Multi-view face segmentation using fusion of statistical shape and appearance models

Constantine Butakoff\textsuperscript{a,b,*}, Alejandro F. Frangi\textsuperscript{a,b,c,*}

\textsuperscript{a}Center for Computational Imaging \& Simulation Technologies in Biomedicine, Department of Information and Communication Technologies, Universitat Pompeu Fabra, Barcelona, Spain
\textsuperscript{b}Networking Research Center on Bioengineering, Biomaterials and Nanomedicine (CIBER-BBN), Barcelona, Spain
\textsuperscript{c}Institució Catalana de Recerca i Estudis Avançats (ICREA), Barcelona, Spain

\begin{abstract}
This paper demonstrates how a weighted fusion of multiple Active Shape (ASM) or Active Appearance (AAM) models can be utilized to perform multi-view facial segmentation with only a limited number of views available for training the models. The idea is to construct models only from frontal and profile views and subsequently fuse these models with adequate weights to segment any facial view. This reduces the problem of multi-view facial segmentation to that of weight estimation, the algorithm for which is proposed as well. The evaluation is performed on a set of 280 landmarked static face images corresponding to seven different rotation angles and on several video sequences of the AV\textregistered{}CAR database. The evaluation demonstrates that the estimation of the weights does not have to be very accurate in the case of ASM, while in the case of AAM the influence of correct weight estimation is more critical. The segmentation with the proposed weight estimation method produced accurate segmentations in 91\% of 280 testing images with the median point-to-point error varying from two to eight pixels (1.8--7.2\% of average inter-eye distance).
\end{abstract}

\section{1. Introduction}

Estimation or identification of the head's pose, or its separation from other facial information have attracted much research interest for quite some time. The 3D nature of head rotation poses a great challenge for any 2D face recognition or segmentation algorithm and, if not accounted for, can cause significant performance drops. Yet there are applications where it is important to be able to efficiently process different facial views. For instance, consider surveillance applications \textit{(e.g., in an airport)}, where facial pose is usually impossible to be kept under control, or intelligent human-machine interfaces where head tracking can play an important role for interaction.

There are a number of strategies to handle multiple facial views, although most of them are applied to face recognition problems. For example, Gong et al.\textsuperscript{[1]} represented faces by either normalized intensities or using the composite face representation scheme based on the Gabor wavelet transform. The authors investigated the possibility of identifying facial pose using the facial manifold. They have shown that images corresponding to pose changes of a continuous face rotation form a smooth curve in pose eigenspace. The authors also argue that it should be possible to construct a simple but generic face pose eigenspace, which can be used to estimate poses of unknown faces. The same idea was considered by Shih et al.\textsuperscript{[2]}, where multi-view face sequence is represented as a B-spline manifold. The Euclidean distance to the manifold is used to estimate the pose of the face in question. Another work that uses two-dimensional Gabor wavelet features for pose invariant face recognition is that by Gokberk et al.\textsuperscript{[3]}. Support Vector Machines (SVM) were proposed for the problem of facial pose discrimination by Huang et al.\textsuperscript{[4]}. Although, instead of estimating the head rotation angle, SVM is used to classify any given image as belonging to one of several available views (three views are considered: frontal, 33\textdegree{} rotation to the left and to the right). A similar approach was taken by Li et al.\textsuperscript{[5]} but using a multi-class kernel support vector classifier instead of SVM and adding one extra class to represent non-faces. Another SVM-based pose estimation strategy can be found in Li et al.\textsuperscript{[6]}. A pose differentiation by \textit{k}-means clustering was proposed by Lee et al.\textsuperscript{[7]}. Okada et al.\textsuperscript{[8,9]} proposed a model, coined PCMAP. It computes bidirectional mappings between facial images and physical parameters (3D head rotation angles), via parameterized manifold representations of faces in the PCA subspace. The model is subsequently used for view-independent face recognition. Finally, in a recent publication by Sanderson et al.\textsuperscript{[10]}, non-frontal views are artificially synthesized from the frontal
ones using methods based on maximum likelihood linear regression and standard multi-variate linear regression.

