AN EXTENDED MULTIRESOLUTION APPROACH TO MOUTH SPECIFIC AAM FITTING FOR SPEECH RECOGNITION

Craig Berry, Anil Kokaram and Naomi Harte

Department of Electronic and Electrical Engineering,
Trinity College, Ireland.
email:cberry@tcd.ie web: www.sigmedia.tv

ABSTRACT

Active Appearance Models (AAMs) are a widely used technique for face tracking. They work by minimising the difference between an unobserved image and a synthetically generated image created by a statistical model of the deformable object, e.g. a face. The Fixed Jacobian algorithm is the most widely used algorithm for fitting AAMs. A Gaussian Image Pyramid fitting technique is used in order to make this process more robust and computationally faster. This paper presents a new image pyramid fitting structure specifically developed for application to an Audio-Visual Speach Recognition (AVSR) system in which the area described by the AAM is reduced as the iterations progress. This allows the fitting technique to be more accurate as the fitting progresses. The new fitting structure is implemented with the Fixed Jacobian algorithm and then compared to a standard approach where the mouth shape is extracted from a full face AAM. The test is performed using images from the CUAVE database. The new structure is shown to be more accurate and robust than the full face approach, with a 14.10% increase in the convergence of the mouth points to within a 4 pixel average difference, while also achieving a 8.53% improvement in the accuracy of the fit.

1. INTRODUCTION

Humans understand speech through both audio and visual cues. Audio-Visual Speech Recognition (AVSR) aims to exploit the bimodal nature of speech in automatic speech recognition systems to increase noise robustness. The extraction of mouth features, be it in terms of shape or appearance, is of primary importance in AVSR. The Active Appearance Model (AAM), proposed by Cootes et al[1], is a popular algorithm used to model deformable objects, e.g. faces.

The goal of an AVSR system is to extract visemic information from the video stream. A viseme is a unit that represents a particular sound in the visual domain. It describes the mouth positions and movements corresponding to that speech sound. In AVSR, AAMs are generally created from full face annotated images. The fitting is then performed on the full face and mouth parameters are extracted from the model[2][3]. The disadvantage to this method is that much of the model's energy describes variations that are not important from a mouth perspective. A mouth only AAM would allow more of the model to describe important variations in the mouth region. The parameters of the model would be specific to the mouth. This is important as it has been shown that mouth specific AAM parameters are reliable visual features in AVSR systems[4]. However it has been noted that the mouth area is not a reliable area for AAM tracking and the projection of face parameters into a mouth AAM space is a common solution[2].

The Fixed Jacobian fitting procedure proposed by Cootes is the most widely known method for AAM fitting. Many authors have looked at ways in which this can be improved[5][6]. However a recent study by Saragih et al. [7] showed that it is a robust method of fitting unobserved face images compared to some of the well known extensions to the original fitting technique. The Fixed Jacobian fitting procedure is traditionally made more robust through a multiresolution approach where a Fixed Jacobian is trained on multiple resolutions of the training images. Typically three levels are used, ranging from full resolution to a quarter of the image resolution. Levels are created by smoothing and subsampling the original image. This multiresolution approach is known as Gaussian Image Pyramid (GIP) approach. This paper proposes a new Extended MultiResolution Approach (EMRA) in which the area described by the AAMs reduces as the iterations progress. This allows higher levels of the pyramid to be more accurate at describing the mouth shape and appearance which is of ultimate importance to an AVSR system.

This EMRA was then tested on 432 annotated images, encompassing 36 different individuals from the CUAVE database[8]. This is then compared to the standard full face GIP and results are also given for intermediary stages between the two. The test involved the creation of 540 AAM pyramids, with each AAM pyramid tested on 600 different fitting initialisations. This large test clearly demonstrates the full range of performance of the EMRA for AVSR mouth fitting.

The novel contribution of this paper is therefore an AAM fitting scheme targeted specifically to the AVSR problem that achieves a significant improvement in accuracy of fit. Furthermore EMRA is experimentally assessed on a speaker independent task, making it suited for a wide range of AVSR systems.

This paper is organised as follows: Section 2 introduces AAMs and the Fixed Jacobian. Section 3 details our Extended MultiResolution Approach. Section 4 lays out the experimental procedure while in Section 5 our results are presented before some conclusions are drawn.

