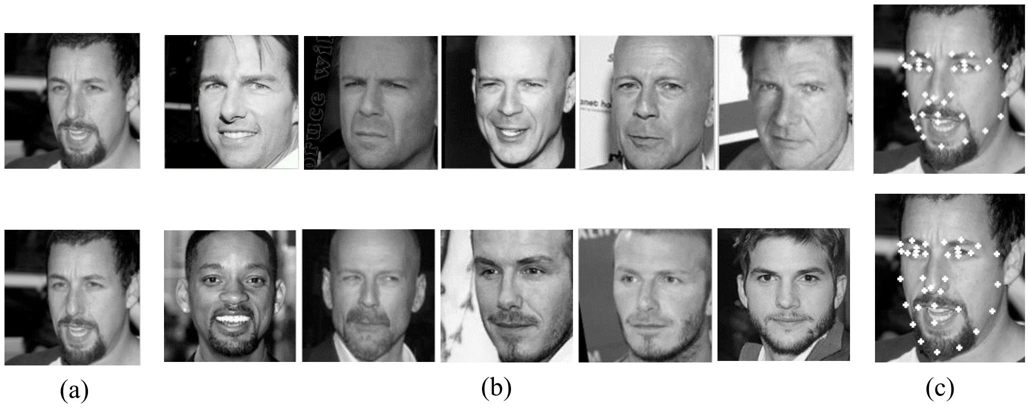


Fig. 1. Comparison of the faces selected by LC-AAM and our proposed scheme: (a) an input face is cropped and normalized based on the position of the two eyes; (b) the faces in the upper row are selected by LC-AAM, which only exhibit similar poses to the input face, while – as shown in the lower row – the faces selected using our scheme have similar poses and texture appearance to the input face; and (c) the final fitting results based on LC-AAM (upper row) and on our proposed scheme (bottom row).

2.2 Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) is a learning method which seeks basis vectors for the sets of two variables, $\boldsymbol{x}$ and $\boldsymbol{y}$, such that their projections onto the basis vectors have a maximized correlation [28]. $\mathbf{X}$ and $\mathbf{Y}$ denote the matrices whose columns are the sets of variables $\boldsymbol{x}$ and $\boldsymbol{y}$ with zero mean, respectively. Suppose that $\mathbf{M}$ and $\mathbf{N}$ are the respective direction matrices of $\mathbf{X}$ and $\mathbf{Y}$, and the corresponding canonical variation matrices of the projection coefficients are denoted by **U** and **V**, i.e., $\mathbf{U}=\mathbf{M}^T\cdot\mathbf{X}$ and $\mathbf{V}=\mathbf{N}^T\cdot\mathbf{Y}$. Then, CCA maximizes the following correlation:

$$\rho = \frac{E[\mathbf{UV}]}{\sqrt{E[\mathbf{U}^2]\mathrm{E}[\mathbf{V}^2]}} = \frac{\mathbf{M}^T\mathbf{C}_{XY}\mathbf{N}}{\sqrt{\mathbf{M}^T\mathbf{C}_{XX}\mathbf{M}\cdot\mathbf{N}^T\mathbf{C}_{YY}\mathbf{N}}}, \tag{8}$$

where *T* represents the transpose operation; $\mathbf{C}_{XX}$ and $\mathbf{C}_{YY}$ denote the within-set covariance matrices of **X** and **Y**, respectively; and $\mathbf{C}_{XY}$ denotes the covariance matrix of **X** and **Y**. It can be shown that the optimal direction matrices **M** and **N** are the eigenvectors of $\mathbf{R}_1=\mathbf{C}_{XX}^{-1}\mathbf{C}_{XY}\mathbf{C}_{YY}^{-1}\mathbf{C}_{YX}$ and $\mathbf{R}_2=\mathbf{C}_{YY}^{-1}\mathbf{C}_{YX}\mathbf{C}_{XX}^{-1}\mathbf{C}_{XY}$, respectively. It should be noted that two issues should be considered for analyzing data efficiently using CCA. The first is that two sets of data should have an intrinsic correlation, which is further increased. The second is that the number of images (i.e. the number of columns in the data matrices) should be larger than the size of an image (i.e. the number of rows in the matrices) in order to obtain feasible results such that the data can be reconstructed.

CCA and its variants have been applied to many different face-analysis applications, such as face and facial-expression recognition [32] [33], 3D and infrared face reconstruction [34], face super-resolution [35] [29], etc. One of the previous AAM works [36] applied CCA to efficiently model the dependency between texture residuals and model parameters in the searching step, which improves the convergence speed.

In [29], the original CCA is extended to orthogonal CCA (oCCA). The orthogonality property is crucial for information reconstruction, and can make the PCA projections more consistent. Therefore, in our proposed AAM framework, we also apply oCCA, which imposes extra constraints on the original CCA and obtains the orthogonal direction matrices **M** and **N** in an iterative way as follows:

$$\underset{\boldsymbol{m}_k,\boldsymbol{n}_k}{\operatorname{argmax}}\, \boldsymbol{m}_k^T\mathbf{C}_{XY}\boldsymbol{n}_k$$

$$\text{subject to } \begin{cases} \boldsymbol{m}_1^T \boldsymbol{m}_k = \boldsymbol{m}_2^T \boldsymbol{m}_k = \cdots = \boldsymbol{m}_{k-1}^T \boldsymbol{m}_k = 0, \\ \boldsymbol{n}_1^T \boldsymbol{n}_k = \boldsymbol{n}_2^T \boldsymbol{n}_k = \cdots = \boldsymbol{n}_{k-1}^T \boldsymbol{n}_k = 0, \\ \boldsymbol{m}_k^T \mathbf{C}_{XX} \boldsymbol{m}_k = 1, \\ \boldsymbol{n}_k^T \mathbf{C}_{YY} \boldsymbol{n}_k = 1, \end{cases} \tag{9}$$

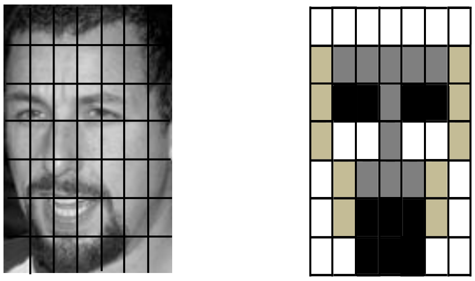
where $\boldsymbol{m}_k$ and $\boldsymbol{n}_k$ are the *k*th column vectors (direction vectors) of the direction matrices **M** and **N**, respectively. The first two constraints are to impose orthogonality on the direction vectors, while the last two are additional constraints on the norms of $\boldsymbol{m}_k$ and $\boldsymbol{n}_k$. The details of deriving the direction vectors can be found in [29]. Having computed the orthogonal-direction matrices, we can further normalize them to become orthonormal.

3. Shape-Appearance-Correlated Active Appearance Model

In this section, we will present our proposed Shape-Appearance-Correlated Active Appearance Model (SAC-AAM) in detail. Our method follows the concept in the previous work shown in [10], which reformulates the conventional AAM as a sparsity-regularized AAM problem. However, in this paper, we propose a more efficient initialization scheme to approximate sparsity regularization by retrieving the *K* nearest neighbors in terms of both pose and texture. Then, oCCA is employed to enhance the correlation between the shape features and the appearance features represented by using PCA. We will show that this can generate more correlated principal components for the shape and the appearance features, which allows optimization to be solved efficiently by using the fast simultaneous inverse compositional algorithm.

3.1 Efficient face-model initialization scheme

In the literature of face detection, recognition and facial-expression analysis, various types of facial features are employed. In our proposed framework, we use two efficient and effective features, namely the Histogram of Oriented Gradients (HOG) [37] and Local Binary Patterns (LBP) [38], for searching example face images with a similar pose and texture appearance to the query face, respectively.



(a) (b)

Fig. 2. (a) A cropped face partitioned into 7×7 windows for extracting the LBP features. (b) The weights used in the Chi square distance measure, where black, dark gray, light gray, and white represent the weights of 4, 3, 2, and 1, respectively.