Less work has been carried out in the area of pose-invariant 2D face segmentation. The most obvious approach to handle that problem is to train the segmentation algorithm with the data extracted from all the possible facial views, as for example was done by Gonzalez-Jimenez et al. [11], and then use it to segment any face. Two other solutions were proposed by Cootes et al. [12]. The first one is to construct several models corresponding to different facial views. Subsequently, during segmentation, the one that best corresponds to the image is chosen. The second approach is to create a Coupled-View Appearance Model using PCA on pairs of opposite views, but this approach requires that the views represent exactly the same expression of the same face rotated by the same angle in opposite directions, in other words, ideally, they have to be captured simultaneously (the authors circumvented this difficulty using a mirror). The work by Gross et al. [13] proposed a modification of AAM to handle occlusions. The AAM was trained on faces with artificially generated occlusions. The head rotation was treated as partial occlusion of the face. Some work has been done in the area of ASMs as well. Wan et al. [14] proposed to decouple ASM of facial features from the ASM of facial contour and use geometric algorithm to match the model to an image. The method was evaluated on the ORL face database featuring left-right rotations of up to 45°. Buxton and Dias [15] used projective geometry to adapt ASM to different viewpoints. Restricting the method to affine imaging conditions the pose variation is removed based on two reference views. Only the contours of facial features (without the contour of the face) are considered by the authors. In another work, Zhou et al. [16] consider one of the profile and the frontal views and use the Generalized Procrustes Analysis to estimate the two clusters in the shape space. The local texture models (which is similar to ASM) are learned for each cluster separately. During segmentation the updates to the shape are computed using each model and summed with appropriate weights to yield the final segmentation. The parameters of the shape model and shape regularization are performed using EM algorithm.

The strategy we propose here bears some resemblance to the approach of Zhou et al. [16]. The majority of aforementioned segmentation approaches require as many facial views as possible in the training set. What we suggest is a way to reduce the training set needed for multi-view face segmentation by applying a recently proposed multiple ASM and multiple AAM fusion algorithm [17]. The idea is to construct a number of models from some predefined facial views, and then segment any view using a model obtained by the weighted fusion of these pre-built models. Note that this is different from the approaches of Lee and Okada [7–9], who decompose the whole range of head motion into a number of sub-ranges, which are in turn approximated by linear subspaces.

In this paper we will mainly focus on the horizontal head rotations due to limitations of the landmarked databases we had access to, but we will show how the framework can be extended to the vertical rotations as well. Following this idea, the models for left, right and frontal views are constructed. Left and right head rotations must not exceed approximately 60° to avoid significant occlusions that induce topology changes in the facial shape (defined by landmarks). Then, given any view of a face, the models are fused with appropriate weights and the face is segmented by the fused model. In this fashion we limit our training set to only three views.

To evaluate the proposed approach, in the first place, we investigate the ideal case when the optimal fusion weights are known, thus allowing us to measure the potential of the method independently of the weight estimation techniques. To estimate the weight itself, any method from those available in the literature (using 3D head models, or facial pose classification or tracking) can be used. Here we simply estimate the weight by minimizing the segmentation error. The experiments demonstrate that the fused model has higher segmentation accuracy than the pre-built models corresponding to the fixed views and the model constructed from all available views. The weight estimation method is tested on the set of manually landmarked images as well as on video sequences. Finally in Section 4 we extend the method to up-down and evaluate on the CAS-PEAL-R1 database [18].

The remainder of the paper is organized as follows. Section 2 provides an algorithm for weight estimation for the problem of multi-view face segmentation. Section 3 evaluates the proposed method in terms of segmentation and weight estimation accuracy. The paper is concluded by a discussion of some aspects of the approach and conclusions in Sections 4 and 5, respectively.