2. ACTIVE APPEARANCE MODELS

Active Appearance Models describe both the shape and appearance of a deformable model such as a face. Given a set of annotated face images a statistical model of both the shape, $\mathbf{x}$ and the texture, $\mathbf{g}$, is constructed as follows:

$$\mathbf{x} = \bar{\mathbf{x}} + \mathbf{P_s}\mathbf{b_s} \tag{1}$$

$$\mathbf{g} = \bar{\mathbf{g}} + \mathbf{P_g}\mathbf{b_g} \tag{2}$$

where $\bar{\mathbf{x}}$ is the mean shape of the annotated data, $\bar{\mathbf{g}}$ is the mean texture, $\mathbf{P_s}$ and $\mathbf{P_g}$ are sets of orthogonal modes of variation of shape and texture respectively, and $\mathbf{b_s}$ and $\mathbf{b_g}$ are the sets of shape and texture parameters.

Fitting AAMs is based on minimising the difference, $\mathbf{r}(\mathbf{p})$, between the pixels of a model generated face, $\mathbf{g_m}$, and the pixels sampled below the current estimate of shape in the image being searched, $\mathbf{g_s}$.

$$\mathbf{r}(\mathbf{p}) = \mathbf{g_s} - \mathbf{g_m} \tag{3}$$

In [1] it is proposed that linear adjustments can be made in order to minimise the cost function.

$$\delta\mathbf{p} = \mathbf{R}\mathbf{r}(\mathbf{p}) \tag{4}$$

where $\delta\mathbf{p}$ is the change required to the model parameters that minimise the difference between the images and $\mathbf{R}$ is the gradient

matrix between $\delta\mathbf{p}$ and $\mathbf{r}$. $\mathbf{R}$ is assumed to be fixed and is termed the Fixed Jacobian. The computational cost of the fitting procedure is significantly reduced as it is not necessary to compute this at each stage of the gradient descent, but instead it can be learnt once in the training phase.

In order to make this linear assumption more robust to larger initial displacements a Gaussian Image Pyramid is constructed. At coarser resolutions the linear assumption is more robust but the finer detail is removed from the model. Therefore, larger movements are made at the coarser resolutions before being refined in the higher resolution levels of the pyramid.

3. MOUTH SPECIFIC AAM FITTING - EMRA

The standard approach to extracting mouth feature points using AAMs is to use a full face model and then to extract mouth feature parameters from this fit[2][3]. This can be in the form of the mouth shape points, the pixels within a Region of Interest around the mouth, or the shape and appearance of the mouth region fitted can be projected into a trained mouth area AAM which is not used during the fitting procedure[2]. This is not optimal as the Principal Component Analysis (PCA) model used in the fitting describes features such as the eyes and nose that are not of interest in terms of the mouth positions. These features are important for AAM fitting and result in a model that is robust when faced with larger perturbations. Mouth only AAMs fail more often than full face due to their limited size which reduces their robustness to large deviations[2].

If a mouth-only AAM could be reliably fit, it would offer a significant advantage in that all of the model's PCA energy would be used to describe the mouth, and variations in its position and appearance. Therefore more subtle movements required to distinguish between different visemes could be represented and the overall accuracy of the mouth positions could be increased. More recently the use of AAM parameters has been shown to perform better than shape positions or appearance alone for distinguishing the different visemes[4]. Therefore having AAM parameters that more specifically describe the structure of the mouth is advantageous. For the Full Face Model such parameters can be obtained by projecting into a mouth only AAM space after the full face fitting procedure is complete, however in this paper it is shown that using mouth area AAMs during the fitting phase gives an overall improvement in accuracy.

This paper proposes a new Extended MultiResolution Approach (EMRA) for further developing the pyramid structure proposed in [1]. It extends the fitting pyramid to include further levels that increase the mouth fitting accuracy of AAMs. The Fixed Jacobian method is used as the fitting method as it has been shown to be quite successful in terms of its convergence rates and its accuracy for subject independent cases[7].

EMRA is a refinement technique whereby a full face AAM is used to find the optimum position and then the model is refined by further reducing the section of the face that the AAM describes. It consists of five levels each requiring its own AAM to be trained. A comparison of the standard 3-level Full Face fitting pyramid and the first four levels of the proposed refinement technique is shown in Figure 1. Figure 2 shows the composition of the fifth level of the pyramid. The 5 levels of EMRA can be summarised as follows:

Level 1: Full Face AAM fitting with a PCA energy of 75% and an image resolution of 0.25.

