Evaluating AAM Fitting Methods for Facial Expression Recognition

Akshay Asthana[1] Jason Saragih[2] Michael Wagner[3] Roland Goecke[1,3]
[1]RSISE, Australian National University, Australia
[2]Robotics Institute, Carnegie Mellon University, USA
[3]Faculty of Information Sciences and Engineering, University of Canberra, Australia
aasthana@rsise.anu.edu.au , jsaragih@andrew.cmu.edu , Michael.Wagner@canberra.edu.au , roland.goecke@ieee.org

Abstract

The human face is a rich source of information for the viewer and facial expressions are a major component in judging a person's affective state, intention and personality. Facial expressions are an important part of human-human interaction and have the potential to play an equally important part in human-computer interaction. This paper evaluates various Active Appearance Model (AAM) fitting methods, including both the original formulation as well as several state-of-the-art methods, for the task of automatic facial expression recognition. The AAM is a powerful statistical model for modelling and registering deformable objects. The results of the fitting process are used in a facial expression recognition task using a region-based intermediate representation related to Action Units, with the expression classification task realised using a Support Vector Machine. Experiments are performed for both person-dependent and person-independent setups. Overall, the best facial expression recognition results were obtained by using the Iterative Error Bound Minimisation method, which consistently resulted in accurate face model alignment and facial expression recognition even when the initial face detection used to initialise the fitting procedure was poor.

1. Introduction

Facial expressions are an important component of interpersonal communication. Despite their non-verbal nature, they convey a lot of information about the person and the person's affective state, intention and personality. Particularly for the recognition of the affective state, humans rely heavily on analysing facial expressions [10, 18]. Facial expressions also support verbal communication due to their complementary nature to the acoustic side of the spoken words. Unlike humans, current computer systems can hardly recognise the affective state of a human user. In fact, even the problem of recognising facial expressions is still largely unsolved although some progress has been made in recent years (Section 2). Providing human-machine interfaces with the capability of recognising facial expressions (and subsequently the affective state) of a user would allow computer systems to monitor a person's state and to react in a suitable way.

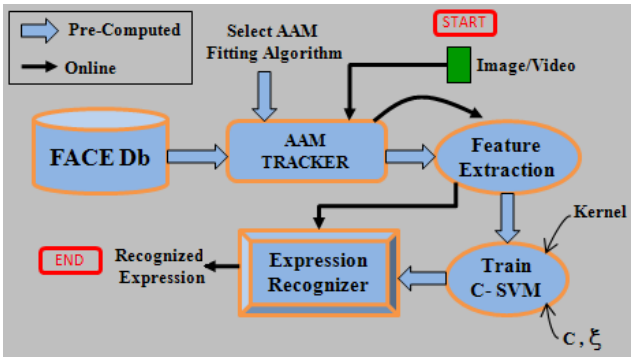

Figure 1: System Overview - Facial Expression Recogniser

While much progress has been made on the issue of classification of facial expressions, for example via Support Vector Machines (SVM) [13], one of the open questions is on the problem of how to extract useful features from the face in an image or video frame. In this paper, we compare the performance of six Active Appearance Model (AAM) fitting methods for the task of automatic facial expression recognition, which serves two purposes. Firstly, it gives the reader a practical guide to the usefulness of particular AAM fitting methods for facial expression recognition in realistic conditions. Secondly, it provides an (indirect) solution to the problem of objectively evaluating the performance of AAM fitting methods.[1] The methods were tested on facial expression images from the Cohn-Kanade database [14] in both a person-dependent (PDFER) and person-independent

[1]While it is possible and common practice to evaluate the performance via a manually obtained ground truth, the required manual annotation is in itself error-prone and subjective.

978-1-4244-4799-2/09/$25.00 ©2009 IEEE

(PIFER) setup. The system uses a real-time facial feature tracker based on the AAM to extract the shape vector in an image. This shape vector is further processed and a compact feature vector, representing the facial features, is obtained. This feature vector, along with the label of the expression associated with it, is used for training an expression classifier that utilises the SVM for classification into six universal facial expressions as well as a neutral expression. After the system has been trained for recognising a set of expressions, it accepts an image as input, followed by the same process of tracking facial features and extracting a feature vector. The SVM-based expression recogniser then uses this extracted feature vector to classify the expression as *Neutral* or one of six universal expressions (*Anger*, *Disgust*, *Fear*, *Joy*, *Sorrow*, *Surprise*). This procedure is illustrated in Figure 1. The remainder of this paper is structured as follows. Section 2 provides an overview of related work. The overview of AAM and various AAM fitting methods compared in this paper is given in Section 3. Our face region-based intermediate representation is presented in Section 4. The SVM classifier employed in FER experiments is described in Section 5. Section 6 details the experiments and discusses the results. Finally, Section 7 provides the conclusions and an outlook on future work.