Each face image is cropped to the size of 160×160 and normalized based on the positions of two eyes' centers [39]. Then, a *K*-NN search based on the HOG features and Euclidean distance is used to select the most similar faces from a dataset. As shown in Fig. 1.(b), the *K* nearest neighbors have similar poses to the test face, due to the fact that HOG captures the edges' orientations and hence an object's shape. However, retrieving faces with similar poses only is insufficient, because some parts of a test face image may be occluded by facial hair or hair shading, in particular when the test face has no images in the training dataset. In order to achieve a more efficient and accurate subspace learning based on the selected samples, the weighted LBP features are also considered in the search, which aims to select faces having a similar texture to the test face. In the search, faces are cropped and normalized in the same way as the training faces, and are also divided into 7×7 windows, which can achieve the best performance by experiment. Because each of the windows has a different degree of importance, different weights are set for them, as shown in Fig. 2.

With the LBP feature histogram for each block of a face image, the weighted Chi square distance is used to measure the similarity between the test face and all the faces in the training dataset as follows:

$$\chi_w^2(f,g)=\sum_{j,i} w_j \frac{(f_{i,j}-g_{i,j})^2}{f_{i,j}+g_{i,j}}, \tag{10}$$

where *f* and *g* are the normalized histograms of the test and training face images, respectively; *i* and *j* are the indices representing the *i*th bin in the histogram of the *j*th block; and w_j is the weight predefined for block *j*.

Fig. 1(b) shows the top five faces selected from the training dataset using the LBP feature. We can see that the selected face images have a similar appearance around the mouth regions. However, these selected faces may have poses that are different from the test face. Having retrieved the similar-pose faces and similar-texture faces using the HOG and LBP features, respectively, the mean shape of the similar-pose faces is computed. In order to use the similar-texture faces more efficiently in learning, they are wrapped to the mean shape by using Procrustes warp. In this way, the initial face model is more similar to the test face in terms of shape and appearance, and thence helps to establish a more locally linear subspace for representation. It should also be noted that, for normal faces without large occlusion or pose variation, our initialization scheme does not affect the efficiency and can achieve a slightly better performance compared to using either one of the two features to search the dataset, as shown in Fig. 10. The overall initialization scheme is illustrated in Fig. 3.

In experiments, we have found that using only the top five pose faces and the top twenty texture faces is sufficient to achieve a good performance. More experimental results will be shown in subsequent sections.

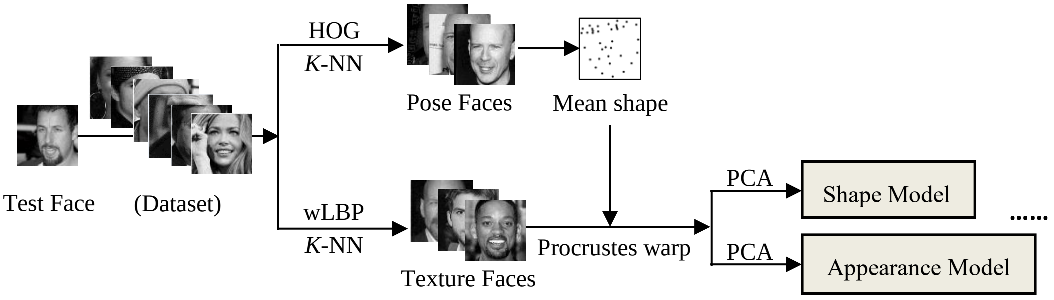


Fig. 3. Illustration of the proposed fast initialization scheme (FIS).

3.2 Orthogonal CCA for SAC-AAM

With the *K* texture faces selected from training samples (after wrapping to the mean shape of the retrieved pose faces), PCA is applied to both the shape matrix $\mathbf{S}=[\boldsymbol{s}_1,\boldsymbol{s}_{2,}...,\boldsymbol{s}_K]$ and the appearance matrix $\mathbf{R}=[\boldsymbol{r}_1,\boldsymbol{r}_{2,}...,\boldsymbol{r}_K]$, where the columns of the matrices are the landmark coordinates and gray-level intensities, respectively, within the shape hull of the respective training faces. We compute the mean shape vector $\bar{\boldsymbol{s}}$ and the mean appearance vector $\bar{\boldsymbol{r}}$. Then, matrices $\mathbf{P}_s$ and $\mathbf{P}_r$ are composed of the orthonormal eigenvectors of the shape and appearance training vectors, respectively. The corresponding projection coefficients of the shape and appearance vectors are denoted by $\mathbf{A}=[\boldsymbol{a}_1,\boldsymbol{a}_{2,}...,\boldsymbol{a}_K]\in\mathbb{R}^{m\times K}$ and $\mathbf{B}=[\boldsymbol{b}_1,\boldsymbol{b}_{2,}...,\boldsymbol{b}_K]\in\mathbb{R}^{n\times K}$. 95% of the total energy of both the shape and appearance information is retained. The number of eigenvectors used for shape and appearance are denoted as *m* and *n*, respectively, of which both are smaller than *K* (*m* and *n* are usually less than 10, while *K* is set at 20 in our algorithms). Similar to Eq. (2) and Eq. (4), the projection coefficients can be computed by:

$$\boldsymbol{a}_i=\mathbf{P}_s^T(\boldsymbol{s}_i-\bar{\boldsymbol{s}}), \text{ and}$$

$$\boldsymbol{b}_i=\mathbf{P}_r^T(\boldsymbol{r}_i-\bar{\boldsymbol{r}}). \tag{11}$$

Since the shape and appearance information of a person possesses an intrinsic correlation, it can be explored and enhanced by applying oCCA to the demeaned coefficients matrices $\hat{\mathbf{A}}=[\hat{\boldsymbol{a}}_1,\hat{\boldsymbol{a}}_{2,}...,\hat{\boldsymbol{a}}_K]$ and $\hat{\mathbf{B}}=[\hat{\boldsymbol{b}}_1,\hat{\boldsymbol{b}}_{2,}...,\hat{\boldsymbol{b}}_K]$. In fact, it is proven that this is equivalent to applying oCCA to matrices **A** and **B** which are already demeaned. By solving the optimization problem in Eq. (9), we can obtain two projection matrices with orthonormal column vectors $\mathbf{W}_s$ and $\mathbf{W}_r$, and two canonical variate matrices $\mathbf{C}_s=\mathbf{W}_s^T\mathbf{A}$ and $\mathbf{C}_r=\mathbf{W}_r^T\mathbf{B}$, where the correlation coefficient $\rho=\frac{E[\mathbf{C}_s\mathbf{C}_r]}{\sqrt{E(\mathbf{C}_s^2)E(\mathbf{C}_r^2)}}$ is maximized. Then, we rewrite $\mathbf{C}_s$ and $\mathbf{C}_r$ as follows:

$$\mathbf{C}_s=\mathbf{W}_s^T\mathbf{A}=\mathbf{W}_s^T\mathbf{P}_s^T(\mathbf{S}-\bar{\mathbf{S}})=\tilde{\mathbf{P}}_s^T\hat{\mathbf{S}} \text{ and}$$

$$\mathbf{C}_r=\mathbf{W}_r^T\mathbf{B}=\mathbf{W}_r^T\mathbf{P}_r^T(\mathbf{R}-\bar{\mathbf{R}})=\tilde{\mathbf{P}}_r^T\hat{\mathbf{R}}, \tag{12}$$

where $\hat{\mathbf{S}}=\mathbf{S}-\bar{\mathbf{S}}$ and $\hat{\mathbf{R}}=\mathbf{R}-\bar{\mathbf{R}}$ are the demeaned shape and appearance matrices, respectively, and $\tilde{\mathbf{P}}_s=\mathbf{P}_s\mathbf{W}_s$ and $\tilde{\mathbf{P}}_r=\mathbf{P}_r\mathbf{W}_r$ are the corresponding eigen-matrices after the oCCA transformation.

As shown in Eq. (12), the multiplication of an original eigen-matrix ($\mathbf{P}_s$ or $\mathbf{P}_r$) and the corresponding oCCA projection matrix ($\mathbf{W}_s$ or $\mathbf{W}_r$) forms a new eigen-matrix ($\tilde{\mathbf{P}}_s$ or $\tilde{\mathbf{P}}_r$), which can project the shape or appearance vector on to a more correlated subspace. In addition, these two new eigen-matrices are orthonormal, which can be proven as follows:

$\tilde{\mathbf{P}}_s^T\tilde{\mathbf{P}}_s = (\mathbf{P}_s\mathbf{W}_s)^T(\mathbf{P}_s\mathbf{W}_s) = \mathbf{I}$ and

$$\tilde{\mathbf{P}}_r^T\tilde{\mathbf{P}}_r = (\mathbf{P}_r\mathbf{W}_r)^T(\mathbf{P}_r\mathbf{W}_r) = \mathbf{I}, \tag{13}$$

where **I** is an identity matrix. Therefore, the new eigen-matrices are applied in the model-fitting process of the feature points. Since the matrices of PCA projection coefficients **A** and **B** are both small, applying oCCA will only increase the computational cost slightly, but can improve the accuracy of final fitting as shown in the experimental results.

