Providing High Social Presence for Mobile Systems via an Unobtrusive Face Capture System

Miguel A. Figueroa-Villanueva[1], Frank A. Biocca[2], Chandan K. Reddy[1],
Jannick P. Rolland[3], George C. Stockman[1]
[1]Computer Science and Engineering Department, Michigan State University, East Lansing, USA,
[2]Department of Telecommunications, Michigan State University, East Lansing, USA,
[3]College of Optics and Photonics, University of Central Florida, Orlando, USA
{[1]miguelf@ieee.org, [2]biocca@msu.edu, [3]jannick@odalab.ucf.edu}

Abstract

During face-to-face collaboration people frequently monitor the other's facial expressions to determine their current state of attention, mood, and comprehension. Capturing a frontal view of the face of mobile users in multi-user collaborative environments has been a challenge for several years. A mobile social presence system is proposed that captures two side views of the face simultaneously and generates a frontal view in real-time. The face is modeled using an active appearance model (AAM) and a mapping of the side model to the frontal model is constructed from training. Frontal views are then generated by applying this mapping to the fitted side model during collaboration. Only a few model coefficients are transmitted for the synthesized facial frames, providing a highly compressed stream. The virtual frontal videos are of good subjective quality and the fitted estimate retains a high fidelity to the true model, with peak signal to noise ratio of about 40DB.

Keywords--- **Mobile face capture, head mounted display (HMD), active appearance model (AAM)**.

1. Introduction

One key motivation for the creation of mobile communication systems and advanced collaborative environments is the increase in real-time communication between mobile and distributed partners. During face-to-face collaboration, users frequently monitor the other's facial expressions to determine their current state of attention, mood, and comprehension. Although we may have face-to-face interactions with workmates or others, many of our social interactions include an increasing number of purely virtual interactions; we rarely or never meet face-to-face. When it comes to communications from remote places, the human face is the most important communicative part of the human body. It has great expressive ability that provides a continuous stream of cues that are used to modulate and tailor interpersonal communication. Mediated communications now include many cases where facial non-verbal information can be critical: negotiation, complex training, emergency communication, stressful or tense interactions, communication of positive affect, and group coordination and motivation. The facial expressions of a remote collaborator or a mobile user can convey a sense of urgency, emotional congruency, lack of understanding or confidence in action, or other nonverbal indicators of communication success or breakdown. With an increase reliance on telecommunication systems for group interaction, there is increased research in advanced social presence technologies. Current advanced teleconferencing and telepresence systems transmit frames of video. These frames are nothing but 2D images from a particular point of view. In order to get additional views, designers use either a panoramic system or interpolate between a set of views.

New and enhanced forms of remote collaboration through sophisticated environments such as those presented in [1-4] provide augmented reality features for a higher degree of *presence* of the remote collaborator in the communication channel and, potentially, free movement and unlimited views of the shared augmented reality environment.

The *Teleportal System* [5] is an augmented reality environment for remote communication and collaboration among multiple users. This effort envisions a Teleportal room such as in [4] that allows single or multiple users to enter a room sized display and use a broadband telecommunication link to engage in face-to-face interaction with other remote users in a 3D, augmented reality environment, hence providing a simultaneous interaction with virtual objects, real objects and models while supporting object interposition. It also allows for unobstructed 3D face-to-face capture and display. This unique feature is designed to support interaction between fully mobile virtual representations of user's faces in 3D space so that their position relative to other participants and objects under discussion is preserved. The goal is to support all the non-verbal and position cues of side-by-side collaboration including attentional cues (e.g., "where is the person looking now"), turn taking and other conversation modulation cues, and situated cues regarding emotional and comprehension states. Figure 1 shows a representation of the kind of interaction that the Teleportal System allows.

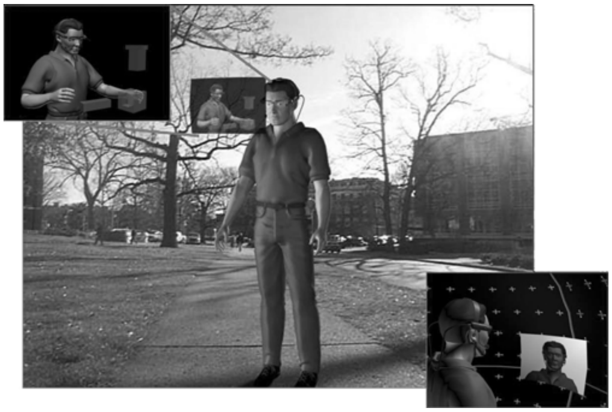

Figure 1 Conceptual drawing of the application scenario enabled by the Teleportal System. Two distant users are interacting on a task. One user is instructing how to proceed from a mobile location, while the other is executing the task. Both have visual feedback of the other's environment.