2. Multi-view face segmentation using optimally fused single-view models

In our previous publication [17] we have proposed a framework for the fusion of Active Shape and Appearance Models. Here we would like to concentrate on its application to multi-view facial analysis. The goal of this work is to show how fusion can be used to reduce the training set of the models and how to extend the capabilities of classical 2D ASMs and AMMs to handle views absent in the training set. In a typical scenario these models have to be trained with many views of a face in order to segment any facial view. In this study we would like to investigate whether it would be possible to limit the required views to frontal and lateral only while still be able to perform the segmentation of any intermediate view. The proposed method relies on the fusion of models corresponding to these three views with an optimal weight.

In this section we would like to describe one of the possible approaches for weight estimation based on iterative segmentation. The idea is to find such a weight that, after segmenting an image with the fused model, the difference between the modeled and the sampled textures is minimized (in the least squares sense). A similar approach has been adopted by Cootes et al. [12] to determine which AAM is best suited for each particular image. We will treat only the case of AAM, for ASMs, it will be demonstrated, are not very sensitive to accurate weight estimation.

Let us construct three AAMs from three sets of views: frontal view, left and right views with 60° head rotation each. They will be referenced by frontal, left and right views and models, respectively.

Let \( w \in [-1, 1] \). Let \( M_f \), \( M_l \) and \( M_r \) denote the frontal, left and right models. Since during head rotation the head rotates either to the left or to the right there is no need to fuse all three models simultaneously, therefore let us formulate the fused model as follows:

\[
M(w) = \begin{cases} 
\left[ (-w) \otimes M_f \right] \oplus \left[ (1 - |w|) \otimes M_f \right], & w < 0 \\
\left[ (1 - |w|) \otimes M_l \right] \oplus \left[ |w| \otimes M_r \right], & w \geq 0
\end{cases}
\]  

(1)

where \( w \) is the weight, “\( \otimes \)” represents weighting the model and “\( \oplus \)” represents fusion. As we can see the problem of finding the optimal model for segmenting a specific facial pose is reduced to optimization of a function of one parameter, varying from \(-1\) to \(1\). When the fused model is used to segment a specific image, the result is the shape of the face (the contours, defined by landmarks) and two texture vectors: one is the real texture inside the shape, sampled from a given image and normalized, and the other is the texture estimated by the model as the best matching facial appearance. The objective of optimization is to find a weight that minimizes Root Mean Squared Error (RMSE) between these two texture vectors. It is not an easy task to formulate the gradient for such a function, so it was decided to use an optimization algorithm without deriva-
left and right rotations and the number stands for the corresponding rotation angle. We will use the same names for the models constructed from the corresponding datasets (e.g., frontal model, left model). And thirdly, we are always going to fuse the frontal model with either r60 or r20, so $M_R$ in (1) corresponds to $r60$ and $M_f$ to $160$.

Whenever the manually landmarked shapes are available, the accuracy of segmentation is evaluated using the point-to-point error:

$$E = \frac{d}{n} \sum_{i=0}^{n} \sum_{j=1}^{d} (\hat{x}_{i,j} - x_{i,j})^2$$  \tag{2}

where $x_i$ is the $i$th element of the fitted shape, $n$-vector $\mathbf{x}$, and $\hat{x}_i$ is the $i$th element of the manually defined shape $\mathbf{X}_d$ and $d$ is the dimensionality of the space where the shape is defined (in the case of 3D facial images $d = 2$).

In the following sections we will evaluate the fusion framework in itself, when the optimal weight is known (to separate the error introduced by the weight estimation), and altogether with the weight estimation scheme.