Level 2: Full Face AAM Fitting with a PCA energy of 85% and an image resolution of 0.5.

Level 3: Mouth/Chin AAM Fitting with a PCA energy of 95% on images of resolution of 0.5.

Level 4: Mouth/Chin AAM Fitting with a PCA energy of 95% using the full resolution of the images.

Level 5: Sampled Mouth Points AAM with a PCA energy of 95% using full resolution images.

3.1 Levels 1-2: Full Face AAM:

In the GIP technique large perturbations are performed in the lower resolution levels. By using images with lower resolutions much of the finer detail is removed thus reducing the complexity of the model and of the optimisation process for that level. In EMRA these low resolution face AAM levels are used to create a starting position for more accurate mouth AAMs that can further refine the shape returned.

The standard approach in GIP is to apply the same PCA energy at each level (typically 95%). Hence the same amount of shape variation is allowed for each level, the complexity reduced only by the decreasing resolution and the loss of finer detail in the appearance frame. In EMRA the complexity of the levels is further adjusted by varying the PCA energies at each level. By reducing the PCA energies at these levels the fitted shape is constrained, as is the appearance. This is similar to the work of Nguyen et al.[9] who showed that lower PCA energies were more reliable at fitting from larger perturbations. However in our approach, the PCA energies are increased along with increasing levels of the pyramid, the fitting technique allowing for further refinement as it approaches the optimum solution.

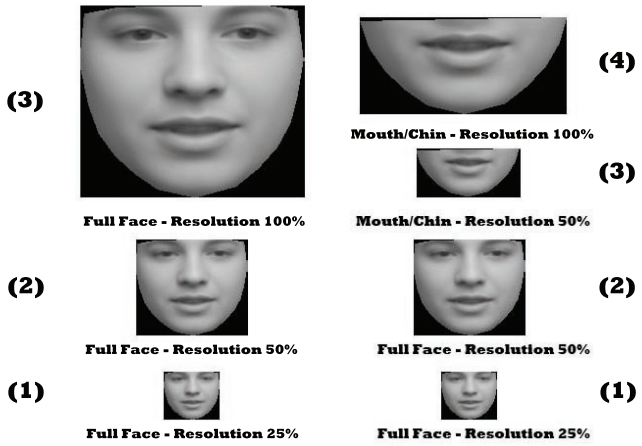

Figure 1: A comparison of the 3-level full face model (left) and a 4-level mouth/chin refinement of a full face model. The images in the diagram are to scale. Even though the Mouth/Chin refinement method has more levels, there are less pixels overall in the 4 levels than in the 3 levels of the full face pyramid.

3.2 Levels 3-4: Mouth/Chin AAM:

The convex hull of the Mouth/Chin AAM is shown at the top right of Figure 1. It includes both the mouth and skin pixels around the mouth area, down to the chin. By using skin pixels around the mouth, levels 3 and 4 are more robust at finding the exact contours of the mouth than a mouth-only AAM. The Mouth/Chin AAMs are used to further refine the mouth positions returned by the fitting of the full face AAMs. The Mouth/Chin has only approximately 40% of the appearance pixels of a Full Face AAM at the same resolution. It also only has 44% of the shape points. This offers significant computational advantages.

The resolution of level 3 is 50% while level 4 uses images at full resolution. Unlike a typical multiresolution approach, level 3 has the same resolution as level 2. Between these levels the model is projected from a Full Face space to the Mouth/Chin Space and fitting is performed. This proved to be more accurate and robust than increasing the resolution of the image while also projecting between the spaces. The PCA energies for levels 3 and 4 are 95%, the refinement process requiring enough variability to accurately describe the mouth shape.

3.3 Level 5: Sampled Mouth Points:

One of the limitations of the Fixed Jacobian fitting technique is that minimising the image difference does not necessarily minimise the

shape difference. In order to place more emphasis on the shape points, which are in general chosen to signify points of interest (significant pixel variation), pixels are sampled around these shape points.

In this final stage further refinement is made by sampling pixels along the boundaries of the mouth points. When creating an AAM a normalised frame is found that all images are warped into. This normalised frame is constrained, in that for all the training images the pixels associated with certain features occupy the same pixels. Hence the mouth corners, the edges of the lips, or the tubercle, will always be in the same position. This is the basis for the appearance model's creation.