3.3 Fitting scheme for SAC-AAM

Fitting an AAM usually involves estimating the model parameters so that the distance between the model instance and the given image is minimized. Typically, this process is presented as the optimization of a least-square problem, as shown in Eq. (5). In our framework, with training faces resembling the test face in terms of shape and appearance and being used for initialization, and the use of more correlated eigen-matrices and the corresponding projection coefficients, we refine the optimization in Eq. (6) with far fewer shape and appearance eigenvectors. In the related work [10], the optimization problem is solved by using the project-out inverse compositional (POIC) algorithm. However, as illustrated in [24], the POIC algorithm is efficient but does not work well for unseen variations, so it is unsuitable for generic situations. In contrast to POIC, the simultaneous inverse compositional (SIC) algorithm [40] has been proven to perform robustly in the case of generic fitting but is extremely complex computationally. To tackle this problem, the fast simultaneous inverse compositional (Fast-SIC) algorithm is employed to achieve relatively accurate fitting results while greatly reducing the computation time. Instead of concatenating the warped shape parameters and appearance parameters, and optimizing them as a whole, Fast-SIC first optimizes the fitting with respect to the appearance parameters, and the solution is then used for optimization with respect to the warped parameters in each iteration. The cost required is slightly more than that of POIC, which is only an approximation to Fast-SIC (and hence to SIC), but achieves better fitting results.

Algorithm 1: Fast-SIC for SAC-AAM

```
Pre-compute:
        (3) Evaluate the gradients $\nabla\bar{\boldsymbol{r}}$ and $\nabla\mathbf{P}_{ri}$ for $i = 1, \ldots, K$
        (4) Evaluate the Jacobian $\frac{\partial W}{\partial \boldsymbol{\alpha}_K}$ at $(\boldsymbol{x};0)$
    Iterate:
        (1) Warp $\boldsymbol{I}$ with $W(\boldsymbol{x};\boldsymbol{\alpha}_K)$ to compute $\boldsymbol{I}(W(\boldsymbol{x};\boldsymbol{\alpha}_K))$
        (2) Compute the error image $E_{Fsic}(\boldsymbol{x}) = \bar{\boldsymbol{r}} + \tilde{\mathbf{P}}_r \cdot \boldsymbol{\beta}_K$ - $\boldsymbol{I}(W(\boldsymbol{x};\boldsymbol{\alpha}_K))$
        (5) Compute the steepest descent image $\mathbf{J} = \nabla\bar{\boldsymbol{r}}\frac{\partial W}{\partial \boldsymbol{\alpha}_K}$
        (6) Project out appearance from J to obtain $\mathbf{J}_{Fsic} = \boldsymbol{\alpha}_K[\mathbf{P}_{rx}\boldsymbol{\beta}'_K, \mathbf{P}_{ry}\boldsymbol{\beta}'_K]\frac{\partial W}{\partial \boldsymbol{\alpha}_K}$
        (7) Compute the Hessian Matrix $\mathbf{H}_{Fsic} = \mathbf{J}_{Fsic}^T\mathbf{J}_{Fsic}$ and invert it
        (8) Compute $\mathbf{J}_{Fsic}^T E_{Fsic}(\boldsymbol{x})$
        (9) Compute $\Delta\boldsymbol{\alpha}_K = \mathbf{H}_{Fsic}^{-1}\mathbf{J}_{Fsic}^T E_{Fsic}(\boldsymbol{x})$
        (10) Update $W(\boldsymbol{x};\boldsymbol{\alpha}_K) \leftarrow W(\boldsymbol{x};\boldsymbol{\alpha}_K) \circ W(\boldsymbol{x};\Delta\boldsymbol{\alpha}_K)^{-1}$ and $\boldsymbol{\beta}_K \leftarrow \boldsymbol{\beta}_K + \Delta\boldsymbol{\beta}_K$
    Until $\|\Delta\boldsymbol{\alpha}_K\| \leq \varepsilon$
```

In our algorithm, we also fit our model into the Fast-SIC optimization framework which firstly linearizes the appearance model and then projects it out. However, in contrast to Fast-SIC, which directly uses the raw pixel intensities as the features without applying any priors, our algorithm can show a further improvement on the fitting performance. Another advantage of our algorithm is that we solve the standard generic AAM dilemma (the number of appearance parameters is at least one order of magnitude greater than the number of shape parameters, i.e. $n \gg m$ for matrices **A** and **B**) using a simple fitting process, where both the numbers of shape and appearance parameters are small, and also smaller than the number of nearest neighbors. We refine the fitting model as in Eq. (14):

$$\left\{\boldsymbol{\alpha}_0,\boldsymbol{\beta}_0\right\} \approx \arg\min_{\{\boldsymbol{\alpha}_K,\boldsymbol{\beta}_K\}} \left\{\boldsymbol{I}(W(\boldsymbol{x};\boldsymbol{\alpha}_K)) - \bar{\boldsymbol{r}} - \tilde{\mathbf{P}}_r \cdot \boldsymbol{\beta}_K\right\}, \tag{14}$$

where $\boldsymbol{\alpha}_K$ and $\boldsymbol{\beta}_K$ are the PCA projection parameters in the more correlated shape and appearance eigen-spaces constructed by using the retrieved *K* nearest face neighbors. The Fast-SIC algorithm used in our proposed fitting model is summarized in Algorithm 1. More details of Fast-SIC can be found in [24].

4. Experiments and analysis

To evaluate the performance of our proposed generic AAM framework, we compare it with several state-of-the-art methods on different datasets, namely the IMM dataset [41] under controlled variations of pose and expression, the Bosphorus dataset [42] with cropped face images under semi-controlled variations of pose and expression, and the labeled faces in the wild datasets LFPW [22] and PubFig [43] with uncontrolled variations of pose and expression, as well as with occlusion. Some faces of these datasets are shown in Fig. 4. All experiments were conducted under Matlab R2010b environment on an Intel i7 3.5 GHz CPU with 16GB RAM PC.

Numerous measurement metrics are proposed for different face analysis methods, such as ROC curves for face recognition and F1-score, precision and recall for face classification and retrieval etc. To measure the performances of facial feature localization, we use the ground-truth-based localization error, as in [9] [26], which is the point-to-point error (PtP Error), normalized by the eye distance. Given the ground-truth landmarks, the localization error e_i^k is computed as follows:

$$e_i^k = \frac{d[(x_i^k, y_i^k), (\tilde{x}_i^k, \tilde{y}_i^k)]}{IOD}, \tag{15}$$

where $d(.,.)$ is the Euclidean distance between the *k*-th landmark of the ground-truth face and the corresponding detected landmark of the *i*-th test face; and *IOD* is the Inter-Ocular Distance, which is the distance between the two eye pupils. According to the measurement in [26], $e_i^k < 0.1$ can be taken as an acceptable error criterion under a controlled environment. In other words, a landmark is considered to be detected correctly if its normalized error is below the threshold. In our paper, we employ the cumulative curve corresponding to the percentage of test images for which the mean localization error of all the landmarks (also called normalized root-mean-squared error (NRMSE)) is less than a specified threshold. In the following subsections, we will elaborate on the experimental setup on each dataset, and compare our proposed method with several state-of-the-art methods both statistically and visually.

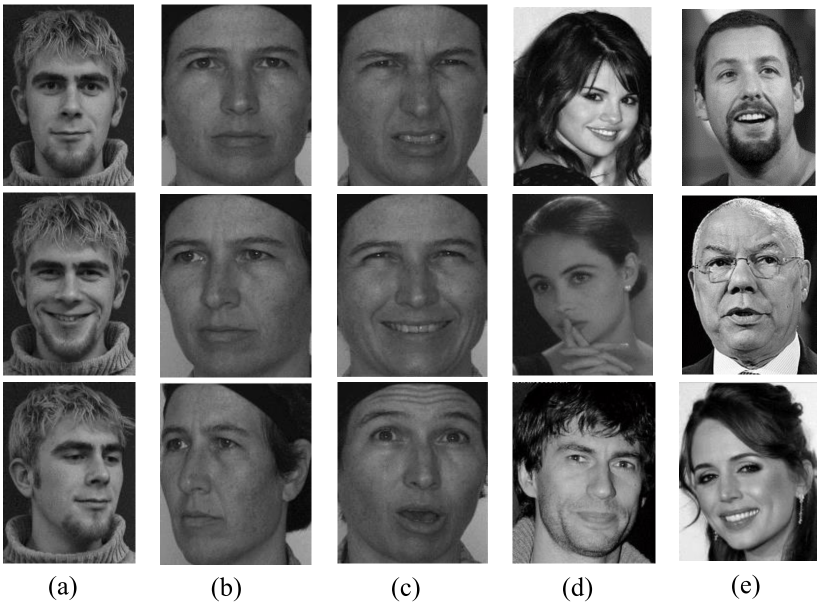


Fig. 4. Sample face images from the selected datasets: (a) IMM dataset, (b) Bosphorus dataset with pose variations, (c) Bosphorus dataset with expression variations, (d) LFPW dataset, and (e) PubFig dataset.

