RUNNING HEAD: Mapping Facial Expression

Mapping and Manipulating Facial Expression

Barry-John Theobald[1], Iain Matthews[2], Michael Mangini[3], Jeffrey R. Spies[5], Timothy R. Brick[5], Jeffrey F. Cohn[4], and Steven M. Boker[5]

[1]School of Computing Sciences, University of East Anglia, Norwich, UK

[2]Weta Digital Ltd., 9-11 Manuka Street, Miramar, New Zealand

[3]Department of Psychology, University of Notre Dame, Indiana, USA

[4]Department of Psychology, University of Pittsburgh, Pittsburgh, USA

[5]Department of Psychology, University of Virginia, Charlottesville, USA

Keywords: active appearance models, expression cloning, expression mapping.

Acknowledgements: Preparation of this manuscript was supported in part by grants NSF BCS–0527485, NSF HSD–0(49)527444, and EPSRC EP/D049075/1. Any opinions, findings, conclusions, or recommendations expressed herein are those of the authors and do not reflect the views of NSF or Carnegie Mellon University.

Contact: Dr. Barry-John Theobald, School of Computing Sciences, University of East Anglia, Norwich, UK, NR4 7TJ. Tel: +44 (0)1603 592574. Email: bjt@cmp.uea.ac.uk

ABSTRACT

Non-verbal visual cues accompany speech to supplement the meaning of spoken words, signify emotional state, indicate position in discourse, and provide back-channel feedback. This visual information includes head movements, facial expressions and body gestures. In this paper we describe techniques for manipulating both verbal and non-verbal facial gestures in video sequences of people engaged in conversation. We are developing a system for use in psychological experiments, where the effects of manipulating individual components of non-verbal visual behaviour during live face-to-face conversation can be studied. In particular, the techniques we describe operate in real-time at video frame-rate and the manipulation can be applied so both participants in a conversation are kept blind to the experimental conditions.

This manuscript has been submitted for the special issue on Audiovisual Prosody.

MAPPING AND MANIPULATING FACIAL EXPRESSION

Audiovisual conversation is information rich, comprised of both the linguistic message content and non-linguistic information. This non-linguistic information might include facial expressions to convey emotional state, head nods to indicate emphatic stress, and posture and eye-gaze to signal turn-taking in discourse. There is increasing interest in understanding the relationship between verbal and non-verbal conversational cues. In large part, this interest comes from the entertainment industry, which requires life-like animation of embodied characters. If interactive dialogue between animated characters is to be believable, it must seem realistic and lifelike.

In psychology and speech perception, the ability to generate experimental stimuli using computer-generated characters is attractive as individual aspects of verbal and non-verbal cues can be manipulated to arbitrary precision and in complete isolation from each other (Massaro, 1998). In this paper, we describe the development of a system for automatically capturing, encoding and manipulating facial expressions and global head movements, and transferring this information from video sequences of one person to video sequences of another. In particular, we describe techniques that allow these manipulations to be accomplished in real-time at video frame-rate during face-to-face conversations and with both participants in the conversation being kept blind to the processing and manipulation of the visual cues accompanying the speech. This system is being developed to provide experimenters with the means for controlled investigation to better understand the mechanisms of communication and social interaction.

BACKGROUND AND RELATED WORK

Computer generated models of faces are useful in studying speech perception because the experimenter has precise control over the stimuli. For example, the use

of a model allows the amplitudes of expressive face movements to be exaggerated or attenuated. Likewise the timing of facial expressions can be manipulated to create arbitrary durations. Many of these effects would be impossible even for highly accomplished actors. An effective face model can make them possible.

In some ways the freedom of manipulation provided by face models presents a problem for researchers; models can extrapolate beyond what is plausible or acceptable. For experimental validity, it is important that the stimuli presented to the viewer be acceptably realistic. Constraining the appearance of generated facial expressions is a non-trivial challenge. The changes in the appearance of a face that we interpret as a facial expression are the product of the complex interaction of neuromuscular signals, anatomical structures (including bone, muscle, subcutaneous fat, and skin), and light (Terzopoulos & Waters, 1993). The difficulty of ensuring that only plausible facial expressions are generated is compounded by the fact that people are expert at detecting and recognising facial expressions and are acutely sensitive to even minute discrepancies from normal behaviour. As the realism of computer generated models increases, the viewer generally becomes more reactive to subtle errors in dynamic sequences – errors are more readily detected and the face may appear "zombie-like". This occurrence often is referred to as the *Uncanny Valley* effect (Mori, 1970).

Generally, the most pleasing, natural, and life-like facial animation is achieved using *performance-driven* techniques. To implement this approach, feature points are located on the face of an actor and the displacement of these feature points over time is used either to update the vertex locations of a polygonal model (Guenter, Grimm, Wood, Malvar & Pighin, 1998; Noh & Neumann, 2001; Williams, 1990), or are mapped to an underlying muscle-based model (Choe, Lee & Ko, 2001).

Performance-driven facial animation enables subtle facial movements to be captured and transferred to the animated face model, but there are several inherent limitations that make this approach unsuitable in many applications. Extensive time is required to setup the capture environment. The motion capture markers (e.g., LEDs or reflective tape or ink) must be applied to the actor, and they may inhibit ease or naturalness of movement. Finally, consistent placement of these markers must be maintained between capture sessions.

An alternative is image-based animation, in which changes in the position of the facial features observed in images or video sequences of one person are used to warp images of another person (Chang and Ezzat, 2005; Chuang & Bregler, 2002; Liu, Shan & Zhang, 2001; Zhang, Liu, Quo, Terzopoulos & Shum, 2006). In this way, there is no longer need for motion capture, which considerably reduces setup time. On the down side, the range of convincing facial expressions that can be transferred is limited to those that actually occurred. It also is difficult to manipulate facial expressions, such as exaggerate their intensity or qualitative appearance, as the face is represented only as raw pixel values. There is no underlying parametric model available to provide quantitative guidance in constructing realistic image manipulation.

