Pools of AAMs: Towards Automatically Fitting any Face Image

Julien Peyras Adrien Bartoli Samir Khoualed
LASMEA, Clermont-Ferrand, France
name.surname@gmail.com

Abstract

Fitting a single generic AAM on an unseen face (that is not in the training set) under any pose and expression is very difficult. The variability of the data is so high that the fitting process usually gets stuck into one of the numerous local minima. We show that a solution to this problem consists to separate the variability sources. We build a pool of *specialized* AAMs. Each AAM is trained over multiple identities, all shown under the same pose and expression. We then retain the AAM that shows the smallest residual error when fitted to the input image. The fitting obtained in this manner is very accurate on unseen faces. The ultimate goal is to automatically train a person-specific AAM. In addition, the pool of specialized AAMs allows us to recognize the face pose and expression at each frame of the video with good performances. The proposed method has potential applications in Human Computer Interaction and driving surveillance, to name just but a few.

1 Introduction

The problem of face analysis in still images and videos has been extensively studied for years. This intense research activity finds its motivation in the possibility to set up a large range of applications in the medical, psychological and linguistic fields (cognitive studies, expression transfer on an avatar, *etc.*). Face analysis is a difficult topic since face images vary in identity, pose and expression. The sought-after model should be able to automatically and reliably describe previously unseen faces under any pose and expression. We describe the two most promising approaches.

The first one is Bartlett *et al.*'s machine learning based expression analysis solution proposed in [1]. Several classifiers are trained for face and eye detection, as well as for the presence and intensity of particular Action Units. These are the elementary deformations occuring on a face, as described by Ekman's Facial Action Coding System [6]. This method is probably the best performing one in the literature for expression analysis on *unseen faces* (faces that are not explicitly learnt by the classifiers). The method is non model-based. This makes it difficult to retrieve the shape, and so restrain the range of possible applications.

The second established approach is the Active Appearance Model (AAM) proposed by Cootes *et al.* [3]. An ad hoc face AAM is trained on manually labeled images, so as to learn the shape and appearance bases. An optimization process is used to fit the

BMVC 2008 doi:10.5244/C.22.44

AAM on an input image: the shape and appearance coefficients of the model are tuned until the model instance matches the input picture. Retrieving the face shape is important for many video post processing systems. Obviously the performance of such systems is directly related to the quality of the face shape description, *i.e.*, the fitting accuracy is crucial.

As Gross *et al.* [7] first pointed out, it is important to distinguish between two situations, providing two different kinds of achievable fitting accuracy:

- the *person-specific* context, where the fitted face has been explicitly learnt by the model. The fitting accuracy is usually very good in this context, and reliable for post processing systems. In [8], Lucey *et al.* use person-specific AAMs to retrieve the face shape and successfully classify facial deformations into Action Units.

- the *person-generic* context, where the fitted face is not in the training set. As first shown by Gross *et al.* in [7], the fitting process is much harder than in the person-specific context. In [10], Peyras *et al.* showed with carefully chosen experiments that fitting an unseen face with an AAM is much less accurate than fitting a face that belongs to the set of images used to train the model. They explained the reason for this: in the generic context, the appearance counterpart of the model cannot fully explain the appearance of the face in the input image. As an unfortunate consequence, the minimum error of the cost function corresponds to a biased position of the model. Even when initialised in the best possible position (the ground-truth shape), the AAM drifts away.

The problem of fitting unseen faces is a corner-stone for an extended amount of applications. As of today, no method have proven able to accurately fit previously unseen faces under a wide range of poses and expressions. AAMs appear to provide an interesting basis to face this problem. One could think that adding more training data would increase the ability of the model to generalize to unseen faces. Indeed, this ability increases with the amount of training data. In practice however, the higher complexity of the AAM makes its fitting unreliable because this induces numerous local minima in the cost function. In other words, the model is so flexible that it 'explains' spurious non face solutions in the image. As a consequence, the solution for reliable and accurate fitting must combine these two contradictory conditions:

- the complexity of an AAM must be kept as low as possible so as to preserve a large convergence basin and be able to find the global cost minimum,

- the range of face images that the AAM can explain must be large, so that the global cost minimum matches the sought after solution.

