Learning Active Appearance Models from Image Sequences

Jason Saragih[1] **Roland Goecke**[1,2*]

[1]Research School of Information Sciences and Engineering, Australian National University
[2]National ICT Australia, Canberra Research Laboratory
Canberra, Australia
Email: jason.saragih@rsise.anu.edu.au, roland.goecke@nicta.com.au

Abstract

One of the major drawbacks of the Active Appearance Model (AAM) is that it requires a training set of pseudo-dense correspondences. Most methods for automatic correspondence finding involve a groupwise model building process which optimises over all images in the training sequence simultaneously. In this work, we pose the problem of correspondence finding as an adaptive template tracking process. We investigate the utility of this approach on an audio-visual (AV) speech database and show that it can give reasonable results.

Keywords: AAM, automatic model building.

1 Introduction

Active appearance models (AAM) are a powerful class of generative parametric models for non-rigid visual objects which couple a compact representation with an efficient alignment method. Since its advent by Edwards *et al.* in (Edwards, Taylor & Cootes 1998) and their preliminary extension (Cootes, Edwards, Taylor, Burkhardt & Neuman 1998), the method has found applications in many image modelling, alignment and tracking problems, for example (Lehn-Schiøler, Hansen & Larsen 2005) (Stegmann & Larsson 2003) (Mittrapiyanuruk, DeSouza & Kak 2005).

The main drawback of AAM is that it requires pseudo-dense annotations for every training image to build its statistical models of shape and texture. Each of these images may require hundreds of corresponding points. Manual annotation for large databases, therefore, are tedious and error prone. The process is especially difficult for objects which exhibit only a small number of corner like features (i.e. the human face contains mostly edges). A process which automates the annotation process is, hence, highly desirable and may encourage a more widespread utilisation of the AAM.

In this paper, we discuss the automatic annotations (finding physically corresponding points across images) of audio-visual (AV) speech databases which consist of sequences of talking heads. As a test case, we investigate its utility on the AVOZES (Goecke & Millar 2004) database. This scenario for automatic annotations is more constrained than the general problem as the changes in shape and texture between consecutive frames in a sequence is relatively small. Nonetheless, we show that this problem is still a challenging one, mainly due to the high dimensionality of the problem which makes it difficult to optimise and avoid spurious local minima.

We approach the automatic annotation process through a tracking perspective, where the annotations in a reference image are propagated through the sequence by virtue of an adaptive template. We begin with an overview of related work in Section 2. The problem of image based correspondences is discussed in Section 3. An outline of our approach to the automatic annotations of image sequences is then presented in Section 4. In Section 5, we describe the results of applying this approach to the AVOZES database. Section 6 concludes with discussions of the results and future directions.

2 Related Work

There has been significant research over the years to automatically find semi-dense correspondences across images of the same class for building AAMs. These methods can be broadly categorised into either feature or image based approaches.

Feature based methods (Chui, Win, Schultz, Duncan & Rangarajan 2003) (Walker, Cootes, & Taylor 1999) (Hill & Taylor 1996) find correspondences between salient features in the image by examining the local structure of the features. The advantage of this method is that feature comparisons and calculations are relatively cheap. The downside however is twofold. Firstly, there may be insufficient salient features in the object to build a good appearance model. Secondly, as the feature comparisons generally consider only local image structure, the global image structure for which the AAM is then modelled is ignored, and hence, the model may be suboptimal.

Image based methods (Cootes, Marsland, Twining, Smith & Taylor 2004) (Baker, Matthews & Schneider 2004) (Jebara 2003) usually find dense image correspondences by finding a nonlinear warping function which minimises some type of error measure between the intensities of the images. The main advantage of these methods is that the global structure of the image is taken into account, better mimicking the AAM for which the correspondences will be used later. The main drawback of this approach is that to accurately represent the shape variations of the visual object, the warping function will generally need to be parametrised using a large number of parameters (generally as set of landmark points). This results in a very large optimisation problem which is slow to optimise and prone to terminating in local minima.

*National ICT Australia is funded by the Australian Government's *Backing Australia's Ability* initiative, in part through the Australian Research Council

3 Image Based Correspondence

The heart of image based methods for correspondence consists of finding a warping function between a set of images such that every location in one image is warped to the same physically meaningful (corresponding) location in all other images. However, as there is no true sense of the physical correspondence of un-annotated images, the quality of a set of warping functions is usually quantified by some measure of model *compactness* built from the warped images. Examples of these measures include MDL (Cootes, Twining, Petrovic, Schestowitz & Taylor 2005), specificity/generalisation (Schestowitz, Twining, Petrovic, Cootes, Crum & Taylor 2006) and minimum volume PCA (Jebara 2003).

Apart from the measure of quality there is a large amount of variation of image based correspondence methods at the implementation level. These variations include, but are not limited to, model and warp parametrisation, model fitting methods and the landmark selection process. In this section, we describe the choices we made on these factors for the experiments presented in Section 5. In most cases, we follow the convention of most AAM implementations.

3.1 Linear Appearance Models

Active appearance models assume the visual phenomenon being modelled takes the form of a degenerate Gaussian distribution, where the shape and texture can be modelled by a compact set of linear modes of variation. The texture is generated as follows:

$$t(\mathbf{x}) = \bar{t}(\mathbf{x}) + \sum_{k=1}^{m_t} q_k t_k(\mathbf{x}), \tag{1}$$

where $t(\mathbf{x})$ is the generated model texture at pixel location $\mathbf{x}$, $\bar{t}(\mathbf{x})$ is the mean texture at that location, $t_k(\mathbf{x})$ is the k^{th} mode of texture variation and q_k is the magnitude of variation in that direction. Similarly, a novel instance of the model's shape can be generated using:

$$\mathbf{s} = \bar{\mathbf{s}} + \sum_{k=1}^{m_s} p_k \mathbf{s}_k, \tag{2}$$

where $\mathbf{s} = [\mathbf{x}_1; \ldots; \mathbf{x}_n]$ is the shape vector of concatenated landmark locations, $\bar{\mathbf{s}}$ is the mean shape, $\mathbf{s}_k$ is the k^{th} mode of shape variation and p_k is the magnitude of variation in that direction.

These models are usually obtained by applying PCA to a set of annotated images, retaining only the m_t and m_s largest modes of variation in shape and texture respectively. The resulting model is a compact representation of a high dimensional visual object by a small set of parameters.

