Facial Performance Transfer via Deformable Models and Parametric Correspondence

Akshay Asthana, *Student Member, IEEE,* Miles Delahunty,
Abhinav Dhall, *Student Member, IEEE,* and Roland Goecke, *Member, IEEE*

Abstract—The issue of transferring facial performance from one person's face to another's has been an area of interest for the movie industry and the computer graphics community for quite some time. In recent years, deformable face models, such as the Active Appearance Model (AAM), have made it possible to track and synthesise faces in real-time. Not surprisingly, deformable face model based approaches for facial performance transfer have gained tremendous interest in the computer vision and graphics community. In this paper, we focus on the problem of real-time facial performance transfer using the AAM framework. We propose a novel approach of learning the mapping between the parameters of two completely independent AAMs, using them to facilitate the facial performance transfer in a more realistic manner than previous approaches. The main advantage of modelling this parametric correspondence is that it allows a "meaningful" transfer of both the non-rigid shape and texture across faces irrespective of the speakers' gender, shape and size of the faces, and illumination conditions. We explore linear and non-linear methods for modelling the parametric correspondence between the AAMs and show that the sparse linear regression method performs the best. Moreover, we show the utility of the proposed framework for a cross-language facial performance transfer that is an area of interest for the movie dubbing industry.

Index Terms—Active Appearance Models, Facial Performance Transfer, Face Modelling and Animation.

1 Introduction

Motion capture and the transfer of facial performance from an actor to a CGI-generated character have long been a focus of much research in the movie industry and computer graphics community. The aim is to copy the source's facial movements as truly as possible, while presenting them in a realistic manner on the animated target character. This is extremely difficult as humans are extremely sensitive to any unnatural occurrence on the face and can easily spot even the minutest misalignment or texture irregularity on the face.

While a fully-automatic solution capable of producing realistic looking facial performance transfer has not been realised yet, recent technological advances in the last decade have enabled markerless solutions for face modelling and tracking resulting in a wide selection of possibilities for semi-automatic and semi-supervised solution to this problem. Therefore, rather than expecting an animator to generate realistic looking facial animation on frame by frame basis from scratch in a painstaking and tedious manual manner, these methods have provided them with a tool to manipulate the facial features and obtain close to realistic looking results in a semi-automatic setting, enabling them to spend more quality time on improving the results and post-processing to achieve excellent results.

Recently, various approaches such as [1], [2], [3], [4], [5], [6] have been proposed for facial performance transfer/cloning. A method for creating photorealistic 3D facial models from pictures of human subjects was proposed in [1]. They employ manual facial pose marking and use a 3D shape morphing algorithm among the face models, but the computational cost is high due to expensive 3D calculations. Others have used an expression ratio image (ERI) [3], the limitation of which is in dealing with illumination changes. A 3D morphable model (3DMM) based approach [4] has been used for animating novel faces by transferring mouth movements and expressions. It achieves good accuracy for both pose and illumination, but at the expense of higher computational complexity. In a different approach to transfer just the lip movements [5], a multidimensional morphable model [7], trained on a large dataset of a speaker, is adapted to animate the lips of another speaker that requires only a small dataset for training.

A different application of facial performance transfer has also been used to improve the facial attractiveness of a target face based on a nearest matching example face [8]. In expression cloning [2], the vertex motion vectors are transferred from the source face to the target model. In this, the dense correspondences among the source and target models are created based on an initial manual selection of corresponding vertices. Then, the 3D motion vectors of the source model are used to create similar animations on the target model. Vlasic et al. [9] used a 3D multilinear framework but their models did not include a texture component or movement of the eyes, teeth, tongue, chin or cheeks, which we model in our method.

In recent years, statistical approaches, such as the Active Appearance Model (AAM) [10] and 3D Morphable Model (3DMM) [4], have been widely and successfully used for building non-rigid deformable models. Their power lies in the combination of a compact parametric representation and

- *A. Asthana and A. Dhall are with the College of Engineering & Computer Science, Australian National University, Canberra, ACT, Australia. Email: aasthana@rsise.anu.edu.au, abhinav.dhall@anu.edu.au*
- *M. Delahunty is with the College of Business & Economics, Australian National University, Canberra, ACT, Australia. Email: miles.dlh@gmail.com*
- *R. Goecke is with the Faculty of Information Sciences and Engineering, University of Canberra, Canberra, ACT, Australia. Email: roland.goecke@ieee.org*

an efficient alignment method. Recently, Theobald et al. [6] proposed to compute a linear mapping between the basis vectors of the shape and texture of two AAMs and used this mapping for facial performance transfer.

Here, we approach the task of facial performance transfer by directly modelling the parametric correspondence between shape and texture of two completely independent AAMs.

- We propose a novel regression based approach to model the relationship between the shape and texture parameters of two completely independent face models, enabling "*meaningful transfer*" of the variation in both shape and texture. "Meaningful" here means taking the personal characteristics into account. For example, if the source subject exhibits large facial movements while the target subject normally shows little, it would look unrealistic if the source's large movements were transferred *verbatim* to the target face. Our approach facilitates realistic facial performance transfer between two subjects, irrespective of their gender, shape of their face or their skin tone.
- We explore several regression strategies to find the most effective way of modelling the relationship between the shape and texture parameters of the two face models. To evaluate the proposed approach, we perform experiments on the subjects of the AVOZES data corpus [11].
- We demonstrate the utility of the approach for *Cross-Language Facial Performance Transfer* by transferring the facial performance from a female Indian subject speaking *Hindi* (national language of India) to a male Australian subject present in the AVOZES data corpus.

2 Motivation and Approach

In this paper, some familiarity with Active Appearance Models (AAM) [10] is assumed. Refer to the supplementary material[3] for an overview of the AAM framework.

The central problem we address in this paper is that of automatically transferring subtle changes in facial features, such as those induced by speech or affect, from one person to another, irrespective of gender, shape of face and skin tone. At the same time, we aim to preserve person-specific qualities (e.g. the amount and manner in which the subject opens and closes the lips while speaking) in order to transfer the facial performance with a high degree of realism.

The advantages of using deformable face models (such as AAMs) to approach this problem are

- Deformable face models provide a compact framework to model and manipulate both shape and texture together and, hence, are well suited for the task of synthesising realistic looking results.
- Person-specific deformable face models, if trained on a sufficient number of images, can generalise well to unseen expressions and can also capture person-specific mannerisms in great detail.