The first condition is satisfied by limiting the size of the training set while the second one requires to expand the training set. To bypass such a contradiction, we propose to separate the sources of variability within the training data. Instead of considering the face as an object that varies in identity, pose and expression, we see it as a collection of objects that vary in identity only: each object has a constant pose and expression. In this view, an AAM must model only one of the three sources of variability: identity, so as to fit a variety of unseen faces under the same pose and expression. We say that such an AAM is *specialized* to a particular pose and expression pair. To deal with many poses and facial deformations, we train a *pool of specialized AAMs*.

Contribution. We showed in [10] that fitting an unseen face with local models increases the generalization ability and the fitting accuracy in comparison to global models covering all facial features. The fitting bias is reduced to a point where the fitting accuracy on unseen faces is equivalent to the accuracy of manual labels. Following this insight, we design two categories of specialized AAMs that locally model the face: the *upper AAMs*, built to fit the eyes and eyebrows, and the *lower AAMs*, designed to fit the mouth. This also presents the advantage to model separately the possible combinations of facial deformations. Our strategy consists to run all upper and lower AAMs on one input picture. For each category we keep the AAM presenting the smallest residual error. This AAM is expected to be the most accurately fitted on the face, and should represent the current pose and expression of this face. Consequently, we expect our method to automatically provide accurate labels on unseen faces under varying expression and pose, and also to correctly classify the pose and expression at any frame of a video. The process is presumably slow and costly. This is often not a limitation: the long off-line training is performed only once, on a video of a person who frequently uses the device at hand. As an example, communication with personal-computers and car driver monitoring systems can be equipped with this technology. As two important contributions, we show that:

- good fitting accuracy, good robustness to position perturbation and high classification rates are obtained,
- the obtained labels can be used to automatically train a person-specific AAM, which is able to fit the face and classify its expression in real-time.

Organization. Section 2 reviews the literature and introduces the AAMs. Section 3 presents the specialized AAMs and the pose and expression database we have used to perform our experiments. In section 4 we show experimental results on still images in a leave-one-identity-out fashion, and on a video where an unseen person displays a series of poses and expressions. We compare the performance of the specialized AAMs against the classical AAM learning all data. Section 5 gives a conclusion and our perspectives. The good fitting results of the specialized models will allow us to build a person-specific AAM for real-time tracking and pose and expression classification on the just-learnt person.

2 Background

2.1 Previous Work

The concept of fitting several models is not new: Cootes *et al.* used one model for each face pose in [4]. However, despite the advantages it presents, this solution were not pursued afterward.

The AAM is not the unique face fitting solutions in the literature. We review some others. Cristinacce *et al.* proposed a competitive template matching solution called *Constrained Local Models* in [5], which were further studied by Wang *et al.* in [11]. This solution exhibits better fitting results than AAMs. Note that these methods can be embedded as the specialized models in our framework. Indeed, pools would increase the discriminability between correct and wrong alignments, which is an important ability when aligning objects with a very high and complex range of variability.

The 3DMM (3D Morphable Model) presented by Vetter *et al.* in [2] can recover the 3D structure of a face from a single picture. This model is too heavy to automatically and reliably fit faces under any pose and expression. Here too, the specialization of multiple 3DMMs could be help to improve the results.

2.2 Background on the AAM

An AAM combines two linear subspaces, one for the shape and one for the appearance. They are learnt from a labeled set of training images [3]. A certain percentage of the whole training set shape and appearance variance is kept. As a rule of thumb, [10] showed that keeping 60% shape variance and 100% appearance variance is 'optimal' in the person-generic context. We therefore keep 60% shape and 95% appearance variance, so as to keep the AAM size reasonable.