Although these separate models of variation (called independent appearance models) have shown to adequately represent the variations exhibited by many visual objects, they fail to take into account the correlations between shape and texture. In some cases, where there is a strong correlation between shape and texture, failing to take these correlations into account may result in a model capable of generating unrealistic instances of the object class. Furthermore, the resulting model may not be as compact as it could be, if these correlations are considered in the model building process. An example of this is a person-specific AAM. In these cases, it is beneficial to perform a second level of PCA, this time on the concatenation of the shape and texture parameters:

$$\mathbf{a} = \begin{bmatrix} \mathbf{W}_s \mathbf{p} \\ \mathbf{q} \end{bmatrix}, \tag{3}$$

where $\mathbf{W}_s$ is a weighting matrix which normalises the difference in units between shape and texture. A common choice for this matrix is an isotropic diagonal matrix representing the ratio between the total variations of shape and texture in the training set. By applying PCA to a set of these training vectors, a combined appearance model is obtained, for which novel instances can be generated as follows:

$$\mathbf{a} = \sum_{i}^{m_a} c_k \mathbf{a}_k, \tag{4}$$

where $\mathbf{a}_k$ is the k^{th} mode of combined appearance variation and c_k is the magnitude of variation in that direction. The combined appearance model can be used to generate novel instances of shape and texture directly as follows:

$$\mathbf{s} = \bar{\mathbf{s}} + \mathbf{Q}_s \mathbf{a} \tag{5}$$

$$\mathbf{t} = \bar{\mathbf{t}} + \mathbf{Q}_t \mathbf{a}, \tag{6}$$

where

$$\mathbf{Q}_s = \mathbf{S}\mathbf{W}_s^{-1}\mathbf{A}_s \tag{7}$$

$$\mathbf{Q}_t = \mathbf{T}\mathbf{A}_t \tag{8}$$

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_s \\ \mathbf{A}_t \end{bmatrix} \tag{9}$$

and

$$\mathbf{S} = [\mathbf{s}_1, \ldots, \mathbf{s}_{m_s}]$$
$$\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_{m_t}]$$
$$\mathbf{A} = [\mathbf{a}_1, \ldots, \mathbf{a}_{m_a}]$$

are matrices of concatenated modes of variations of shape, texture and appearance, respectively. For visual objects exhibiting strong correlations between shape and texture, the resulting combined appearance model is usually more compact than the independent appearance model, exhibiting a smaller number of modes of variation.

3.2 Model Quality

The quality of a model is usually quantified by some measure of compactness. In our work, we follow the method in (Jebara 2003) which estimates compactness of Gaussian distributed models through an approximation of the volume of the variations of the model. The approximation used here is the determinant of the model's covariance matrix, which is equivalent to the sum of the eigenvalues of the model:

$$Q = \sum_{i}^{m} \lambda_i \tag{10}$$

In the AAM, variations in pixel values in the image frame are generated from variations in both shape and texture, each of which is modelled by a Gaussian distribution. Therefore, a measure of compactness of an appearance model must take into account the compactness of both models which may disagree with each other. For example, for the same database, a model which exhibits a compact shape distribution may result in a non-compact texture as it needs to accommodate pixel intensity variations which are not accounted for by the shape. On the other hand, if the texture is evaluated in a reference frame (as opposed to the image frame as is done in an MDL formulation (Cootes et al. 2005)), the shape may be chosen

such that the texture is compact at the cost of a non-compact shape distribution. In (Jebara 2003), only the texture compactness is used as a measure of quality, which may result in a non-compact shape distribution which in turn may result in a model which can generate implausible shapes. Although it is easy to have a single measure of model quality through a weighting of the compactness of shape and texture, this weighting is usually chosen heuristically based on the intuition of good results from manual analysis of example models. In this work, we investigate the trends of the shape and texture compactness measures for different settings of the training parameters.

As a final note, in our implementation the sum in Equation (10) is performed over *all* non-zero eigenvalues of the system rather than only the most significant ones. This is because we want to measure the model quality by considering the total amount of variation in the training set. Since the total variation may differ depending on the implementation details, common methods used in PCA such as retaining only a certain percentage of the total variation may not give a discriminative measure as different amounts of variations may be discarded as noise.

3.3 Landmarks and the Warping Function

The shape of an AAM is defined through a set of landmarks which in turn parametrise the warping function used to project the texture from the image to the reference frame.

3.3.1 Landmark Selection Scheme

The choice of these landmarks is crucial to the compactness of the model. As a rule of thumb, for a given number of landmark points, the set which, under the warping function, accounts for the most amount of shape variation within the object class should be chosen. This way, the variation exhibited in the texture model accredited to shape variation is minimised. However, in the problem of automatic model building, parts of the object which exhibit the most variation in shape are not known a-priori. Therefore, a choice must be made regarding the contribution of each location in the image to the variation in texture due to unaccounted variations in shape.

In general, locations with high texture contribute more to the variation in texture due to unaccounted shape variations than do flat regions. Therefore, we propose using a sequential selective process where landmarks are chosen iteratively based on their saliency, measured by the cornerness of that point in a reference image. This method was adopted in (Cootes et al. 2005), where it was demonstrated that using landmarks on strong edges, and ignoring flat regions, gave the best performance as it allowed more control over the boundary regions in the image. Our method differs however in the way the landmarks are chosen. In their approach, the landmarks are initialised on an equally spaced grid, then moved to the closest strong edge. In our work, we sequentially select the most salient pixel location, then zero-out a small region around that point in the saliency image. This process guarantees that the most salient locations are selected, but prevents trivial landmarks (i.e. those which are too close to represent adequate shape variations) from being selected.

Apart from these salient landmarks, we also add a fixed number of border landmarks, equally spaced around the image border, such that the whole image is encoded into the texture model. As the domain of the texture of an AAM is usually defined within the convex hull of the reference shape only, adding these border landmarks allow the background to be incorporated into the model's texture which may allow a more accurate model building process as the boundary between the object and the background can give strong cues for model fitting.

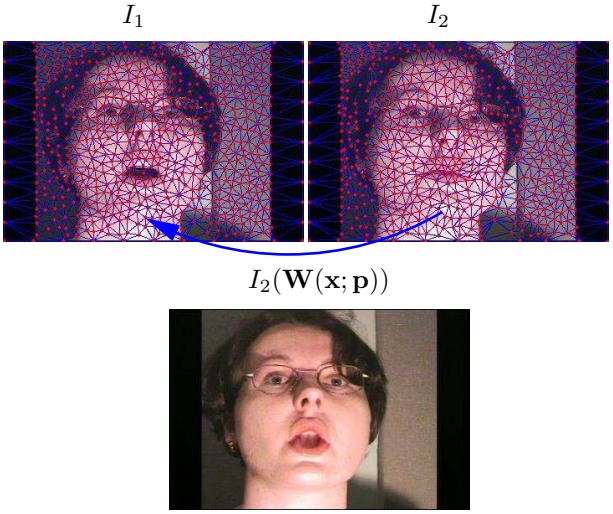

Figure 1: Piecewise-affine warping. Top row: pseudo-dense landmark triangulation. Bottom: I_2 warped onto I_1 using piecewise affine warp defined by triangulation.