For the experiments, we used the Simultaneous Inverse Compositional (SIC) fitting method due to its robustness in a person-specific scenario [12]. The central problem can, therefore, be modified as follows: *Given two completely independent, person-specific AAMs of two subjects, say $\mathcal{M}_1$ and $\mathcal{M}_2$, how can we convincingly transfer the facial performance observed in the face of the first subject onto the model of the second subject?*

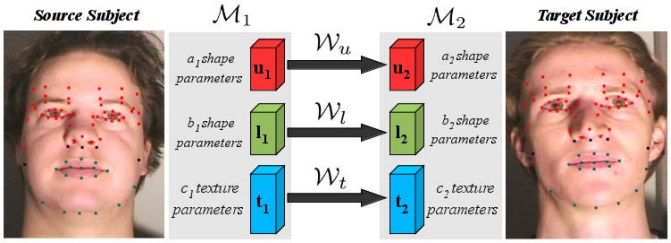


Fig. 1. Overview (Points in Red - Upper Face shape, Points in Green - Lower Face shape, Points in Blue - Shared by both.)

One intuitive way is to use model $\mathcal{M}_1$ to track the facial performance of the first subject, extract shape and texture parameters, and transfer these parameters to the model $\mathcal{M}_2$. However, the transfer of these parameters from $\mathcal{M}_1$ to $\mathcal{M}_2$ is non-trivial [6]. If the relationship between the shape and texture parameters of $\mathcal{M}_1$ and $\mathcal{M}_2$ was one-to-one, then we could directly apply the parameters from $\mathcal{M}_1$ to $\mathcal{M}_2$. However, in reality, it is highly unlikely that two independent AAMs, trained on completely different sets of images, will have a direct one-to-one parametric correspondence. This is due to the application of PCA to different, independent image sets in the model building phase. For example, the first and second shape parameters of $\mathcal{M}_1$ might control the landmark points representing the mouth region, whereas the same region in $\mathcal{M}_2$ might be controlled by the 3^{rd}, 4^{th} and 5^{th} shape parameters. We propose to model this parametric correspondence between the shape and texture parameters of $\mathcal{M}_1$ and $\mathcal{M}_2$ in a data driven regression framework.

Our approach is illustrated in Figure 1. Since a major focus of this work is on the transfer of accurate lip movements induced by speech, we consider two separate shape models; one for the upper part of the face ($\mathbf{u}$ represents the shape parameter vector) and one for the lower part of the face ($\mathbf{l}$ represents the shape parameter vector). This process is based on the *hypothesis* that the majority of movement induced by speech is limited to the lower part of the face, while changes in the upper part of the face are more subtle and can be treated independently. On the other hand, we consider the entire texture of the face as a whole and, hence, use a single texture model ($\mathbf{t}$ texture parameters). Both $\mathcal{M}_1$ and $\mathcal{M}_2$ can have any number of shape and texture parameters, depending on their training sets. The main goal here is to find a mapping function between these sets of shape and texture parameters and to use it to approach the central problem. Given a set of correspondence images, we pose this problem in a data-driven regression framework and compute the mapping functions between the upper-face shape parameters ($\mathcal{W}_u$), the lower-face shape parameters ($\mathcal{W}_l$) and the texture parameters ($\mathcal{W}_t$). Using these learnt regressors, we can then directly transfer the facial performance from $\mathcal{M}_1$ to $\mathcal{M}_2$.

3 Facial Performance Transfer Method

Let $\mathcal{M}_1$ and $\mathcal{M}_2$ be two completely independent AAMs. Let $\mathbf{u}_1$ represent the upper-face shape parameter vector (length a_1)

of $\mathcal{M}_1$; $\mathbf{u}_2$ represent the upper-face shape parameter vector (length a_2) of $\mathcal{M}_2$; $\mathbf{l}_1$ represent the lower-face shape parameter vector (length b_1) of $\mathcal{M}_1$; $\mathbf{l}_2$ represent the lower-face shape parameter vector (length b_2) of $\mathcal{M}_2$; $\mathbf{t}_1$ represent the texture parameter vector (length c_1) of $\mathcal{M}_1$; and $\mathbf{t}_2$ represent the texture parameter vector (length c_2) of $\mathcal{M}_2$.

Given a small set of correspondence images of both subjects, the goal here is to learn the mapping between their shape and texture parameters. We compute the mapping functions $\mathcal{W}_u : \mathbf{u}_1 \rightarrow \mathbf{u}_2$, $\mathcal{W}_l : \mathbf{l}_1 \rightarrow \mathbf{l}_2$ and $\mathcal{W}_t : \mathbf{t}_1 \rightarrow \mathbf{t}_2$ (Figure 1).

Let M be the number of correspondence images, $\mathbf{p}$ be the parameter vector of length a (shape or texture) of the source subject and $\mathbf{p}'$ be the parameter vector of length b (shape or texture) of the target subject. In the following subsections, we briefly discuss different regression strategies of finding the mapping function $\mathcal{W} : \mathbf{p} \rightarrow \mathbf{p}'$ explored in this paper.

Standard Linear Regression Method: Considering it as a baseline method, we solve the problem by assuming $\mathbf{p}'$ to be a linear function of $\mathbf{p}$

$$\mathbf{Y}_L = \mathcal{W}_L \mathbf{X}_L + \nu_L \tag{1}$$
$$\mathbf{X}_L = [\mathbf{p}_1 \quad \dots \quad \mathbf{p}_M] \text{ and } \mathbf{Y}_L = [\mathbf{p'}_1 \quad \dots \quad \mathbf{p'}_M]$$

$\mathcal{W}_L$ is the unknown mapping function that we wish to find and ν_L is the noise term. Hence, this becomes a standard L_2-regularised least squares problem, which we solve for $\mathcal{W}_L$ by

$$\text{minimizing} \left\{ \|\mathbf{Y}_L - \mathcal{W}_L \mathbf{X}_L\|_2^2 + \lambda_L \|\mathcal{W}_L\|_2^2 \right\} \tag{2}$$

where $\|g\|_2 = (\sum_i g_i^2)^{1/2}$ is the L_2-norm of g and $\lambda_L > 0$ is a regularisation factor used to avoid over-fitting. Solving Eq. 2 for $\mathcal{W}_L$, we get

$$\mathcal{W}_L = \mathbf{Y}_L \mathbf{X}_L^T (\mathbf{X}_L \mathbf{X}_L^T + \lambda_L \mathbf{I})^{-1} \quad . \tag{3}$$

Hence, the mapping function $\mathcal{W}_L$ is a matrix of size $b \times a$.

Sparse Linear Regression Method: Assuming $\mathbf{p}'$ to be a linear function of $\mathbf{p}$, the mapping function $\mathcal{W}$ is computed by solving a L_2-regularised least squares problem (Eq. 2). This gives a very dense solution for $\mathcal{W}_L$ (Eq. 3), i.e. typically all the elements of matrix $\mathcal{W}_L$ are non-zero [13]. The goal here is to obtain a sparse solution for $\mathcal{W}$, i.e. a very few number elements of matrix $\mathcal{W}$ should be non-zero.