2. Related Work

For many decades, developing a fast, accurate and robust automated system for recognising a face and facial expressions has been a goal in computer vision. In [11], the Facial Action Coding System (FACS) that defines the human face by a number of Action Units (AUs) and represents the facial expressions by different combinations of these AUs was proposed. Since the classification into AUs is based on facial anatomy, practically all expressions can be represented by this coding scheme. Hence, FACS is by far the most widely used method for facial expression recognition. However, one of the inherent difficulties with the FACS coding scheme is that it requires a highly trained human expert to manually score each frame of a video. As well as being an extremely tedious process, manual FACS scoring suffers from inconsistencies between scorers. In [8], a comprehensive comparative study of various approaches for an automatic facial action recognition system was presented, where techniques such as optical flow analysis, local feature analysis, Gabor wavelets, principal component analysis (PCA), linear discriminant analysis (LDA), and independent component analysis (ICA) were employed. More recently in [3], various machine learning techniques were compared, coupled with appearance based features for facial expression and action recognition. However, one of the major drawbacks for all these is that they ignore the spatial arrangement and motion of the anatomical features, such as eyes, mouth, eyebrows and chin. As a result, these methods are highly susceptible to changes in pose, illumination and other sources of variation regularly encountered in a real world environment [16].

In recent years, a powerful technique based on deformable models has become popular for non-rigid object tracking and has started to make its way into the field of real-time face and facial expression recognition. In this deformable model based approach, the non-rigid shape and visual texture (intensity and color) of an object are statistically modelled using a low dimensional representation obtained by applying PCA to a set of labelled training data. After these models have been created, they can be parametrised to fit a new image of the object, which might vary in shape or texture or both. One of the deformable model based approaches, known as the Active Appearance Model [9], has become very popular for tracking non-rigid objects such as the human face.

The utility of AAM tracking in the context of real-time analysis of facial expressions has previously been demonstrated in a number of works. In [19], the authors present an approach for gender-based expression recognition based on AAM tracking followed by the classification via SVM into 4 basic expressions (happy, sad, angry and neutral). The experiments were performed on still images and a maximum accuracy of 79.9% for gender based expression classification and 76.4% for gender independent expression classification was reported. In [16], the authors compare 3 different feature representations and subsequently utilise SVM for the classification of different expressions and AUs. The authors also state that the Nearest Neighbor (NN) classifier based on PCA or LDA can be used, although no improvement in the performance was reported when using NN instead of SVM. The three types of evaluated features were S-PTS (similarity normalised shape), S-APP (similarity normalised appearance) and C-APP (canonical appearance). (S-PTS)+(C-APP) features performed better than S-PTS and S-APP. (S-PTS)+(C-APP) features are obtained by concatenating the similarity normalised shape and the shape normalised (canonical) appearance. That work validates the assertion that features based on AAMs can be used for accurate expression recognition. However, the authors used a person-dependent AAM tracker to extract the feature vectors for the experiment and also reported a major difficulty in tackling the problem of subject head movement. In [7], a real-time approach for expression recognition in video by utilising AAM tracking and spectral graph clustering was presented. However, the tracking was limited to the mouth region only. In contrast, a template-based facial feature tracker was used in [17], followed by a SVM-based expression classification. An accuracy of 71.8% for person-independent expression recognition and 87.5% for person-dependent expression recognition was reported. In other work, different intermediate representations of AAM

tracked shape and appearance vectors for training the expression classifiers have been investigated (see [6], for example) and the application of rough set theory for AAM based expression recognition has also been pursued [4].

3. Active Appearance Model (AAM)

The AAM is a powerful generative class of methods for modelling and registering deformable visual objects which has been very popular in recent years due to its excellent performance. The power of this generative model stems both from its compact representation of appearance (comprising shape and texture) as well as its rapid fitting to unseen images.

For constructing the AAM [9], each annotated training image is aligned into a common coordinate frame by Procrustes analysis. The modes of shape variation are obtained by applying PCA to the set of aligned shapes. The texture variation is similarly modelled by applying PCA to a set of images, warped to a canonical frame defined using the mean shape of the aligned shapes. As a result, a parametrised model is formed that is capable of representing large variation in shape and texture by a small set of parameters.

The process of finding the model parameters $\mathbf{p}$ that best fit the given image $\mathcal{I}$ is known as AAM fitting and is performed by updating the model parameters $\mathbf{p}$ iteratively: [2]

$$\Delta p = \mathcal{U}(\diamond; \mathbf{p}) \circ \mathcal{F}(\mathcal{I}; \mathbf{p}) \quad \text{where} \quad \mathbf{p} \leftarrow \mathbf{p} + \Delta\mathbf{p} \tag{1}$$

where, $\mathcal{F}$ is a feature extraction function that represents image $\mathcal{I}$ at its current parameter settings, $\Delta\mathbf{p}$ are the updates to be applied to the current parameters and $\mathcal{U}$ is the vector valued update function. The accuracy of prediction for updating the parameter $\mathbf{p}$ generally depends on a good coupling between $\mathcal{F}$ and $\mathcal{U}$. The AAM fitting algorithms can be broadly classified into two categories [21]:

Generative fitting deals with the problem of fitting as minimisation/maximisation of some measure of fitness between the model's texture and the warped image region. This approach is attractive as it has a clear intuitive basis for its formulation. However, it suffers from a number of drawbacks, such as limited generalisability as well as difficulties in attaining rapid fitting. Some examples of generative fitting, compared in the paper, are FJ, POIC, SIC and RIC (see below).

Discriminative Fitting deals with the problem of fitting by directly learning a fixed relationship between the features and the parameter updates, by using the features extracted from parameter settings which are perturbed from their optimal setting in each training image. Although this approach lacks the elegance of the generative approach, it has been shown to overcome some of the limitations of its generative counterpart. Some discriminative AAM methods, compared in the paper, are IEBM and HFBID (see below). Other methods also exist, e.g. [15].