This final level in EMRA exploits this attribute as the sampled pixels are acquired by applying a mask to each image in the normalised frame in order to select equivalent pixels. The mask is shown in Figure 2. It consists of a 5x5 square of pixels around each mouth point and a line of 5xL between two connected points, where L is the number of pixels between these points. This line of pixels between points encloses the lips/skin boundary and the lip/-mouth boundaries. This AAM is therefore focussed primarily on minimising the difference between areas of high shape importance. It emphasises important elements of the shape, i.e. that points occur on boundaries.

The mouth area, especially around the shape points, is highly variable. This makes fitting such an AAM quite difficult especially when large perturbations are required. However by using this as a further refinement to the levels introduced earlier, the method will only need to make small changes.

Other authors have looked at training AAMs using sampled pixel regions. In [10] a certain subset of the pixels were chosen by Cootes et al. during the fitting stage depending on their value. The highest value pixels were deemed most important for the multivariate regression and other lower value pixels were removed in order to reduce the dimensionality. Nguyen's technique in [9] is similar to the method presented here. The AAM models are trained using a subset of the pixels in the face region by sampling areas around each of the points. However the size of the region around each box is much larger than considered here. In fact the mouth area is fully contained within their sampling region. In EMRA the sampled area is highly specific to the mouth points themselves and boundaries between them. This makes it inherently less computationally expensive to train and fit than with other AAM levels at the same resolution.

4. EXPERIMENTAL VALIDATION:

An extensive experiment was designed to compare the performance of the original GIP to the EMRA proposed here. In order to show the value of each additional component of the EMRA, intermediary experiments are also performed. The different pyramid structures tested were as follows:

Case A: *3 Level Full Face AAM Pyramid:* Resolutions of .25, .5 and 1 with a PCA Energy of 95% at each level.

Case B: *4 Level Face and Mouth/Chin AAM Pyramid:* The first two levels of case A followed by 2 Mouth/Chin AAM levels at resolutions .5 and 1 also with PCA Energy of 95%.

Case C: *3 Level Full Face AAM Pyramid with variable PCA Energies:* 3 Full Face levels with image resolutions of .25, .5 and 1 with PCA energies of 75%, 85% and 95% respectively.

Case D: *4 Level Face and Mouth/Chin AAM Pyramid with variable PCA Energies:* The first two levels of case C followed by 2 Mouth/Chin AAM levels at resolutions of .5 and 1 both with a PCA energy of 95%.

Case E: *5 Level Face, Mouth/Chin and Mouth Sample AAM Pyramid:* The 4 levels of case D followed by one level of a mouth sample AAM with a PCA energy of 95% at a resolution of 1.

The PCA energies above describe how much of the mouth and texture variation are retained and also how much of the variation is kept when these are combined into the appearance model. The

Figure 2: The final level of the pyramid is an AAM trained on sample pixels around the shape points in the mouth region. The image above shows the pixels that are sampled by the red mask placed on a mouth/chin region. The limited amount of pixels within the red area results in a much less computationally expensive fitting level than the four levels that precede it.

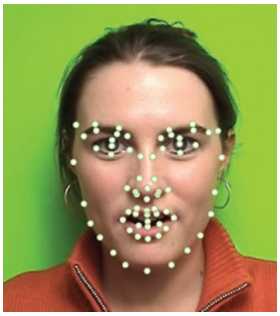

Figure 3: The annotated points used for the Cuave database.

Jacobians of the full face levels were trained using perturbations of 3 pixels in translation, 10% in scale and .1 rads in rotation. The Mouth/Chin Jacobians, being used solely for refinement, were trained using smaller perturbations of .5 pixels in translation, 1.3% in scale and .014 rads in rotation. The Mouth Sample Jacobains are trained using perturbations of .5 pixels in translation, 2.6% in scale and .03 rads in roation.

Each of these AAM pyramids were trained and tested on the CUAVE database[8]. The CUAVE database consists of 36 speakers, 19 male and 17 female, and 20 pairs of speakers speaking both connected and continuous speech. A wide variety of subjects with different skin tones and visual features such as spectacles, hats and facial hair are included.

432 frontal images were hand annotated representing 12 images from each of the 36 subjects speaking individually. For each speaker two visemes were described, each having 6 images equally spaced over the frames of that particular viseme. The annotations consist of 68 points as shown in Figure 3, with 19 of these points corresponding to mouth locations.