In [14], it has been shown that if there exists an optimal *sparse solution*, it can be efficiently computed by convex optimisation. Hence, we recast the problem as a L_1-regularised least squares problem, which can be reformulated as a convex quadratic program and then solved efficiently to give a sparse solution $\mathcal{W}_S$. With the mapping function $\mathcal{W}_S \in \mathbf{R}^{a \times b}$

$$\mathbf{X}_S = [\mathbf{p}_1^T; ....; \mathbf{p}_M^T] \text{ and } \mathbf{Y}_S = \left[\mathbf{p'}_1^T; ....; \mathbf{p'}_M^T\right]$$

Consider the linear model

$$\mathbf{y}_i = \mathbf{X}_S \mathbf{w}_i + \nu_{Si} \qquad i = 1, \dots, b \tag{4}$$

where vectors $\mathbf{y}_i$ and $\mathbf{w}_i$ are the i^{th} columns of matrices $\mathbf{Y}_S$ and $\mathcal{W}_S$, respectively, and ν_S is the noise term. We determine $\mathbf{w}_i$ by solving a L_1-regularised least squares problem

$$\min \|\mathbf{X}_S \mathbf{w}_i - \mathbf{y}_i\|_2^2 + \lambda_S \|\mathbf{w}_i\|_1 \tag{5}$$

where $\|g\|_1 = \sum_j |g_j|$ is the L_1-norm of a and $\lambda_S > 0$ is a regularisation factor. Eq. 5 can be reformulated as a convex quadratic program [13] with linear inequality constraints:

$$\min \|\mathbf{X}_S \mathbf{w}_i - \mathbf{y}_i\|_2^2 + \lambda_S \sum_{j=1}^{a} u_j \tag{6}$$
$$\text{subject to } -u_j \leq w_j \leq u_j \text{ , where } j = 1, \dots, a \text{ and } u \in \mathbf{R}^a.$$

For the experiments presented in this paper, we solve this convex quadratic program (Eq. 6) by a specialised interior-point method [13] that uses the preconditioned conjugate gradients algorithm to compute the search direction. This results in a sparse solution for $\mathbf{w}_i$. Hence, the mapping function $\mathcal{W}_S = [\mathbf{w}_1 \, \mathbf{w}_2 \dots \mathbf{w}_b]$ is a sparse matrix of size $a \times b$.

Non-Linear Regression Method: Assuming $\mathbf{p}'$ to be a non-linear function of $\mathbf{p}$, the goal is to compute the mapping function $\mathcal{W}$ so that it allows more accurate prediction, better handling of noisy data and improves overall generalisability. To this end, we pose the problem in a multivariate input-output regression framework.

$$\mathbf{X}_N = [\mathbf{p}_1 \quad \dots \quad \mathbf{p}_M] \text{ and } \mathbf{Y}_N = [\mathbf{p}'_1 \quad \dots \quad \mathbf{p}'_M]$$

where $\mathbf{X}_N \in \mathbf{R}^{a \times M}$, $\mathbf{Y}_N \in \mathbf{R}^{b \times M}$, $\mathbf{x}_j$ is the j^{th} column of matrix $\mathbf{X}_N$ and $y_{i,j}$ is the $(i,j)^{th}$ element of matrix $\mathbf{Y}_N$. Let the training set

$$\mathcal{T}_i = \{(\mathbf{x}_j, y_{i,j})\}_{j=1}^{M} \qquad i = 1, \dots, b \tag{7}$$

where $\mathbf{x} \in \mathcal{X}_N$ (the set of multivariate inputs) and $y \in \mathcal{Y}_N$ (the set of outputs/targets). Here, a simple approach is to learn a non-linear mapping function $w_i : \mathcal{X}_N \rightarrow \mathcal{Y}_N$, where $i = 1, \dots, b$, that results in the mapping function $\mathcal{W}_N = [w_1 \, w_2 \dots w_b]$.

We use Gaussian Process Regression (GPR) [15] to compute $\mathcal{W}_N$. However, it should be noted here that this approach, also known as multi-kriging [15], works on the assumption of an independent model for each output dimension and, hence, cannot capture the relationship between the outputs. In order to avoid this *loss of information*, [16], [17] extended the GPR framework to Multiple-Output Gaussian Process Regression (MGPR) that uses the latent function framework and convolution process to model the dependencies between the output dimensions. Therefore, we use the MGPR framework to compute $\mathcal{W}_N$ in this paper.

3.1 Managing Global Shape Parameter

In the previous section, we addressed the problem of optimally transferring the local shape parameters, i.e. the parameters that represent observed non-rigid shape variations. We now address the problem of determining the global shape parameters (scale, rotation and position) for aligning the synthesised face with the rest of the head in the target scene. This must be robust enough to avoid visual oddities even in continuous video sequences, which is difficult due to the human sensitivity to the smallest inconsistencies in the positioning of key features, such as the eyes, and even the slightest change in face scale not in proportion with the rest of the head.

We start by extracting the global shape parameters of the upper and lower face shape models for the target head (Figure 4) separately. Using these, we align the synthesised upper face shape model with the target head and, then, we align the synthesised lower face shape model with the synthesised upper face shape model aligned previously.

To align the synthesised upper face shape model with the upper part of the target head, we directly apply the global shape parameters of the target head, extracted from the upper face shape model, to the synthesised upper shape. This is mainly due to the negligible scaling effect that we encounter in the upper face.[1] Note that the boundary of the upper face consists of the landmark points that are mostly rigid, unlike the landmark points on the eyebrows and eyelids that may rise and fall quite independently of any overall head motion.

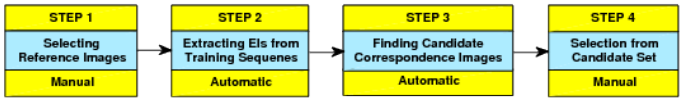
Once the upper face has been aligned, we focus our attention on the alignment of the lower part of the face, which is a difficult task because of the lack of landmark points in the lower face that have correlation with the rigid motions of the head. Moreover, ignoring the *scaling* factor can result in visual oddities because, unlike the upper face, the contour of the lower face consists of the landmark points on the cheek and the jaw line that may rise and fall due to the lip and jaw movements, thereby inducing a scaling effect to the lower face shape model. To deal with the *scaling* issue, we extract the scaling factors from the global shape parameters of the lower face shape model, computed while tracking the sequence of the source subject that we wish to synthesise. We apply this scaling factor[1] directly to the synthesised lower face shape and generate the final synthesised lower face shape that we can align with the previously aligned upper face. The upper and lower face shape models have three common landmark points just below the nose, marked in blue in Figure 1. To align the final synthesised lower shape with the previously aligned upper face, we use these three correspondence points and apply a similarity transformation [18] to the final synthesised lower shape, keeping the *unit* scaling factor and applying only the rotation and translation factors.