### 3.1. Fusion framework evaluation with a known optimal weight

In the first experiment we want to demonstrate how the fusion of models can be used to segment the views absent in the training set. To that end, the optimal fusion weights for 120, r20, 140 and r40 datasets have been determined. To determine these weights all the images have been segmented by the model $M(w)$ with $w$ varying from $-1$ to $1$ in steps of 0.05. To reduce the effects of bad initialization, the segmentation began by centering the mean shape at the centroid of the ground truth shape and rescaling it to the average face size of the training set. For each image, the weight resulting in the smallest point-to-point error with respect to the average face size of the training set. For each image, the weight resulting in the smallest point-to-point error with respect to the manual landmarks was declared optimal.
presented in Fig. 2. The error is plotted against the absolute value of
the weight to verify the similarity of the graphs corresponding to
left and right views. It can be noted that ASM appears to be much
less sensitive to the accuracy of weight estimation. The AAM in its
turn demonstrates no significant reduction in segmentation accu-

racy within ±0.05 interval around the optimal weight. As it can
be seen from that figure (in the case of AAM) the optimal weights
are ±0.7 for 140 and r40 sets and ±0.25 for 120 and r20 sets.

Fig. 3 presents a comparison of accuracy of segmentation per-
fomed according to various strategies with 95% confidence inter-
vals. The meaning of the labels depends on the testing set and is
presented in Table 1. Each cell of this table explains how the model
is constructed for each particular testing set.

The point-to-point errors are computed with respect to man-
ual landmarks. It is worth to note that normal is the typical ap-
proach used when all the available views are used to train the
model (in this case only right and frontal or left and frontal; using
all three of them significantly distorts the mean shape and the
segmentation error is quite large), and closest is the same ap-
proach as that of Cootes et al. [12]. As a reference, the figure
shows the baseline segmentation results obtained by models
constructed from the test sets themselves, thus providing the best
possible results. It can be seen that, in all the cases, the model ob-
tained by the fusion performed better or in the case of ASM equally well as the best of the other models (except the
baseline). An example of segmentation using each of the men-
tioned models can be seen in Fig. 4.

Considering the presented results we can conclude that:

- Using different weights the fusion provides a way to linearly
  “interpolate” active shape and appearance models. The
  approach when the closest model is chosen could be consid-
ered as zero-order or nearest neighbor interpolation.
- The segmentation of any facial view, corresponding to a hori-
  zontal head rotation, can be improved by fusion of the two closest
  views, provided that the weights are estimated correctly. In this
  particular case, having only three models corresponding to fron-
  tal, left and right views, fusion could be used to interpolate these
  models and use the result to segment any other view.
- Active Appearance Models are much more sensitive to the cor-
  rect weight estimation. Which means that while for ASM it
  would be enough to fuse frontal and 160 models with equal
  weights to segment any left view, AAM requires more accurate
  weight estimation.

3.2. Fusion framework evaluation with unknown optimal weight

In this section we would like to evaluate how the fusion benefits
the accurate segmentation of face in static images and videos,
when the weight has to be estimated, using the approach proposed

<table>
<thead>
<tr>
<th>Model type</th>
<th>Test set</th>
<th>140</th>
<th>120</th>
<th>r20</th>
<th>r40</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal</td>
<td>Single model trained on frontal and 160 sets</td>
<td>Single model trained on frontal and 160 sets</td>
<td>Single model trained on frontal and 160 sets</td>
<td>Single model trained on frontal and 160 sets</td>
<td></td>
</tr>
<tr>
<td>Fused</td>
<td>Fusion of frontal, 160 and 160 models with the weight equal −0.7</td>
<td>Fusion of frontal, 160 and 160 models with the weight equal −0.25</td>
<td>Fusion of frontal, 160 and 160 models with the weight equal 0.25</td>
<td>Fusion of frontal, 160 and 160 models with the weight equal 0.7</td>
<td></td>
</tr>
<tr>
<td>Closest</td>
<td>160 model</td>
<td>Frontal model</td>
<td>Frontal model</td>
<td>r20 model</td>
<td></td>
</tr>
<tr>
<td>Baseline</td>
<td>140 model</td>
<td>120 model</td>
<td>r20 model</td>
<td>r60 model</td>
<td></td>
</tr>
</tbody>
</table>

Fig. 4. An example of segmenting one of the 140 images using different AAM models. (a) Manual segmentation; (b) best possible segmentation with AAM built from the
testing set; (c) fused AAM with the optimal weight; (d) AAM built from the frontal and 160 sets; (e) 160 AAM; (f) frontal AAM.