Fitting an AAM consists to find the shape and appearance instances that make the residual error between the image and the synthesized model as small as possible. We use Baker and Matthews' optimization framework [9] with the *Simultaneous Inverse Compositional Algorithm*.

3 A Pool of Specialized AAMs

3.1 The Concept

In [10], both global and local models are specialized on the frontal pose and neutral expression. Since stuffing various poses and expressions into a single AAM spoils its fitting performance, we extend here the concept of *specialized AAM*. The idea is to build a pool of AAMs, each being specialized on a particular pose and expression pair. The whole pool would then encompass a continuum of poses and expressions.

Each specialized AAM is built over N different identities, giving the AAM a certain ability to explain unseen faces. Unfortunately, none of the publicly available face databases present a large range of facial deformations under several head poses and an homogeneous illumination. For this reason, we had to build our own pose and expression database that we present in the rest of the section.

3.2 The Pose and Expression Database

Our current database has 15 identities taken under 3 views (frontal, 10° and 20° in azimut) displaying 21 facial (upper or lower) deformations. We kept the illumination homogeneous. All pictures (63 per identity) were manually labeled thoroughly to maximize the label accuracy. Taking pictures and labeling them represents about 3 hours of work per identity. The facial deformations we use are showed in figure 1. Figure 2 shows a sample of people from the database.

It is obvious that more people, more poses and more deformations could be included in the database to fit more unseen people under a less restricted amount of poses and expressions. However one faces several difficulties:

- it is *time consuming and tedious* to label images with high accuracy, as this present study requires.

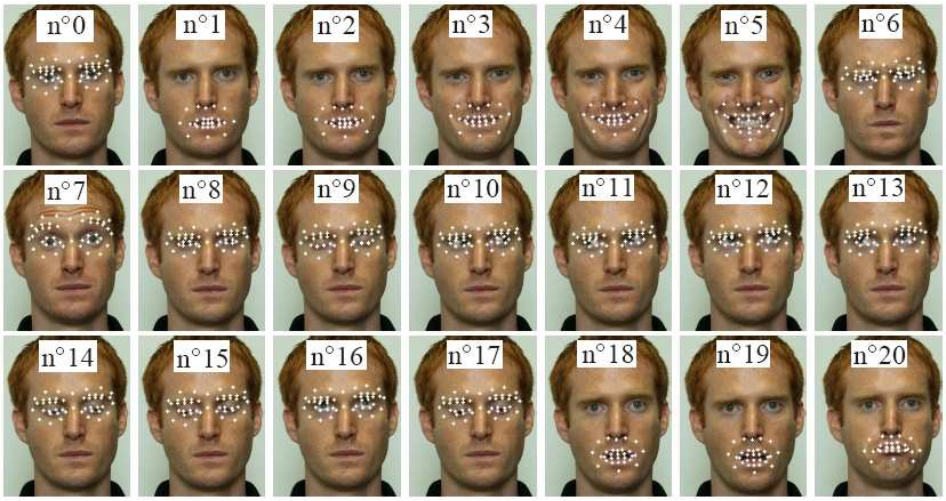

Figure 1: Facial deformations represented in the database. The manually placed landmarks represent the vertices used for training or fitting (for testing purposes). The deformation number is indicated on top of each of the thumbnails. Each deformation is meant to represent some Action Units or a particular combination of them [6].