[2] **Notation:** Vectors are written in lowercase bold. Functions are written in upper case calligraphic font with $\circ$ denoting their composition, for example: $\mathcal{A}(\mathcal{B}(x); y) = \mathcal{A}(\diamondsuit; y) \circ \mathcal{B}(x)$.

3.1. Fixed Jacobian Method (FJ)

Proposed in [5], it is one of the original algorithms developed for AAM fitting and deals with the problem of fitting as the minimisation of the least squares fit between the model's texture and the warped image region, where it is assumed that the Jacobian of the error is fixed for all settings of the model parameters. This enables a linear update model to be pre-computed through the pseudo-inverse of the fixed Jacobian, estimated offline through numerical differentiation, averaging over the training set. Since the assumption of a fixed Jacobian holds only loosely, the method requires the use of an adjustable step size, where at each iteration the predicted parameter updates are halved until a reduction in the texture difference between the model and the cropped image is attained. The result is a reasonably efficient and accurate fitting procedure. However, if the object exhibits large variation in shape and texture, its performance deteriorates because of the assumption of a fixed linear update model which can be too restrictive.

3.2. Project-out Inverse Compositional Method (POIC)

Proposed in [2], it is one of the fastest AAM fitting algorithms to date and belongs to the class of fitting methods using the inverse compositional approach, where the roles of image and model in the error function are reversed. In this adaptation of inverse-compositional image alignment, the fitness function, which measures the difference between the model's appearance and the cropped image region, is grouped into two components: one which lies within the subspace of appearance deformations and another which is orthogonal to it. This procedure requires optimisation over the shape parameters only, assuming the optimal choice (in a maximum likelihood sense) of the texture parameters is chosen at each iteration. Since the minimisation of the fitness function depends only on the subspace orthogonal to the texture variation, a fixed linear update model can be computed analytically over the shape parameters only. This better justifies the assumption of a linear update model as compared to FJ and is also extremely fast. However, this approach does not work well when the object exhibits large variation in shape and texture with respect to the mean shape and texture, limiting its usage to relatively simple applications (e.g. PDFER) [12].

3.3. Simultaneous Inverse Compositional Method (SIC)

Proposed in [1], it is an another adaptation of inverse compositional image alignment for AAM fitting that addresses the problem of the significant shape and texture variability by finding the optimal shape and texture parameters simultaneously. Although the derivative of the warping function can be pre-computed, the linear update model has to be recomputed at each iteration as it depends on the current appearance parameters. However, rather than recomputing the linear update model at every iteration using the current estimate of appearance parameters, it can be approximated by evaluating it at the mean appearance parameter values, allowing the update model to be pre-computed, thus dramatically improving the computational efficiency. Also, since the work presented in this paper deals with the problem of AAM fitting, building the linear update model based on the mean appearance parameters, which on average are closer to the true parameter values, is an optimal choice.

3.4. Robust Inverse Compositional Method (RIC)

In [1], the idea of the inverse compositional method for AAM fitting is extended further by using an M-estimator (robust penaliser) instead of the least squares fitting criterion, resulting in an iteratively reweighed least squares fitting scheme. This method requires the normalisation of the mean subtracted cropped image (error image) w.r.t. the direction of appearance variability [1]. For this purpose, the error image is first projected onto the subspace of appearance variability. This projected error image is used for generating the model's appearance that is later subtracted from the error image to get the measure of the fitness function. An assumption of spatial coherence of the outliers helps in reducing the computational complexity to a certain extent, but it is still slower than the efficient approximation of SIC.

3.5. Iterative Error Bound Minimisation Methods (IEBM)

A novel linear update scheme is proposed in [20], which is based on reducing the error bounds over the data, rather than the typical least squares criterion. It uses the optimality property of Support Vector Regression (SVR), i.e. each sample is adjusted to achieve its respective parameter setting where the error is minimised, giving priority to those samples that produce maximum error. Combined with an iterative scheme, all samples in the training set are guided towards their solution, placing a higher priority on samples with large errors. IEBM focuses on building the update model by utilising the information from various combinations of parameter settings. Since this update model is learnt offline, the method is extremely efficient.

Action Units	**Regional Description**	**Feature IDs**
AU1, AU2, AU4.	Intra-feature movement between Eyebrows, Inter-feature movement between Eyebrow and Eyes	V1, V2, V3, H1
AU5, AU7, AU41, AU42, AU43, AU44, AU45, AU46.	Intra-feature movement of the Eyes	V6, V7, H2, H3
AU6	Intra-feature movement of the Cheeks	H8
AU9	Intra-feature movement of the Nose	V8, H4
AU10	Inter-feature movement between Nose and Mouth	V13
AU12, AU15, AU16, AU18, AU20, AU22, AU23, AU24, AU25, AU28.	Intra-feature movement of the Mouth	V9, V10, V11, V12, H5, H6, H7
AU27	Inter-feature movement between Nose and Mouth	V5
AU17, AU26	Inter-feature movement between Nose and Chin	V4

Table 1: Features Description

3.6. Haar-like Feature Based Iterative-Discriminative Method (HFBID)

The IEBM method was extended further in [21] by using a nonlinear update model for AAM fitting that uses multi-modal weak learners, based on Haar-like features, which allow efficient online evaluation using the integral image. To avoid overlearning, the boosting procedure is embedded into an iterative framework with an intermediate resampling step. This process affords well regularised update models through limiting the ensemble size and indirectly increasing the sample size. As with IEBM, this method was shown to exhibit high fitting speed and accuracy. However, its performance was not compared with IEBM. Implementation details are described in [21].