A hybrid approach based on Blanz and Vetter's (1999) three-dimensional morphable model (3DMM) transfers facial expressions from images of one person's face to images of another person. This expression transfer is achieved by *copying* to the new face the changes in facial geometry and facial appearance that result from displaying a facial expression (Blanz, Basso, Poggio & Vetter, 2003). 3DMMs have the advantage of a parameterised face model that provides control over the deformation of the facial geometry, and an image-based texture-map that provides

near-photorealistic rendering of the facial appearance. 3DMMs are capable of producing highly convincing results, but there are several practical limitations that restrict their use. Firstly, this approach requires a collection of 3D laser range scan images to build the face model. Secondly, changes in face shape and appearance that result from movements of the facial features as measured in images of one person are *copied* directly to images of another person. This makes it difficult to model and apply person-specific characteristics of facial expressions. Thirdly, because the inside of the mouth is difficult to model, the same generic inner mouth must be used for all subjects. And fourthly, the algorithms are computationally expensive, which currently rules out real-time applications. Approaches that address these limitations are needed.

The lack of facial expression mapping in the system proposed by Blanz, Basso, Poggio and Vetter (2003) was addressed recently in Vlasic, Brand, Pfister and Popovic (2005) using multi-linear face models to represent changes in images that are due to identity, age, and facial expression. The learned multi-linear model can be used to predict the appearance of common facial expressions in images of new, unseen faces. However, to construct the multi-linear model, a 3D laser range scanner is required to capture facial images of a number of people displaying the same pre-specified facial expressions (e.g., smile, frown, surprise, and so on). While this approach remains computationally expensive, it can be approximated efficiently in 2D (Macedo, Vital Brazil & Velho, 2006) using Active Appearance Models (AAMs) (Cootes, Edwards & Taylor, 2001), but this approximation still requires the used of a common set of posed facial expressions from which the multi-linear model is learned.

We have developed an AAM-based system that captures, manipulates, and transfers facial expressions, visual speech gestures, and changes in head pose, from video sequences, all in real-time at video frame-rates. Our approach requires no

marker-based tracking, offline processing, or high-level semantic descriptors. The full face, including the eyes and inside of the mouth, can also be transferred. This makes it possible for this approach to be used for analysing and manipulating real-time conversation in an interactive two-way videoconference experimental paradigm.

LINEAR DEFORMABLE FACE MODELS

The key to our approach is a linear deformable model that both can analyze and synthesize images of faces. The idea of representing an ensemble of images of faces as a compact statistical model was first proposed by Sirovich and Kirby (1987). They noted that humans store and recognize a large number of faces almost instantaneously, which suggests the information required to encode faces may be represented in only a few tens of parameters (Kirby & Sirovich, 1990). They proposed a compact linear face model of the form,

$$\mathbf{F} = \mathbf{F}_0 + \sum_{i=1}^{m} \mathbf{F}_i p_i, \qquad (1)$$

where $\mathbf{F}$ is an instance of a face, $\mathbf{F}_0$ is a reference face, $\mathbf{F}_i$ are basis faces that characterise the allowed deformation from the reference face, and each p_i defines the contribution of the corresponding basis face in representing $\mathbf{F}$. Thus, faces are not coded directly, but rather indirectly as displacements from the reference.

Turk and Pentland (1991) developed the idea proposed by Sirovich and Kirby (1987) into a system for recognising faces in images. Thus, $\mathbf{F}$ is a (vectorized) face image, $\mathbf{F}_0$ is the mean (vectorised) image computed from a training set of face images, and each $\mathbf{F}_i$ is referred to as an *Eigenface*. The model is usually obtained by applying principal components analysis (PCA) to a set of training images containing faces. Linear models have since become one of the most popular techniques for modelling flexible objects, such as faces, and they have a wide range of applications,

which include: Low-bandwidth transmission of face images (Koufakis & Buxton, 1999; Mogghaddam & Pentland, 1995; Theobald, Kruse, Bangham, & Cawley, 2003), face recognition and verification (Blanz & Vetter, 2003; Edwards, Cootes, & Taylor, 1998; Shackleton & Welsh, 1991; Yang, Zhang, Frangi, & Yang, 2004), automated lip-reading (Luettin, Thacker, & Beet, 1996; Matthews, Cootes, Bangham, Harvey, & Cox, 2002), and face synthesis (Cosker, Marshall, Rosin, Paddock, & Rushton, 2004; Fagel, Bailly, & Elisei, 2007; Govokhina, Bailly, Bretin, & Bagshaw, 2006; Hong, Wen, Huang, & Shum, 2002; Theobald, Bangham, Matthews & Cawley, 2004; Theobald & Wilkinson, 2007).

Active Appearance Models: AAMs

Perhaps the most common form of linear deformable face model is the AAM (Cootes, Edwards & Taylor, 2001), which provides a compact statistical representation of the shape *and* appearance variation of the face as measured in 2D images. A recent extension, the so called 2D+3D AAM (Xiao, Baker, Matthews, & Kanade, 2004), provides a convenient method for constructing models of 3D shape variation from 2D images, overcoming the need for an expensive 3D laser range scanner required for a 3DMM.

The *shape* of an AAM is defined by a 2D triangulated mesh and in particular the vertex locations of the mesh. Mathematically, the shape, **s**, of an AAM is defined as the concatenation of the *x* and *y*-coordinates of *n* vertices that make up the mesh:

$$\mathbf{s} = \left(x_1, y_1, x_2, y_2 ..., x_n, y_n\right)^{\mathrm{T}}.$$

For face models, vertices typically are defined to delineate the eyes, eyebrows, nose, mouth and the outline of the face. A compact model that allows a linear variation in the shape is given by,

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{m} \mathbf{s}_i p_i , \qquad (2)$$

where the coefficients p_i are the shape parameters that allow the mesh to undergo both rigid (head pose) and non-rigid (expression) variation. Such a model is usually computed by applying PCA to a set of meshes that have been aligned with a corresponding set of training images. Typically this alignment is done manually, although automated approaches have shown some success (Wang, Lucey & Cohn, 2008). The base shape $\mathbf{s}_0$ is the mean shape, and the vectors $\mathbf{s}_i$ are the (reshaped) eigenvectors corresponding to the m largest eigenvalues. Care is required when hand-labelling images to ensure that all example shapes have the same number of vertices, and vertices are labelled in the same order in all images. Hand-labelling errors can result in significant sources of variation in the shape component of the AAM.