4.1 Performance on a controlled dataset

In this experiment, 156 gray-scale face images of 39 distinct subjects in the IMM dataset [41] are selected. Each subject is sized 640×480 and has 4 images with neutralfrontal, smiling-frontal, neutral-left, and neutral-right views, respectively. We use the re-annotated faces with 58 landmarks similar to our previous work [9]. Since it is a relatively small and simple dataset, we select one subject for testing and others for training each time. We mainly examine the efficiency of our proposed SAC-AAM framework together with each of the contributions of our proposed framework, i.e. the fast initialization scheme (denoted by SAC-AAM without oCCA) and using oCCA to increase the correlation between shape and appearance (denoted as SAC-AAM without FIS). We compare them with the recent locality-constrained AAM (LC-AAM) [10] and the adaptive-profile ASM (APASM) [9]. For the IMM dataset, we retrieve the top five pose faces and twenty texture faces using the *K*-NN search for our proposed SAC-AAM, and the top twenty nearest neighbors for LC-AAM as described in [10]. The experimental setup is the same for all the methods compared, and the experiment results are presented in Fig. 5 in terms of the cumulative curves.

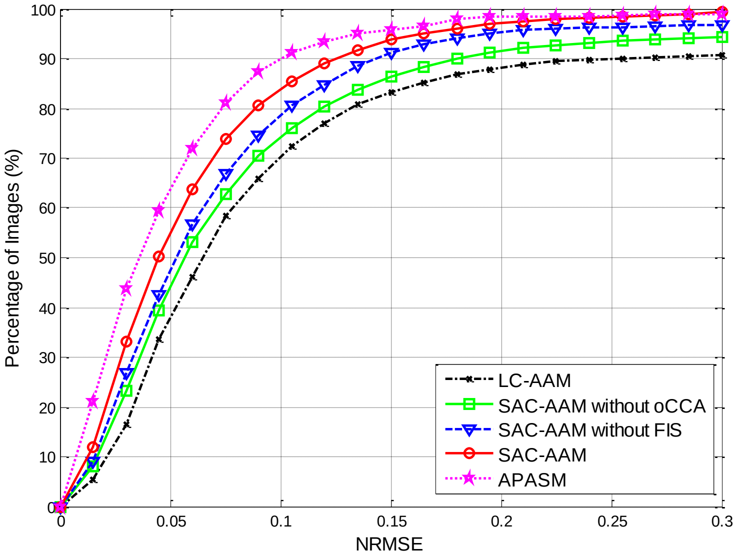


Fig. 5. Fitting results of different methods on the IMM dataset.

From the results, we can see that by using the proposed face-model initialization scheme and oCCA to improve the correlation of appearance and shape, our proposed method can achieve a more accurate fitting performance than LC-AAM. Each of the contributions can make some improvements to our proposed framework, and our SAC-AAM achieves detection accuracy of higher than 80 percent when the error criterion equals 0.1, and is at least 10 percent higher than LC-AAM. However, for the IMM dataset, our previous work, APASM, achieves the best performance because it is based on the ASM model which locally searches for the best position for each landmark and works well under controlled environments. Nevertheless, it is the slowest one among the methods compared. For our proposed SAC-AAM, it takes 10-12s to process and localize each query image in the IMM dataset. LC-AAM requires a similar runtime, but APASM needs about 20s.

4.2 Performance on a semi-controlled dataset

To further examine each of the contributions of our proposed framework, and to compare them with the methods mentioned in Section 4.1, the Bosphorus dataset [42] is employed. It contains the high-resolution images of 105 people with larger variations in pose and expression than the IMM dataset, as illustrated in Figs. 4(b) and 4(c). These face images have been cropped to include only the face, such that the landmarks on their face contours cannot be localized. We further divide the dataset into faces with pose variations and faces with expression variations. For those faces with pose variations, each subject has four poses: frontal, right10, right20, and right30, respectively, with a total of 32 landmarks. For those with expression variations, each subject has five different expressions: angry, happy, disgusted, surprised, and eye-closed, with a total of 22 landmarks. The image size is reduced to 280×340 for faster computation. The same experimental settings and parameter selection are used as in Section 4.1, and one subject is selected for testing, while the others are selected for training each time. Fig. 6 and Fig. 7 show the corresponding cumulative curves with pose variations and expression variations.

From the results, we can see that refining the initial face model, by adding local constraints, can improve the overall performance of the AAM models on the cropped face images from the semi-controlled dataset. Even if those points lying along the face contour are excluded, our proposed AAM framework with each contribution achieves better performance than LC-AAM and our previous work APASM, with about 5-10 percent higher in detection accuracy. For APASM, we can also observe that it deals with pose variations better than expression variations due to the use of the HoG feature being able to select training images with similar poses. Compared with our proposed AAM framework, which increases the correlation between shape and texture under pose and expression variations, when the variations become larger, the performance of ASM-based methods deteriorates because they determine the final location of each feature point separately. The average runtime of the proposed SAC-AAM method on the Bosphorus dataset is about 5.5s. To give a better illustration, some of the visual fitting results based on APASM, LC-AAM, and our proposed SAC-AAM on the IMM and Bosphrus datasets are shown in Fig. 8.

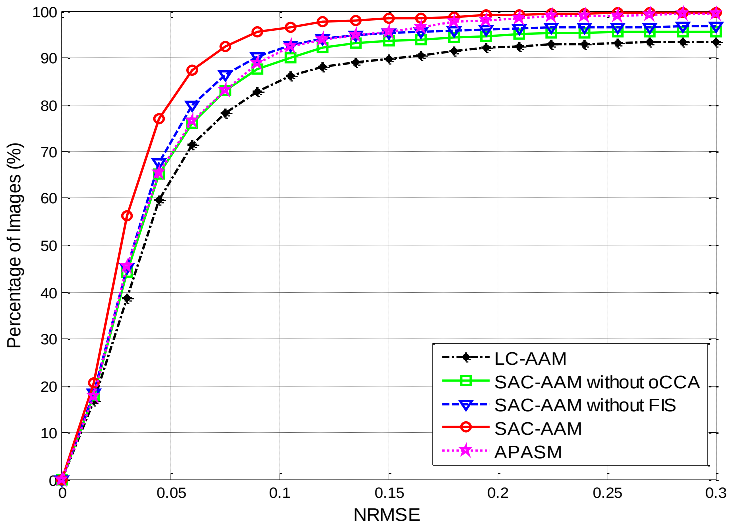


Fig. 6. Fitting results of the different methods on the Bosphorus dataset with pose variations.