1.1. Main objectives

Capturing a clear, detailed frontal view of the face of mobile users in multi-user collaborative environments has been a challenge for several years. Technologies that occlude the user's field-of-view are not practical and potentially dangerous in full mobile outdoor settings. Other applications of facial capture systems include teleconferencing, wearable computing, and collaborative mixed reality environments. The Mobile Face Capture System (MFCS) is responsible for obtaining and transmitting a quality frontal face video of a remote user involved in the communication. The MFCS proposed here captures the two side views of the face simultaneously and generates the frontal view. This face capture equipment consists of two miniature video cameras and convex mirrors [5]. Figure 2 shows a conceptual drawing that illustrates the face-capture cameras and the mirrors with respect to the user's head. Each of the cameras is pointed towards the respective convex mirror, which is angled to reflect an image of one side of the face. The convex mirrors produce a slight distortion of the side view of the face. The left and right video cameras capture the corresponding side views of the human face in real-time. The goal of the work in this paper is to synthesize a frontal view facial image from the two side views recorded by the head mounted display (HMD) side cameras.

1.2. Advantages

Consider the contrast with conventional capturing techniques, where either the face capture system is static within the environment, for example a single camera mounted on a display, or the capture system is bulky, costly, and computationally expensive, for example a room instrumented with a sea of cameras [1]. The MFCS system is static with respect to the user's head movements, uses only two cameras to produce a wide range of views of a user's head including a possible stereoscopic view, can capture the face regardless of location, and works on any basic processor.

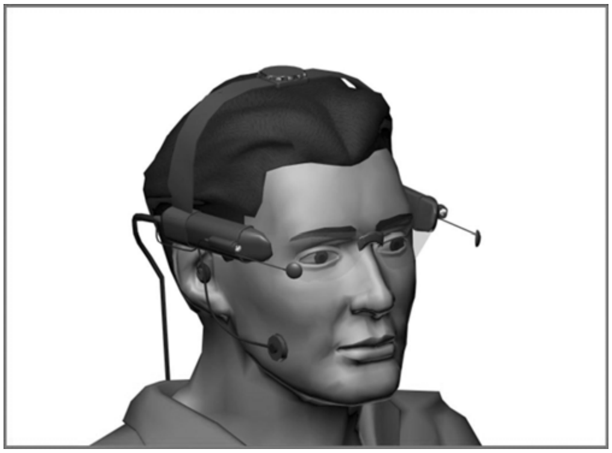

Figure 2 Mobile Face Capture System (MFCS) concept with two convex mirrors and two lipstick cameras.

Most previous systems have been built using a highly instrumented fixed indoor environment, while our work is motivated by a need to be mobile. In its current implementation it relies on the use of a head mounted display (HMD) that includes a projective display and mobile face capture system (MFCS).

The HMD will ultimately allow all participants to: (a) view 3D images of the face of remote collaborators, (b) view unobstructed the real local participants and objects, and (c) view the blending of physical and virtual objects. Although the MFCS system can be used with any display, combining it with an augmented reality (AR) HMD allows the user to see the 2D or 3D faces of collaborators in appropriate locations for interpersonal communication relative to their body or the environment.

Our current MFCS prototype consists of two side cameras and front mirrors as depicted in Figure 2. The basic requirement of the MFCS is that it must produce quality video of the wearer's face without interfering with the ability to perform other required tasks such as object manipulation and the 3D visualization of a remote communicator or of shared data or objects with that communicator. However, around one's office, a participant may reach out to data or files to share with others and any obstruction of any one participant's direct view would prohibit executing those tasks. We also anticipate the use of MFCS with outdoor, fully mobile AR systems or next generation mobile phones. In this demanding setting, the MFCS approach can minimize visual occlusions so as to not interfere with simple walking, driving, object manipulation or non-mediated face-to-face interaction. This current work forms a stepping-stone for the creation of a complete 3D augmented reality based face-to-face communication system that can produce stereoscopic views of the users via a real-time augmented reality display.

1.3. Organization of the paper

The remainder of this paper is organized as follows. Section 2 describes the relevant background for the MFCS. Section 3 describes the hardware system design. The equipment used and the optics issues are discussed. The algorithms and methods used in the MFCS are explained in Section 4. Section 5 illustrates some results of using the prototype MFCS. Section 6 presents conclusions of the work with the MFCS and suggests ideas for future work.

2. Related background

To synthesize a frontal view facial image, a model of the face is created during the training stage. This model is used to characterize the input streams at run time. The model is also used to create or instantiate the desired views (i.e., the frontal view, but potentially a wide range of views).

Face modeling has been used as a tool to aid in a large number of applications such as person identification, face surveillance, face animation, expression cloning, etc. As a result, there are a number of techniques employed to model the face. They can be categorized into 2D and 3D techniques. This paper follows a 2D analysis by synthesis approach: namely, *Active Appearance Models*, for its robustness and computational efficiency.

Active Appearance Models (AAMs) [6-8], which first appeared in [9], are non-linear, generative, and parametric statistical models of a certain deformable object in the 2D image plane. In particular, face modeling has been one of the most popular applications of AAMs [9].

The typical application scenario of AAMs involves a training phase, where the model is built, and a fitting phase, where a search is made to find the optimal model parameters that minimize the distance from the generated model instance and the input image. A detailed and comprehensive survey on the subject of AAMs and the closely related concepts of Active Blobs, Direct Appearance Models, and Morphable Models can be found in [10].

Other approaches to image synthesis have been reported [11] that could be used to synthesize images by interpolating some reference views of a static scene. In [12, 13] some extensions were made to be able to handle dynamic scene interpolation. These techniques rely on estimating the epipolar geometry of the scene and having a set of reliable correspondences. Also, care has to be taken to properly blend or interpolate between images to obtain a visually pleasing result. Typically, AAMs are less sensitive and error-prone than such approaches, at the expense of having to provide a database of training samples.