The *appearance*, $A(\mathbf{x})$, of an AAM is an image that exhibits variation under the control of the model parameters and is defined as the pixels, $\mathbf{x} = (x, y)^{\mathrm{T}}$, that lie within the base mesh, $\mathbf{x}$ $\mathbf{s}_0$. AAMs allow linear appearance variation. That is, the appearance can be expressed as a base appearance, $A_0(\mathbf{x})$, plus a linear combination of l appearance images, $A_i(\mathbf{x})$:

$$A(\mathbf{x}) = A_0(\mathbf{x}) + \sum_{i=1}^{l} A_i(\mathbf{x}) \quad i \qquad \mathbf{x} \quad \mathbf{s}_0 , \qquad (3)$$

where the coefficients $_i$ are the appearance parameters. The appearance component of the AAM is usually computed by first warping the manually annotated training images to the base shape, then applying PCA to the shape-normalised images (Cootes, Edwards & Taylor, 2001). The base appearance $A_0(\mathbf{x})$ is the mean shape normalised image, and the vectors $A_i(\mathbf{x})$ are the (reshaped) eigenvectors corresponding to the l largest eigenvalues. The advantage of this two-stage approach to modelling the face is

that the resulting modes of variation are more compact and cleaner, in the sense that they model changes in the facial features rather than change in the images, as is the case with *Eigenfaces*. An example AAM is shown in Figure 1.

[INSERT FIGURE 1 ABOUT HERE]

Facial image synthesis. Near-photorealistic images of faces are synthesised using an AAM by first applying the shape parameters, Equation (2), to generate the shape of the AAM, then applying the appearance parameters, Equation (3), to generate the AAM image, $A(\mathbf{x})$. The synthesised image is created by warping $A(\mathbf{x})$ from the base shape, $\mathbf{s}_0$, to the model-synthesised shape using a piece-wise affine warp. Example images synthesised using an AAM are shown in Figure 2. We note here that the AAMs used in this work require a common structure. Each person-specific model requires the equivalent number of landmarks and pixels and the ordering of the landmarks must be consistent across all models. The shape of the model is constrained to include information in the region of the face only to the eyebrows. The forehead region is not included in our current work as there are no identifying features that can consistently be located across all individuals. Occlusion of the forehead by hair, for instance, is a problem. Solutions have been proposed (e.g., Gross, Matthews and Baker, 2006), however these solutions significantly increase the computational expense.

[INSERT FIGURE 2 ABOUT HERE]

Multi-segment Active Appearance Models. The standard approach to constructing the appearance component of the AAM is to warp the images onto $\mathbf{s}_0$ and concatenate *all* pixels bound by the mesh before applying PCA. The assumption is that the probability distribution of the pixel intensities is Gaussian. However, this assumption often is not met for face images. Consequently, some important information might be

considered noise and discarded. In addition, if the distribution of the data is multimodal or non-linear, then linearly interpolating the basis vectors to re-synthesize face images can result in parameters that appear valid, in that they are within the limits of the training examples, but lie in invalid regions of the global parameter space (Bregler & Omohundro, 1995). When assumptions about the distribution of the data are not met, some blurring in the rendered images may occur. This consequence is most striking in the eyes and inner mouth, in particular the teeth. These are important areas of the face as they are ones on which viewers tend to focus their attention.

To address this issue, we apply a relatively straightforward extension to improve the quality of rendering by constructing a piece-wise PCA model. Independent appearance models are constructed for each region of the face (skin, eyes, nose, and inner-mouth). This can be done in the coordinate frame of $\mathbf{s}_0$, so the pixel indices for the different regions of the face are constant across all images. The appearance for individual segments then can be regenerated and copied into the appearance vector $A(\mathbf{x})$ before warping to the shape **s**. This notion could be taken further by clustering the regions of the face according to the shape parameters, for example building separate models of appearance to account for an open and closed mouth. This also allows different model segments to be encoded with more or less resolution, giving more significance to regions of the face to which a viewer is likely to pay particular attention.

MANIPULATING FACIAL EXPRESSIONS

In this section we illustrate the usefulness of AAMs for studies of social interaction. Individual verbal and non-verbal conversational cues obtained from the face in a video sequence can be manipulated independently so that their separate and combined effects can be studied. For example, one might wish to decouple the

apparent sex of an individual from his/her behaviour to determine each factor's influence during conversation. Or, one might wish to attenuate or exaggerate the degree of expressiveness to examine its influence on perceived personality traits. Exaggerated expressions and head movements are believed to increase a sense of liveliness and lead to greater rapport (Grahe & Bernieri, 1999). Attenuating expressiveness and head movement may create a sense of disinterest, sadness, or depressed affect (Gehricke & Shapiro, 2000; Schwartz, Fair, Salt, Mandel & Klerman, 1976). Using AAMs, the behavior and identity of each person can be manipulated during actual conversation without the individual's awareness that a change has occurred.

The following descriptions of mapping and manipulating facial expressions consider only the *intensity* of those expressions. Focus to date has been on developing techniques that operate in real-time at video frame-rate, allowing conversational cues to be manipulated during live face-to-face interaction in a videoconferencing environment. Manipulating other factors, such as the frequency of occurrence of expressions, and the temporal properties of facial expressions will be the focus of future work. This is difficult to achieve during live conversation, as *a priori* knowledge of the temporal properties of the facial expressions over the conversation is required.

Manipulating the Intensity of Facial Gestures

Facial expressions can be represented in terms of the parameters of an AAM, which in turn represent displacements from the origin in the space spanned by the model. It follows that multiplying the parameters by a scalar greater than unity increases the distance from the mean, thereby exaggerating the facial expression represented by the parameters. Conversely, scaling by a value less than unity

decreases the distance from the origin, thereby attenuating the intensity of the facial expression. Equations (2) and (3) can easily be extended to allow the intensity of facial expressions to be manipulated, as follows:

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{m} \mathbf{s}_i p_i \quad , \tag{4}$$

$$A(\mathbf{x}) = A_0(\mathbf{x}) + \sum_{i=1}^{l} A_i(\mathbf{x}) \quad _i \qquad \mathbf{x} \quad \mathbf{s}_0 . \tag{5}$$

To illustrate the effect of manipulating the intensity of facial expressions in this way, an AAM was constructed and used to track a face during a conversation in a video sequence. The tracking algorithm used is the *inverse compositional* algorithm described by Matthews and Baker (2004), which runs far in excess of video frame-rate (>200 fps). The goal is to minimize the sum of squares difference between a model instance $A(\mathbf{x}) = A_0(\mathbf{x}) + \sum_{i=1}^{m} {}_i A_i(\mathbf{x})$ and the input image warped back onto the base mesh $I(\mathbf{W}(\mathbf{x};\mathbf{p}))$. Following (Gross, Matthews and Baker, 2006), this is achieved as follows:

Pre-compute:

- Compute the gradient of base appearance $A_0(\mathbf{x})$.
- Compute the Jacobian of the warp at $(\mathbf{x};\mathbf{0})$, $\frac{\partial \mathbf{W}}{\partial \mathbf{p}}$.
- Compute steepest descent images $\mathbf{S}(\mathbf{x}) = A_0(\mathbf{x}) \frac{\partial \mathbf{W}}{\partial \mathbf{p}}$.
- Project out appearance variation $\mathbf{S}_p(\mathbf{x}) = \mathbf{S}(\mathbf{x}) \sum_{i=1}^{m} A_i(\mathbf{x}) \mathbf{S}(\mathbf{x}) \; A_i(\mathbf{x}) \quad \mathbf{x} \quad \mathbf{s}_0$.
- Compute the Hessian $\mathbf{H} = \sum_{\mathbf{x}} A_0 \frac{\partial \mathbf{W}}{\partial \mathbf{p}}^T A_0 \frac{\partial \mathbf{W}}{\partial \mathbf{p}}$.

Iterate:

- Warp the image onto the base shape $I(\mathbf{W}(\mathbf{x};\mathbf{p}))$.
- Compute the error image $E(\mathbf{x}) = I(\mathbf{W}(\mathbf{x};\mathbf{p})) \quad \mathbf{A}_0(\mathbf{x})$.
- Compute the warp update $\mathbf{p} = H^{1} \quad A_0 \frac{\partial \mathbf{W}}{\partial \mathbf{p}}^{T} E(\mathbf{x})$.
- Update the warp $\mathbf{W}(\mathbf{x};\mathbf{p}) \quad \mathbf{W}(\mathbf{x};\mathbf{p}) \circ \mathbf{W}(\mathbf{x};\ \mathbf{p})^{1}$.

The fitting algorithm iterates until convergence, i.e., until the change in the parameters between iterations falls below some threshold. The key to the efficiency is error is computed in the coordinate frame of model and not the image. Thus the Hessian matrix required to minimize the error function is constant and can be pre-computed. Also, since the appearance variation is *projected-out* from the steepest-descent images, the appearance parameters can be ignored during the fitting process. The AAM search is local and so requires a reasonable starting estimate for the parameters $\mathbf{p}$. To track the face throughout a video sequence, the first frame can be manually hand-labelled, and then each subsequent frame can use the fit from the previous frame as the starting estimate. The parameters measured from the video can then be scaled using Equations (4) and (5), and the resultant facial expressions re-synthesized by applying the parameters to the model. Figure 3 shows example frames from a video sequence re-rendered using the AAM both with and without adjustment to the intensity of the facial expressions.

[INSERT FIGURE 3 ABOUT HERE]

Varying degrees of exaggeration/attenuation can be applied to the different parameters of the AAM. However, to ensure the resultant facial expressions are valid, we generally apply the same degree of scaling to all parameters. The scaling parameter in Equations (4) and (5) is different to indicate the shape and appearance

components of the model can be manipulated independently. To ensure any manipulated facial expression is valid, in the sense that it appears genuine, constraints are placed on the parameter values to ensure that they lie within chosen limits of the original training data. Typically the parameters are constrained to lie within ±3 standard deviations from the mean (Cootes, Edwards & Taylor, 2001). This prevents artifacts such as the mouth opening abnormally wide, or other violations of facial topology, for example the upper and lower lips intersecting when the mouth closes.

An advantage of using an AAM for this task is that it takes only in the order of 150ms to capture a video frame, extract the AAM parameters automatically, apply the manipulation, re-render the face, and display the resultant image. Thus, AAMs allow face images to be manipulated during face-to-face interaction within a videoconferencing environment. Thus two people can communicate with one another over the videoconference, but the face(s) displayed are re-rendered using the AAM, and not the image captured directly from the camera. The manipulation of the parameters can be applied prior to re-synthesis, thus both interlocutors are free to converse naturally and are kept blind to any manipulation of the facial expression. A further advantage is the scaling of the parameters allows facial expressions to be manipulated *precisely* (e.g., smile exactly half as much), which is impossible using an actor.

Copying Facial Expressions between Faces

The previous section described how the *intensity* of facial expressions can be manipulated using AAMs. However, it might also be useful to present facial expressions *independently of identity*. This would be particularly beneficial in experiments contrasting the influence of social expectation and behaviour on natural conversations. This *could* be achieved using an actor, whereby the actor behaves

differently in different conversations, thereby effectively dissociating identity from behaviour. However, the conversations are no longer natural, and the reliability is largely dependent on the skill of the actor.

Alternatively, this can be achieved by transferring facial expressions from images of the face of one person (the source) to images of the face of another (the target). This is achieved by measuring the difference between an image containing a reference facial expression (say, neutral) and an expressive image, then *imposing* this difference onto an image of the same reference facial expression for the target identity. This difference can be measured in terms of shape and appearance information (Blanz, Basso, Poggio & Vetter, 2003) or in terms of raw pixel intensity changes (Liu, Shan & Zhang, 2001).

This form of facial expression mimicry fits naturally within our AAM framework. The parameters of an AAM are, by definition, offsets from the mean. Provided that the base shape and appearance for two models represent approximately the same facial expression, then swapping the mean shape and appearance from one AAM into another will effectively recreate the same facial expression (behaviour) on a different face (identity). Equation (2) for reconstructing the shape can be easily modified to take this transfer into account as follows:

$$p_i = \mathbf{s}_i^T(\mathbf{s}_0 \quad \mathbf{s}) \qquad \mathbf{t} = \mathbf{t}_0 + \sum_{i=1}^{m} \mathbf{s}_i p_i \tag{6}$$

where the left hand side of Equation (6) uses the source model to compute the parameters from a known shape, and the right hand side applies the parameters to the source modes of variation and the target mean shape. The structure of the equation for transferring appearance is exactly the same. Thus facial expressions can be simultaneously copied to an image of the target face and manipulated in terms of the intensity. The computational cost of this mimicry is exactly as outlined in the

previous sections, so it again can be used to manipulate live face-to-face conversations over a videoconference without either conversant being aware of the manipulation. Example facial expressions rendered using AAMs both before and after substituting the mean shape and appearance from a target model into a source model are shown in Figure 4, where the case for both within-gender and between-gender mimicry are shown.