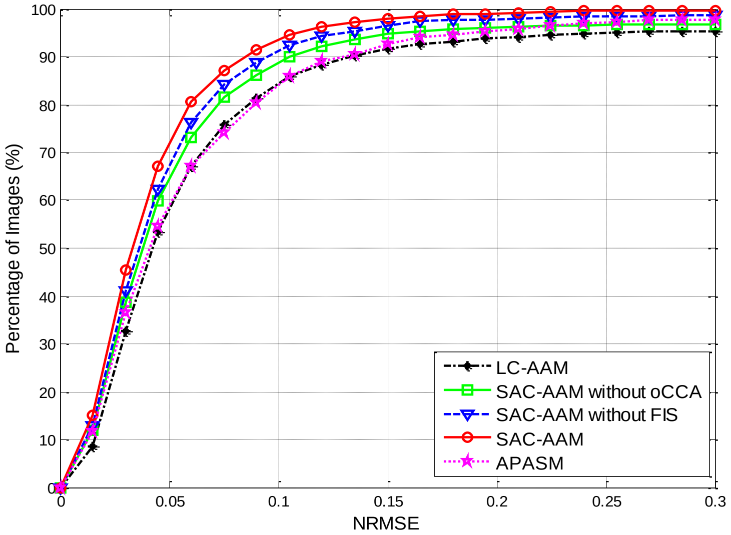
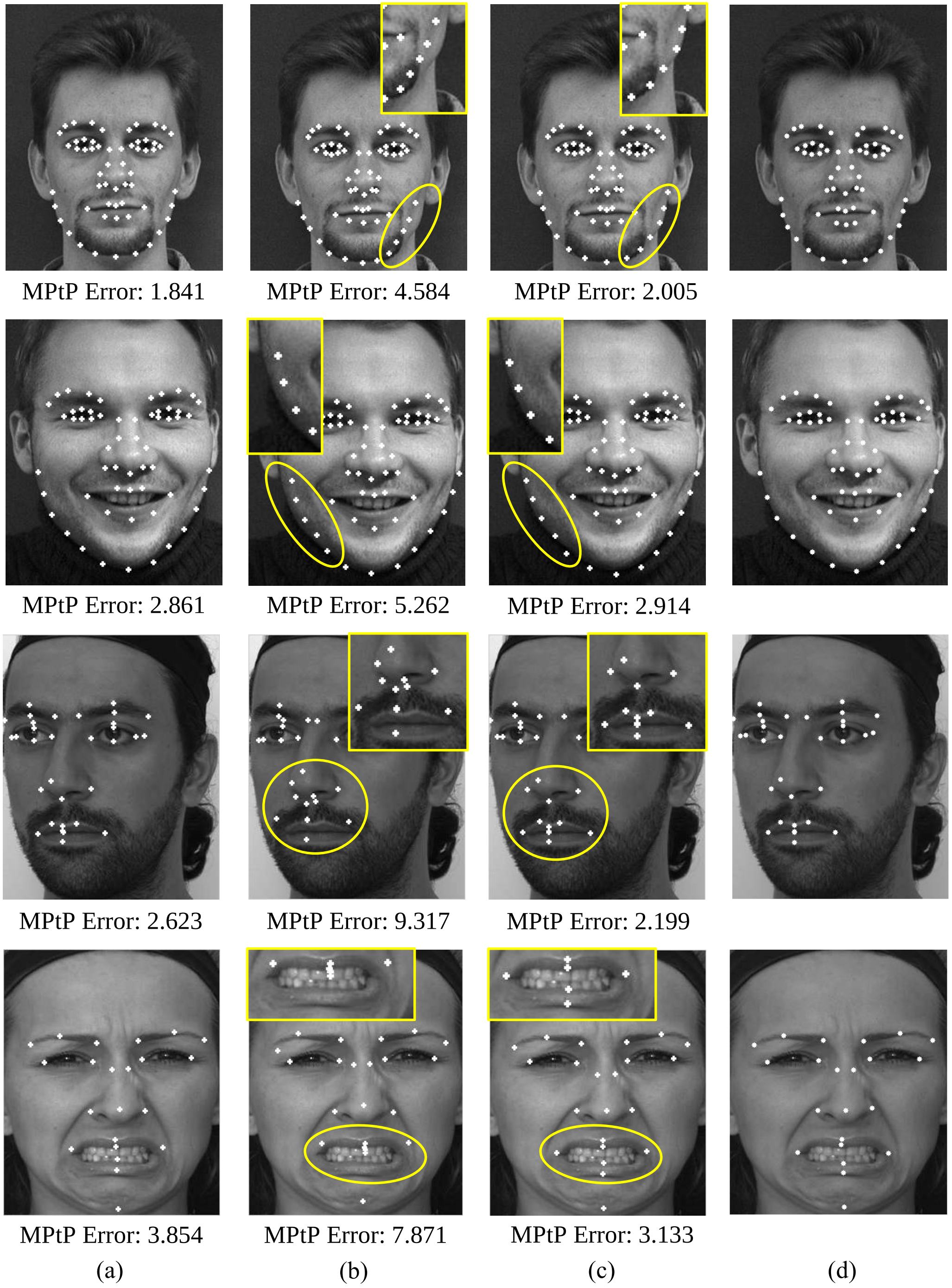


Fig. 7. Fitting results of different methods on the Bosphorus dataset with expression variations.



MPtP Error: 1.841 MPtP Error: 4.584 MPtP Error: 2.005

MPtP Error: 2.861 MPtP Error: 5.262 MPtP Error: 2.914

MPtP Error: 2.623 MPtP Error: 9.317 MPtP Error: 2.199

MPtP Error: 3.854 MPtP Error: 7.871 MPtP Error: 3.133

(a) (b) (c) (d)

Fig. 8. Visual fitting results and the corresponding mean point-to-point errors of different methods on the IMM dataset (the first two rows) and Bosphorus dataset (the last two rows): (a) APASM, (b) LC-AAM, (c) our proposed SAC-AAM framework, and (d) the corresponding face image with ground-true landmarks.

For each individual face image, we calculate the mean point-to-point error (MPtP error) between the estimated landmarks and the ground-true landmarks for all the feature points. This measurement shows the overall localization performance in a straightforward way. Although all the methods can achieve a good performance on this semi-controlled dataset, we can still observe an obvious improvement using our proposed SAC-AAM framework compared to LC-AAM, as highlighted by the yellow circles and the corresponding enlarged regions shown in the yellow rectangles.

4.3 Performance on an in-the-wild dataset

Nowadays, with the rapid improvement of facial-feature localization techniques, as well as the availability of new face datasets, the ultimate goal of recently proposed methods is to localize facial points accurately on faces in the wild, especially under unseen variations. Therefore, in this experiment our proposed AAM framework is evaluated on a famous in-the-wild dataset, namely the re-annotated LFPW dataset [44]. To have a fair evaluation, we compare our method (together with each of the contributions) with several state-of-the-art AAM methods, namely LC-AAM, Active Orientation Models (AOMs) [23], and AAM with the fast simultaneous inverse compositional algorithm (AAM-FSIC) [24].

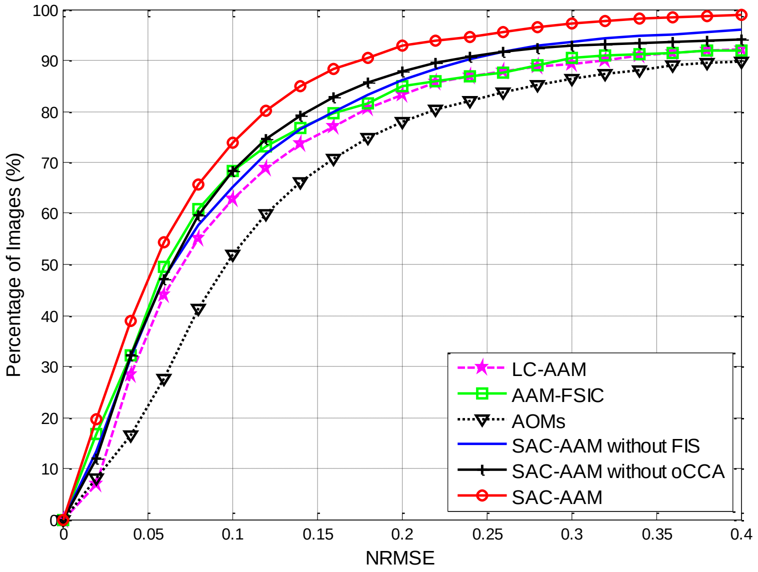


Fig. 9. Fitting results of the different methods on the LFPW dataset in the wild.

The re-annotated LFPW dataset is an improved version of the original LFPW dataset [22], where each face is labeled with 68 points. All images in the dataset were downloaded from the web with large variations in pose, expression, and lighting conditions, as shown in Fig. 4(c). The resolutions of the images also vary hugely from

100×100 to 400×400. We select 800 face images for training and 222 face images for testing, without the subjects in the training set included in the testing set. The fitting results of the different methods are shown in Fig. 9. We can observe that the performance of all methods decreases with the in-the-wild faces, while our methods, together with each contribution, achieve superior results compared to other state-of-art methods, with an average of 10 percent higher detection accuracy.