Reddy et. al. [14] proposed a method for synthesizing a frontal face image using a similar HMD. The proposed approach was to calibrate the system to obtain a set of warping functions to map pixels from the side images to virtual frontal image coordinates. A structured light grid was projected onto the face from the front and the deformation in the side images recorded to be used for warping during operation. Problems included use of structured light in the field, the blending of the two side images at their seam in the frontal image, and image distortion created by facial expressions not modeled well by the static warp. In contrast, the approach of this paper produces consistently smooth images and high compression. The costs are several minutes of training and fitting and the need to store the models at both the sender and receiver [15].

3. Face capture system design

3.1. Current hardware performance

The current HMD prototype cameras, Sony DXC-LS1 with Fujinon YF12B-7 lenses, are tethered by cables to a P4 1.7 GHz PC with 496 MB RAM. A Panasonic GP-KR202 video camera is positioned on the desk to take a real frontal video during training. The subject puts on the MFCS and minor adjustments may be made to the orientation of the mirrors. The subject then faces the Panasonic camera and speaks and gestures using a standard script. Standard office lighting is used. The system records synchronized video from the side MFCS and frontal observing cameras. Current storage resources limit us to recording 70 frames per session. The longest step in the training process is the manual identification of face feature points in the side (46 points) and frontal images (95 points), which may take 15 minutes. Fitting the AAM models to the side and front images and fitting the mapping from the side images to the frontal images is done in real-time. Thus, the entire training time for a single subject is currently about 15 minutes. Future improvements, including more automation in face point identification and sharing of data between subjects should reduce the training to 2 to 10 minutes. During the user task, generation of the virtual frontal video can be done in real time. Matthews and Baker [10] have shown that similar computations can be performed at over 260 frames per second.

Notice that there exist mobile and wireless counterparts for all the equipment in this prototype version, which can be replaced with off-the-shelf and dedicated hardware to obtain a mobile system.

3.2. Optical System Layout

The general layout of the system is shown in Figure 2. The calculations for estimating the variable parameters are simplified by unfolding the overall system. When the system is unfolded, the mirror can be represented as a negative lens (see Figure 3). The main components of this system are the (a) human face, (b) camera, and (c) mirror. The various parameters that are involved in the calculations are as follows.

1. *Human face:* The main parameters of the face that affect the geometry of the system are height and width. Other factors, such as skin color and illumination, affect the performance of the system but have no effect on the geometry. The dimensions of an average face are:
    - H - Height of the head to be captured (~ 250mm).

- W-Width of the head to be captured (~ 175mm).

2. *Camera:* The main parameters are the size, the weight, the minimum working distance, the field of view, and the depth of field. Based on the approximate values of these parameters, we have obtained the off-the-shelf lipstick camera, Sony DXC-LS1. The two cameras are color balanced using their built-in hardware capabilities. The 12mm focal length lens has the following values:
   - Sensing area: 1/4", or equivalently 3.2mm(y) x 2.4mm(x).
   - Pixel Dimensions: the image sensed has a resolution of 768 x 494.
   - Focal Length (F_c): The focal length of the lens selected is 12 mm (VCL - 12UVM).
   - Field of View (FOV): The field of view of the camera with the above mentioned lens is 15.2° x 11.4°.
   - Diameter (D_c): The diameter of the lens and the camera is 12mm.
   - *f*-number (N_c): The *f*-number for this camera lens is 1. Although in practice, we adjust the iris according to illumination, we consider an *f*-number of 1 in the estimation of the other parameters.
   - Minimum Working Distance (MWD): The minimum working distance for the selected lens is 200mm.
   - Depth of Field (DOF): This parameter is dependent critically on the lens *f*-number which will vary with various illuminations. The higher the *f*-number the larger the DOF. This system requires however to consider the DOF of the camera and mirror combined. If the system has large DOF then it will be more portable and can accommodate many users without much change in the position and focus of the cameras. The DOF computation for the camera and mirror combined will be treated elsewhere with an in depth development of the optical layout and design.
3. *Mirror:* This is the most flexible component of the system. Hence, all the parameters of this component are estimated and the component is custom made. The various parameters of the mirror that will affect the geometry of the system are:
   - Diameter (D_m) / *f*-number (N_m)
   - Focal Length (F_m) or Radius of Curvature (R_m)
   - Magnification Factor (M_m)
4. *Distances*: Between these three components, we have the following distances:
   - D_{cm} - Distance between the camera and the mirror.
   - D_{mf} - Distance between the mirror and the face.

3.2.1. Estimation of the Variable Parameters (D_{mf} and D_m). From the theory of pupils and windows, the camera is the limiting aperture from the intermediary image plane located behind the mirror. Hence, the camera acts as the pupil of the system and the mirror is the window.

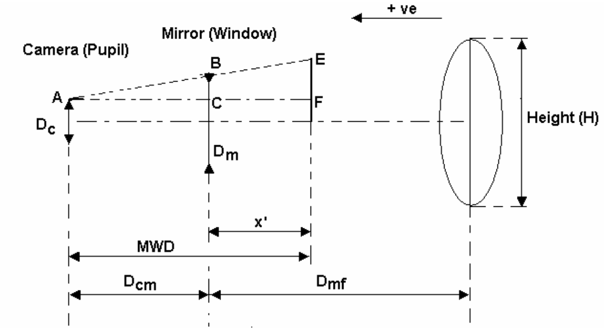

Figure 3 Optical system diagram for the estimation of the variable parameters D_{mf} and D_m.