3.3.2 Warping Functions

The most common warping function used for AAMs is the piecewise affine warp. This type of warp utilises a triangulation of landmarks in the reference image, where pixels within the domain of each triangle are warped using an affine function. Although there are many other warping function which can be used, such as thin-plate splines or B-Splines, the piecewise affine warp is simple and efficient. Furthermore, it allows the inverse of the warp to be computed efficiently, which is beneficial in an image generation process where the texture in the reference frame is projected onto the image frame.

Although the piecewise affine warp has the disadvantage that it is discontinuous at the boundaries of the triangles, we find that a sufficiently dense set of landmarks chosen according to the scheme described in Section 3.3.1 usually results in a triangulation where the edges in the image correspond to edges of the triangles, minimising the effect of this discontinuity. An example of a pseudo-dense landmark selection with its triangulation and warping process is shown in Figure 1.

3.4 Alignment

Regardless of the model building process used, automatic AAM construction generally involves a non-rigid registration to align the model to an image. The alignment process essentially finds the model parameters which best describe the image. This process usually involves minimising some measure of fitness between the model and the image which contains a data term and a smoothness term:

$$C = C_d + \eta C_s, \tag{11}$$

where C_d is the data term, C_s is the smoothness term and η is a regularisation parameter which trades off the contribution of the data and smoothness terms to the total cost.

3.4.1 The Data Term

The data term is usually defined as a function of the difference between the model's texture and the image texture warped back to the reference frame:

$$C_d = \sum_{\mathbf{x}\in\Omega} \rho\left(E(\mathbf{x}); \sigma\right) \tag{12}$$

$$E(\mathbf{x}) = t(\mathbf{x}; \mathbf{q}) - I(\mathbf{W}(\mathbf{x}; \mathbf{p})), \tag{13}$$

where Ω is the domain over which the model's texture is defined (i.e. the convex hull of the landmark points), $t(\mathbf{x}; \mathbf{q})$ is the model's texture, $I(\mathbf{W}(\mathbf{x}; \mathbf{p}))$ is the image texture warped back to the reference frame, and ρ is some type of function over the residuals, parametrised by σ.

A common function used in AAM alignment is the L2-norm (Baker & Matthews 2002), in which case, the data term takes the least squares form. However, in some cases it may be beneficial to use a robust error function to minimise the effects outliers in the data. This is particularly important in model building as regions which are not yet accounted for by the texture model may deteriorate the estimate of the shape in other parts of the image, leading to a non-compact model. For the experiments presented in Section 5, we use the *Geman-McClure* function:

$$\rho(r; \sigma) = \frac{r^2}{\sigma^2 + r^2}, \tag{14}$$

which has been used extensively for optical flow estimation (Black & Anandan 1993) (Blake, Isard & Reynard 1994).

The choice of the scale parameter σ for robust error functions is always problematic as it depends on the distribution of the residuals. One approach is to use the assumption that the corresponding error functions model the underlying distribution of residuals, and find σ which best fits that distribution. However, this usually leads to a complex non-linear estimation process. Therefore, in our work, we assume a contaminated Gaussian distribution for the residuals. In this framework, the estimate of σ can be derived from the median value of the absolute residuals:

$$\sigma = 1.4826 \text{ med}\left(E(\mathbf{x})\right) \tag{15}$$

which has been claimed to have excellent resistance towards outliers, tolerating almost 50% of them (Sawhney & Ayer 1996).

3.4.2 The Smoothness Term

In automatic model building the landmarks should be allowed to move freely to minimise the data term. However, as the AAM's shape consists of a pseudo-dense set of landmarks, the dimensionality of the optimisation process is very large, which if not constrained is likely to get trapped in spurious local minima. These minima usually correspond to implausible shapes. As such, a smoothness term is required to encourage the model to deform smoothly.

The form of the smoothness constraint is dependent on the visual object being modelled. The most common of which is to penalise the magnitude of the deformation of every landmark from a reference shape, as was adopted in (Baker et al. 2004). The problem with this approach is that it does not take into account the spatial relationship between the deformation of landmarks. In this work, we penalise only the *difference* between the deformation of landmarks, similar to the smoothness constraint in variational optical flow estimation (Brox, Bruhn, Papenberg & Weickert 2004). The differences are weighted by a smooth function of the landmark distances in a predefined shape:

$$C_s = \sum_{i,j}^{n} k_{ij} \|\mathbf{d}(i,j)\|^2, \tag{16}$$

where

$$k_{ij} = \frac{\exp\left(-\frac{\|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\|^2}{2\sigma_s^2}\right)}{\sum_j^n \exp\left(-\frac{\|\hat{\mathbf{x}}_i - \hat{\mathbf{x}}_j\|^2}{2\sigma_s^2}\right)} \tag{17}$$

is a smoothing factor and

$$\mathbf{d}(i,j) = [\mathbf{W}(\mathbf{x}_i; \mathbf{p}) - \hat{\mathbf{x}}_i] - [\mathbf{W}(\mathbf{x}_j; \mathbf{p}) - \hat{\mathbf{x}}_j] \tag{18}$$

is the difference between landmark displacements, with $\hat{\mathbf{x}}_\mathbf{k} = \mathbf{W}(\mathbf{x}_k; \mathbf{p}_0)$ being the location of the k^{th} landmark in the predefined shape, parametrised by $\mathbf{p}_0$. In most works utilising a smoothness measure, the predefined shape is always set to the reference shape (i.e. $\mathbf{p}_0 = \mathbf{0}$). The problem with this is that it assumes the deformations are isotropic for all landmarks. This type of smoothing does not fit the notion of a linear shape class which is modelled by a degenerate Gaussian. In contrast, we set the predefined shape as the initial shape in the alignment process. Smoothing the deformations in an isotropic manner starting from this shape better suits the form of the shape model as it does not over constrain the overall highly anisotropic shape deformations whilst still encouraging the landmarks to deform smoothly.

3.4.3 Optimisation

To optimise the cost function in Equation (11) we adopt the Gauss-Newton method which is commonly used for image alignment. To allow the use of the robust error function in the Gauss-Newton optimisation procedure, the data term must be reformulated. Since it contains no *squared* term, the derivation of the parameter update requires a second order Taylor expansion, akin to the Newton algorithm. Therefore, following (Baker, Gross & Matthews 2003), we replace the data term in Equation (12) with:

$$C_d = \sum_{\mathbf{x}\in\Omega} \varrho\left(E(\mathbf{x})^2; \sigma\right) \tag{19}$$

and the reformulated robust error function:

$$\varrho(r; \sigma) = \frac{r}{\sigma^2 + r} \tag{20}$$

This requires only that the error function is symmetric, which is satisfied by the Geman McClure function.