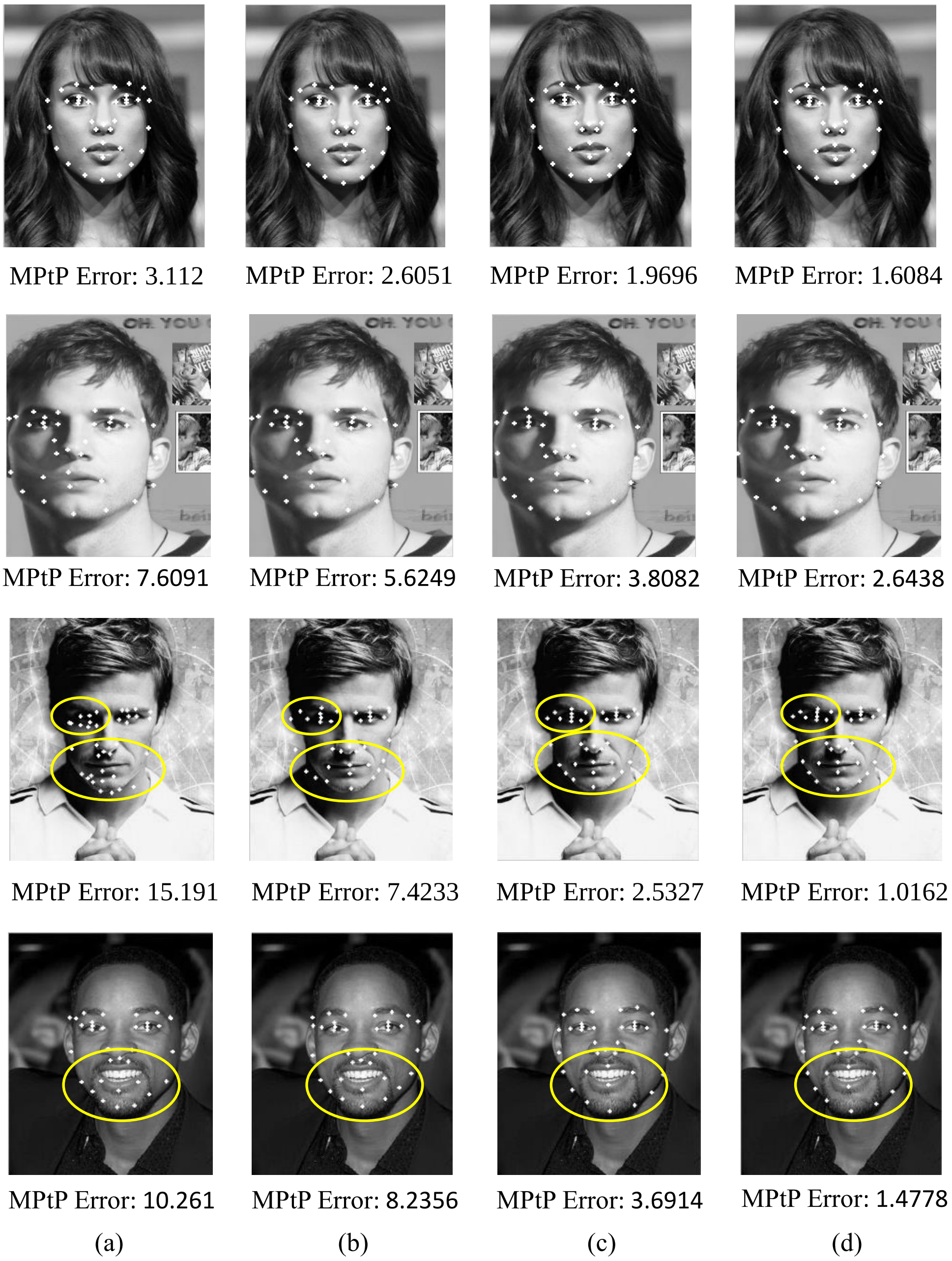
4.4 Visual performance on faces in the wild

In this section, we conducted two experiments based on face images in the wild. The fitting results are illustrated visually, based on the PubFig dataset [43]. This dataset is similar to the LFPW dataset, but each face image has 36 feature points.

In the first experiment, the PubFig dataset was used for both training and testing. Because our framework is based on LC-AAM, we visually compare the fitting results based on LC-AAM, SAC-AAM with one of the two contributions, i.e. SAC-AAM without oCCA and SAC-AAM without FIS, and SAC-AAM with both oCCA and FIS. Some selected fitting results, as well as their corresponding mean point-to-point errors (MPtP errors), are shown in Fig. 10.

We can observe that, for those generic cases under simple pose and illumination variations (shown in the first two rows), all methods can work well. However, for the faces with strong variations in illumination or with occlusion, our proposed SAC-AAM framework, as well as SAC-AAM with one of the two contributions, achieves much better performance, which are highlighted with yellow circles in Fig. 10.

For better visualization, we have also illustrated in Fig. 11 the improvements of the fitting results by enlarging the regions marked by the yellow circles in Fig. 10. With both of the proposed contributions, our SAC-AAM method can achieve much better localization performances in the occluded mouth regions, facial contours, and eye regions, where most existing AAM methods cannot achieve accurate results.



MPtP Error: 3.112 MPtP Error: 2.6051 MPtP Error: 1.9696 MPtP Error: 1.6084

MPtP Error: 7.6091 MPtP Error: 5.6249 MPtP Error: 3.8082 MPtP Error: 2.6438

MPtP Error: 15.191 MPtP Error: 7.4233 MPtP Error: 2.5327 MPtP Error: 1.0162

MPtP Error: 10.261 MPtP Error: 8.2356 MPtP Error: 3.6914 MPtP Error: 1.4778

(a) (b) (c) (d)

Fig. 10. Visual fitting results and the corresponding mean point-to-point errors of different methods on the PubFig dataset: (a) LC-AAM, (b) SAC-AAM without oCCA, (c) SAC-AAM without fast initialization scheme, and (d) our proposed SAC-AAM framework.

In the second experiment, we evaluated the generalization capability of different methods using training and testing data from two different datasets. Similar to Section 4.3, we compare our proposed SAC-AAM framework with LC-AAM, AOMs, and AAM-FSIC. For AOMs, the source code provided uses the Multi-PIE dataset as the training set. For the other methods, the LFPW dataset is the training dataset, while PubFig is the testing dataset. Some visual results are illustrated in Fig. 12, and some highlighted regions (e.g. the mouth region, facial contour, and eye region) are also enlarged and illustrated in Fig. 13 for better visualization. Our proposed SAC-AAM again generalizes better for unseen faces than other recent AAM variants.

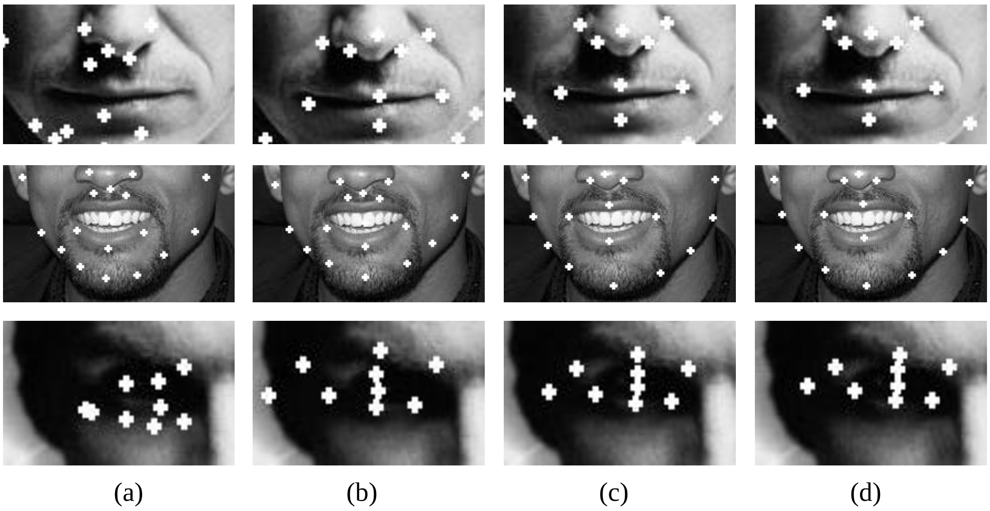


Fig. 11 Enlarged visual fitting results of selected results from Fig. 10. The first row is the mouth region, the second row is the face contour, and the third row is the eye region: (a) LC-AAM, (b) SAC-AAM without oCCA, (c) SAC-AAM without fast initialization scheme, and (d) our proposed SAC-AAM.