In the unfolded configuration, the mirror is represented as a negative lens with image focal length f'_m equal in magnitude to that of the mirror with an opposite sign. The imaging equation for the equivalent lens to the mirror yields

$$\frac{1}{x'} = \frac{1}{D_{mf}} + \frac{1}{f'_m} \tag{1}$$

where *x'* is negative because the values D_{mf} and f'_m are negative. Hence, the image in the unfolded case is virtual and thus it is always between the lens and the human face. A study was made of estimated values for D_m as a function of the *f*-number and, based on the practical values for the size of the mirror (D_m) and the distances (D_{mf} and D_{cm}), the mirror was customized. A convex mirror of radius of curvature 155.04 mm was made corresponding to the *f*-number of 2. The convex side of the mirror was coated for the visible light spectrum. Figure 4 shows two sample images obtained from this optical system specification.

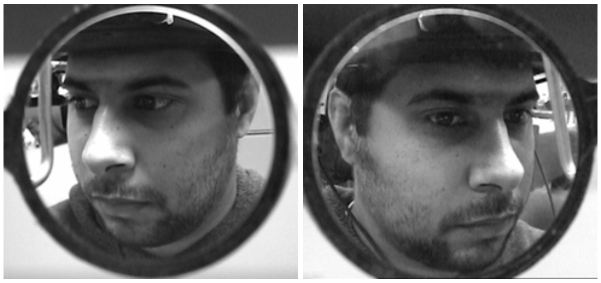

Figure 4 Sample images acquired from the current MFCS prototype with the optical specifications above.

4. Virtual view synthesis

4.1. System design

We present a generative and parametric method for face video synthesis. We build an AAM model from training data and use a regularization technique to determine the mapping between the AAM parameters for the side view model and the parameters for the front view model. Figure 5 depicts the training process where the goal is to build the corresponding AAM models and estimate the linear operator that describes the forward mapping.

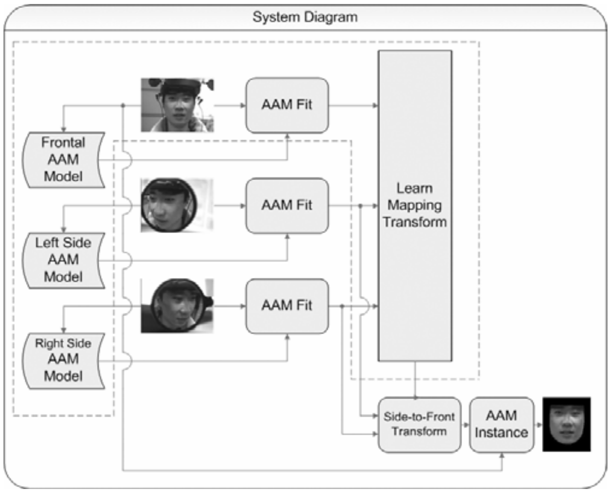

Figure 5 ***System diagram.*** **From training, the AAM model is learned along with the transformation between side view parameters and frontal view parameters. At runtime the learned transformation is used to estimate frontal view parameters that instantiate the frontal view AAM.**

After the forward linear operator is estimated, it can then be used to predict the frontal parameters for the respective AAM, as shown in Figure 5.

4.2. AAM modeling

This section describes the basic formulation of the AAM technique that provides the basis of our design. It is divided into the AAM model creation, the model instantiation, and the fitting process and it follows the notation presented in [10].

4.2.1. Model definition. AAMs model the shape, which accounts for the rigid form as well as the possible deformations, and texture (i.e., lighting intensity) of an object.

The shape is defined as a closed triangulated mesh, which can be represented as a vector containing the concatenation of vertex locations:

$$\mathbf{s} = (x_1, y_1, x_2, y_2, \cdots, x_v, y_v)^T \tag{2}$$

where v is the number of vertices of the mesh.

If there are n training shape vectors, then we can assume (provided that enough samples are given) that any new shape can be explained as a linear combination of those given as training:

$$\mathbf{s} = \mathbf{s}_0 + \sum_{i=1}^{n} p_i \mathbf{s}_i \tag{3}$$

where $\mathbf{s}_0$ is the mean shape and the $\mathbf{s}_i$'s are the variations or deformations from the mean. p_i's are the shape parameters.

The texture or appearance can be defined as the pixel intensities relative to the mean shape $\mathbf{s}_0$. Let $\mathbf{x}$ be the pixel locations in $\mathbf{s}_0$, then

$$A(\mathbf{x}) = A_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i A_i(\mathbf{x}) \quad \forall \mathbf{x} \in \mathbf{s_0} \tag{4}$$

is the appearance function.