[INSERT FIGURE 4 ABOUT HERE]

The requirement that the mean shape and appearance for both the source and the target AAM represent similar facial expressions is generally not a problem. The mean shape and appearance are computed from a number of images displaying a range of facial expressions, so averaging over many examples usually results in the means converging to approximately the same facial expression, even for different identities and gender – see Figure 5.

[INSERT FIGURE 5 ABOUT HERE]

Mapping Facial Expressions between Faces

The method for copying facial expressions between faces described previously substitutes the mean (shape and appearance) from one AAM to another. The modes of variation (the basis shape and appearance vectors) of the source model still are used to generate the expression itself. Hence *the cloned expression appears as the source expression imposed onto the target identity*. See, for example, the teeth in Figure 4. Whilst this is useful for presenting facial expressions independently of identity, we might wish to *map* gestures from facial images of one person to facial images of another such that the characteristics of the target person are retained.

If the basis vectors between two AAM spaces could be interpreted to have the same meaning, say the first vector opens and closes the mouth, the second the eyes,

and so on, the entire model for the target person could be substituted for the source, rather than substituting only the mean shape/appearance. The source model could be used to analyse the video and determine the AAM parameters, which could then be applied directly to the target model. However, basis vectors computed using PCA simply maximise the projected variance of the input data – the source and target basis vectors do not correspond in any physically meaningful way. Simply applying the parameters directly from one model to another will produce incorrect facial expressions. The problem is illustrated in Figure 6.

[INSERT FIGURE 6 ABOUT HERE]

The manner in which the features of the face can vary is highly constrained. Individual features generally do not move independently of one another (in a global sense), and features generally move in the same way between faces – the mouth can open and close, the lips can round, the eyes can blink, and so on. It follows therefore that the information contained within one AAM space should be embedded in some way within another. The question is: How does one map the meaning of a basis vector of one AAM to the space spanned by another? One approach to solving this problem is to capture many individuals displaying the same facial expressions, then use multi-linear models to separate expression from identity (and any other source of variation of interest). Expressions can then be predicted and synthesised on new faces (Vlasic, Brand, Pfister & Popovic, 2005; Macedo, Vital Brazil & Velho, 2006). Our approach to *mapping* between AAM spaces avoids the need for this complex, controlled, data capture.

The basis of our approach is not to think about the problem as mapping facial expressions between images of different people, but rather to consider it a geometric problem. A change in the facial features captured by a basis vector in one model-

space ought to be captured by a combination of the basis vectors in another model-space (because the movement of the facial features is highly constrained, as stated). The basis vectors themselves are computed when the model is trained and remain constant thereafter, so we can pre-compute the alignment of a basis **vector** from one model (the source model) with the basis **vectors** of another model (the target model) by computing the inner products

$$\langle \mathbf{s}_i, \mathbf{t}_j \rangle \quad j = 1 \ldots m.$$

These inner products are then used to weight the parameter values computed using the source model so they correctly represent, as far as possible, the same deformation on the target face. Thus, Equation 2 rewritten in matrix form and extended to incorporate the mapping can be written as:

$$\mathbf{t} = \mathbf{t}_0 + \mathbf{T}(\mathbf{R}\mathbf{p}_s), \tag{7}$$

where the columns of $\mathbf{T}$ are the basis vectors spanning the *target* space, $\mathbf{R}$ is a $q \quad r$ matrix of inner products (the target space is of dimension q and the source of dimension r), and $\mathbf{p}_s$ are the parameters representing the expression in the *source* space. Note: because $\mathbf{R}$ does not depend on expression and can be pre-computed, the cost of mapping an expression is only a matrix-vector product. Equation (3), which is used to generate the appearance images, takes exactly the same form as Equation (7). Cloning facial expressions in this way has the advantage that we do not explicitly need to define anything about the facial expressions. This information is implicit from the basis vectors in the source and target spaces. Also, we are not concerned with the direction of the eigenvectors. For example, an increasingly positive value for a source parameter might, say, open the mouth, while the same action could be defined by an increasingly negative value in the target space. In this instance the inner product for that combination of vectors is negative (the displacements from the mean are largely

in opposite directions), so the sign of the parameter value is flipped when the parameter is mapped. The parameters are also constrained to lie within the chosen ± 3 standard deviation limits of the *target* model, so the appearance of the resulting synthesised faces are valid given the original training data of the target person. This ensures the characteristics of the target face are preserved after mapping. The parameters can also be scaled after mapping to manipulate the intensity of the facial expression, as described previously.

Example images illustrating the mapping between various facial expressions are shown in Figure 7. Although there is some degradation in image quality, images can be generated at video frame-rate. If greater image quality is desired, multi-segment modelling can be used to improve the rendering and reduce the effects of blurring (see Figure 8) at the additional cost of a pre-computed look up from pixels within a segment to pixels within the face image. Further effort can be applied to improve the quality of static images: however this would increase the computational cost and rendering time. In a previous study (Theobald, Kruse, Bangham, & Cawley, 2003) we found that the image degradation such as that introduced by an AAM was not statistically significant in terms of the perceived naturalness of the *speech dynamics* in comparison with, say, temporally distorted video. The temporal properties of facial gestures are unaffected by the manipulation we have described in this paper.

[INSERT FIGURE 7 ABOUT HERE]

[INSERT FIGURE 8 ABOUT HERE]

SUMMARY AND DISCUSSION

In this paper we have described techniques for manipulating and transferring facial expressions from video sequences containing one face to video sequences containing other faces. AAMs are constructed for a source face and one or more target

faces. The source face is tracked in a video sequence using the inverse compositional AAM fitting algorithm and the model parameters representing the facial expressions are mapped to the target face(s) and re-rendered, all in real-time at video frame-rate. The advantages of using AAMs are: 1) the mapping is simple and intuitive and requires no manual specification of the type or degree of facial gestures. 2) The model can account for a high degree of variability in the images, which offers more flexibility than purely image-based approaches. 3) Near-videorealistic avatars for new subjects can be created without the cost of recording and processing a complete training corpus for each – visual speech and expression information can be synthesized for one face and transferred to another. 4) The model allows the mapped expressions to be constrained, so that they best match the appearance of the target producing the expression. 5) The model parameters can be manipulated easily to exaggerate or attenuate the mapped expression, which would be difficult using comparable image-based approaches.

The methods described here will be used in perceptual studies of dyadic conversation in a videoconferencing environment. The facial expressions of one conversant will be manipulated as described and the effect of these adjustments on behaviour during real-time natural conversations will be investigated. The advantage of this approach is both interlocutors are blind to any manipulation, either in terms of the expressiveness of their actions, or the identity or apparent gender of the person they appear to be.