5. Conclusions

In this paper, we have proposed a shape-appearance-correlated Active Appearance Model (SAC-AAM) for generic facial-feature localization. Based on the idea of approximating the local appearance of feature points with locality constraints to improve face-model initialization, we have proposed an efficient initialization scheme which retrieves *K* nearest neighbors with similar poses and textures to a test face from a training set. With a small number of representative samples, the correlation between the shape and the appearance models can be learned more efficiently and this can better represent the test face images. To further improve the fitting performance of AAM, we have applied oCCA to increase the correlation between the shape features and the appearance features represented by PCA. With these two main contributions, we have devised our AAM model and solved the optimization using the recently proposed fast simultaneous inverse compositional (Fast-SIC) algorithm. With only a small number of training images selected for learning and the fast optimization algorithm used, our proposed framework is efficient and accurate. Experimental results on different datasets have shown the better performance, in terms of the statistical and visual performances, of our proposed framework, as well as each of the two contributions i.e. fast initialization scheme and oCCA to increase correlation between shape and appearance. The fitting results have also demonstrated that our method can achieve superior performance compared to other state-of-the-art AAM models, especially under generic environments.

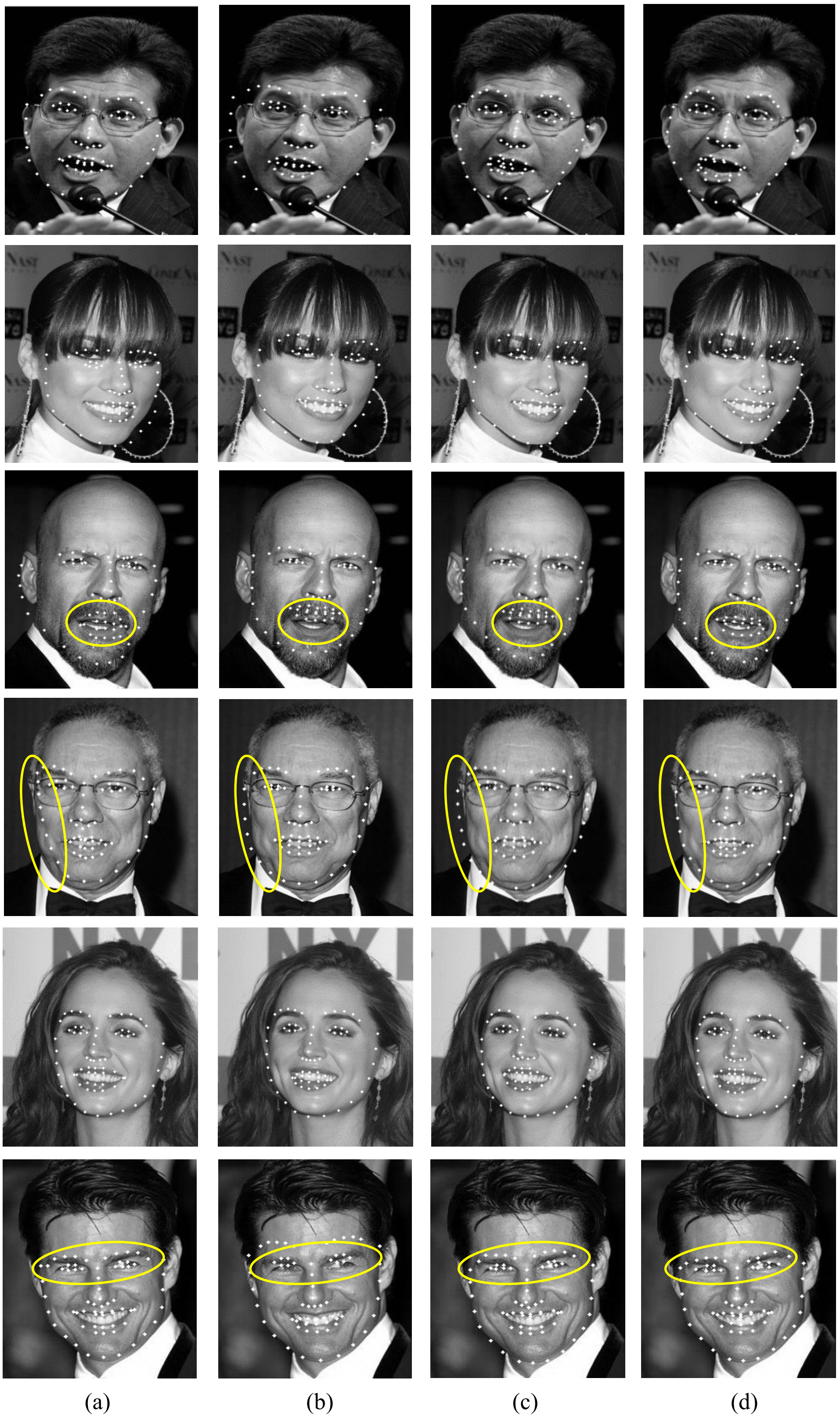


Fig. 12. Visual fitting results of different methods training on the LFPW dataset and testing on the PubFig dataset: (a) LC-AAM, (b) AOMs, (c) AAM-FSIC, and (d) our proposed SAC-AAM.

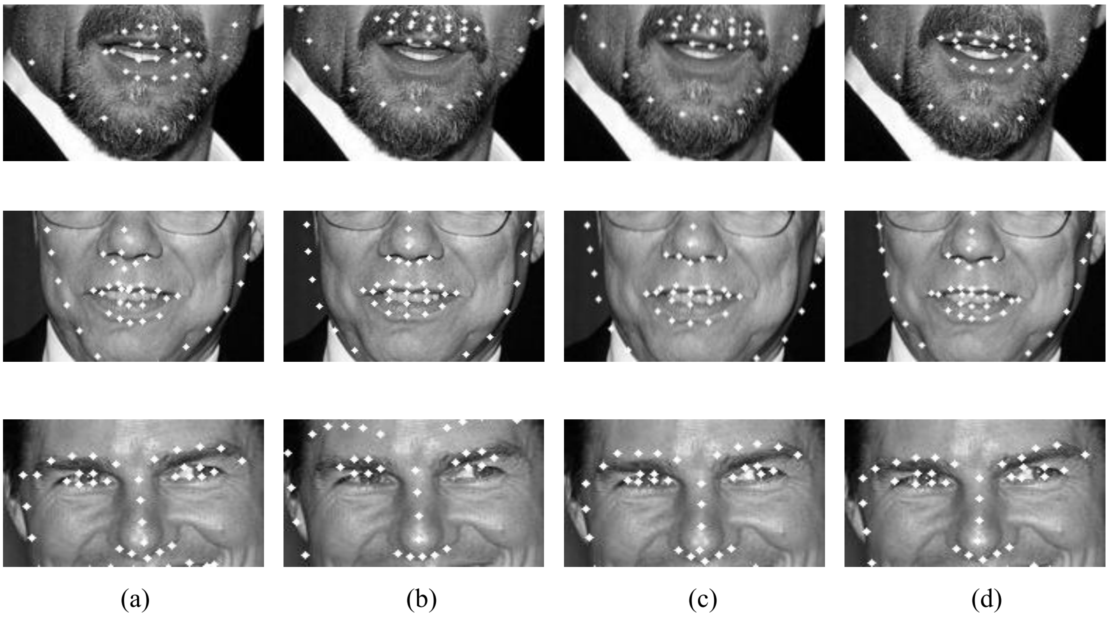


Fig. 13. Enlarged visual fitting results of selected results from Fig. 12. The first row is the mouth region with mustache, the second row is the face contour, and the third row is the eye region: (a) LC-AAM, (b) AOMs, (c) AAM-FSIC, and (d) our proposed SAC-AAM.

References

[1] Chin R. T. and Dyer C. R., "Model-based recognition in robot vision," *ACM Computing Surveys (CSUR)*, vol. 18, no. 1, pp. 67-108, 1986.

[2] Choi W. P., Lam K. M. and Siu W. C., "An adaptive active contour model for highly irregular boundaries," *Pattern Recognition*, vol. 34, no. 2, pp. 323-331, 2001.

[3] Cootes T. F., Taylor C. J., Cooper D. H. and Graham J., "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38-59, 1995.

[4] Cootes T. F., Edwards G. J. and Taylor C. J., "Active appearance models," *TPAMI*, vol. 23, no. 6, pp. 681-685, 2001.

[5] Yan S., Liu C., Li S. Z., Zhang H., Shum H. Y. and Cheng Q., "Face alignment using texture-constrained active shape models," *Image and Vision Computing*, vol. 21, no. 1, pp. 69-75, 2003.