With this reformulation, the Gauss-Newton Hessian of the data term is given by:

$$\mathbf{H}_d = \sum_{\mathbf{x}\in\Omega} \varrho'(E(\mathbf{x})^2) \mathbf{J}_d(\mathbf{x})^T \mathbf{J}_d(\mathbf{x}) \tag{21}$$

where $\varrho'(E(\mathbf{x})^2)$ is the derivative of the reformulated robust error function and

$$\mathbf{J}_d(\mathbf{x}) = \left[-\nabla I(\mathbf{W}(\mathbf{x}; \mathbf{p})) \frac{\partial \mathbf{W}(\mathbf{x}; \mathbf{p})}{\partial \mathbf{p}}, \frac{\partial \mathbf{t}(\mathbf{x}; \mathbf{q})}{\partial \mathbf{q}}\right] \tag{22}$$

is the Jacobian of the data term. It should be noted here that since we allow the landmark points to move freely, the warping function $\mathbf{W}$ is directly parametrised by the location of the landmarks (i.e.

$\mathbf{p} = [\mathbf{x}_1; \ldots; \mathbf{x}_n]$). Therefore, the distance measure in Equation (18) is equivalent to:

$$\mathbf{d}(i,j) = (\mathbf{x}_i - \hat{\mathbf{x}}_i) - (\mathbf{x}_j - \hat{\mathbf{x}}_j) \tag{23}$$

This is in contrast to the usual AAM formulation where the warp is parametrised by the magnitudes of the modes of shape variation.

The Gauss-Newton Hessian of the smoothness term is given by:

$$\mathbf{H}_s = \sum_{i,j}^{n} k_{ij} \left[\mathbf{J}_x(i,j)^T \mathbf{J}_x(i,j) + \mathbf{J}_y(i,j)^T \mathbf{J}_y(i,j) \right] \tag{24}$$

where the $2k^{th}$ entry of the x smoothness term's Jacobian $\mathbf{J}_x(i,j)$ is given by:

$$\mathbf{J}_x(i,j)^{2k} = \begin{cases} 1 & \text{if } k = i \\ -1 & \text{if } k = j \\ 0 & \text{otherwise} \end{cases}, \tag{25}$$

and similarly for the $(2k+1)^{th}$ entry of $\mathbf{J}_y$. For $\mathbf{J}_x$, entries at the $(2k+1)^{th}$ locations are all zero, and similarly for the $2k^{th}$ locations of $\mathbf{J}_y$. This simple form, which affords a fast calculation of the Hessian and gradient, is a result of optimising directly over the landmark locations.

The parameter updates of the Gauss-Newton optimisation of Equation (11) then takes the following form:

$$\begin{bmatrix} \Delta\mathbf{p} \\ \Delta\mathbf{q} \end{bmatrix} = - \left[\mathbf{H}_d + \eta \begin{bmatrix} \mathbf{H}_s & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \right]^{-1} \left[\mathbf{g}_d + \eta \begin{bmatrix} \mathbf{g}_s \\ \mathbf{0} \end{bmatrix} \right] \tag{26}$$

where

$$\mathbf{g}_d = \sum_{\mathbf{x} \in \Omega} \varrho'(E(\mathbf{x})^2) \mathbf{J}_d(\mathbf{x})^T E(\mathbf{x}) \tag{27}$$

$$\mathbf{g}_s = \sum_{i,j}^{n} k_{ij} \left[\mathbf{J}_x(i,j)^T \mathbf{d}_x(i,j) + \mathbf{J}_y(i,j)^T \mathbf{d}_y(i,j) \right] \tag{28}$$

are the gradients of the data and smoothness term respectively.

The optimisation process can usually be sped-up by using the inverse compositional formulation (Matthews & Baker 2003). By reversing the roles of the model and the image in the data term, the gradients of the data term can be precomputed and hence a large proportion of computation needs to be done only once. The extensions of the inverse compositional image alignment (ICIA) algorithm to robust error norms was proposed in (Baker et al. 2003). With this formulation, the Hessian of the data term cannot be precomputed, despite the fixed gradients, as the derivative of the robust error terms cannot be precomputed. Although an efficient approximation has been derived by assuming spatial coherence of the outliers, this implementation is not particularly effective for automatic model building from databases as the images are generally occlusion free, with outliers stemming mainly from misalignment, image noise, changes in texture not yet accounted for by the texture model, and interlacing effects. The presence of the smoothness term means that the Hessian needs to be updated and inverted at every iteration which is the most costly part of the optimisation when there is a large number of landmark points. Furthermore, for the methods described in Section 4, the model is updated after every image, requiring the gradients to be recomputed. Due to these factors, we predict that using the ICIA formulation will not give dramatic improvements in efficiency and hence it was not incorporated into our implementation.

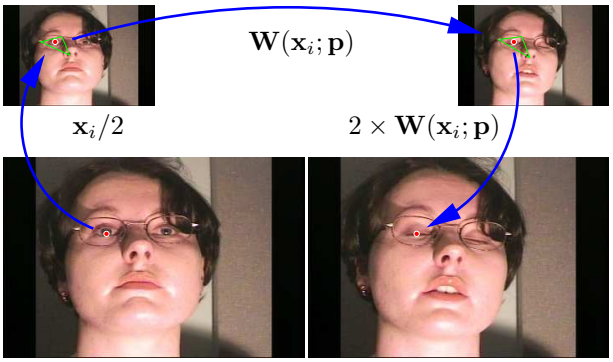

Figure 2: Initialising points in lower levels of the Gaussian pyramid. Top row: Warps at higher pyramid level. Bottom: Landmarks at current pyramid level.

3.4.4 Gaussian Pyramid

Despite the use of the smoothing term, the optimisation process may still converge to a local minimum due to the high dimensionality of the problem. This problem can be partially alleviated by optimising on a Gaussian-pyramid.

There are issues however with regards to how the shape is parametrised between the levels of the pyramid. A pseudo-dense correspondence at the lowest level of the pyramid may result in an over parametrised model at the highest level of the pyramid, which results both in a slow alignment process as well as the higher likelihood of getting stuck in local minima. Instead, in this work we build a separate model for each level. Starting at the highest pyramid level, a set of landmarks is chosen as described in Section 3.3.1. With this, an automatic model building process described in Section 4 is performed. Moving down the pyramid, a new set of landmarks is chosen from the reference image.

The propagation of these landmarks to other images is illustrated in Figure 2. First the landmarks are downscaled to the previous pyramid level (bottom to top left in Figure 2). Then the landmarks are warped using the found correspondence for that level (top row), and finally up-scaled back to the current pyramid level (top to bottom right).

With the smoothness term described in Section 3.4.2, the use of the Gaussian pyramid allows a stiff regularisation parameter η to be used as the movements of points at every level will be relatively small. This in turn allows the optimisation process to better avoid local minima.

4 Incremental Model Building

Most approaches to automatic model building can be classed as groupwise, where a model is iteratively refined from an initial estimate by first fitting it to each image, followed by a reconstruction of a new model from the fitted images. One of the drawbacks of this approach is that it does not take into account the sequential nature of images in video. As such, its initial estimate of the model may be far from the optimum, which may cause the algorithm to converge slowly or get stuck in local minima.