REFERENCES

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3d faces. In *Proceedings of SIGGRAPH*, pages 187–194.

Blanz, V., Basso, C., Poggio, T., & Vetter, T. (2003). Reanimating faces in images

and video. *Eurographics*, pages 641–650.

Blanz, V., & Vetter, T. (2003). Face recognition based on fitting a 3d morphable model. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *25*(9), 1063–1074.

Bregler, C., & Omohundro, S. (1995). Nonlinear manifold learning for visual speech recognition. *Proceedings of the International Conference on Computer Vision*, 494–499.

Chang, Y., & Ezzat. T. (2005). Transferable videorealistic speech animation. *ACM SIGGRAPH/Eurographics Symposium on Computer Animation,* pages 29–31.

Choe, B., Lee, H., & Ko. H. (2001). Performance-driven muscle-based facial *Animation. Journal of Visualization and Computer Animation, 11*(2),67–79.

Chuang, E., & Bregler, C. (2002). Performance driven facial animation using blendshape interpolation. *Technical Report CS-TR-2002-02*, Stanford University.

Cootes, T., Edwards, G., & Taylor. C. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 23*(6), 681–685.

Cosker, D., Marshall, D., Rosin, S., Paddock, P., & Rushton, S. (2004). Towards perceptually realistic talking heads: Models, metrics, and McGurk. *Proceedings of Symposium on Applied Perception in Graphics and Visualization*, pages 2474–2477.

Edwards, G., Cootes, T., & Taylor, C. (1998). Face recognition using active appearance models. *Proceedings of the European Conference on Computer Vision,* pages 581–595.

Fagel, S., Bailly, G., & Elisei, F. (2007). Intelligibility of natural and 3D-cloned German speech. *Proceedings of Auditory Visual Speech Processing*, 126–131.

Gehricke, J., & Shapiro, D. (2000). Reduced facial expression and social context in

major depression: Discrepancies between facial muscle activity and self-reported emotion. *Psychiatry Research, 95*(2), 157–167.

Govokhina, O., Bailly, G., Bretin, G., & Bagshaw, P. (2006). TDA: A new trainable trajectory formation system for facial animation. *Proceedings of Interspeech*, pages 2474–2477.

Grahe, J., & Bernieri, F. (1999). The importance of nonverbal cues in judging rapport. *Journal of Nonverbal Behavior, 23*(4), 253–269.

Gross, R., Matthews, I. & Baker, S. (2006). Active appearance models with occlusion. *Image and Vision Computing, 24*(6), 593–604.

Guenter, B., Grimm, C., Wood, D., Malvar, H., & Pighin. F. (1998). Making faces. In *Proceedings of SIGGRAPH*, pages 55–66.

Hong, P., Wen, Z., Huang, T., & Shum, H. (2002). Real-time speech-driven 3D animation. *Proceedings of 3D Data Processing, Visualization and Transmission*, pages 713–716.

Kirby, M., & Sirovich, L. (1987). Application of the Karhunen-Loeve procedure for the characterization of human faces. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 12*(1), 103–108, January.

Koufakis, I., & Buxton, B. (1999). Very low bit-rate face video compression using linear combination of 2D face views and principal components analysis. *Image and Vision Computing, 17*, 1031–1051.

Liu, Z., Shan, Y., & Zhang, Z. (2001). Expression mapping with expressive ratio images. *SIGGRAPH*, pages 271–276.

Luettin, J., Thacker, N., & Beet, S. (1996). Speechreading using shape and intensity information. *Proceedings of the International Conference on Spoken Language Processing,* pages 58–61.

Macedo, I., Brazil, E., & Velho, L. (2006). Expression transfer between photographs through multilinear AAMs. *Brazilian Symposium on Computer Graphics and Image Processing,* pages 239–246.

Massaro, D. (1998). *Perceiving talking faces*. MIT Press.

Matthews, I., & Baker, S. (2004). Active Appearance Models Revisited. *International Journal of Computer Vision, 60*(20), 135–164.

Matthews, I., Cootes, T., Bangham, A., Harvey, R., & Cox, S. (2002). Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 24*(2), 198–213, February.

Mori, M. (1970). The uncanny valley. *Energy*, 7(4), 33–35.

Mogghaddam, B., & Pentland, A. (1995). An automatic system for model-based coding of faces. *Proceedings of the Data Compression Conferences*, pages 362–370.

Noh, J., & Neumann, U. (2001). Expression Cloning. *SIGGRAPH*, pages 277–288.

Schwartz, G., Fair, P., Salt, P., Mandel, M., & Klerman, G. (1976). Facial expression and imagery in depression: An electromyographic study. *Psychosomoatic Medicine, 28*(5), 337–347.

Shackleton, M., & Welsh, W. (1991). Classification of facial features for recognition. *Proceedings of Computer Vision and Pattern Recognition*, pages 573–579.

Sirovich, L., & Kirby, M. (1987). Low-dimensional procedure for the characterization of human faces. *Journal of the Optical Society of America (A), 4*, 519–524.

Terzopoulos, D. & Waters, K. (1993) Analysis and synthesis of facial image sequences using physical and anatomical models. *IEEE Transactions of Pattern Analysis and Machine Intelligence, 15*(6), 569–579.

Theobald, B., Kruse, S., Bangham, A., & Cawley G. (2003). Towards a low-

bandwidth talking face using appearance models. *Image and Vision Computing*, *21*(13–14), 1117–1124.

Theobald, B., Bangham, A., Matthews, I. & Cawley, G. (2004). Near-videorealistic Synthetic Talking Faces: Implementation and evaluation. *Speech Communication, 44*, 127–140, 2004.

Theobald, B. & Wilkinson, N. (2007). A real-time speech-driven talking head using appearance models. *Proceedings of Auditory Visual Speech Processing*, 264–269.

Turk, M., & Pentland, A. (1991). Eigenfaces for recognition. *Journal of Cognitive Neuro-Science*, 3(1), 71–86.

Vlasic, D., Brand, M., Pfister, H. & Popovic, J. (2005). Face transfer with multilinear models. *ACM Transactions on Graphics, 24*(3), 426–433.

Wang, Y., Lucey, S., & Cohn, J. (2008). Enforcing convexity for improved alignment with constrained local model. *In Proceedings of Computer Vision and Pattern Recognition*.