A cross-validation test is performed on a leave one out basis, that subject being used for testing. For the remaining 35 speakers, 6 images from each subject are chosen for training. 5 perturbations per PCA mode were adequate to generate the Fixed Jacobian at each level.

This test is repeated three times for each subject, requiring a total of 108 AAM pyramids for each case. Testing is performed on the 12 annotated images of the subject not included in the training of the corresponding AAM. For each image a total of 50 perturbations are made (a total of 64,800 test examples for each case above). These perturbations are made on the mean model, where the mean shape has been optimally aligned to the annotated points using Procrustes analysis.

The 50 starting positions for each image in the test are created by randomly perturbing this mean model in translation, rotation and scale. The maximum translation is $\pm$10pixels in both the x and y coordinates, the maximum rotation is $\pm$.1rads and the maximum

scaling is by $\pm 10\%$.

5. ANALYSIS OF RESULTS:

Convergence is defined as an average RMS difference between each mouth point in the fitted shape less than 4 pixels from the annotated points. The convergence accuracy is then only considered on examples that have converged below this threshold. It is worth noting that other authors use different definitions of convergence which make it difficult to compare results on the Fixed Jacobian. Cootes et al [11], for example, defined convergence to occur when the mean point position error was less than 7.5 pixels per point.

Table 1 summarises the results for each of the pyramid structures. The full face Fixed Jacobian (case A) is shown to be quite reliable with 67.4% of the examples converging below a 4 pixel average difference. To verify the performance of the Full Face Fixed Jacobian AAM the annotated images from CUAVE were then tested using the original testing method noted in [11]. For this test the mean model was perturbed up to ± 15 pixels in both x and y and 10% in scale. There was no rotation of the mean model prior to the fitting process. The test returned a convergence rate of 77.89%. This is comparable to the 81% quoted in [11], especially given the differing datasets that the tests were performed on.

Case	Convergence Rates	Improvement	Pixel Error	Improvement
A:	67.40%	-	2.93	-
B:	73.07%	5.67%	2.78	5.12%
C:	75.97%	8.57%	2.92	0.34%
D:	79.84%	12.44%	2.78	5.12%
E: (EMRA)	81.50%	14.10%	2.68	8.53%

Table 1: Table showing the convergence rates and pixel convergence errors of Cases A-E. A is equivalent to obtaining mouth positions from a full face model whilst Case E is the EMRA approach. Cases B, C and D are intermediary stages.

The converged error of the mouth points extracted from the Full Face Fixed Jacobain is 2.93 pixels, taken as an average of the point to point errors from each individual. The use of a Mouth/Chin refinement AAM in case B increases the accuracy of the average fit by 5.12%, while also increasing the convergence rate by 5.67%.

Case C and D are equivalent to cases A and B but with a stepped increase in the PCA accuracy of the lower full face levels. There is no significant improvement in the average accuracy, but there is a significant 8.57% increase in the convergence rates as expected. With less variance contained in the first two levels of the fitting structure, larger movements can be reliably made with less likelihood of falling into local minima. The final level for case C and case D does not differ from that of case A and B. Thus it is logical that the average fit would be the same, as there is no more variability in movement for the final fit.

Case E shows that the addition of the extra mouth sample AAM level to case D improves the convergence rate by 1.66%. More importantly for a refinement method, there is a 3.41% improvement in convergence error over case D. The mouth sample AAM level has a lower computational cost in fitting, with the Fixed Jacobian being of much lower dimensionality to previous AAM levels. Hence its addition does not significantly increase computation time.

Comparing the overall structure proposed against the standard 3 level Full Face pyramid approach (case A) there is a 8.53% improvement in fitting accuracy coupled with a 14.1% improvement in the number of images that converge within our limit of an average difference of 4 pixels. Figure 4 is a convergence accuracy plot: a histogram showing the distribution of the convergence accuracy of the EMRA and Full Face Fixed Jacobian approaches. As is clearly shown a greater proportion of the EMRA samples have lower convergence errors.

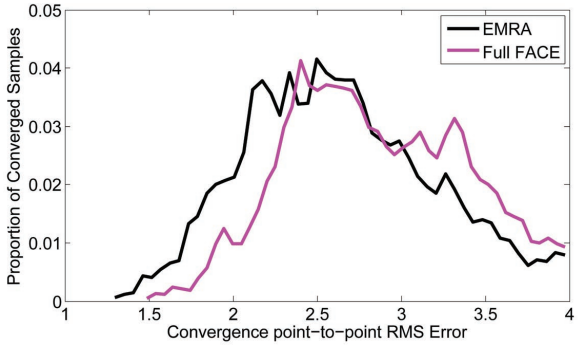

Figure 4: A histogram of convergence accuracy comparing the EMRA to the Full Face Fixed Jacobian approach.