4. Interpreting Action Units (AUs)

The design of AUs is anatomically motivated and makes FACS highly adaptable and capable of representing almost every expression [11]. However, since the scope of most of the vision based expression recognition systems is based on changes in appearance, it might not be possible to extract the information needed to indicate the activation of AUs based on anatomical facts. For example, it is the contraction of the *corrugator* muscle that activates AU4 and the contraction of the *frontalis* muscle that activates AU1 [22]. To overcome this problem, we group the AUs together on a regional basis (Table 1) and use a set of rules to extract the information within the region, regarding the appearance changes, that can be utilised for expression recognition (Figure 2). V_η and H_η are the normalisation distances used to normalise the feature vector with respect to the varying

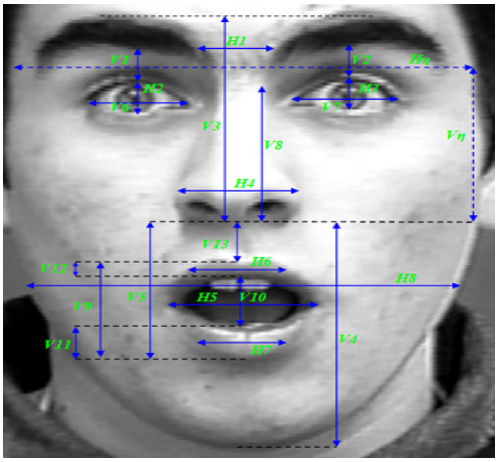

Figure 2: Features for regional scheme

size of faces of different people. The advantage of using such a scheme is that we no longer require information on individual muscles that activate an AU - for example AU25, which is activated by 3 different muscles [11] -, but that we nevertheless extract the necessary information for the appearance changes from the region concerned.

5. Classification of Facial Expressions

Inherently, the support vector machine (SVM) is a two-class problem classifier. Since our implementation of FER is a 7-class problem (Neutral and 6 basic emotions [10]), the 'one-against-one' method [13] is adopted to construct a multi-class SVM to handle this problem. Another issue concerned with designing an SVM based classifier is the choice of the kernel. We used the Linear kernel for PDFER and Radial Basis Function (RBF) kernel for PIFER. The problems of PDFER and PIFER differ in the level of complexity of the data to be classified. For PDFER, a single classifier needs to be trained on the features of a single person, which makes it a much simpler classification problem. The mapping of data in a plane using a linear kernel will suffice, whereas for PIFER, the single classifier needs to be trained on the features extracted from all the diverse people in the database. Hence, the RBF kernel is used for PIFER, as it has the capacity to map the features in the higher dimension and provide better classification for this complex data.

Since we treat PDFER and PIFER as separate problems, the use of different kernels does not affect the performance evaluation for various AAM fitting methods. The important point to note here is that a linear kernel is used to evaluate the performance of all fitting methods for PDFER, and a RBF kernel is used to evaluate the performance of all fitting methods for PIFER. Hence, the consistency in the evaluation process is maintained as far as the classification of expressions is concerned.

6. Experiments

The proposed PDFER and PIFER have been successfully implemented using each of the fitting algorithms discussed in Section 3. Previously in [12], it was shown that POIC does not work well for objects varying significantly in shape and texture, i.e. it is only suited for tracking of simpler objects. Hence, POIC was excluded from our person independent experiments (PIFER), where adaptability to changes in size, shape and texture is the essence of the problem. The linear regressor of IEBM and non-linear regressor of HFBID were trained with the initial bounds of $\pm 10^o$, ± 0.1, ± 20 pixels, and ± 1.5 standard deviations of rotation, scale, translation and non-rigid shape parameters, respectively. Both PDFER and PIFER can be used to recognise *any arbitrary set of expressions*, but are evaluated here in a 7-class setup.

Our experimental dataset contained 3424 images of 30 subjects (15 females / 15 males) chosen randomly from the Cohn-Kanade Database (CKDb) [14], with each speaker expressing 6 basic expressions starting from a Neutral expression. Overall, the dataset contained 992 images for *Neutral*, 448 images for *Anger*, 296 images for *Disgust*, 346 images for *Fear*, 532 images for *Joy*, 423 images for *Sorrow* and 387 images for *Surprise*. Further, the arbitrary selection of speakers from CKDb ensured that the diversity of the database with respect to the gender of the speaker and shape, size and texture of the faces was maintained.

The experiments, for the results presented in this paper, have been independently conducted for a person dependent scenario (PDFER) and for a person independent scenario (PIFER). For PDFER, 30 real-time AAM trackers, one for each subject, were trained separately, whereas, for PIFER, a single real-time AAM tracker was used that was trained to track the facial features across the 30 speakers in the database. It should be noted here that 30 images per person were used to train the AAM trackers. The shape vector of length 138, representing 69 landmark points tracked by the AAM was further processed using the strategy presented in Section 4 and a feature vector of length 21 was extracted using the scheme given in Figure 2. This feature vector is used throughout our experiments for expression recognition. A 5-fold cross-validation scheme is used to evaluate the performance and utility of each fitting algorithm for PDFER and PIFER.