4 Locating Correspondence Images

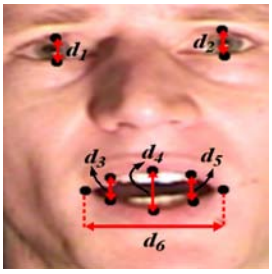
In this section, we turn to finding a set of correspondence images across the source and target subjects (i.e. the facial expression of the source and target subjects should be similar), so that the mapping functions between them can be learnt efficiently. Methods such as [19] allow establishing the correspondence between the meshes of any two arbitrary objects. However, our problem of finding the corresponding face images is a lot simpler owing to the structural similarity between faces and the constrained nature of the face models. Methods, such as [2], [5], [9], can be used for this purpose with some manual intervention. We propose to use a Structural SIMilarity (SSIM) [20] based method [21], [22].

An *Expression Image* (EI) [21], the basis for our method, is extracted from the mouth and eye region, since they are

1. It is important to note that, as a part of pre-processing, all frames from the videos were similarity normalised in order to eliminate the scaling effect induced by the head movement.



(a) Overview: Locating the correspondence images



(b) Distance vector

(c) Selected reference images

(d) Image and Extracted EI

(e) Image and Extracted EI

Fig. 2. (d),(e) Left: Sample image. Right: Extracted EI.

the main regions of interest. The EI is a visual descriptor and represents a distance vector $\mathbf{d} = \{d_1, \ldots, d_n\}$, where $n = 6$, in a normalised frame (Figure 2(b)). The overview of the proposed approach is given in Figure 2(a). We begin by selecting a pair of reference images for source and target subject manually (Section 4.1). EIs are automatically extracted from each frame of the training sequences (Section 4.2) and are used to find a set of candidate correspondence images using a SSIM based distance measure automatically (Section 4.3). This candidate set is then used for manually selecting a final set of correspondence images (Section 4.4).

4.1 Selecting Reference Images

We manually select reference images, exhibiting the same expression, for the region of interest of the source and target speakers (Figure 2(c)). We used the reference images that exhibit an open mouth expression with the tongue and teeth visible. These reference images are similarity normalised before generating the reference distance vector, $\mathbf{d}^R = \{d_1^R, \ldots, d_n^R\}$.

4.2 Extracting EIs from Training Sequences

As a pre-processing step, shapes of all the frames are aligned into a common coordinate frame w.r.t. the reference images via similarity normalisation. We extract the *normalised distance vector*, $\bar{\mathbf{d}}$, from each of these frames

$$\bar{\mathbf{d}}_m = \left[\frac{d_1^m}{d_1^R}; \ldots; \frac{d_n^m}{d_n^R}\right]^T = \left[\bar{d}_1^m; \ldots; \bar{d}_n^m\right]^T, m = 1 \ldots M \quad (8)$$

where M is the total number of frames extracted from the training sequences. Then, the EI is generated from each of the normalised distance vectors. For example, $\mathcal{E}_p^{\text{Source}}$ represents the EI for the p^{th} source frame (Figure 2(d)) and $\mathcal{E}_q^{\text{Target}}$ represents the EI for the q^{th} target frame (Figure 2(e)).

4.3 Finding Candidate Correspondence Images

Once we have computed the EIs for the entire training sequence, we compare each of the source subject's EIs with all the EIs of the target subject based on the SSIM distance metric. The SSIM [20] similarity distance metric performs three different similarity measurements of luminance, contrast and structure, and thereafter combines them to obtain a single number. The SSIM metric between two windows w_1 and w_2 of the same size $N \times N$ is given by

$$\text{SSIM}(w_1, w_2) = \frac{(2\mu_{w_1}\mu_{w_2} + c_1)(2\sigma_{w_1 w_2} + c_2)}{({\mu_{w_1}}^2 + {\mu_{w_2}}^2 + c_1)({\sigma_{w_1}}^2 + {\sigma_{w_2}}^2 + c_2)} \quad (9)$$

where μ_{w_1} and μ_{w_2} are the average of w_1 and w_2, respectively. $\sigma^2_{w_1}$ and $\sigma^2_{w_2}$ are the variance of w_1 and w_2, respectively. $\sigma_{w_1 w_2}$ is the covariance between w_1 and w_2. $c_1 = (k_1 L)^2$ and $c_2 = (k_2 L)^2$ are regularisation variables to stabilise the division with weak denominator. L is the dynamic range of the pixel-values. For details on the derivation of SSIM please refer to [20]

The SSIM distance between the p^{th} source frame and the q^{th} target frame is represented by $\mathcal{SD}_{pq}$ where

$$\mathcal{SD}_{pq} = \text{SSIM}(\mathcal{E}_p^{\text{Source}}, \mathcal{E}_q^{\text{Target}}) \quad (10)$$

The target frame with maximum $\mathcal{SD}$ is the best correspondence image for the source frame in consideration. Hence, for every source frame, we select target frames with an SSIM distance greater then the threshold value (δ), i.e. the $\mathcal{SD} \geq \delta$, as the candidate correspondence images (Figure 3).

4.4 Selection from Candidate Set

The EIs used for selecting the candidate correspondence images were generated from the shape of the face, whereas the texture information was completely ignored. Computing the SSIM score based on the texture is a difficult problem because, for example, different people have different amounts of oral cavity and teeth/tongue visible, while otherwise exhibiting similar expressions. Therefore, we manually select the visually best looking correspondence image from the candidate set.

5 Experiments and Discussion

In this section, we evaluate various regression strategies (see Section 3) to model the parametric correspondence between AAMs and use them for facial performance transfer of unseen sequences. We compare the proposed approach with the work of Theobald et al. [6]. Note that, similar to the method used for generating the result videos for the proposed method, the global shape parameters were transferred from the source to the target subject and the pre-processing steps[2] were used to generate the result videos for Theobald et al. method. We also demonstrate the utility of the system for *Cross-Language Facial Performance Transfer*.

We conducted experiments on the AVOZES data corpus [11] as it is a rich source for audio-video speaking-face data. To demonstrate the proposed *Facial Performance Transfer* approach, we transfer the performance from source subject ***A1*** to two separate target subjects ***A2*** and ***A3*** present in AVOZES (Figure 4). To demonstrate our *Cross-Language Facial Performance Transfer* framework, we separately recorded a video of a female subject of *Indian* origin (see supplementary material[3]) repeating the sequences present in AVOZES along with some more complex (and new) sequences in both English and Hindi (national language of India). Using this, we show that the proposed framework can effectively transfer the facial performance from the source subject ***B1*** to the target subject ***A1*** (Figure 4), so that the target subject appears to be speaking Hindi, although the face model was only trained on the sequences spoken entirely in English. This approach has significant potential for the movie and TV dubbing industry.