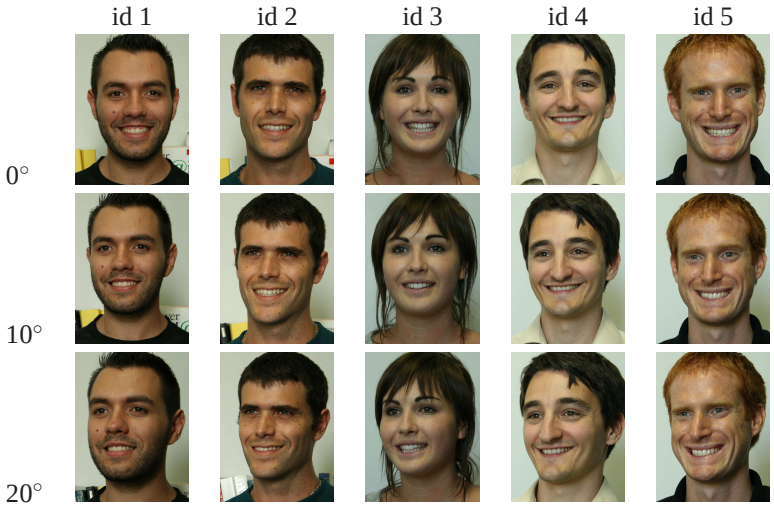

Figure 2: 5 of the 15 identities of the database for all poses and deformation n°5.

- the *appearance and deformation* of faces are *wide-ranging*. The set of people forming the database must capture this diversity, in quality and quantity.
- the *quality of the deformations is very important* to prevent from badly defined deformation classes and their possible overlaps. The selected people composing the database should therefore be actors or possess some particular talents to perform facial deformations on demand.

4 Evaluation and Tests

4.1 Leave-One-Identity-Out Test

The test consists to train a pool of specialized AAMs on N identities and to operate the fitting on one of the remaining faces. In this way, the identity we fit is unknown from the AAMs. We perform this leave-one-identity-out test 15 times. N can at most be 14. For each test identity, 63 images (21 expressions under 3 poses) must be fitted with all upper or all lower specialized AAMs. For each image, we run all AAMs and keep as the winner the one that makes the smallest residual error at convergence, after 30 iterations. Our goal is to assess the two following points:

- the *fitting accuracy*, *i.e.*, the quality of each label position on the face at convergence: we measure it by comparison with manual labels taken as a reference,
- the *basin of convergence*, *i.e.*, the ability to cope with perturbed initializations,
- the *classification rate*, *i.e.*, the frequency of correct correspondence between the pose and expression of the winning AAM and the true pose and expression.

From the first and second observations we will see whether our solution allows one to automatically label a sequence showing an unseen face. From the third observation, we will see if such a sequence can also be automatically labeled in terms of pose and expression.

4.1.1 Fitting Accuracy and Convergence Basin

To score the fitting accuracy of the winning AAM on a test image, we normalize the coordinates of the manual labels (taken as a reference) in scale. This makes the external eye *resp* mouth corners distant by 100 *resp* 80 pixels for the upper *resp* the lower model. The scale value used to normalize the reference shape is also applied to the winning AAM's shape. We consider the accuracy error to be the mean euclidean distance between the labels of the reference and the fitted shape.

The top of figure 3 shows that the fitting accuracy of specialized models increases with the number N of identities learnt by the specialized models. This accuracy is however lower than the accuracy of the single learn-all-data AAM. This holds when the models are very well initialized, using the manual labels. On the other hand, bottom of figure 3 shows that the specialized models are more robust than the single model to initial perturbations and more often converge to the correct solution. This is a suitable property for tasks like tracking: in video sequences, the geometric discrepancy between the frames can be large and the fitting function must present a steep and large basin of convergence to ensure the model to find a correct solution in any frame. A test on a video is performed in §4.2.