In the experiments, the AAM parameters for each image in the database were perturbed for initialisation to simulate the misalignments, in the initialisation, that might be encountered by the use of any generic face detector under real world conditions. This testing strategy helps us to test the robustness of the system based on each of the fitting algorithm. For PDFER, the AAM parameters were per-

Person Dependent FER									
Fitting	**Init.**	**Neutral**	**Anger**	**Disgust**	**Fear**	**Joy**	**Sorrow**	**Surprise**	**Overall**
HFBID	**+/- 5**	94.96	94.42	95.95	96.82	94.17	95.27	94.32	**95.01**
	+/- 10	94.05	93.30	94.93	93.93	94.17	93.62	92.51	**93.81**
	+/- 20	91.03	91.07	92.57	92.77	92.48	93.85	91.21	**91.94**
	+/- 30	90.12	80.80	89.53	92.20	90.79	80.85	90.18	**88.03**
IEBM	**+/- 5**	95.56	95.76	97.97	97.11	95.30	95.51	94.06	**95.74**
	+/- 10	96.17	95.54	97.64	97.69	94.92	95.98	93.80	**95.88**
	+/- 20	89.31	95.54	97.64	96.82	93.98	96.69	93.02	**93.66**
	+/- 30	92.84	95.09	97.64	96.82	92.67	95.04	92.76	**94.19**
FJ	**+/- 5**	96.77	93.08	93.92	94.51	94.17	98.35	93.02	**95.18**
	+/- 10	90.93	85.27	88.18	88.15	91.73	89.60	90.70	**89.60**
	+/- 20	76.11	71.21	75.34	71.97	77.82	85.58	74.68	**76.26**
POIC	**+/- 5**	96.07	93.97	93.92	95.09	93.05	96.93	92.25	**94.71**
	+/- 10	91.13	84.38	95.27	87.28	85.90	72.34	86.82	**86.59**
	+/- 20	74.70	68.75	67.91	78.03	71.43	60.76	70.03	**70.91**
SIC	**+/- 5**	96.07	95.09	95.95	93.93	93.61	98.35	93.02	**95.27**
	+/- 10	83.57	89.06	88.51	87.57	93.42	76.36	75.97	**84.90**
	+/- 20	74.29	78.35	61.15	78.32	73.87	78.96	65.63	**73.63**
RIC	**+/- 5**	94.46	91.52	93.24	95.95	93.80	95.98	90.96	**93.81**
	+/- 10	95.16	93.08	93.24	94.80	92.67	85.34	87.86	**92.26**
	+/- 20	76.11	75.22	75.68	77.75	81.02	64.30	65.89	**74.27**

Person Independent FER									
Fitting	**Init.**	**Neutral**	**Anger**	**Disgust**	**Fear**	**Joy**	**Sorrow**	**Surprise**	**Overall**
HFBID	**+/- 5**	97.98	89.96	93.92	94.22	91.17	94.09	94.57	**94.28**
	+/- 10	95.67	88.84	93.24	93.64	89.10	90.31	92.25	**92.29**
	+/- 20	96.07	88.62	90.20	89.02	88.35	88.89	90.18	**91.12**
	+/- 25	87.10	78.57	85.81	89.60	85.34	74.00	89.66	**84.52**
IEBM	**+/- 5**	94.46	94.64	87.84	93.64	88.53	96.69	92.76	**92.99**
	+/- 10	93.15	91.96	88.85	92.77	87.59	95.27	91.99	**91.85**
	+/- 20	92.54	88.62	87.50	91.04	87.97	92.91	91.21	**90.63**
	+/- 25	91.23	88.62	86.82	90.46	87.78	90.07	89.92	**89.60**
FJ	**+/- 5**	87.30	83.48	57.09	69.08	83.27	28.37	89.92	**74.74**
	+/- 10	74.40	72.32	52.03	70.23	75.00	31.44	84.50	**67.70**
	+/- 20	49.29	61.83	32.77	59.25	47.37	41.84	61.50	**50.67**
SIC	**+/- 5**	66.83	39.29	37.84	49.42	42.48	19.39	41.86	**46.50**
	+/- 10	57.86	37.72	28.04	32.37	40.79	24.35	42.38	**41.53**
RIC	**+/- 5**	83.47	66.74	43.58	76.88	65.79	27.42	72.09	**66.21**
	+/- 10	72.48	59.15	41.22	70.81	52.26	17.73	65.63	**57.18**