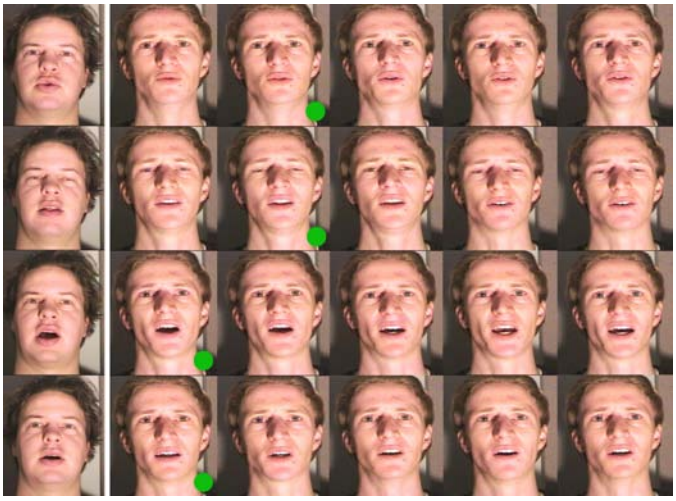


Fig. 3. Sample Results: Column 1 show source images (***A1***). Column 2-6 show top 5 correspondence images for target (***A2***). Selected correspondence image : Marked with green circle.

Fig. 4. A1, A2, A3 used in Sec. 5.1; B1, A1 used in Sec. 5.2.

5.1 Performance Evaluation

In this set of experiments, we used the sequences from AVOZES *Module 6 - Application Sequences - Continuous Speech* for training and validation purposes. We used sequences from AVOZES *Module 5 - Application Sequences - Digits* for testing. AVOZES *Module 6* contains three *4s* videos (30fps) of continuous speech sequences

- "*Joe took father's green shoe bench out.*"
- "*Thin hair of azure colour is pointless.*"
- "*Yesterday morning on my tour, I heard wolves here.*"

These sequences were designed to contain almost all phonemes and visemes of Australian English [11] and, hence, are suitable for the training and validation purposes here. In

addition, AVOZES *Module 5* contains one *2s* video (30fps) per digit (0-9) enclosed in the carrier phrase "*You grab /DIGIT/ beer*" and is suitable for testing the facial performance transfer for unseen sequences.

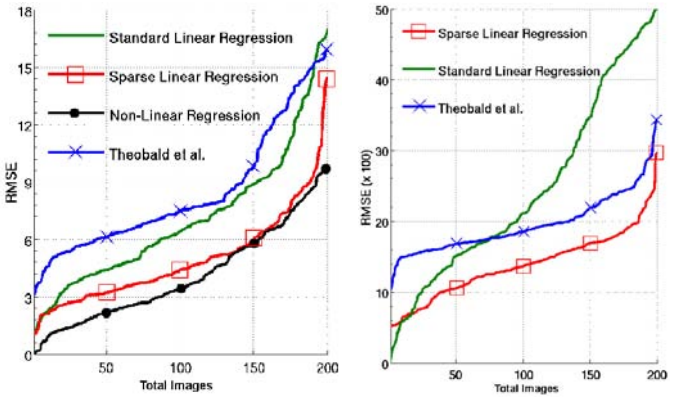
In the first experiment, we transferred the facial performance from the source subject ***A1*** to the target subject ***A2***. We trained AAMs (see Section 3) for ***A1*** and ***A2*** independently, so that each AAM captures the person-specific mannerisms and expressiveness. For this, each model was trained on 100 images from AVOZES *Module 6*. 95% of the variation of shape and texture were retained. Once the models were trained, we extracted 100 correspondence images (Section 4) for ***A1*** and ***A2*** from AVOZES *Module 6* and computed the mapping functions $\mathcal{W}_u : \mathbf{u}_1 \rightarrow \mathbf{u}_2$, $\mathcal{W}_l : \mathbf{l}_1 \rightarrow \mathbf{l}_2$ and $\mathcal{W}_t : \mathbf{t}_1 \rightarrow \mathbf{t}_2$ via different regression strategies (Section 3).

To verify whether the parametric correspondence was successfully learnt by the mapping functions, we synthesised the entire AVOZES *Module 6* sequences by transferring the facial performance from ***A1*** to ***A2*** (see supplementary material[2,3]). To further experimentally evaluate the performance of different regression strategies, a leave-one-out cross-validation scheme was adopted to synthesise all the correspondence images, i.e. the mapping functions were learnt from 99 correspondence images and the remaining single image ***A1*** was used to synthesise the correspondence image ***A2***. Then, we computed the error, using an RMS error measure, between the shape and texture of the synthetic and original correspondence images of ***A2***. We repeat the same set of experiments for the facial performance transfer from source subject ***A1*** to target subject ***A3***. Figures 5(a) and 5(b) show these error distributions.

The AAM texture vector in our experiments is a large vector of length 50000-55000, which has been trained to capture a sufficient amount of variation in texture, so that it can produce realistic looking results. This yielded texture parameter vectors $\mathbf{t}_1$ and $\mathbf{t}_2$ (Section 3) of dimensions in the range of 60-70. MGPR [16], used in the *Non-Linear Regression Method*, uses a convolution processes (CP) framework to model the relationship between all the output dimensions that demands significant computational and storage capabilities [17]. Hence, dealing with high-dimensional texture parameters is a complex and expensive problem. For this reason, we excluded the *Non-Linear Regression Method* from modelling the mapping functions between the texture parameters. On the other hand, the shape parameter vectors $\mathbf{u}_1$, $\mathbf{u}_2$, $\mathbf{l}_1$ and $\mathbf{l}_2$ (Section 3) have dimensions in the range of 5-9, which is a much simpler and more inexpensive problem that suits MGPR framework well.

From the experimental results in Figure 5, we can infer that the proposed approach using the *Sparse Linear Regression Method* and the *Non-Linear Regression Method* outperformed [6] significantly for both shape and texture. Also, the proposed approach using the *Standard Linear Regression Method* outperformed [6] for shape, but lacked for the texture. For the latter, the problem lies in the complex nature of the texture parameter vector. Given limited training data for the texture vectors, a standard linear framework is too restrictive and cannot model the complex mapping function accurately.

2. After the synthetic face has been generated by our system, Gaussian blurring of an alpha-mask was employed to smooth the boundaries of the synthesised face with that of the target frame in order to generate the final synthesised video results.



(a) Shape RMSE (b) Texture RMSE

Fig. 5. RMSE between the original and synthetic images.