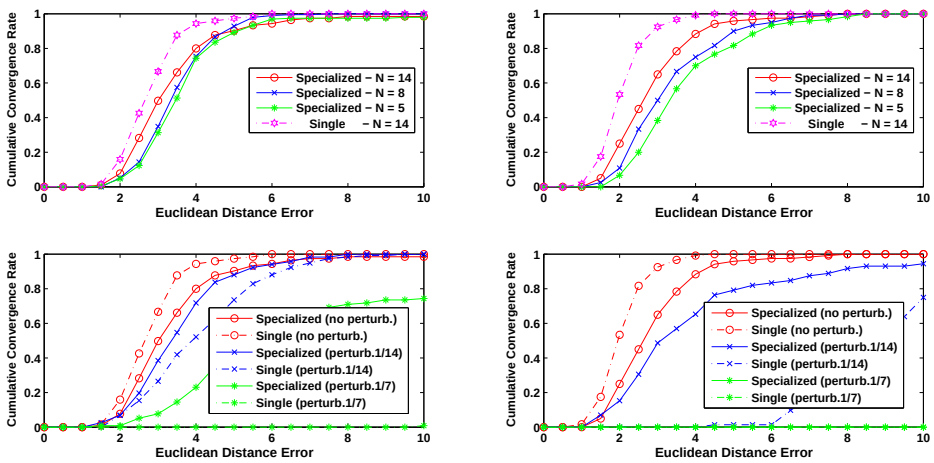

Figure 3: (Top) the fitting accuracy is illustrated for various sizes N of the training dataset. The behavior is similar for specialized upper AAMs (top left) and lower AAMs (top right). For comparison purposes with the $N = 14$ case, the dashed curve represents the accuracy obtained with a single AAM classically learning all deformations, identities and poses. For these tests, manual labels are used to perfectly initialize all model fittings. (Bottom) the fitting robustness to the initial position is compared for single and specialized upper AAMs (bottom left) and lower AAMs (bottom right) for different perturbation intensities. The case of perfect initialization is repeated for comparison sake. For the perturbed cases, the models are initialized in a shifted position along the horizontal and vertical axes. Two shifting magnitudes are tested: $1/14$ and $1/7$ of the distance separating the two external eye corners. For the latter, there is initially no overlap between the eyes in the image and the eyes in the model.

4.1.2 Classification Results

Table 1 shows the percentage of correct classification obtained on all test images for different training set sizes and for two position perturbation magnitudes in the $N = 14$ case. The classification rate increases with the number of identities learnt by the AAMs. For $N = 14$, 71% correct expression classification is reached on unseen identities when no perturbation is introduced in the initial position. This percentage of correct expression classification is high, and most of the classification confusion is obviously made between classes that are very close. The three different intensities for the smile (deformations n°2, 3 and 4), and the two mouth apertures (deformations n°18 and 19) can easily be confused and the classification failure mainly comes from this reason. Coarser clusters of expressions would improve the classification results, whereas further detailing the range of expressions in the database would increase the possible overlap between classes.

4.2 Test on a Sequence

We compare the tracking ability of our pool of specialized AAMs and the single learn-all-data AAM. To do so, we select, manually label the most informative frames of the

↓ Correct Classification of ↓	N=5	N=8	N=14	N=14 (perturb.1/14)	N=14 (perturb.1/7)
Expression	64%	68%	71%	67%	23%
Angle	68%	73%	74%	72%	40%
Expression + Angle	48%	52%	54%	50%	10%

Table 1: Percentage of correct classification of expression only, angle only and both expression and angle. Results are given for various sizes of the training set and for the case where the model trained over 14 identities is perturbed in position.

sequence and train an AAM able to fit the whole sequence with high accuracy. This provides the reference face labels at each frame for the test described below.

The test consists to perform the long multiple specialized AAM fitting process on each frame and score the fitting accuracy with respect to the reference labels. A similar test is done with the single, learn-all-data upper and lower AAMs. The fitting result on one frame is used to initialize the model on the following frame. Results are normalized in shape as in §4.1.1. Figure 6 compares the tracking accuracy of specialized and single AAMs, whereas figures 4 and 5 shows some fitting results obtained with specialized AAMs and with the single AAMs. For the upper face, the single AAM is often more accurate when it starts close from the solution, confirming our previous observations. On the other hand, more robustness to sudden facial deformations and pose changes is observed for the specialized models. The single model is not able to track eye blinks and pose variations. This tasks is however well accomplished by the specialized models. For the lower face, the single AAM is unable to track the mouth and remains in its initial position along the whole video. The model expressivity is too high to be sensitive to the displacement of low gradient area such as the mouth. Indeed, it can express with low residual error any area surrounding the mouth. The specialized models manage to correctly fit all frames of the sequence.