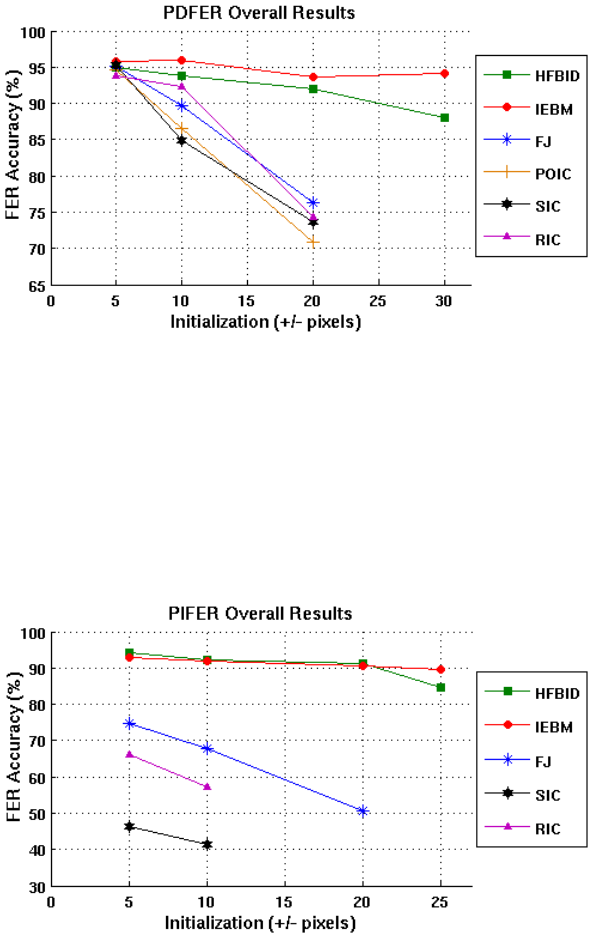

Figure 3: Overall Results for FER

turbed to ± 5, ± 10, ± 20 and ± 30 pixels for initialisation from their optimal settings, whereas for PIFER, the AAM parameters were perturbed to ± 5, ± 10, ± 20 and ± 25 pixels for initialisation from their optimal settings. Overall results for PDFER and PIFER in terms of percentage accuracy are given in Figure 3.

For PDFER, all fitting algorithms perform well for ± 5 pixels perturbation for initialisation with IEBM exhibiting the highest accuracy at 95.74% and RIC the poorest at 93.81%, whereas for PIFER, HFBID and IEBM perform well for ± 5 pixels perturbations for initialisation with HFBID showing the highest accuracy at 94.28%, while FJ, SIC, and RIC struggle to adapt to the person-independent scenario. It is rare in real world applications to get such an accurate initialisation, but nevertheless, these results can act as a good indicator of the fact that AAM tracking based expression recognisers can give accurate results, provided a good fitting is achieved by virtue of accurate initialisation. As the perturbation is increased to ± 10 pixels for initialisation, a slight dip in accuracy for FJ, POIC and SIC is observed, however, HFBID, IEBM and RIC are able to maintain almost the same accuracy for PDFER, whereas for PIFER, the accuracy of FJ, SIC and RIC deteriorates even further. In contrast, HFBID and IEBM are able to maintain almost the same accuracy.[3]

As the perturbation increases to ± 20 pixels for initialisation, for PDFER, a significant decrease in accuracy for FJ, POIC, SIC and RIC is observed, while HFBID and IEBM show an impressive accuracy of 91.94% and 93.66%, respectively. For PIFER, SIC and RIC are unable to converge

[3]It should be noted here that while working on a real world problem such as Facial Expression Recognition, getting an absolute result is highly improbable (if not impossible). There are many expressions that cannot be categorised absolutely into any of the 6 basic expressions, for example, say the transition images between a neutral and a smiling expression, and are prone to be misclassified. Hence, in this paper, a small variation of 1% in the accuracy of the expression recognition is considered insignificant.

and a significant fall in the accuracy of FJ is observed, while HFBID and IEBM show an impressive accuracy of 91.12% and 90.63%, respectively. These results clearly show the efficacy of HFBID and IEBM to maintain high accuracy for both PDFER and PIFER consistently, even when initialisation is poor, which is quite often the case when a generic face detector is used in a real world environment. Going a step further for PDFER, on increasing the perturbation to ± 30 pixels for initialisation, FJ, POIC, SIC and RIC are unable to converge, whereas, HFBID and IEBM still maintain high accuracy with IEBM performing better at 94.19% accuracy compared to the HFBID's which achieves 88.03% accuracy. In comparison for PIFER, on increasing the perturbation to ± 25 pixels for initialisation, FJ, SIC, and RIC are unable to converge, however, both HFBID and IEBM still maintain high accuracy. Here, IEBM once again performs reasonably better, with 89.60% accuracy compared to HFBID's accuracy of 84.52%, as it did for PDFER.

It is worth noting here that the regressors for HFBID and IEBM were trained in the initial bound of ± 20 pixels for translation, so when the testing is done at perturbations larger than ± 20 pixels for initialisation, i.e. ± 30 pixels in case of PDFER and ± 25 pixels in case of PIFER, the extrapolation capacity of the regressor is being tested instead of the interpolation capacity which was previously tested for perturbations of ± 5, ± 10 and ± 20 pixels. Although nonlinear regressors of HFBID can potentially provide more accurate predictions than their linear counterpart used in IEBM, their training procedure is generally more complicated. In the case of HFBID, only a local solution for the predictor can be attained due to the greedy learning properties of the boosting procedure used to train the method. Furthermore, a true implementation of the boosting procedure that requires the evaluation of all possible features before appending one to the ensemble is not computationally feasible, due to the large number of possible Haar-like features. Due to these training complexities, in practice, the simpler linear models can be expected to extrapolate better and show higher accuracy. This result shows that both HFBID and IEBM are capable of handling extremely poor initialisation and, hence, are better suited for both PDFER and PIFER systems, with IEBM slightly outperforming HFBID under extreme conditions. The confusion matrices for IEBM are given in Table 2. Refer to the *supplementary material* [4] for details.