Please cite this article in press as: C. Butakoff, A.F. Frangi, Multi-view face segmentation using fusion of statistical shape and appearance models, Comput.
in Section 2. The goal is to find a weight for the model $M(w)$ in (1) such that it provides the best possible reconstruction of the facial texture after segmentation. In other words, the RMSE between the sampled and generated textures should be minimal. An example of this objective function for all the sets has been plotted in Fig. 5 on a logarithmic scale. The error is computed for each weight sampled from the interval $[1,1]$ in steps of 0.05. It can be seen that these functions are not strictly unimodal, nevertheless the minimization algorithm is robust to certain local minima [19].

Since in many of the cases considered in the following experiments, as in real life, the landmarks are unavailable, the model matching will start from the mean model instance rescaled to fit into the smallest rectangle containing the face. The rectangle is defined manually but could be estimated as output of any face detection algorithm [21,22].

Firstly, to evaluate the accuracy of the weight estimation, the algorithm was applied to all the landmarked sets: frontal, 120, 140, 160, r20, r40, and r60. For every image the estimated weight was compared to the optimal value, estimated by exhaustive search. The distributions of the absolute differences between the optimal and estimated weights are presented in Fig. 6. Table 2 shows how many images from the datasets (40 images in each) had the weight estimation error smaller than 0.05 and 0.10. The last row shows the number of images where the model diverged completely due to incorrect weight estimation. The boxplot of the point-to-point errors corresponding to the Table 2 is shown in Fig. 7a, where crosses (outliers) correspond to the cases of divergence and circles correspond to the median error obtained by the AAM constructed from the frontal, 160 and r60 sets (in other words, without fusion). In general it can be seen that the fusion was always beneficial, especially for the extreme rotations, where the specific left or right model took over the responsibility for the segmentation and outperformed the generic model (partially this improvement also comes from the better initialization of the fused model, as its initial shape had correct rotation while the generic model always started from the frontal view). The most successful estimations were achieved for the frontal images, mostly because of the same reason – the minimization also started from a pose close to frontal. We would also like to note that in some cases, when the weight estimation error was slightly larger than 0.1, the model still was able to correctly segment the images.

![Fig. 5. Texture errors of segmenting AV@CAR database by a fused AAM (1), computed for the weight varying from –1 to 1 in steps of 0.05. Plotted are the 1st, 2nd (median) and 3rd quartiles for all the sets. The ordinate is scaled logarithmically.](image)

![Fig. 6. Histograms of weight estimation errors per testing set of the AV@CAR database for AAM fusion. The errors are computed with respect to the optimal weights, estimated by exhaustive search.](image)

<table>
<thead>
<tr>
<th>Set</th>
<th>120</th>
<th>r20</th>
<th>140</th>
<th>r40</th>
<th>160</th>
<th>r60</th>
<th>Frontal</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.05</td>
<td>24</td>
<td>34</td>
<td>22</td>
<td>32</td>
<td>30</td>
<td>35</td>
<td>4</td>
<td>220</td>
</tr>
<tr>
<td>0.10</td>
<td>32</td>
<td>34</td>
<td>27</td>
<td>33</td>
<td>36</td>
<td>30</td>
<td>38</td>
<td>230</td>
</tr>
<tr>
<td>Diverged</td>
<td>4</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>5</td>
<td>9</td>
<td>0</td>
<td>25 (8.9%)</td>
</tr>
</tbody>
</table>

Table 2 Percentages of correctly estimated weights for the AV@CAR database.