By assuming that the appearance of the visual objects varies slowly between consecutive frames in a

sequence, the model building process can be posed as a tracking problem. Although the complexity of the warping function is much higher than most tracking problems, which generally solve only for a similarity or affine transform, the same mechanisms apply. We start with an initial template, without loss of generality taken as the first image in the sequence, and propagate the landmark positions to the other images in the sequence through a consecutive alignment process. Unlike typical tracking problems however, due to the high dimensionality of the parameter space, the alignment process must generally utilise gradient based approaches as non-gradient methods such as a particle-filters will be too computationally expensive to evaluate.

One of the main difficulties associated with template tracking is due to the changes in the object's texture throughout the sequence. Although this problem can be partially alleviated by using a robust error function, as the sequence progresses the object's texture may undergo significant changes such that treating them as outliers may lead to misalignment. One solution to this problem is to update the template using the texture from the previous frame. However, simply replacing the texture with the most recent image makes the algorithm prone to drifting. In this work, we investigate the utility of two adaptable template approaches for automatic model building from image sequences.

4.1 Method 1: Grounded Templates

There are a number of approaches to the template update problem which reduce the drifting phenomenon, for example (Matthews, Ishikawa & Baker 2004) (Zhong, Jain & Dubuisson-Jolly 2000) (Loy, Goecke, Rougeaux & Zelinsky 2000). In this work we follow the approach of (Loy et al. 2000), where the new template is defined as a weighted combination of the initial template and the texture from the most recent image:

$$T_t(\mathbf{x}) = \alpha T_0(\mathbf{x}) + (1-\alpha)T_{t-1}(\mathbf{x}) \tag{29}$$

The parameter $\alpha \in (0 \dots 1)$ is a grounding factor which reduces the drifting effect whislt allowing the template to adapt to the current object's texture.

As the template is updated once before the alignment process in the next image, the optimisation process needs only be done over the landmark locations. Therefore, the Jacobian of the data term in Equation (22) for this method is simply:

$$\mathbf{J}_d(\mathbf{x}) = -\nabla I(\mathbf{W}(\mathbf{x};\mathbf{p}))\frac{\partial \mathbf{W}(\mathbf{x};\mathbf{p})}{\partial \mathbf{p}} \tag{30}$$

and the Gauss-Newton update in Equation (26) is now given by:

$$\Delta \mathbf{p} = -\left[\mathbf{H}_d + \eta \mathbf{H}_s\right]^{-1}\left[\mathbf{g}_d + \eta \mathbf{g}_s\right] \tag{31}$$

The output of the template matching algorithm is a set of corresponding annotations in every image in the sequence, from which an appearance model can be built in the usual manner.

4.2 Method 2: Incremental Texture Learning

One of the weaknesses of the template update approach is that it takes into account only the initial and most recently encountered textures. As such it makes no use of the knowledge of the variations in texture which have been encountered earlier in the sequence. One possibility to incorporate this information is to perform an incremental model building process as the object is tracked throughout the sequence.

For this algorithm we utilise incremental PCA (Li 2004) to update the model, rather than the template, after matching to every new image. Starting with the template of the first image, we match it to the next image using the approach described in Section 4.1. Some of the variations captured as outliers may in fact be intrinsic variations of the object rather than just image noise. The texture of the newly aligned image is then used as a new data instance for the linear model, for which incremental PCA is used to integrate it into the model. The amnesic factor (a weighting between the current model and the new data instance) is set to $\frac{n}{1+n}$, where n is the number of samples used to build the current model, so that every sample integrated into the model is given the same importance. See (Li 2004) for details.

Once the model exhibits some linear modes of variation apart from the mean, matching to the next image should be done by simultaneously updating the landmark locations and the texture parameters $\mathbf{q}$ using the update equations described in Section 3.4.3. This way, images which exhibit texture variations previously encountered in the sequence will be matched better than using a fixed template. Again, the data term is formulated using the robust error function to account for texture variations not yet encountered in the sequence.

5 Experiments

5.1 The AVOZES Database

AVOZES (Goecke & Millar 2004), the Audio-Video Australian English Speech data corpus, is a database of 20 speakers uttering a variety of phrases which was designed for research on the statistical relationship of audio and video speech parameters with an audio-video automatic speech recognition task in mind. Although sparse annotations for the vital mouth points, such as lip corners, are available, these points are chosen manually and represent only a heuristic intuition about their usefulness for automatic speech recognition. A more elaborate set of cues may be useful for audio-video speech recognition which may not be directly obvious. AAMs, which encode both a pseudo-dense set of landmark points as well as texture variations, provide a rich set of features to a speech recognition system which may allow better recognition rates to be achieved. An intensive study of the application of AAMs in this domain can be found in (Neti, Potaminos, Luettin, Matthews, Glotin & Vergyri 2001).

In our experiments we used the continuous speech sequences for each of the speakers exclusively. The continuous speech part of AVOZES consists of three sequences, each with a different phrase. The length of all sequences range from 90 to 150 frames. As the video files in the database consist of a stereo pair, warped to half the height, we used only part of the sequence pertaining to the left camera, which we scaled to the true ratio.

For each of the speakers, we performed both of the image based correspondence methods described in Section 4 on all three sequences together. Since there may be large differences between the start and end of different sequences of the different phrases, we find the image which is most similar in the later two sequences to an image in the first sequence. After tracking through the first sequence the model is tracked in the other sequence starting from the most similar image found previously, initialising the shape estimate to the corresponding image in the first sequence. The

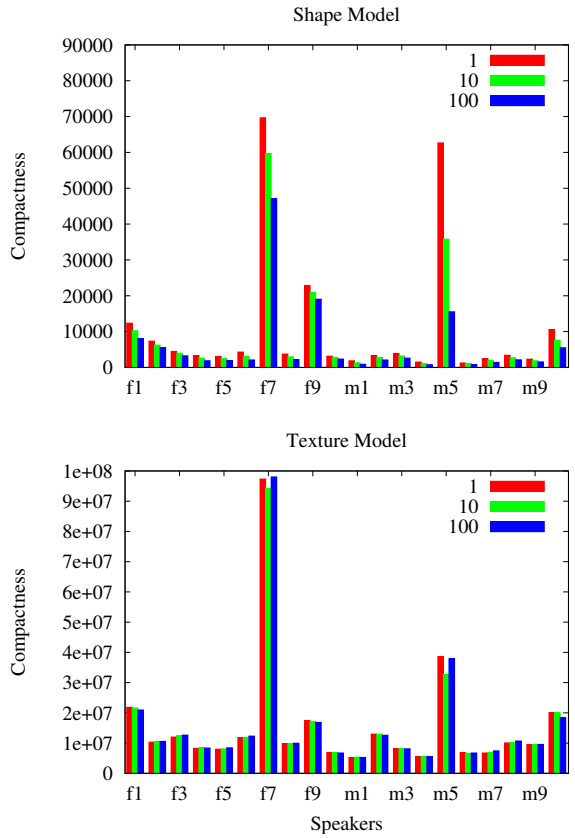

Figure 3: Shape and texture model compactness for every speaker in AVOZES. The models were built from correspondences found using the grounded template method with three settings of the regularisation parameter $\eta = \{1, 10, 100\}$.