[6] Sung J., Kanade T. and Kim D., "A unified gradient-based approach for combining ASM into AAM," *International Journal of Computer Vision*, vol. 75, no. 2, pp. 297-309, 2007.

[7] Milborrow S. and Nicolls F., "Locating facial features with an extended active shape model," in *ECCV*, 2008, pp. 504-513.

[8] Milborrow S. and Nicolls F., "Active shape models with SIFT descriptors and MARS," *VISAPP*, vol. 1, no. 2, 2014.

[9] Sun K., Zhou H. L. and Lam K. M. , "An Adaptive-Profile Active Shape Model for Facial-Feature Detection," in *ICPR*, 2014, pp. 2849-2854.

[10] Zhao X., Shan S., Chai X. and Chen X., "Locality-constrained active appearance model," in *ACCV*, 2013, pp. 636-647.

[11] Smith B. M., Brandt J. and Zhang L. Lin Z., "Nonparametric context modeling of local appearance for pose-and expression-robust facial landmark localization," in *CVPR*, 2014, pp. 1741-1748.

[12] Shen X., Lin Z., Brandt J. and Wu Y., "Detecting and aligning faces by image retrieval," in *CVPR*, 2013, pp. 3460-3467.

[13] Yu K., Zhang T. and Gong Y., "Nonlinear learning using local coordinate coding," in *Advances in neural information processing systems*, 2009, pp. 2223-2231.

[14] Milborrow S., Bishop T. E. and Nicolls F., "Multiview active shape models with SIFT descriptors for the 300-W face landmark challenge," in *ICCV Workshop*, 2013, pp. 378-385.

[15] Brunet N., Perez F. and De la Torre F., "Learning good features for active shape models," in *ICCV Workshop*, 2009, pp. 206-211.

[16] Burgos-Artizzu X. P., Perona P. and Dollár P., "Robust face landmark estimation under occlusion," in *ICCV*, 2013, pp. 1513-1520.

[17] Gross R., Matthews I. and Baker S., "Generic vs. person specific active appearance models," *Image and Vision Computing*, vol. 23, no. 12, pp. 1080-1093, 2005.

[18] Liu X., "Generic face alignment using boosted appearance model," in *CVPR*, 2007, pp. 1-8.

[19] Saragih J. and Göcke R., "Learning AAM fitting through simulation," *Pattern Recognition*,

vol. 42, no. 11, pp. 2628-2636.

[20] Cristinacce D. and Cootes T., "Automatic feature localisation with constrained local models," *Pattern Recognition*, vol. 41, no. 10, pp. 3054-3067, 2008.

[21] Saragih J. M., S. Lucey and J. F. Cohn, "Face alignment through subspace constrained mean-shifts," in *ICCV*, 2009, pp. 1034-1041.

[22] Belhumeur P. N., Jacobs D. W., Kriegman D. and Kumar N., "Localizing parts of faces using a consensus of exemplars," in *CVPR*, 2011, pp. 545-552.

[23] Tzimiropoulos G., Alabort-i-Medina J., Zafeiriou S. and Pantic M., "Generic active appearance models revisited," in *ACCV*, 2013, pp. 650-663.

[24] Tzimiropoulos G. and Pantic M., "Optimization problems for fast aam fitting in-the-wild," in *ICCV*, 2013, pp. 593-600.

[25] Cootes T. F., Edwards G. J. and Taylor C. J., "Comparing Active Shape Models with Active Appearance Models," in *BMVC*, 1999, pp. 173-182.

[26] Çeliktutan O., Ulukaya S. and Sankur B., "A comparative study of face landmarking techniques," *EURASIP Journal on Image and Video Processing*, vol. 1, no. 13, 2013.

[27] Wang N., Gao X., Tao D. and Li X., "Facial Feature Point Detection: A Comprehensive Survey," *arXiv preprint arXiv:1410.1037*, 2014.

[28] Hardoon D., Szedmak S. and Shawe-Taylor J., "Canonical correlation analysis: An overview with application to learning methods," *Neural computation*, vol. 16, no. 12, pp. 2639-2664, 2004.

[29] Shen X. B., Sun Q. S. and Yuan Y. H., "Orthogonal canonical correlation analysis and its application in feature fusion," in *IEEE International Conference on Information Fusion*, 2013, pp. 151-157.

[30] Goodall C., "Procrustes methods in the statistical analysis of shape," *Journal of the Royal Statistical Society*, pp. 285-339, 1991.

[31] Matthews I. and Baker S., "Active appearance models revisited," *IJCV*, vol. 60, no. 2, pp. 135-164, 2004.

[32] Pong K. H. and Lam K. M., "Multi-resolution feature fusion for face recognition. Pattern Recognition," *Pattern Recognition*, vol. 47, no. 2, pp. 556-567, 2014.

[33] Zheng W., Zhou X., Zou C. and Zhao L., "Facial expression recognition using kernel canonical correlation analysis (KCCA)," *IEEE Transactions on Neural Networks*, vol. 17, no. 1, pp. 233-238, 2006.

[34] Reiter M., Donner R., Langs G. and Bischof H., "3D and infrared face reconstruction from RGB data using canonical correlation analysis," in *ICPR*, 2006, pp. 425-428.

[35] Huang H., He H., Fan X. and Zhang J., "Super-resolution of human face image using canonical correlation analysis," *Pattern Recognition*, vol. 43, no. 7, pp. 2532-2543, 2010.

[36] Donner R., Reiter M., Langs G., Peloschek P. and Bischof H., "Fast active appearance model search using canonical correlation analysis," *TPAMI*, vol. 28, no. 10, p. 1690, 2006.

[37] Dalal N. and Triggs B., "Histograms of oriented gradients for human detection," in *CVPR*, 2005, pp. 886-893.

[38] Ahonen T., Hadid A. and Pietikainen M., "Face description with local binary patterns: Application to face recognition," *TPAMI*, vol. 28, no. 12, pp. 2037-2041, 2006.

[39] Zhao X., Chai X., Niu Z., Heng C. and Shan S., "Context modeling for facial landmark detection based on Non-Adjacent Rectangle (NAR) Haar-like feature," *Image and Vision Computing*, vol. 30, no. 3, pp. 136-146, 2012.

[40] Baker S., Gross R. and Matthews I., "Lucas-kanade 20 years on: A unifying framework: part 3," *CMU-RI-TP-03-05*, 2003.

[41] Nordstrøm M. M., Larsen M., Sierakowski J. and Stegmann M. B., "The IMM face database-an annotated dataset of 240 face images," Technical University of Denmark, Tech. Rep. 2004.

[42] Savran A., Alyüz N., Dibeklioğlu H., Çeliktutan O., Gökberk B., Sankur B. and Akarun L., "Bosphorus database for 3D face analysis," , 2008, pp. 47-56.

[43] Kumar N., Berg A. C., Belhumeur P. N. and Nayar S. K., "Attribute and simile classifiers for face verification," in *ICCV*, 2009, pp. 365-372.

[44] Sagonas C., Tzimiropoulos G., Zafeiriou S. and Pantic M., "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *ICCV Workshop*, 2013, pp. 397-403.

[45] Guo G. and Mu G., "A framework for joint estimation of age, gender and ethnicity on a large database," *Image and Vision Computing*, vol. 32, no. 10, pp. 761-770, 2014.

[46] Gao X., Su Y., Li X. and D. Tao, "A review of active appearance models," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 40, no. 2, pp. 145-158, 2010.

[47] Cui Y., Zhang J., Guo D. and Jin Z., "Robust facial landmark localization using classified random ferns and pose-based initialization," *Signal Processing*, 2014.

[48] Wan K. W., Lam K. M. and Ng K. C., "An accurate active shape model for facial feature extraction," *Pattern recognition letters*, vol. 26, no. 15, pp. 2409-2423, 2005.

[49] Shen X., Lin Z., Brandt J., Avidan S. and Wu Y., "Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking," in *CVPR*, 2012, pp. 3013-3020.

[50] Sun T. and Chen S., "Locality preserving CCA with applications to data visualization and pose estimation," *Image and Vision Computing*, vol. 25, no. 5, pp. 531-543, 2007.

[51] Zhao X., Shan S., Chai X. and X. Chen, "Cascaded shape space pruning for robust facial landmark detection," in *ICCV*, 2013, pp. 1033-1040.

[52] Zhu X. and Ramanan D., "Face detection, pose estimation, and landmark localization in the wild," in *CVPR*, 2012, pp. 2879-2886.