4.2.2. Model instantiation. Given a set of parameters, $\mathbf{p} = (p_1, p_2, \cdots, p_n)^T$ for the shape and a set of parameters, $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \cdots, \lambda_m)^T$ for the appearance, an image can be synthesized corresponding to an instantiation of the model. The shape and appearance are generated independently by applying the parameters to Equations 3 and 4, respectively. However, the appearance is defined in terms of the mean shape $\mathbf{s}_0$, which requires warping to the generated shape instance. This process can be represented as:

$$I_M(W(\mathbf{x}; \mathbf{p})) = A(\mathbf{x}) \tag{5}$$

where $W(\mathbf{x;p})$ is a piecewise affine warp from $\mathbf{s}_0$ to $\mathbf{s}$. $\mathbf{x}$ defines the pixel in $\mathbf{s}_0$ to be warped and $\mathbf{p}$ determines the shape $\mathbf{s}$ to be warped to.

4.2.3. Model fitting. In the fitting phase the goal is to search for the model parameters that minimize the error between the current image and the model instance for those parameters. This error can be defined as:

$$E(\mathbf{x}) = A_0(\mathbf{x}) + \sum_{i=1}^{m} \lambda_i A_i(\mathbf{x}) - I(W(\mathbf{x}; \mathbf{p})) \tag{6}$$

where the first term corresponds to the appearance defined by parameters $\boldsymbol{\lambda}$ of the model at pixel $\mathbf{x}$ in the base mesh, $\mathbf{s}_0$, and the second term corresponds to the pixel in the input image as determined by the warp $W(\mathbf{x;p})$. Hence, the problem has been reduced to an optimization problem with cost function $E(\mathbf{x})$ $\forall\, \mathbf{x} \in \mathbf{s}_0$ and parameters $\boldsymbol{\lambda}$ and $\mathbf{p}$ to search for. It should be noted that in practice, Principal Component Analysis (PCA) is applied to the shape and texture vectors, which makes the search more manageable.

4.3. Face modeling

Currently, the side view models are created with a mesh of 46 points and the frontal model with a 95 point mesh. In Figure 6 the contours of the base meshes are presented with two sample deviations along the first principal component direction. Note that this corresponds to the opening and closing of the mouth.

4.4. Frontal parameter estimation

The process of training and fitting an AAM has been briefly described in Section 4.2. After obtaining a synchronized stream of M images of the subject from the MFCS side cameras and a frontal camera, one can use this processing to obtain two row vectors $\mathbf{y}_i$ and $\mathbf{x}_i$ containing the frontal and side parameters for the i^{th} image, respectively. This can be written as:

$$\mathbf{y}_i = (\mathbf{p}_{Fi}^T, \lambda_{Fi}^T) \quad (7)$$

$$\mathbf{x}_i = (\mathbf{p}_{Li}^T, \lambda_{Li}^T, \mathbf{p}_{Ri}^T, \lambda_{Ri}^T) \quad (8)$$

for $i = 1, \cdots, M$. *F, L, R* indicate the front, left, and right side parameters, respectively.

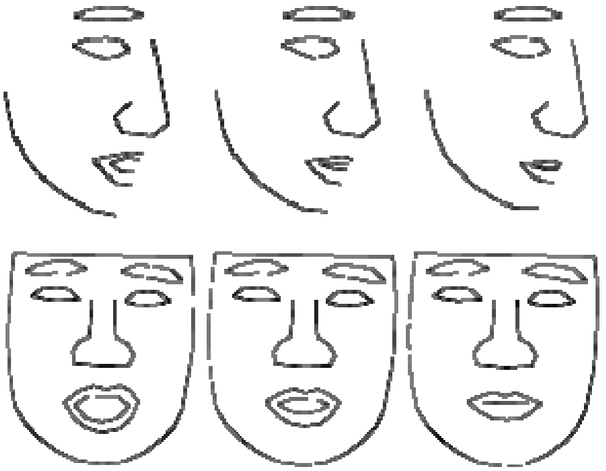

Figure 6 AAM side and front shape mesh contours and two sample variations along the first principle component (mode 1).

To fit a *P*-degree polynomial to the data we can write:

$$y_{ij} = a_{0j} + \mathbf{x}_i\mathbf{a}_{1j} + \mathbf{x}_i^2\mathbf{a}_{2j} + \cdots + \mathbf{x}_i^P\mathbf{a}_{Pj} \quad (9)$$

where x_i^k denotes *element-by-element* exponentiation by *k*, y_{ij} is the j^{th} element of the i^{th} sample vector $\mathbf{y}_i$, and a_{oj}, $\mathbf{a}_{ij}$ are the coefficients that determine the polynomial.

If we let $\mathbf{y}_{\cdot j}$ denote the j^{th} column of the matrix Y, which has *M* rows equal to the stack of $\mathbf{y}_i$ row vectors, then we can write in matrix notation:

$$\begin{pmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{ij} \\ \vdots \\ y_{Mj} \end{pmatrix} = \begin{pmatrix} 1 & \mathbf{x}_1 & \mathbf{x}_1^2 & \cdots & \mathbf{x}_1^P \\ 1 & \mathbf{x}_2 & \mathbf{x}_2^2 & \cdots & \mathbf{x}_2^P \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \mathbf{x}_i & \mathbf{x}_i^2 & \cdots & \mathbf{x}_i^P \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \mathbf{x}_M & \mathbf{x}_M^2 & \cdots & \mathbf{x}_M^P \end{pmatrix} \begin{pmatrix} a_{0j} \\ \mathbf{a}_{1j} \\ \mathbf{a}_{2j} \\ \vdots \\ \mathbf{a}_{Pj} \end{pmatrix}$$

or alternatively,

$$\mathbf{y}_{\cdot j} = X\mathbf{a}_{\cdot j} \quad (10)$$

The least squares (LS) solution to the system in Equation 10 minimizes the residual error $\left\| y_{\cdot j} - Xa_{\cdot j} \right\|$ and is given by:

$$\mathbf{a}_{\cdot j} = (X^TX)^{-1}X^T\mathbf{y}_{\cdot j} \quad (11)$$

for $j = 1, \cdots, B$.