EMRA gives significant gains over the standard approach. Improvements are made on most of the subjects in the database. There are eight individuals who obtain significant ($> 25\%$) improvement in their convergence rates with the new fitting structure. One of the principal reasons for this is that they are better represented by the appearance model trained solely on the mouth area which has more variability within the mouth structure, allowing the appearance model to more accurately describe the mouth shape.

However, certain individuals prove to be much more difficult than others. The fitting ability of the Full Face Fixed Jacobian for a male subject (S10m) and a female subject (S20f) is shown in figure 5(a). For each two cases are shown, one where the image has not been used in training but the subject has (b), and another where the subject is not used at all during AAM training (c). The latter represents the person independent AAM training used in the experiment noted before. For S10m the error for the image unseen fit was 1.93 pixels, whereas for the subject unseen fit the error is 2.67 pixels.

S20f represents the most difficult subject for AAM fitting in the CUAVE database, with very large distinguishing mouth features not present in other subjects. This subject achieves approximately 0% convergence rate for all AAM cases in the experiment, including case A. In both cases of the fit, the scale, translation and rotation perform well. In the image unseen fit, facial features such as the eyes are fitted well, however the mouth detail has clearly not been seen in training. This results in a mouth fit error of 4.32 pixels. This result is important as even when the subject has been used in training, an unseen image of that subject returns a fit that would not pass our convergence criteria. When the subject is unseen the fit is quite bad, the model struggling to fit the distinguishing features of the subject returning a mouth fit error of 8.67 pixels. This prompts the question of how well the 3 level Full Face AAM (case A) could represent the annotated shape. This would represent the optimum fit that the Fixed Jacobian could return.

To quantify the ability of the model to represent the different annotated shapes in CUAVE a leave one out experiment was performed. The annotated shapes of the unseen test subject were then projected into the AAM space to see how accurately the model could describe the mouth shape. However these projections are not absolute bounds, they are optimal in the sense of the Full Face shape and appearance. In some cases the Fixed Jacobian may reduce the mouth pixel error over other parts of the full face shape. This results in a somewhat lower mouth shape error at the expense of the overall face shape error, a process which can also work in reverse. In general the Fixed Jacobian of a Full Face Model will minimise the error over all pixels, and hence all shape points. Therefore the optimal projection of the full shape provides a good measure of how well the PCA space of a Full Face model can represent a given mouth shape.

The average mouth pixel error of the converged examples was 2.6102. This is found by averaging the optimal projection into a

Full Face AAM space for each individual. This number represents the optimal result obtainable for a full face AAM. The result of the proposed EMRA was 2.68 pixels as opposed to 2.93 for the standard GIP approach (see Table 1). The technique therefore reduces the error of the standard GIP approach by 78% when compared to the optimal average pixel error. It shows the benefit of AAM refinement, in cases where small deformable objects need to described. If the region around the object can be described then a refinement process can be more robust at returning an accurate result.

As expected the images upon which the fits were not successful had optimal point to point errors of greater than 4 pixels. The subjects are too far away from the base position. Therefore the Fixed Jacobian cannot succeed in its fit under the targets set. However once the mouth shape can be accurately described by the appearance model it is found that the Fixed Jacobian performs well.

Increasing the PCA energies of the appearance model would theoretically increase its ability to represent the different mouth shapes. However at higher accuracies ($> 95 - 97\%$ of the shape and appearance) there is a tail off in the ability of the Fixed Jacobian to fit. Such a cost curve contains more local minima and is less robust at fitting.

A refinement process, such as EMRA, can focus the modes of the model to describe the areas of interest, while still using more robust lower levels to make larger perturbations. In essence, this approach avoids the pitfalls associated with the Fixed Jacobian at large PCA energies, but offers an improved fit in a specific region of interest.

6. CONCLUSION

In this paper, EMRA, a new Fixed Jacobian fitting structure for extracting mouth positions from unseen images, has been presented. As the levels progress, the PCA energy is increased allowing for more variability, while the area that the model describes decreases. This allows for higher variation within the model with lower computational costs, while still maintaining the robustness of a full face model during the initial iterations where it is most important.