7. Conclusion and Future Work

The aim of the work presented in this paper has been to investigate the utility of different AAM fitting algorithms in the context of real-time facial expression recognition. From the results, it is clear that Iterative-Discriminative (ID) approach adopted in IEBM and HFBID boosts the fitting performance significantly and affords an extremely rapid fitting time (even more so than POIC that is commonly considered the fastest AAM fitting method) by virtue of the predictive capacity of discriminative method and an iterative framework of generative fitting. For the simpler problem of PDFER, generative fitting algorithms (FJ, POIC, SIC and RIC) compete with their ID counterparts (IEBM and HFBID). However, under extremely poor initialisation, all generative fitting algorithms fail to converge, while ID algorithms are able to perform with almost the same accuracy. For the relatively difficult problem of PIFER, ID algorithms come into their own, outperforming generative fitting algorithms in all cases, even under extremely poor initialisation. Under extreme conditions, IEBM marginally outperforms HFBID by taking advantage of its linear regressor, whose predictive domain is much simpler than that of the non-linear regressor used by HFBID. A future direction for this research is to take advantage of accurate ID fitting algorithms to tackle the problem of pose-invariant expression recognition. Also, the ability to adapt to unseen data for PIFER based on ID fitting algorithms will be evaluated.

[4] http://users.rsise.anu.edu.au/~aasthana/ACII09/ACII09DivX.avi
http://users.rsise.anu.edu.au/~aasthana/ACII09/ACII09Supp.pdf

References

[1] S. Baker, R. Gross, and I. Matthews. Lucas-Kanade 20 years on: A unifying framework: Part 3. Technical report, RI, Carnegie Mellon University, USA, Nov. 2003.

[2] S. Baker and I. Matthews. Equivalence and Efficiency of Image Alignment Algorithms. In *Proc. CVPR 2001*, volume 1, pages 1090–1097, 2001.

[3] M. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior. In *Proc. CVPR 2005*, pages 568–573, June 2005.

[4] P. Chen, G. Wang, Y. Yang, and J. Zhou. Facial Expression Recognition Based on Rough Set Theory and SVM. In *RSKT 2006*, pages 772–777, 2006.

[5] T. Cootes, G. Edwards, and C. Taylor. Active Appearance Models. In *Proc. ECCV'98*, pages 484–498, June 1998.

[6] D. Datcu and L. Rothkrantz. Facial Expression Recognition in still pictures and videos using Active Appearance Models. A comparison approach. In *CompSysTech'07*, June 2007.

[7] F. De la Torre Frade, J. Campoy, Z. Ambadar, and J. Cohn. Temporal Segmentation of Facial Behavior. In *Proc. ICCV 2007*, Oct. 2007.

[8] G. Donato, M. Bartlett, J. Hager, P. Ekman, and T. J. Sejnowski. Classifying Facial Actions. *IEEE TPAMI*, 21(10):974–989, Oct. 1999.

[9] G. Edwards, C. Taylor, and T. Cootes. Interpreting Face Images Using Active Appearance Models. In *Proc. FG'98*, pages 300–305, 1998.

[10] P. Ekman. Basic emotions. In T. Dalgleish and T. Power, editors, *The Handbook of Cognition and Emotion*, pages 45–60. John Wiley & Sons, 1999.