Williams, L. (1990). Performance driven facial animation. *Computer Graphics, 24*(2), 235–242.

Xiao, J., Baker, S., Matthews, I., & Kanade, T. (2004). Real-time combined 2D+3D active appearance models. *In Proceedings of Computer Vision and Pattern Recognition*, pages 535–542.

Yang, J., Zhang, D., Frangi, A., & Yang, J. (2004). Two-dimensional PCA: A new approach to appearance-based face representation and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 26*(1), 131–131, January.

Zhang, Q., Liu, Z., Quo, G., Terzopoulos, D. & Shum, H. (2006). Geometry-driven

photorealistic facial expression synthesis. *IEEE Transactions on Visualization and Computer Graphics, 12*(1), 48–60.

FIGURE CAPTIONS

Figure 1. An example AAM derived from fifteen hand-labelled images from a longer sequence of a single individual. Shown are the mean shape and appearance (A and D respectively), and the first two modes of variation of the shape (B-C) and appearance (E-F) components of the model. Note the AAM includes the region of the face only to the eyebrows. There are no defining features that are easily identifiable on the forehead that can consistently be placed manually across individuals. Work is ongoing to overcome this limitation.

Figure 2. Example face images synthesised using AAMs trained on three different individuals.

Figure 3. Facial expressions of varying intensity rendered using an AAM. Left column shows the expressions scaled to 25% intensity, the middle column shows the expressions as measured in original video, and the right column shows expressions exaggerated to 150% of the intensity measured in the video. The effect of scaling the parameters is much more pronounced in more extreme expressions.

Figure 4. Dissociating expression and identity by *copying* changes in the features of the face that characterise expressions. Here expressions on (A) a reference face are copied to other faces of (B) the same gender and (C) the opposite gender. As expressions are copied, the characteristics of the expression are the same on all faces. For example, the nasiolabial furrows, the shape of the smile, the mole on the cheek and the teeth are all characteristic of the person in (A).

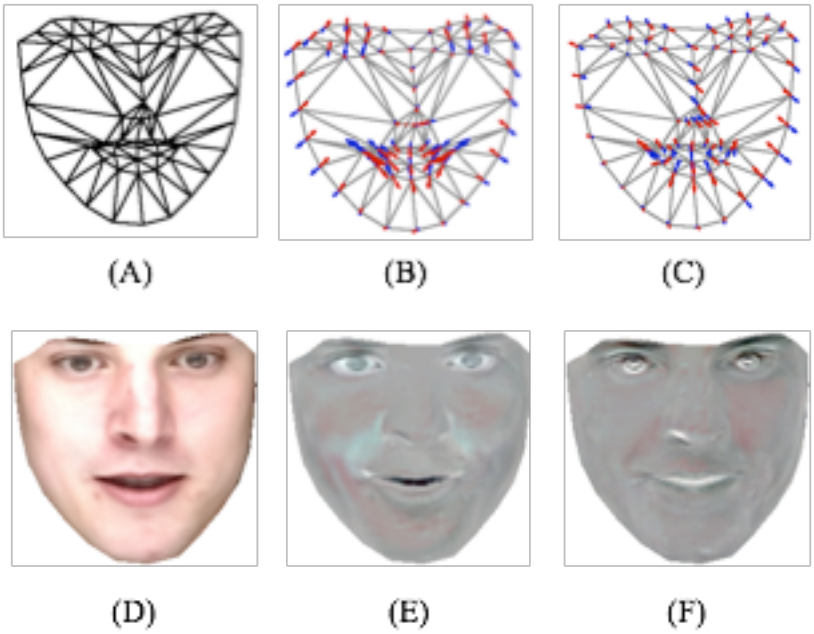
Figure 5. The base appearance of three AAMs each trained on twenty images of the respective individual. The mean appearances converge to approximately the same expression.

Figure 6. Illustration of the mismatch in alignment between model-spaces. The parameters (A) in the source space representing the expression to be mapped can be copied directly to (B) the same location on the 2D plane. However the resulting misalignment means that this point is not in the same location relative to the target basis vectors. Thus, the expression generated does not match the original. Instead, the parameters should be mapped by considering the rotation of the basis vectors between the model-spaces such that (A) is mapped to the point (C) in the same relative location between model-spaces.

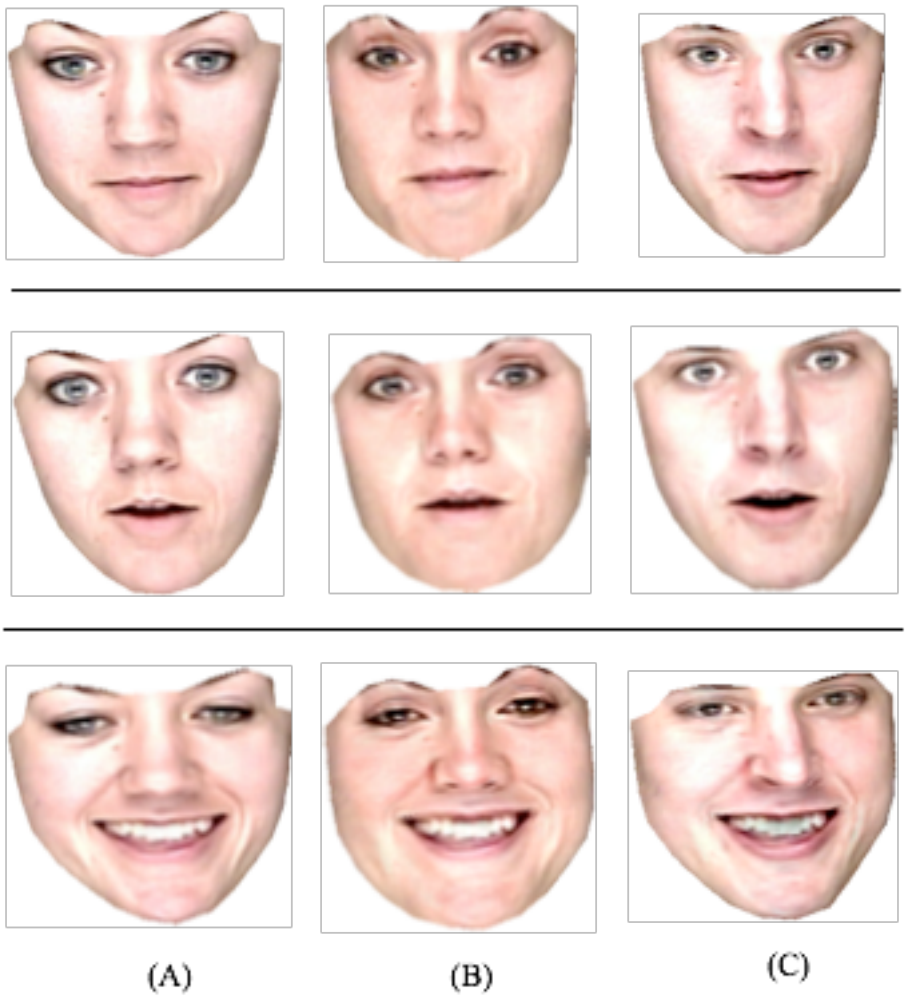
Figure 7. Dissociating expression and identity by *mapping* changes in the features of the face that characterise expressions. Here expressions on (A) a reference face are mapped to another face (B) of the same gender and (C) of the opposite gender. As expressions are mapped the characteristics of the expression are, as far as possible, consistent with expression changes on the respective people. These are the same expressions as those shown in Figure 5. Notice the mole on the cheek no longer appears on the cloned faces, and the teeth for each person appear as their own. In Figure 5, the same teeth as is visible in (A) are copied into the inner mouth for each cloned person.