It is shown that EMRA performs significantly better than a standard 3 level full face model in both convergence rates and convergence accuracy of the mouth region. The convergence rate is increased by 14.1% where a boundary indicating convergence is set at 4 pixel, while the accuracy of these converged examples is increased by 8.53%.

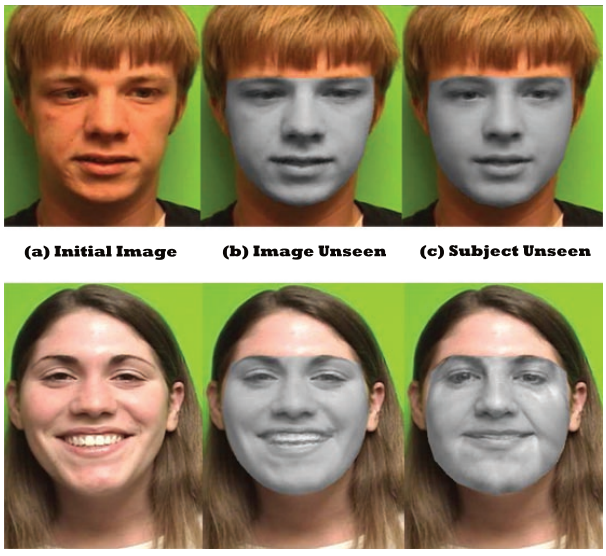

Figure 5: The fitting ability of the Full Face Fixed Jacobian for a male subject (S10m) and a female subject (S20f).

Though the new fitting structure improves the overall performance of the Fixed Jacobian technique, it does not correct some of the major issues with it. It was found that once the appearance model was able to represent a shape within the 4 pixel boundary, then the fixed jacobian was quite effective at fitting. However for certain individuals the model cannot represent the shape within this boundary.

Though the results quoted here show an improvement in terms of average point to point error over a sizeable database, the next stage is to compare the EMRA's visemic recognition accuracy against that of the standard 3-level Full Face Fixed Jacobian. This will establish whether a better mouth fit in terms of pixel accuracy can be translated into higher recognition performance in an AVSR system.

7. ACKNOWLEDGEMENTS

The authors gratefully acknowledge the contribution of Irish Research Council for Science Engineering and Technology and the financial support of Science Foundation Ireland under Grant Number 09/RFP/ECE2196.

REFERENCES

[1] T. Cootes, G. Edwards, and C. Taylor. Active Appearance Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(6):681–685, 2001.

[2] A. Katsamanis, G. Papandreau, and P. Maragos. Face Active Appearance Modeling and Speech Acoustic Information to Recover Articulation. *IEEE Transactions on Audio, Speech and Language Processing*, 17(3):411–422, 2009.

[3] R. Goecke and A. Asthana. A Comparative Study of 2D and 3D Lip Tracking for AV ASR. *Int. Conf. on Auditory-Visual Speech Processing*, pages 235–240, 2008.

[4] Y. Lan, R. Harvey, B. Theobald, Ong E., and R. Bowden. Comparing Visual Features for Lipreading. *AVSP*, pages 102–106, 2009.

[5] I. Matthews and S. Baker. Active Appearance Models Revisited. *International Journal of Computer Vision*, 60(2):135–164, 2004.

[6] R. Donner, M. Reiter, G. Langs, P. Peloschek, and H. Bischof. Fast Active Appearance Model Search Using Canonical Correlation Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1690–1694, 2006.

[7] J. Saragih and R. Goecke. Learning AAM fitting through simulation. *Pattern Recognition*, 42(11):2628–2636, 2009.

[8] E. Patterson, S. Gurbuz, Z. Tufekci, and J. Gowdy. A new audio-visual database for multimodal human-computer interface research. *In Proceedings of the International Conference on Acoustics, Speech and Signal Processing - ICASSP*, pages 2017–2020, 2002.

[9] M. Nguyen and F. de la Torre. Metric Learning for Image Alignment. *International Journal of Computer Vision*, 88:69–84, 2010.

[10] T. Cootes, G. Edwards, and C. Taylor. A comparative evaluation of active appearance models algorithms. *British Machine Vision Conference*, 2:680–689, 1998.

[11] T. Cootes, G. Edwards, and C. Taylor. Active Appearance Models. volume 2, pages 484–498, 1998.