Please cite this article in press as: C. Butakoff, A.F. Frangi, Multi-view face segmentation using fusion of statistical shape and appearance models, Comput. Vis. Image Understand. (2009), doi:10.1016/j.cviu.2009.11.001
The first one is landmark placement. The principal role of landmarks is to define contours of a face and its features (eyes, nose, mouth, etc.). It can be noted that, as the head rotates, some landmarks, if maintained fixed, would become occluded. At the same time the visible contours of the face changes. Therefore, during head rotations the landmarks must stay on the contour of the features they outline (just as it is done in many other studies [13,12]). Since both ASM and AAM model the displacements of points, we tried to keep the number of irrelevant displacements (along the contour) to the minimum by keeping the landmarks that do not correspond to the prominent facial features approximately in the same relative position with respect to the prominent ones. Fig. 11 illustrates the correspondence between landmarks in two facial views.

Another problem is the absence of texture information in occluded areas. Since there is only one camera, there is no way of getting information about the occluded cheek of the rotated head. Therefore, when the texture is warped from the rotated face to the mean shape of the fused model, there will be texture distortions. The views with more occlusions will have their texture stretched depending on the rotation angle between that view and the view corresponding to the mean shape. In spite of this problem, AAMs recover the texture quite well and most facial features are still distinguishable as demonstrated in Fig. 12. Of course
the effect can be reduced by including more views of faces but that is contrary to our goal, which was to reduce the number of training views.

Now we would like to put our proposed method in the context of related work. Among the existing approaches to pose-independent 2D face segmentation based on AAM and ASM there are several approaches that are worth noting. Cootes et al. [12] constructed separate models for a number of viewpoints and segmented a given face by the model corresponding to the closest viewpoint. Similarly Vogler et al. [23] proposed to construct several models for different pose angles and the relevant model was picked based on the pose estimated by a 3D deformable model. These cases were considered in Section 3.1, where it was shown that these approaches can be improved by our fusion without requiring additional training data. In the context of our fusion they can be interpreted as a nearest neighbor model interpolation and by using our fusion the interpolation can be made linear. On the other hand, since they already handle the angle estimation, adding our fusion would not require any additional weight estimation and precomputing the fused models would avoid increasing the computational complexity of those algorithms. Cootes et al. [12] also proposed another approach, the Coupled-View Appearance Models, where the model is constructed from pairs of different facial views, but this approach requires that all the facial views are captured simultaneously. The papers by Gross et al. [13] and Hu et al. [24] proposed alternative approaches, one based on a modified AAM matching algorithm and the other using an AAM based on wavelets. All of these approaches require more than just frontal and two lateral views for training. Our goal was to develop a strategy for model interpolation that allows creating facial appearances unavailable during training. As a consequence the training set can be reduced to only a minimum set of views. Another paper by Zhou et al. [16] reported all the errors as difference between the methods giving the percentages of images where their method outperformed the other ones.

---

**Fig. 9.** Evaluation of the pose estimation accuracy, in the AV@CAR database, (assuming linear relationship between the weight and pose angle) on testing sets (a) and video sequences (b). The pose errors are computed as the difference between the optimal and estimated weight multiplied by 60. The optimal weight is estimated by exhaustive search. The size of each image and video frame is 768 × 576. Error bars represent 95% confidence interval of the mean.

**Fig. 10.** Sample video frames wherein the AAM has diverged.

**Fig. 11.** Landmark correspondence between the frontal and profile views. The landmarks that do not correspond to prominent facial features stay approximately in the same relative position with respect to the prominent ones.

**Fig. 12.** Texture obtained by AAM segmentation: (a) an original from the 120 set and the matched model; (b) an original from the 140 set and the matched model.
In spite of the aforementioned difficulties, we can compare our method to that of Wan et al. [14] who used the ORL face database [25] for evaluation. Since the results are reported in pixel unit, we expressed the results as percentage of average inter-eye distance (distance between eye centers). In AV@CAR database this distance is approximately 111 pixels and in the ORL – 35 pixels. It should be noted as well that most subjects from the ORL database have only two lateral views (left and right) per face, each corresponding to a rotation angle smaller than 45°. The normalized segmentation accuracy results, presented in Table 3, seem to be comparable for both algorithms, although those corresponding to the frontal view are substantially better in our case. On the other hand it should be noted that our model was trained on 60°-rotated and frontal faces, and tested on other facial views of the same people, while Wan et al. [14] used the same views for both training and testing, but corresponding to different people. Thence, in the latter case it can be concluded that the view-specific information has been learned by the model and it is not clear how it would handle views unavailable during the training.