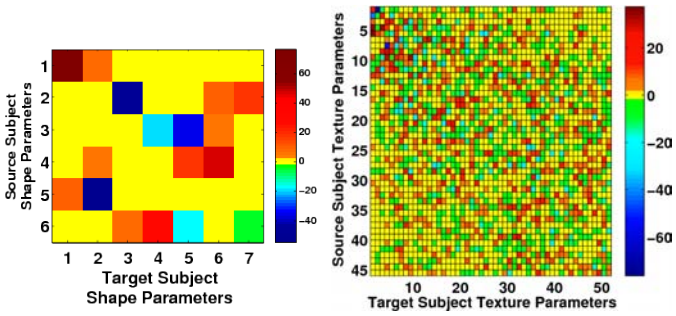
The *Sparse Linear Regression Method*, which uses a sparse representation framework to model the mapping function, shows significant improvement in the performance for both shape and texture. Here, the L_1-regularised least squares problem (Eq. 5) will always have a solution (which may not be unique) that can be efficiently solved as a convex optimisation program and yields a sparse solution [14], i.e. the regression coefficients for irrelevant input features are set to 0. This reduces the model complexity and, hence, avoids over-fitting.

Moreover, in the *Sparse Linear Regression Method*, we used a specialised interior-point method [13] to solve the convex quadratic program (Eq. 6), which has been shown capable of solving large scale and complex sparse problems efficiently. Hence, it is well suited for the task of modelling the mapping functions between the texture parameter vectors. Furthermore, the *Non-Linear Regression Method* shows more empirical improvements in the accuracy of modelling the mapping function for the shape parameter vectors by virtue of the MGPR framework [16], [17] that not only provides a more flexible non-linear regression framework by the use of GPR [15], but also by the use of the CP framework to model the relationship between all the output dimensions.

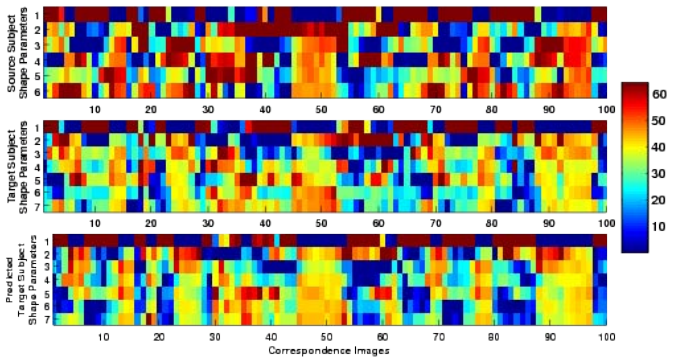
The *quality* of the generated synthetic images has an aesthetic, subjective element to it. While transferring the facial performance from one model to another for a sequence of images, it is important to transfer the changes occurring in the shape and texture. Maintaining the consistency and correlation between frames is equally important for generating realistic looking results. Although the *Non-Linear Regression Method* outperforms the *Sparse Linear Regression Method* experimentally in modelling the mapping function between the shape parameter vectors, the visual quality of the *Sparse Linear Regression Method* is superior. For example, notice the noise (*wave effect*) present in the synthetic output videos (see supplementary material[3]) for the *Non-Linear Regression Method*, especially in the upper part of the face.

The reason for the superior visual quality of the *Sparse Linear Regression Method* is the simplicity of the linear regression framework whose predictive domain is much simpler than that of the non-linear regressor. This observation is consistent

Fig. 6. Results ***A1*** to ***A2*** and ***A3***; **(i)** Sparse Linear Regression; **(ii)** Non-Linear Regression; **(iii)** Theobald *et al.* Method.



(a) Shape Mapping Function (b) Texture Mapping Function



(c) Predicting the shape parameters using $\mathcal{W}_S$. Note that all the values have been scaled for the ease of visualisation.

Fig. 7. Visualisation of the learnt parametric correspondence via Sparse Linear Regression Method.

with *Occam's Razor*, which states that the simplest model, which explains the data, is often the correct model. Non-linear regressors can potentially provide more accurate predictions, but their training procedure is generally more complicated. In practice, with limited training data at hand, the simpler linear models can be expected to extrapolate (i.e. to predict unseen data) better and show more consistent results [23]. Therefore, the *Sparse Linear Regression Method* is well suited for the task of modelling the mapping function and transferring the facial performance from one model to another.

In order to visualise the learnt parametric correspondence via *Sparse Linear Regression Method*, the average of 100 mapping functions ($\mathcal{W}_S$ and $\mathcal{W}_T$) learnt for cross-validation between subject ***A1*** and ***A2*** that produced the RMS errors in Figure 5 are shown in Figure 7(a) and 7(b). In Figure 7(a), the parametric correspondence between the source shape model (represented by 6 shape parameters along Y-axis) and the target shape model (represented by 7 shape parameters along X-axis) shows that the proposed method can easily learn the mapping functions between two independent models, irrespective of the number of parameters. Similarly in Figure 7(b), the parametric correspondence between the source texture model (represented by 46 texture parameters along Y-axis) and the target texture model (represented by 52 texture parameters along X-axis) can be visualised. Notice the pattern of sparsity (represented in yellow) in the mapping function. Moreover, Figure 7(c) visualises the accuracy of $\mathcal{W}_S$ to transfer the shape parameters using the source subject shape parameters (first row). Notice the similar pattern between the ground-truth target subject shape parameters (second row) and the predicted target subject shape parameters (last row).

To test the generalisation capability of the proposed approach, we synthesised the entire AVOZES *Module 5* sequences, which is a set of unseen talking sequences, by transferring facial performance from ***A1*** to ***A2*** and ***A3***. See supplementary material[3] for detailed test result videos. Figure 6 shows some sample facial performance transfer results.

5.2 Cross-Language Facial Performance Transfer

In this experiment, we explore the utility of the proposed approach for a *Cross-Language Facial Performance Transfer*, which is both a challenging and exciting application with many potential application areas, for example in the movie dubbing industry. The goal here is to show whether we can transfer the facial performance and lip movements from one person to another, irrespective of their gender, ethnic background and language. To demonstrate this, we transfer the facial performance from a female subject ***B1*** of Indian origin speaking complex sentences in *Hindi* to a male subject ***A1*** of Australian origin present in AVOZES, using the training data extracted from sequences spoken completely in English.