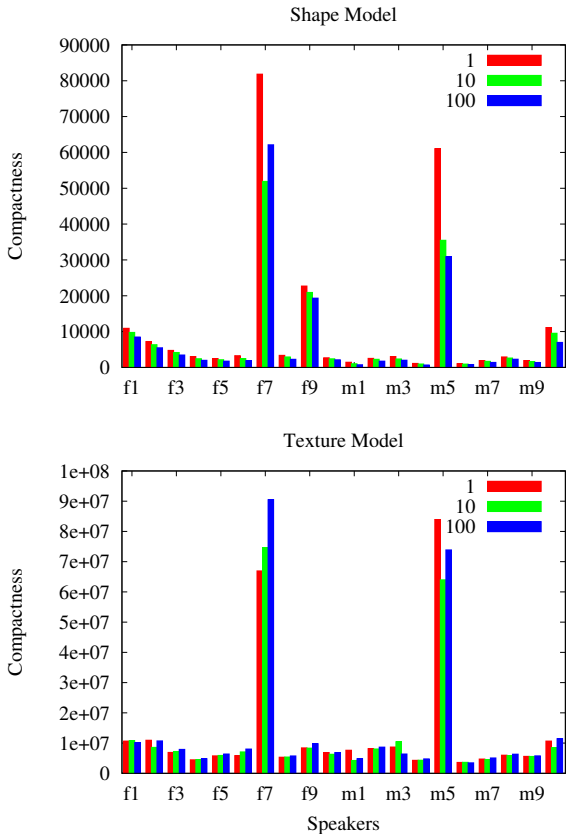

Figure 4: Shape and texture model compactness for every speaker in AVOZES. The models were built from correspondences found using the incremental texture learning method with three settings of the regularisation parameter $\eta = \{1, 10, 100\}$.

tracking process in these other sequences is performed forwards and backwards until the beginning and end of the sequences respectivley. From the resulting correspondences, the compactness of the shape and texture models are calculated as described in Section 3.2. The experiments were repeated for a number of settings of the smoothing parameter η.

5.2 Results

In Figure 3 and 4, histograms of the shape and texture model compactness of each of the speakers in the AVOZES database built from correspondences obtained using the methods described in Section 4.1 are shown for three different settings of the regularisation parameter η. Comparing the two methods, the shape compactness differs little between them. The main difference lies in the texture compactness, where the incremental texture learning method generates models which are around twice as compact for most speakers compared to the grounded template method. As discussed in Section 4.2, this result is expected as the incremental texture learning retains memory of previously encountered texture variations. Also, as the alignment process may contain errors which may accumulate throughout the sequence, this approach is more constrained to valid texture instances rather than just the first and most recently encountered texture, which may be erroneous.

Studying each method independently, as expected the compactness of the shape model improves as η is increased. Perhaps somewhat more surprisingly, the texture model's compactness is effected little by the different settings of the regularisation parameter. We attribute this to the fact that the texture model is evaluated in a reference frame. The effect of this is that for groups of landmarks which correspond to *flat* parts of the image, their movements contribute little to the change in the texture when projected onto the reference frame. As such, shapes with significantly different landmark locations in these flat regions may result in very similar texture. An example of this is shown in Figure 5. Landmarks in flat regions are more likely to be perturbed by image noise and hence, for the same texture compactness, the model with better shape compactness is generally a better model.

From the correspondences in each image, found using the incremental texture learning method with $\eta = 100$, we built a combined appearance model (see Section 3.1) using every 10^{th} image in the sequences. The mean and first mode of variation on all speakers are shown in Figure 7 and 8. Although the correspondences appear to be of high quality in most speakers, observed through the *crispness* of the images, there are a few for which the tracking method seemed to have failed to obtain the correct correspondences across the sequences. In particular, the f7 and m5 speakers are particularly poor, where the first mode of variation seems to entail the presence or disappearance of visual artefacts. Referring to the texture compactness histograms in Figure 3 and 4 it can be seen that these two speakers exhibit the least compact model out of the database by a significant margin.

It is clear that in these cases, the tracking process used to find the correspondences failed significantly in parts of the image, resulting in the texture model needing to account for variations due to misalignment rather than intrinsic texture variations of

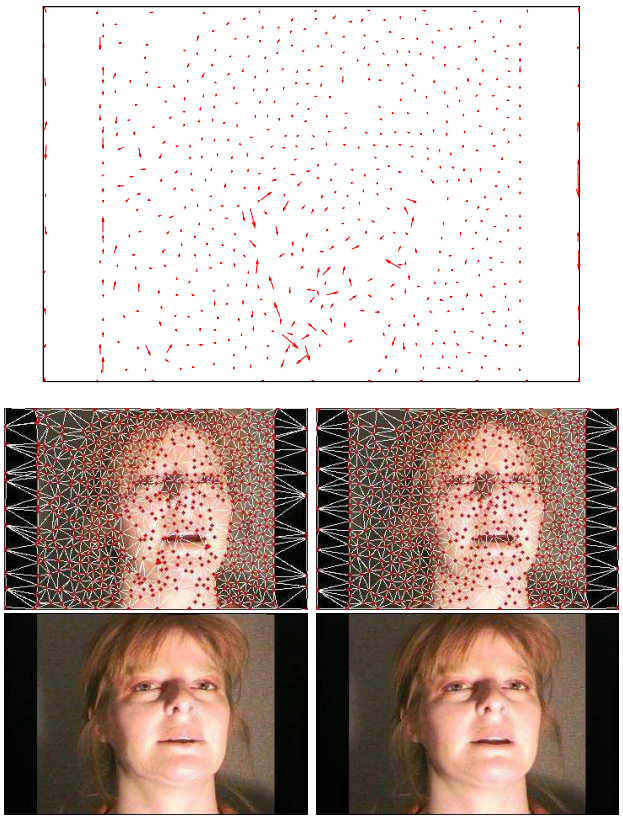

Figure 5: Two shapes with significant landmark differences in *flat* regions exhibiting similar texture when projected to the reference frame. Top: Shape difference. Middle: Shapes of two images. Bottom: Texture projections onto the reference image.

the speaker. On closer inspection, we found that these two speakers exhibited significant motion towards the camera during some parts of the sequences. Example images from these sequences are shown in Figure 6. As such, significant parts of the background are occluded when the speakers are close to the camera, but reappear when they are further from it. Because the background exhibits some strong texture and colour variations (see the white strip behind the speaker's heads), the disappearance/emergence of these areas perturb the alignment process significantly, despite using a robust error function.