Tikhonov regularization can be employed to reduce the effects of noise in the data and numerical instability related to small singular values of A. In essence, we parameterize the solution to the system in Equation 10 obtaining a balance between trying to fit the data (i.e., reduce the residual error of the solution) and constraining the solution to a minimal norm. The *regularization parameter* μ determines this balance and the solution becomes:

$$\mathbf{a}_{\cdot j}^{(\mu)} = \arg\min\{\| \mathbf{y}_{\cdot j} - X\mathbf{a}_{\cdot j} \| + \mu \| \mathbf{a}_{\cdot j} \|\} \quad (12)$$

It can be shown that for a given μ the solution that minimizes Equation 12 is:

$$\mathbf{a}_{\cdot j}^{(\mu)} = (X^TX + \mu I)^{-1}X^T\mathbf{y}_{\cdot j} \quad (13)$$

One common method to choose the value of μ is to select the value that minimizes the *generalized cross-validation* (GCV) defined by:

$$V(\mu) \equiv \frac{\| \mathbf{y}_{\cdot j} - X\mathbf{a}_{\cdot j} \|}{[Tr(I - X(\mu))]^2} \quad (14)$$

where $X(\mu) = XX^T(XX^T + \mu I)^{-1}$.

5. Experimental results

In this section, we first introduce the results of the AAM modeling on each view of the face and then follow with a discussion of the quantitative results for the parameter estimation.

5.1. AAM Models

An AAM of the frontal face and the side view images was built for each subject as described in Section 4. It was built using 8 frames out of a 71 frame video stream per view. The subsets were spaced at 10 frames apart and the AAM was built to capture 99% of the variation when PCA was applied. This reduced the representation of the face to a model parameterized by only 6-7 coefficients (i.e., we are able to synthesize an image of the face for each of the 71 frames with at most 7 floating point numbers). Figure 9 presents samples of the synthesized faces of two subjects as well as the original frames. The synthesized images are very similar to the original images and they properly convey the facial expressions of the subjects.

5.2. Frontal Parameter Estimates

A *leave-one-out* approach was followed to estimate the residual differences reported in this section. This is done to properly estimate the expected error for unseen images (i.e., images that weren't used to find the solution) and avoid overfitting the data. Two basic measures are reported here: the residual differences of the parameters of the AAM models, $\left\| y_{\cdot j} - Xa_{\cdot j} \right\|$, and the *peak signal-to-noise-ratio* (PSNR), which provides a standard measure of similarity between the originally synthesized frontal image and the one synthesized by estimating the parameters using the fitted polynomial.

The PSNR between two $M \times N$ grayscale images, I and $\hat{I}$, is given by:

$$PSNR = 20 \log_{10} \frac{255}{RMSE} \ dB \tag{15}$$

where

$$RMSE = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} (I(i,j) - \hat{I}(i,j))^2}{NM} \tag{16}$$

In Table 1, it is shown how the residual error for the first coefficient tends to zero as the degree of the polynomial used is increased, while the estimated error using the leave-one-out approach starts increasing after polynomial degree 2. This is an indication that for the limited amount of data that we are currently using (i.e., 71 frames) we can not apply a polynomial fit with degree over 2 and care has to be taken not to be misled by the absolute residual difference.

Error Method	Polynomial Degree			
	1	2	3	4
Absolute Error	0.0019	0.0011	0.0009	0.0005
Leave-one-out	0.0024	0.0023	0.0070	0.0160

Table 1 Absolute residual error vs. leave-one-out estimate for parameter one.

In Table 2, a summary of the residual differences for the first two parameters and the PSNRs is presented. The effect of increasing the polynomial degree without providing enough data is clearly observed for the least squares (LS) solution where the error mean (μ) and the standard deviation (σ) steadily increase. It shows that the regularized least squares (RLS) solution does not blindly rely on the data and therefore is more robust to noise (e.g., outliers) and avoids overfitting the data.

Figure 7 shows the PSNR average +/- the standard deviations for the LS solution and RLS solution as a function of the polynomial degree. It can be observed how the regularized approach has slightly higher PSNR values and partially overcomes the over fitting problem, while the LS approach has a faster decreasing average PSNR and increasing standard deviation.

It should be noted that although the differences in PSNR are not substantial, they are very significant. They should not be disregarded as insignificant, given that as more variation is introduced to the AAM model the number of coefficients necessary to parameterize the face will increase and more ambiguity will be present in the mapping, making these gaps larger. Also, notice how in Figure 8 the image generated by the LS approach is highly distorted, while the one synthesized by RLS is much smoother.

Finally, Figure 9 shows, in the last two rows, the originally synthesized frontal image and the one synthesized by estimating the parameters using a polynomial of degree 2.