We repeat the same procedure as in Section 5.1 to train the AAM for ***B1***. We extracted 125 correspondence images (see Section 4) for ***B1*** and ***A1*** from *Module 6* and computed the mapping functions via the Sparse-Linear Regression Method (see Section 3). Note that the mapping functions are learnt on the sequences spoken completely in English. These mapping functions are then used to transfer the facial performance for the complex sequences spoken by ***B1*** in Hindi to ***A1***. Figure 8(a) show sample results. Figure 8(b) show temporal dynamics

3. Videos:http://users.rsise.anu.edu.au/~aasthana/TVCG11/Supplement.tar
Document:http://users.cecs.anu.edu.au/~aasthana/TVCG11/ReadMe.pdf

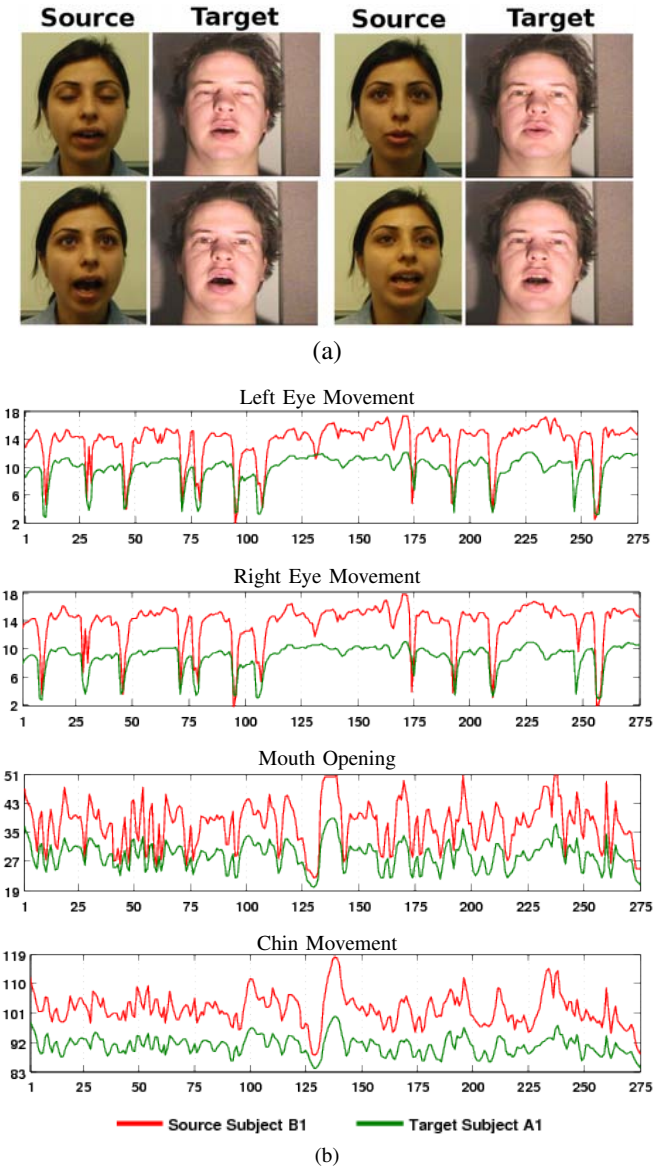


Fig. 8. (a) Facial Performance Transfer from ***B1*** to ***A1*** via Sparse Linear Regression Method. (b) Comparison of temporal dynamics (X-axis: Frame numbers, Y-axis: Facial feature location).

for the selected sample facial features i.e. the distance between the upper and lower eyelid of the eyes signifying the temporal dynamics of the eye movement, the distance between the upper and lower lip signifying the temporal dynamics of the mouth, the distance between the chin and the tip of the nose (stable reference point) signifying the temporal dynamics of the chin movement, respectively. Notice the correlation between the temporal dynamics of the source and target subject that shows the utility of the proposed approach for the *Cross-Language Facial Performance Transfer* application. See the supplementary material[3] for detailed test result videos.

5.3 Empirical Experiments

Judging the quality of a facial performance transfer is ultimately always subjective and has an aesthetic element to it. Hence, we evaluate it by conducting two separate empirical

First empirical experiment: Question 1

Rating / Method	Excellent	Good	Fair	Poor
Theobald et al.	7.8%	33.4%	52.9%	5.9%
Sparse Linear Regression	29.4%	56.9%	13.7%	0.0%
Non-Linear Regression	15.7%	47.0%	35.3%	2.0%

First empirical experiment: Question 2

Method	Theobald et al.	Sparse Linear Reg.	Non-Linear Reg.
Participants	11.0%	62.5%	26.5%

Second empirical experiment

Rating	Excellent	Good	Fair	Poor
Participants	20.00%	53.33%	23.33%	3.33%

Fig. 9. Results from the Empirical Experiments

experiments. Please note that all the participants, involved in both the experiments, were unaware of the sources of the result videos and have never seen the result videos before. In the first experiment, the original and synthetic AVOZES *Module 5* sequences (i.e. the digit sequences) transferring the facial performance from the subject ***A1*** to ***A2*** and ***A3*** were shown to 50 participants and were asked to answer two questions. The first question asked each participant to rate the quality of synthesis ("*Compared to the original digit sequence, how realistic does the facial performance transfer look?*") produced by the three methods (Theobald et al., Sparse linear regression and Non-linear regression methods) on a scale of "Excellent", "Good", "Fair" or "Poor". The second question asked each participant to choose the method that they felt produced the best synthesis ("*Choose the method that produces the most realistic facial performance transfer?*"). In the second empirical experiment, the cross-language facial performance transfer result from the subject ***B1*** to ***A1*** was shown to 30 participants of Indian origin who were fluent in Hindi and were asked to rate the quality of synthesis ("*How realistic does the cross language facial performance transfer look?*") on a scale of "Excellent", "Good", "Fair" or "Poor". Since the second empirical experiment is not comparative in nature, the result should be treated as a qualitative indication.

The first empirical experiment's results (Figure 9) show that 86.3% of participants rated the synthesis via the *Sparse Linear Regression Method* to be either good or excellent, clearly exceeding 41.2% for the *Theobald et al. Method* and 62.7% for the *Non-Linear Regression Method*. A one-way ANOVA found the results to be statistically significant at the $p < 0.000005$ level. Moreover, 62.5% of the participants rated the synthesis via the *Sparse Linear Regression Method* to be the best. For the second empirical experiment, 73.33% of the participants rated the cross-language facial performance transfer to be either good or excellent, while 23.33% gave it a fair rating.

6 Conclusion and Future Work

We propose an approach to learn the mapping between the parameters of two completely independent deformable face models and use them to facilitate the facial performance transfer. The proposed approach facilitates "meaningful transfer" of facial performance by transferring the changes induced both in the shape and texture of the face while preserving person specific qualities and mannerisms. It also shows good generalisation capability and works irrespective of the subject's gender,

ethnic background and language. Moreover, it requires only a small video corpus (we used three 4s videos) for modelling the parametric correspondence. Overall, the *Sparse Linear Regression Method* is best suited for the task of learning the parametric correspondence. The results and the rating provided by the human participants are very encouraging.

One limitation, as with any model-based approach, is that the quality of the synthesised sequence is directly related to the accuracy of the model used to extract the parameters. If the model encounters any unseen expression that cannot be optimally represented by the model, the synthesised face can show some visual oddities. Therefore, our method requires a good quality 2D deformable model for source and target subject. However, some of the visual oddities resulting from common problem areas, such as the lip contact line, teeth and the oral cavity, can be eliminated by augmenting the system with post-processing steps such as [2], [4]. Moreover, the SSIM-based method used for finding the correspondence images is semi-automatic and methods such as [19], [24] are being explored to make it completely automatic.