As models of the other speakers, which exhibit relatively small amounts of head movement, were able to be built compactly, we suspect that databases which exhibit a uniform background to not exhibit this problem. However, in cases where this is not practical, one solution would be to initialise the feature points within the face region exclusively, either using a manual crop in the first image or using some type of skin colour detector. It should be noted however, that the accuracy of the alignment around the peripheral of the face using this approach may be inferior to that which encodes background.

As a final note, although the methods tested here have shown to give reasonably compact models when no significant visual artifacts disappear or emerge throughout the sequence, because the correspondences are obtained in a pairwise manner the model quality may be improved through a groupwise method. In fact, the methods discussed in this work can be used as a good initialisation for groupwise methods which will encourage faster convergence and help avoid local minima.

Figure 6: Images from the f7 and m5 speakers which illustrate the large differences in scale affecting content in the images.

6 Conclusion

In this work, we have investigated the utility of adaptive tracking methods for automatically building pseudo-dense correspondences across a sequence of a deformable object, with an AV database as a test case. We compared two methods, the grounded template and incremental texture learning method, measuring their performance through a shape and texture compactness measure as well as a qualitative analysis of the resulting linear models of variation.

Through extensive experiments we have shown that this approach can be used to build highly compact models of a linearly deforming object which includes the background in the image. We also found that if the background exhibits significant texture, despite being static, movements of the object which causes these textured regions to be occluded or new textured regions to appear later in the sequence, significantly degrades the performance of this method. However, we suspect that this is a problem exhibited by most image based correspondence methods which utilise diffeomorphic warps and do not explicitly model the disappearance or emergence of visual artifacts.

Future work on extending this method might involve investigations into efficiency gains of using the inverse compositional formulation, evaluating alignment error in the image rather than reference frame, and extensions to incremental shape model learning. Although the methods investigated in this paper and their possible extensions allow significant savings on human intervention, requiring only one manual markup per speaker, for large databases containing thousands of sequences this approach may be infeasible. The much more difficult problem of finding correspondences across sequences of different instances of the same object class (different speakers in AVOZES, for example) remains an open problem.

References

Baker, S., Gross, R. & Matthews, I. (2003), Lucas-Kanade 20 years on: A unifying framework: Part 3, Technical report, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

Baker, S. & Matthews, I. (2002), Lucas-kanade 20 years on: A unifying framework: Part 1, Technical Report CMU-RI-TR-02-16, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

Baker, S., Matthews, I. & Schneider, J. (2004), 'Automatic construction of active appearance models as an image coding problem', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(10), 1380–1384.

Black, M. & Anandan, P. (1993), The robust estimation of multiple motions: Affine and piecewise-smooth flow fields, Technical report, Xerox PARC.

Blake, A., Isard, M. & Reynard, D. (1994), Learning to track curves in motion, *in* 'IEEE Conference on Decision Theory and Control', pp. 3788–3793.

Brox, T., Bruhn, A., Papenberg, N. & Weickert, J. (2004), High accuracy optical flow estimation based on theory of warping, *in* T. Pajdla & J. Matas, eds, '8th European Conference on Computer Vision', Vol. 4, Springer-Verlag, Paragues, Czech Republic, pp. 25–36.

Chui, H., Win, L., Schultz, R., Duncan, J. S. & Rangarajan, A. (2003), 'A unified non-rigid feature registration method for brain mapping', *Medical Image Analysis* **7**(2), 113–130.

Cootes, T. F., Edwards, G., Taylor, C. J., Burkhardt, H. & Neuman, B. (1998), Active appearance models, *in* 'European Conference on Computer Vision', Vol. 2, pp. 484–489.

Cootes, T. F., Marsland, S., Twining, C. J., Smith, K. & Taylor, C. J. (2004), Groupwise diffeomorphic non-rigid registration for automatic model building, *in* 'European Conference on Computer Vision', pp. 316–327.

Cootes, T. F., Twining, C. J., Petrovic, V., Schestowitz, R. & Taylor, C. J. (2005), Groupwise construction of appearance models using piece-wise affine deformations, *in* 'British Machine Vision Conference', Vol. 2, pp. 879–888.

Edwards, G., Taylor, C. J. & Cootes, T. F. (1998), Interpreting face images using active appearance models, *in* 'IEEE International Conference on Automatic Face and Gesture Recognition', pp. 300–305.

Goecke, R. & Millar, J. B. (2004), The audio-video australian english speech data corpus avozes, *in* '8th International Conference on Spoken Language Processing INTERSPEECH 2004 - ICSLP', Vol. III, ISCA, Jeju, Korea, pp. 2525–2528.

Hill, A. & Taylor, C. J. (1996), A method of non-rigid correspondence for automatic landmark identification, *in* '7th British Machine Vision Conference', Vol. 2, pp. 323–332.

Jebara, T. (2003), Images as bags of pixels, *in* 'International Conference on Computer Vision', pp. 265–272.

Lehn-Schiøler, T., Hansen, L. K. & Larsen, J. (2005), Mapping from speech to images using continuous state space models, *in* 'Lecture Notes in Computer Science', Vol. 3361, Springer, pp. 136 – 145.

Li, Y. (2004), 'On incremental and robust subspace learning', *Pattern Recognition* **37**(7), 1509–1518.

Loy, G., Goecke, R., Rougeaux, S. & Zelinsky, A. (2000), Stereo 3D lip tracking, *in* '6th International Conference on Control, Automation, Robotics and Vision', Singapore.

Matthews, I. & Baker, S. (2003), Active appearance models revisited, Technical Report CMU-RI-TR-03-02, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA.

Matthews, I., Ishikawa, T. & Baker, S. (2004), 'The template update problem.', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26**(6), 810–815.

Mittrapiyanuruk, P., DeSouza, G. N. & Kak, A. C. (2005), Accurate 3D tracking of rigid objects with occlusion using active appearance models, *in* 'IEEE Workshop on Motion and Video Computing', pp. 90–95.

Neti, C., Potaminos, G., Luettin, J., Matthews, I., Glotin, H. & Vergyri, D. (2001), Large-vocabulary audio-visual speech recognition: A summary of the john hopkins summer 2000 workshop, *in* 'Workshop on Multimedia Signal Processing (MMSP)', Cannes.

Sawhney, H. S. & Ayer, S. (1996), 'Compact representation of videos through dominant and multiple motion estimation', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **18**(8), 814–830.

Schestowitz, R. S., Twining, C. J., Petrovic, V. S., Cootes, T., Crum, B. & Taylor, C. J. (2006), Non-rigid registration assessment without ground truth, *in* 'Medical Image Understanding and Analysis', Vol. 2, pp. 151–155.