Subject I

Method	PSNR (dB)		Coeff 1		Coeff 2	
	μ	σ	μ	σ	μ	σ
LS1	39.55	**2.98**	0.0024	0.0040	**0.0003**	0.0006
LS2	40.26	3.67	0.0022	0.0040	**0.0003**	**0.0005**
LS3	39.70	4.12	0.0069	0.0274	0.0011	0.0060
LS4	39.64	5.02	0.0161	0.0834	0.0021	0.0117
LS5	38.40	6.43	0.2659	1.6369	0.0201	0.1263
RLS1	39.76	3.11	0.0023	0.0037	**0.0003**	0.0006
RLS2	**40.38**	3.39	**0.0019**	**0.0031**	**0.0003**	**0.0005**
RLS3	40.06	3.82	0.0049	0.0214	**0.0003**	**0.0005**
RLS4	40.35	3.99	0.0041	0.0219	**0.0003**	0.0011
RLS5	39.92	4.41	0.0109	0.0539	0.0017	0.0068

Table 2 Results for one subject (other subjects follow similar patterns) of the PSNR and Parameter Residuals for the first two parameters shown for the LS and RLS solutions. The number next to LS and RLS indicates the polynomial degree.

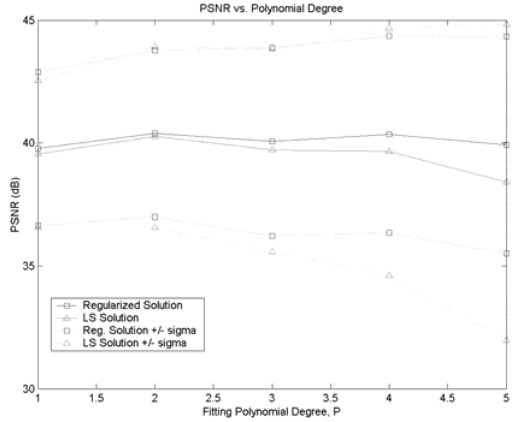

Figure 7 PSNR mean and standard deviation plot.

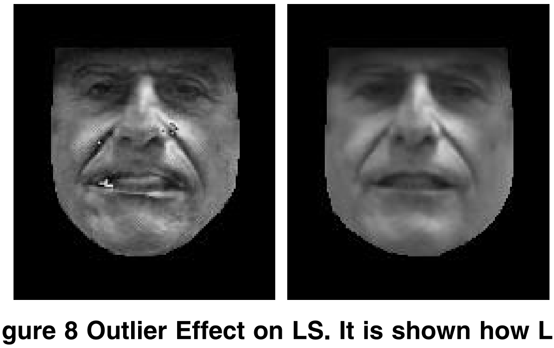

Figure 8 Outlier Effect on LS. It is shown how LS1 (left) is more susceptible to outlier effects than RLS1 (right).

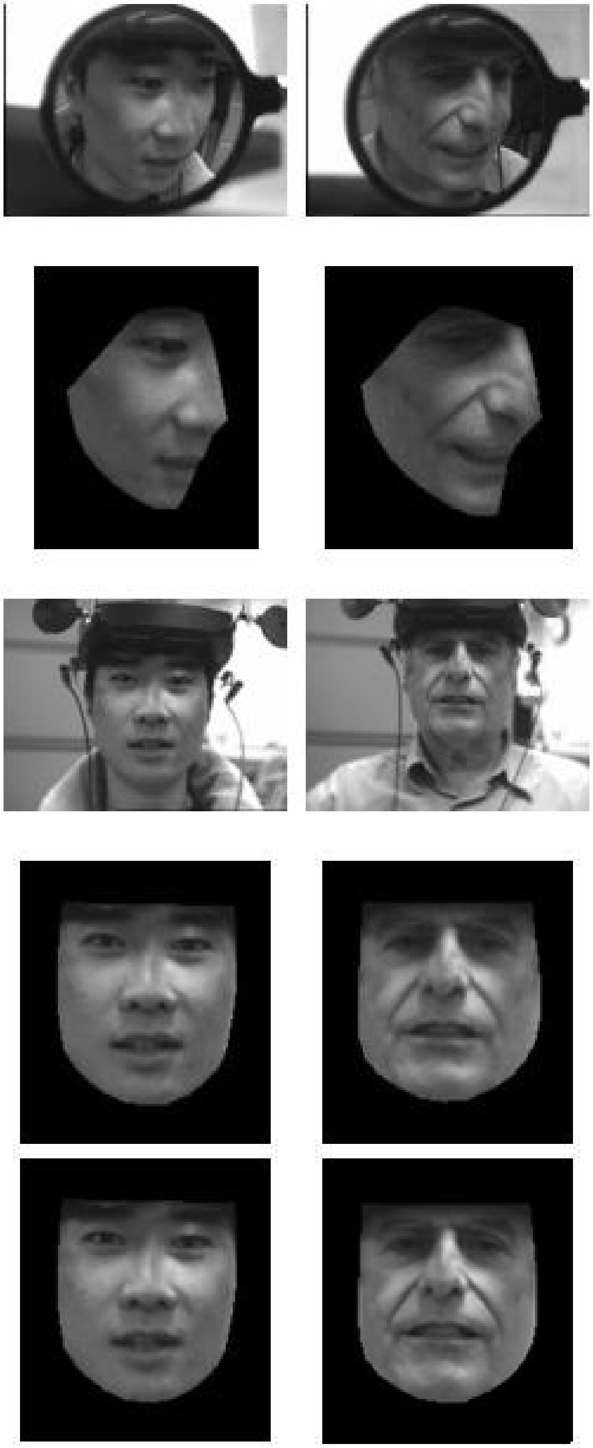

Figure 9 Samples of two side (top) and frontal (center) original views and the AAM model instantiation for each and the respective estimated images (bottom).