Figure 8. Multi-segment modelling and re-rendering. (A) An expression on a source confederate as measured in a video sequence and re-rendered on their model, and the

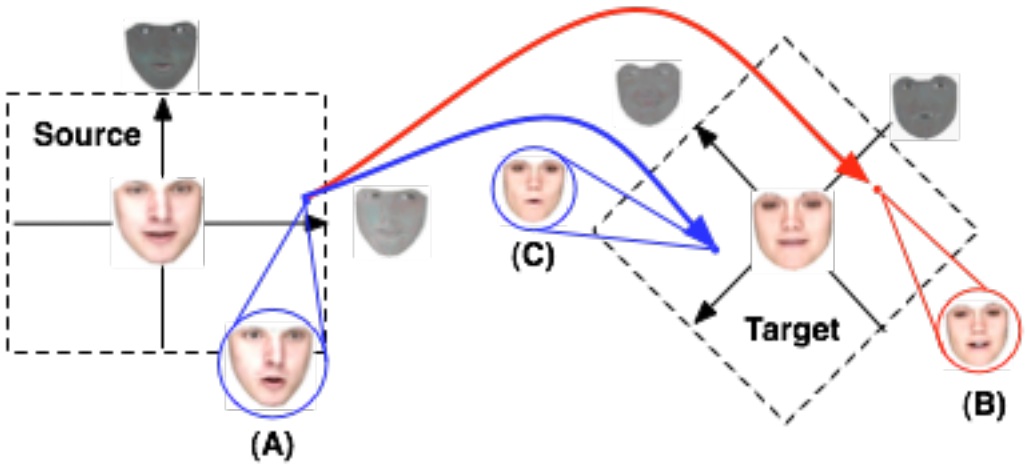
expression mapped to a target face and rendered using (B) a standard AAM and (C) a multi-segment AAM. The multi-segment model reduces some of the blurring artefacts apparent in the standard AAM rendered image.

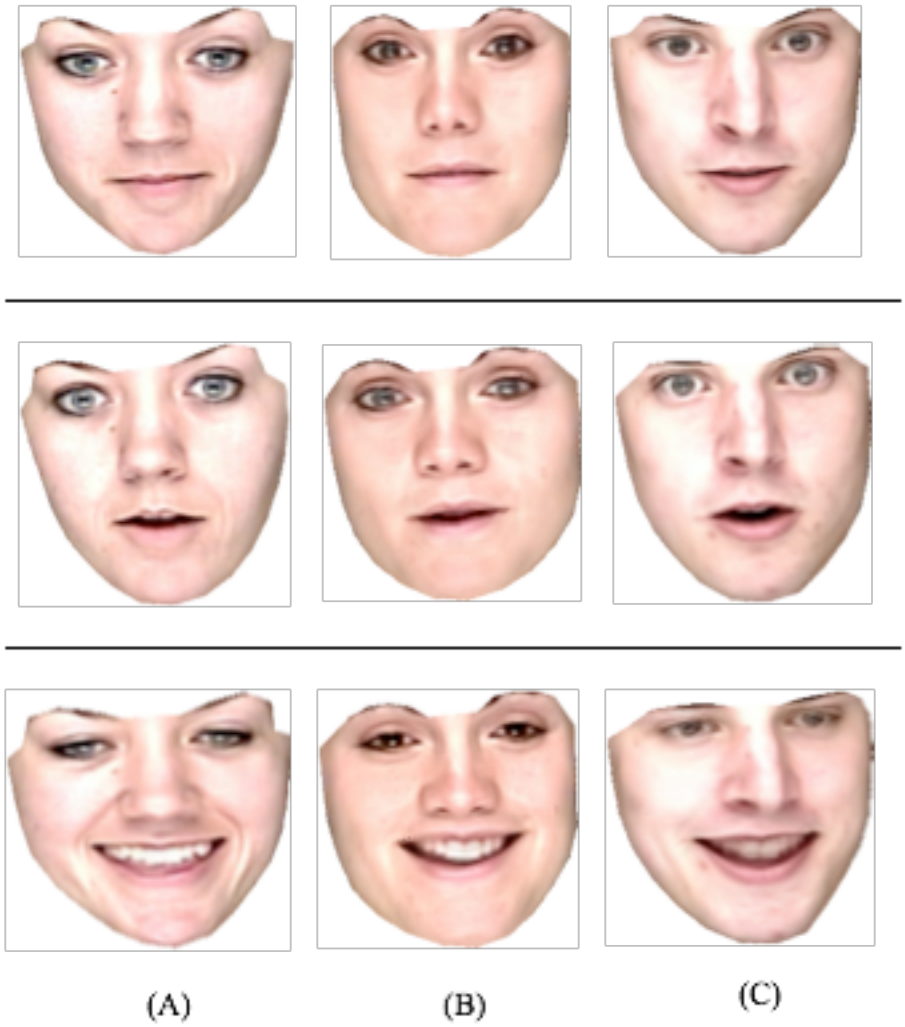

(A)
(B)
(C)
(D)
(E)
(F)

25% 100% 150%



(A) (B) (C)


Source
(A)
(C)
Target
(B)



(A) (B) (C)

(A) (B) (C)