ACKNOWLEDGEMENTS

The authors would like to thank Jason Saragih (CSIRO) for the use of the DeMoLib software. The work presented in this paper was in part supported by the ARC grant TS0669874.

REFERENCES

[1] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. Salesin, "Synthesizing realistic facial expressions from photographs," in *Proc. of SIGGRAPH*, 1998, pp. 75–84. 1

[2] J. Noh and U. Neumann, "Expression cloning," in *Proc. of SIGGRAPH*, 2001, pp. 277–288. 1, 4, 9

[3] Z. Liu, Y. Shan, and Z. Zhang, "Expressive expression mapping with ratio images," in *Proc. of SIGGRAPH*, 2001, pp. 271–276. 1

[4] V. Blanz, C. Basso, T. Vetter, and T. Poggio, "Reanimating Faces in Images and Video," in *Proc. of EUROGRAPHICS*, 2003, pp. 641–650. 1, 9

[5] Y.-J. Chang and T. Ezzat, "Transferable videorealistic speech animation," in *Proc. of ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA)*, 2005, pp. 143–151. 1, 4

[6] B. Theobald, I. Matthews, J. Cohn, and S. Boker, "Real-time Expression Cloning using Appearance Models," in *Proc. Int. Conf. Multimodal Interfaces ICMI2007*, 2007, pp. 134–139. 1, 2, 5, 6

[7] M. J. Jones and T. Poggio, "Multidimensional Morphable Models: A Framework for Representing and Matching Object Classes," *Int. J. Computer Vision*, vol. 29, no. 2, pp. 107–131, 1998. 1

[8] T. Leyvand, D. Cohen-Or, G. Dror, and D. Lischinski, "Data-driven enhancement of facial attractiveness," *ACM Trans. on Graphics*, vol. 27, no. 3, 2008, doi: 10.1145/1399504.1360637. 1

[9] D. Vlasic, M. Brand, H. Pfister, and J. Popović, "Face transfer with multilinear models," *ACM Trans. on Graphics*, vol. 24, no. 3, pp. 426–433, 2005. 1, 4

[10] G. Edwards, C. Taylor, and T. Cootes., "Interpreting Face Images Using Active Appearance Models," in *Proc. of IEEE Int. Conf. Automatic Face and Gesture Recognition FG'98*, 1998, pp. 300–305. 1, 2

[11] R. Goecke and B. Millar, "The Audio-Video Australian English Speech Data Corpus AVOZES," in *Proc. Int. Conf. Spoken Language Processing ICSLP2004*, 2004, pp. 2525–2528. 2, 5

[12] S. Baker, R. Gross, and I. Matthews, "Lucas-Kanade 20 years on: A unifying framework: Part 3," Robotics Institute, Carnegie Mellon University, USA, Tech. Rep. CMU-RITR-03-35, 2003. 2

[13] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An Interior-Point Method for Large-Scale L1-Regularized Least Squares," *IEEE J Selected Topics in Signal Proce.*, vol. 1, no. 4, pp. 606–617, 2007. 3, 6

[14] D. L. Donoho, "For Most Large Underdetermined Systems of Linear Equations, the Minimal L1-norm Solution is also the Sparsest Solution," *Comm. on Pure and Applied Math*, vol. 59, pp. 797–829, 2004. 3, 6

[15] C. E. Rasmussen, *Gaussian processes for machine learning*. MIT Press, 2006. 3, 6

[16] P. Boyle and M. Frean, "Dependent Gaussian Processes," in *Proc. Advances in Neural Information Processing Systems (NIPS 17)*, 2005, pp. 217–224. 3, 6

[17] M. Alvarez and N. D. Lawrence, "Sparse Convolved Gaussian Processes for Multi-output Regression," in *Proc. Advances in Neural Information Processing Systems (NIPS 21)*, 2008, pp. 57–64. 3, 6

[18] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004. 4

[19] R. W. Sumner and J. Popovic, "Deformation Transfer for Triangle Meshes," in *Proc. of SIGGRAPH*, 2004, pp. 399–405. 4, 9

[20] Z. Wang, A. C. Bovik, H. R. Sheikh, S. Member, E. P. Simoncelli, and S. Member, "Image Quality Assessment: From Error Visibility to Structural Similarity," *IEEE Trans. on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004. 4, 5

[21] A. Dhall, A. Asthana, and R. Goecke, "Facial Expression Based Automatic Album Creation," in *Proc. 7th Int. Conf. Neural Information Processing (ICONIP2010)*, 2010, pp. 485–492. 4

[22] A. Dhall, A. Asthana, and R. Goecke, "A SSIM Based Approach for Finding Similar Facial Expressions," in *Proc. EmoSPACE Workshop, IEEE Int. Conf. Automated Face and Gesture Recog. FG2011*, 2011. 4

[23] T. Jan and A. Zaknich, "An adjustable model for linear to nonlinear regression," in *Proc. IEEE Int. Joint Conf. Neural Networks*, 1999, pp. 846–850. 7

[24] P. Vacha, "Texture similarity measure," in *WDS'05 Proc. of Contributed Papers: Part I - Mathematics and Comp. Sciences*, 2005, pp. 47–52. 9

Akshay Asthana is a PhD student in the College of Engineering & Computer Science at the Australian National University. He was awarded the Australian Leadership Award Scholarship by AusAID in 2008. He graduated from the Jaypee Institute of Information Technology University (Noida, India) with a degree of Bachelor of Technology in Computer Science in 2007. His research interests are in computer vision, computer graphics, affective computing, pattern recognition and HCI.

Miles Delahunty received his Bachelors degrees in Economics and Science, majoring in mathematics, in 2010 from the Australian National University where he is currently an Economics Honours student in the College of Business & Economics. He was also an ANU Summer Research Scholar 2009/2010 in the College of Engineering & Computer Science. His current interests include game theory, probability theory, and all things combinatorial.

Abhinav Dhall is a PhD student in the College of Engineering & Computer Science at the Australian National University. He was awarded the Australian Leadership Award Scholarship by AusAID in 2010. He graduated from the DAV Institute of Engineering & Technology (Jalandhar, India) with a degree of Bachelor of Technology in Computer Science in 2006. His research interests are in computer vision, computer graphics, affective computing, pattern recognition and HCI.

Roland Göcke is Assistant Professor of Software Engineering and Head of the Vision and Sensing Group at the Faculty of Information Sciences and Engineering, University of Canberra. He is also an Adjunct Faculty member in the College of Engineering & Computer Science at the Australian National University. He received his Masters degree in Computer Science from the University of Rostock, Germany, in 1998 and his Ph.D. in Computer Science from the Australian National University in 2004. His research interests are in affective computing, pattern recognition, computer vision, human-computer interaction and multimodal signal processing.