6. Concluding discussion

We have designed a system to capture a video stream of two side views of the face using the MFCS and a supplementary view from a third camera used only during the training stage. By modeling the three views of the face using an AAM and finding the regularized solution to the mapping between the two side views and the supplementary view, we can estimate the parameters for this missing view at run time.

By solving this problem using a statistical generative method, we avoid the difficulties associated with blending the two separate images and pose estimation. Our generated videos are of good subjective quality and maintain a high fidelity, about 40 dB PSNR, between the original model and the estimated one. Furthermore, we have a completely automatic system at run time. The AAM's and the linear regularization techniques employed have proven to be efficient maintaining this application in the real time domain. We conclude that our MFCS and mathematical methods support the intended collaborative distributed applications.

Implemented in a full mobile system, this approach offers the possibility of communicating the full facial expression of a mobile user anywhere and anytime when higher levels of social presence are needed for example emergency, affective, or procedural communication. It is important to note that our frontal videos are generated from video frames taken during training. While this is sufficient for communicating the state of mind of the collaborator, it is not a video or telepresence system. The face is reconstructed from an analysis of changing parameters. For example, it cannot communicate aspects of the current environment; for example, the reflection of a fire on a firefighter's face. Future research will investigate methods to blend in the environmental lighting, when needed. Future directions include generating full 3D models of the remote user's face, creating a fully mobile prototype, adding temporal correlation information to the process of estimating the synthetic view parameters, and evaluating the effect of this additional social presence on user behavior during mobile collaboration and communication.

Acknowledgements

The authors would like to thank Timothy F. Cootes for making publicly available his AAM software implementation.

References

[1] H. Fuchs, G. Bishop, K. Arthur, L. McMillan, R. Bajcsy, S. Lee, H. Farid, and T. Kanade. Virtual Space Teleconferencing using a Sea of Cameras. In *Proceedings of the First International Symposium on Medical Robotics and Computer Assisted Surgery (MRCAS)*, Pittsburgh, PA, USA, September 1994, pp. 161-167.

[2] R. Raskar, G. Welch, M. Cutts, A. Lake, L. Stesin, and H. Fuchs. The Office of the Future: A Unified Approach to Image-Based Modeling and Spatially Immersive Displays. In *Proceedings of SIGGRAPH 98*, July 1998, pp. 179-188.

[3] L-Q. Xu, B. Lei, and E. Hendriks. Computer Vision for a 3D Visualization and Telepresence Collaborative Working Environment. In *BT Technology Journal*, 20:64-74, January 2002.

[4] H. Hua, L.D. Brown, C. Gao, and N. Ahuja. A New Collaborative Infrastructure: SCAPE. In *Proceedings of the IEEE Computer Society Conference on Virtual Reality (VR)*, Los Angeles, CA, USA, March 2003, pp. 171-179.

[5] F. Biocca and J.P. Rolland. Teleportal Face-to-Face System. US Patent Number 6,774,869, issued August 10, 2004.

[6] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active Appearance Models. In *Proceedings of the 5th European Conference on Computer Vision (ECCV),* Freiburg, Germany, June 1998, pp. 484-498.

[7] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active Appearance Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI),* 23(6):681-685, June 2001.

[8] T.F. Cootes and P. Kittipanya-ngam. Comparing Variations on the Active Appearance Model Algorithm. In *Proceedings of the British Machine Vision Conference (BMVC),* Cardiff, UK, September 2002.

[9] G.J. Edwards, C.J. Taylor, and T.F. Cootes. Interpreting Face Images using Active Appearance Models. In *Proceedings of the 3rd International Conference on Automatic Face and Gesture Recognition (FG)*, Nara, Japan, April 1998, pp. 300-305.

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

[11] S.M. Seitz and C.R. Dyer. View Morphing. In *Proceedings of SIGGRAPH 96*, 1996, pp. 21-30.

[12] R.A. Manning and C.R. Dyer. Interpolating View and Scene Motion by Dynamic View Morphing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Ft. Collins, CO, USA, June 1999, pp. 1388-1394.

[13] Y. Wexler and A. Shashua. On the Synthesis of Dynamic Scenes from Reference Views. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* Hilton Head, SC, USA, June 2000, pp. 1576-1581.

[14] C.K. Reddy, G.C. Stockman, J.P. Rolland, and F.A. Biocca. Mobile Face Capture for Virtual Face Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW)*, Washington, DC, USA, June 2004, pp. 77-83.

[15] F.G. hamza Lup and J.P. Rolland. Scene Synchronization for Real-Time Interaction in Distributed Mixed Reality and Virtual Reality Environments. *Special issue Collaborative Virtual Environments, PRESENCE*, 13(3):315-327, 2004.