Stegmann, M. B. & Larsson, H. B. (2003), Fast registration of cardiac perfusion MRI, *in* 'International Society of Magnetic Resonance In Medicine', Toronto, Canada, p. 702.

Walker, K. N., Cootes, T. F., & Taylor, C. J. (1999), Automatically building appearance models from image sequences using salient features, *in* D. T.Pridmore, ed., 'British Machine Vision Conference', Vol. 2, pp. 463–562.

Zhong, Y., Jain, A. K. & Dubuisson-Jolly, M. P. (2000), 'Object tracking using deformable templates', *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**(5), 544–549.

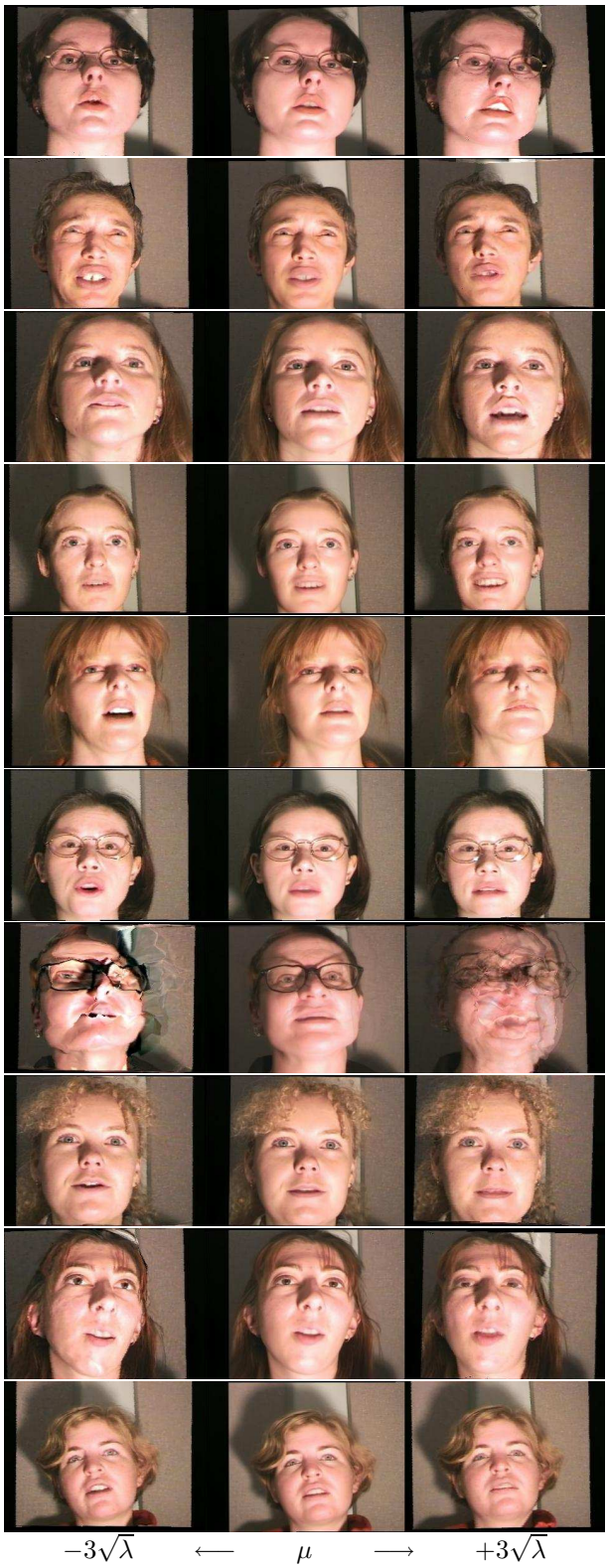

Figure 7: The first mode of variation of the female speakers in AVOZES, varied between ± three standard deviations.

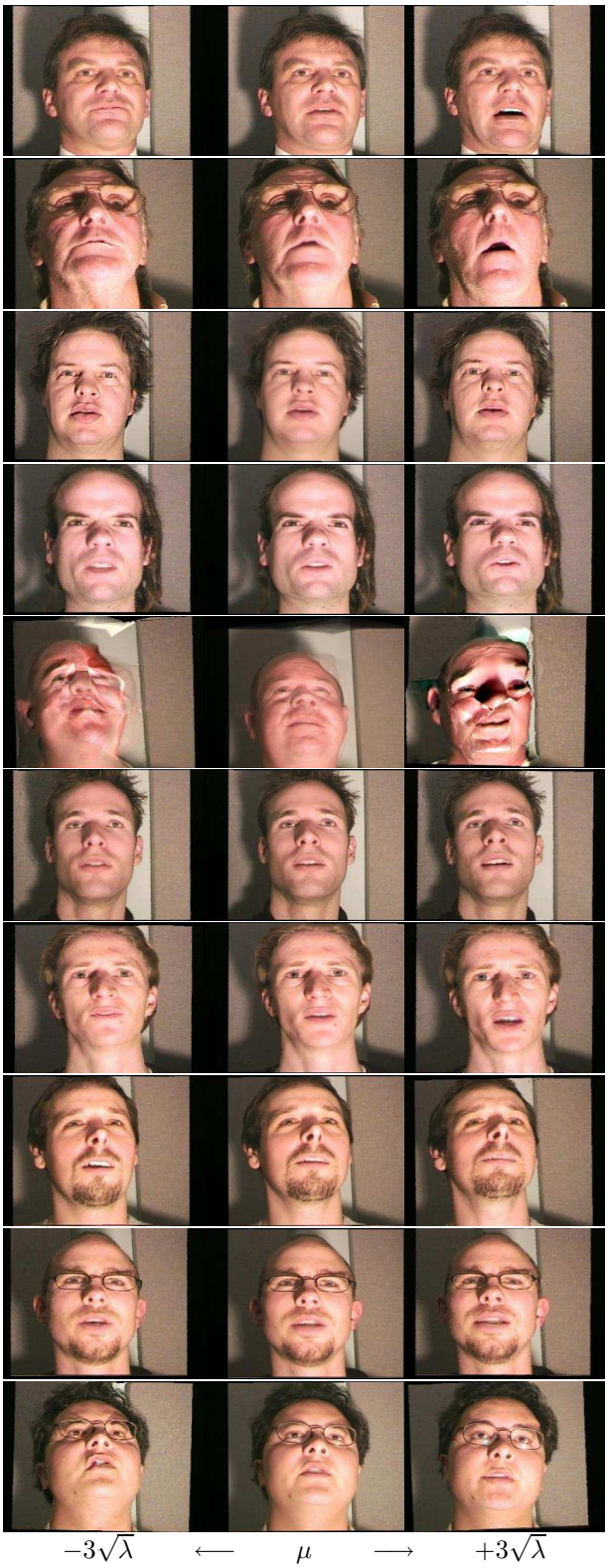

Figure 8: The first mode of variation of the male speakers in AVOZES, varied between ± three standard deviations.