Finally, we would like to comment on performance. As it was mentioned, it takes about eight iterations to converge to the optimal weight and to segment an image. Normally, it takes a couple of minutes to perform these eight matchings and fusions (for 768 × 576 images with the face occupying approximately a rectangular area of 240 × 200 pixels) with our non-optimized matching routines. But if the images are reduced four times to a still reasonable size, when the face occupies approximately 60 × 50 pixels the whole process takes about 5 s (on the Intel Pentium Q6600 2.40 GHz). It is worth to mention that the proposed optimization approach relies on interval partitioning and since all the possible partitionings are easily predictable, all the models can be fused a priori and stored. In this case the whole segmentation takes about 0.4 s (including loading and saving data).

### Table 3
Segmentation accuracy comparison between different algorithms in terms of point-to-point error.

<table>
<thead>
<tr>
<th>Our approach on AV@CAR database</th>
<th>Wan et al. [14]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Set</td>
<td>Mean error (%)</td>
</tr>
<tr>
<td>120</td>
<td>2.6</td>
</tr>
<tr>
<td>140</td>
<td>2.6</td>
</tr>
<tr>
<td>r20</td>
<td>2.6</td>
</tr>
</tbody>
</table>

![Figure 13](https://example.com/image.png)

**Fig. 13.** Segmentation accuracy of 20 CAS-PEAL-R1 individuals measured as a mean squared error with respect to the ground truth eye centers, normalized by the average inter-eye distance of frontal views. Shown are the results for up (a), front (b) and down (c) views and angles from −45° to 45°. Circles represent the segmentation by the AAM constructed from the up, down, frontal and ±45° facial views.
The evaluation presented here is limited to only horizontal head rotations. The problem is twofold. On the one hand, the few available multiple-view databases lack in consistency of pose acquisition and the acquired poses do not correspond to the labeling. On the other hand the big issue is unavailability of landmarked facial contours for these databases. Therefore our approach was adapted to the problem of left–right head rotation, nevertheless it can be generalized to any head pose by minimizing the following function

\[ M(\omega_1, \omega_2) = M_{UD}(\omega_1) \odot M_{LR}(\omega_2) \odot (1 - |\omega_1| - |\omega_2|) \odot M_F(\theta) \]

where

\[ M_{UD}(\omega) = \begin{cases} -\omega \odot M_U, & \omega < 0 \\ \omega \odot M_D, & \omega \geq 0 \end{cases} \]

\[ M_{LR}(\omega) = \begin{cases} -\omega \odot M_L, & \omega < 0 \\ \omega \odot M_R, & \omega \geq 0 \end{cases} \]

and \( M_U, M_D \) are the models constructed from the faces looking upwards and downwards, respectively. To optimize this function, any minimization algorithm can be used. To reduce the number of function evaluations, which involve segmentation, we preferred to use algorithms without derivatives. One possible candidate for minimizing this new function of two variables is Powell algorithm, but any other can be used.

To evaluate the above function we have selected first 10 men and 10 women from the CAS-PEAL-R1 database [18], which had relatively consistent up-down head rotations (user numbers 1–13, 32, 45, 46, 56, 63, 71, and 89). The left–right head rotation was quite consistent because a set of cameras has been positioned around the head. The 67° and 90° angle rotations had been discarded because half the face has been occluded. The remaining angles were ±45°, ±22° and 0° for the faces looking upwards, downwards and straight (300 images in total). It is interesting to note that the 45° pose in this database looks like the 60° pose in AV@CAR (this fact does not influence the evaluation as long as the views are consistent in the training sets). The images corresponding to the up, down and front 0°, as well as front ±45° have been manually landmarked and used for model construction. All other images have been used for evaluation. The model was again initialized within a bounding box of the face. The rectangle in this case was estimated from the centers of the eyes (supplied with the database) using the formulas presented in Table 4. The formulas were derived based on visual observation. Note the symmetry about frontal pose and that the \( y_i \) are always the same except for the “down” pose where the face appears shorter.