Person-Dependent FER										Person-Independent FER								
±5 **Pixel Initialisation**										±5 **Pixel Initialisation**								
Exp.	**Neutral**	**Anger**	**Disgust**	**Fear**	**Joy**	**Sorrow**	**Surprise**	**Overall**	**%**	**Neutral**	**Anger**	**Disgust**	**Fear**	**Joy**	**Sorrow**	**Surprise**	**Overall**	**%**
Neutral	948	3	2	4	5	23	7	948/992	95.56	937	4	16	9	13	5	8	937/992	94.46
Anger	14	429	2	3	0	0	0	429/448	95.76	21	424	2	0	0	1	0	424/448	94.64
Disgust	5	0	290	1	0	0	0	290/296	97.97	30	2	260	0	4	0	0	260/296	87.84
Fear	9	0	1	336	0	0	0	336/346	97.11	17	0	0	324	2	2	1	324/346	93.64
Joy	19	0	1	2	507	1	2	507/532	95.30	41	0	7	8	471	3	2	471/532	88.53
Sorrow	16	0	1	2	0	404	0	404/423	95.51	9	2	0	1	2	409	0	409/423	96.69
Surprise	7	0	0	13	3	0	364	364/387	94.06	14	0	4	3	7	0	359	359/387	92.76
								3278/3424	95.74								3184/3424	92.99
±10 **Pixel Initialisation**										±10 **Pixel Initialisation**								
Exp.	**Neutral**	**Anger**	**Disgust**	**Fear**	**Joy**	**Sorrow**	**Surprise**	**Overall**	**%**	**Neutral**	**Anger**	**Disgust**	**Fear**	**Joy**	**Sorrow**	**Surprise**	**Overall**	**%**
Neutral	954	1	2	8	9	17	2	954/992	96.17	924	9	12	10	9	16	12	924/992	93.15
Anger	20	428	0	0	0	0	0	428/448	95.54	22	412	7	4	0	4	0	412/448	91.96
Disgust	5	0	289	2	0	0	0	289/296	97.64	29	2	263	0	2	0	0	263/296	88.85
Fear	5	0	0	338	3	0	0	338/346	97.69	14	0	0	321	4	0	7	321/346	92.77
Joy	21	0	0	6	505	0	0	505/532	94.92	41	2	0	13	466	8	2	466/532	87.59
Sorrow	17	0	0	0	0	406	0	406/423	95.98	13	2	0	3	2	403	0	403/423	95.27
Surprise	8	0	0	13	3	0	363	363/387	93.80	16	0	3	4	8	0	356	356/387	91.99
								3283/3424	95.88								3145/3424	91.85
±20 **Pixel Initialisation**										±20 **Pixel Initialisation**								
Exp.	**Neutral**	**Anger**	**Disgust**	**Fear**	**Joy**	**Sorrow**	**Surprise**	**Overall**	**%**	**Neutral**	**Anger**	**Disgust**	**Fear**	**Joy**	**Sorrow**	**Surprise**	**Overall**	**%**
Neutral	886	8	9	2	4	22	61	886/992	89.31	918	7	16	18	14	9	10	918/992	92.54
Anger	18	428	0	2	0	0	0	428/448	95.54	32	397	4	3	5	4	3	397/448	88.62
Disgust	7	0	289	0	0	0	0	289/296	97.64	29	3	259	5	0	0	0	259/296	87.50
Fear	6	0	0	335	0	0	5	335/346	96.82	18	0	2	315	4	2	5	315/346	91.04
Joy	21	5	0	0	500	0	6	500/532	93.98	41	0	2	14	468	7	0	468/532	87.97
Sorrow	14	0	0	0	0	409	0	409/423	96.69	21	2	0	4	3	393	0	393/423	92.91
Surprise	13	0	0	11	3	0	360	360/387	93.02	13	0	2	6	9	4	353	353/387	91.21
								3207/3424	93.66								3103/3424	90.63
±30 **Pixel Initialisation**										±25 **Pixel Initialisation**								
Exp.	**Neutral**	**Anger**	**Disgust**	**Fear**	**Joy**	**Sorrow**	**Surprise**	**Overall**	**%**	**Neutral**	**Anger**	**Disgust**	**Fear**	**Joy**	**Sorrow**	**Surprise**	**Overall**	**%**
Neutral	921	4	3	2	7	23	32	921/992	92.84	905	14	12	13	17	18	13	905/992	91.23
Anger	22	426	0	0	0	0	0	426/448	95.09	31	397	13	0	0	7	0	397/448	88.62
Disgust	7	0	289	0	0	0	0	289/296	97.64	23	0	257	11	0	5	0	257/296	86.82
Fear	4	0	0	335	7	0	0	335/346	96.82	23	0	0	313	3	2	5	313/346	90.46
Joy	21	0	2	7	493	0	9	493/532	92.67	39	0	4	13	467	6	3	467/532	87.78
Sorrow	18	0	0	3	0	402	0	402/423	95.04	13	5	17	0	0	381	7	381/423	90.07
Surprise	10	0	0	14	4	0	359	359/387	92.76	13	2	18	2	4	0	348	348/387	89.92
								3225/3424	94.19								3068/3424	89.60

Table 2: Confusion Matrices for IEBM: Left, person-dependent FER. Right, person-independent FER.

[11] P. Ekman, W. Friesen, and J. Hager. The Facial action coding system, Research Nexus eBook, Salt Lake City, UT. 2002.

[12] R. Gross, I. Matthews, and S. Baker. Generic vs. person specific active appearance models. *Image and Vision Computing*, 23(11):1080–1093, 2005.

[13] C. Hsu, C. Chang, and C. Lin. A practical guide to support vector classification. Technical report, Department of Computer Science, National Taiwan University, 2005.

[14] T. Kanade, J. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *Proc. IEEE FG'00*, 2000.

[15] X. Liu. Generic face alignment using boosted appearance model. In *Proc. CVPR 2007*, pages 1079–1088, 2007.

[16] S. Lucey, A. Ashraf, and J. Cohn. Investigating Spontaneous Facial Action Recognition through AAM Representations of the Face. In *Face Recognition Book*. Pro Literatur Verlag, Apr. 2007.

[17] P. Michel and R. El Kaliouby. Real Time Facial Expression Recognition in Video using Support Vector Machines. In *Proc. ICMI 2003*, pages 258–264, Nov. 2003.

[18] R. Picard. *Affective Computing*. MIT Press, Cambridge (MA), USA, 1997.

[19] Y. Saatci and C. Town. Cascaded Classification of Gender and Facial Expression using Active Appearance Models. In *Proc. FG'06*, pages 393–398, 2006.

[20] J. Saragih and R. Goecke. Iterative Error Bound Minimisation for AAM Alignment. In *Proc. ICPR'06*, pages 1192–1195, Aug. 2006.

[21] J. Saragih and R. Goecke. A Nonlinear Discriminative Approach to AAM Fitting. In *Proc. ICCV 2007*, Oct. 2007.

[22] Y.-L. Tian, T. Kanade, and J. Cohn. Recognizing Action Units for Facial Expression Analysis. Technical report, RI, Carnegie Mellon University, USA, Dec. 1999.