5 Conclusion

We proposed a solution to the difficult problem of fitting AAMs on unseen faces, allowing variations in pose and expression. This is done by means of a pool of AAMs, each one being specialized on a particular pose and expression. These AAMs are fitted on the picture in a multiple fitting fashion. We tested our solution and compared it with the classical way that consists to train a single AAM on a large training set, including variability sources in pose, expression and identity. We showed that the inertia of such an AAM is so high that the model easily stays into its initial position. On the other hand, our pool of specialized AAMs present a good ability to discriminate the different pairs of pose and expression, and shows to accurately fit the face despite fast head motion and expression changes often present in a video. We plan to use the accurate fitting of our specialized models to automatically label all frames of a sequence, and train a person-specific AAM for this person. This AAM can be used to reliably track the face in real-time during the following sessions. In addition, the specialization of each AAM may allow us to flag the face pose and expression at each frame of the subsequent videos in real-time.

The main next difficulty we foresee is to make our solution invariant to illumination

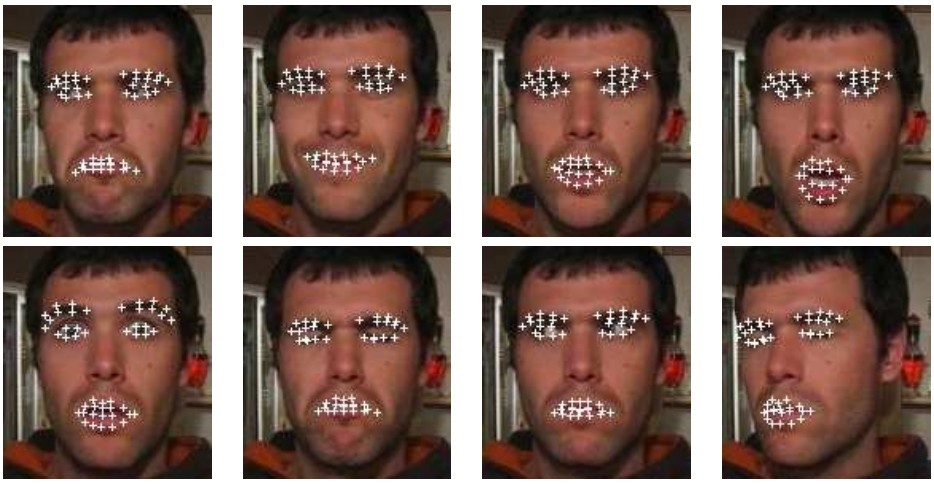

Figure 4: Fitting results obtained with the specialized models on some frames (number 22, 41, 45, 52, 58, 74, 78 and 106) extracted from a video showing identity 2. High fitting accuracy is often observed and difficult facial changes such as eye blinks and sudden pose variations are retrieved reliably. The classification of 5 *resp* 1 upper *resp* lower deformations is missed among 23 *resp* 11 deformations occuring over the complete sequence. An example of failure is showed on the bottom left frame where deformation n°7 is confused with deformation n°0.