The accuracy of segmentation was estimated as the mean squared error of localizing the eye centers with respect to the ground truth (see Fig. 13). Again the result of the fused model is better than the AAM constructed from all the views. Since it is impossible to judge the overall segmentation accuracy only by eye localization, we also checked visually the quality of the segmentation. For minimization we used VNL implementation of Powell algorithm.

In case of the traditional AAM, in a majority of cases, the model did not detect head rotation and failed to converge. As a result 76% of the images were well segmented by the fused model with Powell minimization, while the AAM constructed from all the views succeeded only on 26%. What was visually observed, on both AV@CAR and CAS-PEAL-R1 databases, is that usually the fused model either segmented the images well or the minimization algorithm got stuck in a local minimum and, due to incorrect weights, the
Fig. 15. Examples of almost converged model on CAS-PEAL-R1, with one image per tested pose. The number on top indicates the error in eye center localization in percents of average inter-eye distance. Minimization has been performed using Powell algorithm starting from (0,0) weights.

Fig. 16. Examples of diverged model on CAS-PEAL-R1, with one image per tested pose. No divergences have been observed for the frontal pose. The number on top indicates the error in eye center localization in percents of average inter-eye distance. Minimization has been performed using Powell algorithm starting from (0,0) weights.
segmentation failed to converge to a satisfactory result. The CAS-PEAL-R1 case was more challenging because the parameter search space is much bigger. Some example images, with the eye localization accuracy on top of each, can be seen in Figs. 14–16. In spite of a significant improvement in the segmentation success with respect to the normal AAM, it is clear that the Powell minimization did not recover weights well in all the cases. The principal reason for this is that the minimization always starts in the same way, ignoring the pose. A better alternative could be to start the minimization from better initial weights, using multiple initializations or using an algorithm more robust to local minima. For example, running the minimization, in parallel, using five different initial weights: (0,0), (0.5,–0.5), (–0.5,0.5), (–0.5,–0.5) increases the number of converged cases to 94%.

5. Conclusions

In this work we have presented an application of AAM and ASM fusion to multi-view face segmentation. The fusion can be casted into a model interpolation problem, allowing to obtain a better segmentation for views absent in the training set. The latter leads to a possibility of reducing the amount of manually landmarked facial views required for training and keeping them to a minimum: frontal and two lateral (60°) facial views. Then if the fusion weight is estimated correctly, any facial view can be segmented by the fused model. In Section 2 we presented a simple algorithm for weight estimation. The estimation was successful in 91.1% of 280 testing images. In the remaining cases the algorithm converged to incorrect local minima resulting in incorrect segmentation. Since each image was segmented independently the weight estimation in video sequences could be significantly improved by tracking the weight or the facial pose along the sequence.

Acknowledgments

This work was partially funded by Grants TIC2002-04495-C02 and TEC2006-03617/TIC from the Spanish Ministry of Education & Science, and Grant FIT-360005-2007-9 from the Spanish Ministry of Industry. A.F. was supported by Ramón y Cajal Research Fellowship. The Center for Computational Imaging and Simulation Technologies in Biomedicine (CISTIB) at Universitat Pompeu Fabra is a member of the Biosecure (IST-2002-507534) European Network of Excellence and the CIBER-BBN national center (CB06/01/0061).

References


ARTICLE IN PRESS


Please cite this article in press as: C. Butakoff, A.F. Frangi, Multi-view face segmentation using fusion of statistical shape and appearance models, Comput. Vis. Image Understand. (2009), doi:10.1016/j.cviu.2009.11.001