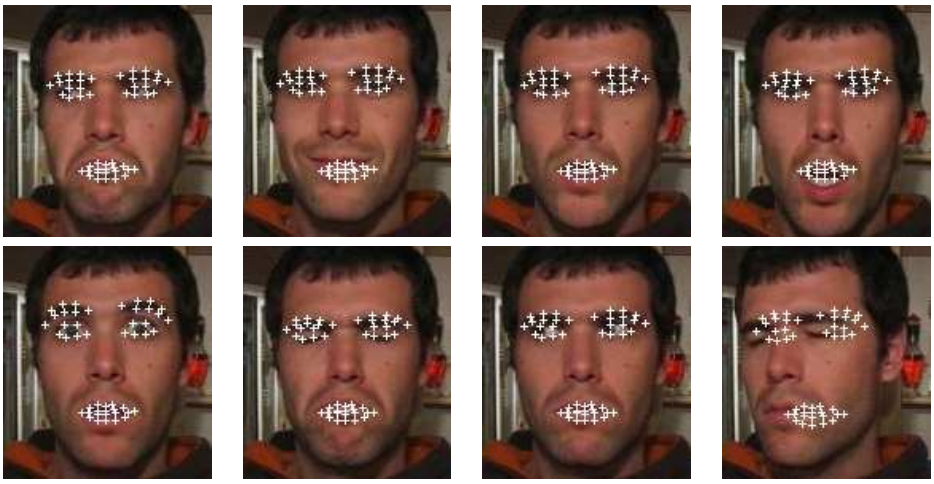

Figure 5: Fitting results of the single learn-all-data upper and lower AAMs showed for the same frames as in Figure 4. The inertia of these models are so high that they cannot handle large feature displacements. This is particularly true for the lower AAM because the mouth is likely to present small image gradients. The upper AAM shows to nicely track the raised eyebrows and provide good fitting accuracy on some of the frames.

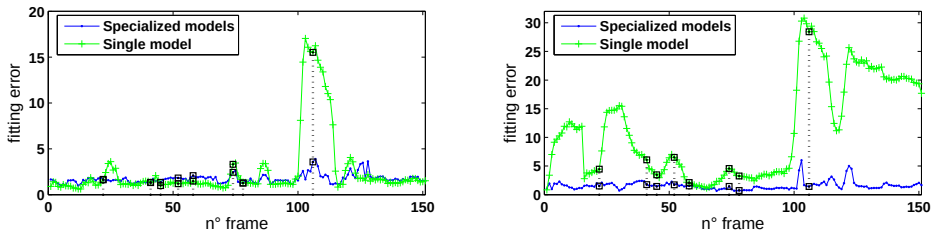

Figure 6: The fitting accuracy error along the frame sequence is compared for the single AAM and the specialized AAMs. Left and right show the results on upper and lower face respectively. Tracking results are accurate and reliable for the specialized models. The tracking can be even more accurate for the single AAM on the upper face, but it is also less reliable when sudden changes occur. On the lower face, the single model results completely unreliable. The fitting error of the frames shown on figures 4 and 5 are indicated by squares.

changes and occlusions. Real-life contexts require the applications to be robust to such changes.

References

[1] M.S. Bartlett, G. Littlewort, M.G. Frank, C. Lainscek, I. Fasel, and J. Movellan. Automatic Recognition of Facial Actions in Spontaneous Expressions. *Journal of Multimedia*, 2006.

[2] V. Blanz and T. Vetter. A Morphable Model For the Synthesis of 3D-faces. *SIGGRAPH*, 1999.

[3] T. Cootes, G. Edwards, and C.J. Taylor. Active Appearance Models. *ECCV*, 1998.

[4] T. Cootes, K. Walker, and C. Taylor. View Based Active Appearance Models. *AFGR*, 2000.

[5] D. Cristinacce and T. Cootes. Feature Detection and Tracking with Constrained Local Models. *BMVC*, 2006.

[6] P. Ekman and W. V. Friesen. *Facial Action Coding System (FACS) : Manual*. Palo Alto: Consulting Psychologists Press, 1978.

[7] R. Gross, I. Matthews, and S. Baker. Generic vs. Person Specific Active Appearance Models. *Image and Vision Computing*, 23(11):1080–1093, November 2005.

[8] S. Lucey, I. Matthews, C. Hu, Z. Ambadar, F De la Torre Frade, and J. Cohn. AAM Derived Face Representation for Robust Facial Action Recognition. *AFGR*, 2006.

[9] I. Matthews and S. Baker. Active Appearance Models Revisited. *IJCV*, 60(2):135–164, November 2004.

[10] J. Peyras, A. Bartoli, H. Mercier, and P. Dalle. Segmented AAMs Improve Person-Independent Face Fitting. *BMVC*, 2007.

[11] Y. Wang, S. Lucey, and J. Cohn. Enforcing Convexity for Improved Alignment with Constrained Local Models. *CVPR*, 2008.