HIGH-PRESENCE, LOW-BANDWIDTH, APPARENT 3D VIDEO-CONFERENCING WITH A SINGLE CAMERA

Timothy R. Brick
Human Dynamics Laboratory
University of Virginia
tbrick@virginia.edu

Jeffrey R. Spies
Human Dynamics Laboratory
University of Virginia
jspies@virginia.edu

Barry-John Theobald
School of Computing Sciences
University of East Anglia
b.theobald@uea.ac.uk

Iain Matthews
Disney Research, Pittsburgh
iain.matthews@disney.com

Steven M. Boker
Human Dynamics Laboratory
University of Virginia
boker@virginia.edu

ABSTRACT

Small digital video cameras have become increasingly common, appearing on portable consumer devices such as cellular phones. The widespread use of video-conferencing, however, is limited in part by the lack of bandwidth available on such devices. Also, video-conferencing can produce feelings of discomfort in conversants due to a lack of co-presence. Current techniques to increase co-presence are not practical in the consumer market due to the costly and elaborate equipment required (such as stereoscopic displays and multi-camera arrays).

To address these issues, this paper describes a real-time, full frame-rate video-conferencing system that provides simulated three-dimensionality via motion parallax in order to achieve a higher level of co-presence. The system uses a deformable 3D face model to track and re-synthesize each user's face using only a single monocular camera, so that only the (few tens of) parameters of the model need be transferred for each frame. This both provides motion-tracking for the simulated 3D experience and reduces the bandwidth requirements of the video-conference to the order of a few hundred bytes per frame.

Bandwidth and processor usage for the algorithms are reported. Possible implications and applications of this technology are also discussed.

1. INTRODUCTION

As small digital video cameras have become less costly and more ubiquitous, showing up on consumer products such as laptops, PDAs, and cellular phones, video-conferencing has seen increasingly widespread usage. In order to gain widespread acceptance, video-conferencing faces two major difficulties. First, full frame-rate transmissions have heavy bandwidth requirements, and average users are often constrained to low frame-rates and poor quality transmission, even with the use of video compression algorithms. Second, basic video-conferencing lacks a feeling of *common presence* and *shared space*, fundamentally changing the dynamics of conversation and potentially causing the users to feel uncomfortable [5].

This material is based upon work supported in part by NSF Grant BCS–0527485. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

More advanced video-conferencing systems create a sense of co-presence by inducing the perception of three-dimensionality. This can be accomplished using binocular disparity technologies such as stereoscopic displays or augmented reality systems. An alternate approach, motion parallax, approximates a three-dimensional experience by rotating a 3D model of the object based on the user's viewing angle. This has been reported to provide a greater feeling of co-presence than the binocular approach [1], but current implementations require the use of expensive motion-tracking technologies such as multi-camera, optical tracking arrays [3]. Regardless of the ultimate display technology, whether binocular or motion-based, imaging the object to be displayed requires at least a two-camera system.

This paper describes a real-time, full frame-rate (30+ fps) video-conferencing system that provides an experience of co-presence without the need for expensive displays, multi-camera arrays, or elaborate motion-capture equipment. This system is capable of using a single, commodity camera and an Active Appearance Model—a statistical model of the face—to capture and display near-photorealistic video of each user's face, while simultaneously tracking each user's movement. The system's representation of each user's appearance results in extremely low bandwidth usage. Tracking allows the use of motion parallax to create a three-dimensional experience.

2. ACTIVE APPEARANCE MODELS

An Active Appearance Model (AAM) is a generative, parametric model that encapsulates a compact statistical representation of the shape and the appearance of an object [2]. AAMs are most commonly used to track [4] and synthesize [6] faces in images and video sequences, where the compact nature of the model allows faces to be tracked, manipulated and rendered all at video frame-rate and during live face-to-face conversation over a video link [9].

The *shape* component of an AAM is represented by n two-dimensional (2D) vertices, $\mathbf{s}_0 = (x_1, y_1, \ldots, x_n, y_n)^T$, connected to form a triangulated mesh, and a set of basis shapes, $\mathbf{s}_i$, that define the allowed variation in the shape. Any particular instance of a shape is generated from a linear combination of basis shapes added to $\mathbf{s}_0$:

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

where the coefficients p_i are the shape parameters that represent the shape $\mathbf{s}$. The shape component of an AAM is typically computed

by applying principal components analysis (PCA) to a set of shapes hand-labeled in a set of images. In this instance $\mathbf{s}_0$ is the mean shape and the vectors $\mathbf{s}_i$ are the (reshaped) eigenvectors corresponding to the m largest eigenvalues. The vertices that define the structure of the shape are typically chosen to delineate the facial features (eyes, eyebrows, nose, mouth, and face outline). To ensure the model is sufficiently generative, i.e. *all* facial expressions of interest can be represented as some combination of the basis vectors, the hand-labeled images must contain a suitably diverse collection of facial expressions.

The *appearance* component of an AAM is defined as an image, $\mathbf{A}_0(\mathbf{x})$ formed of the pixels $\mathbf{x} = (x, y)^{\mathrm{T}} \in \mathbf{s}_0$, and a set of basis images, $\mathbf{A}_i(\mathbf{x})$, that define the allowed variation in the appearance. The appearance component of an AAM is computed by first shape normalizing the images by warping from the hand-labeled vertices to $\mathbf{s}_0$, then applying PCA to the resultant image set. Again, $\mathbf{A}_0(\mathbf{x})$ is the mean image and $\mathbf{A}_i(\mathbf{x})$ are the (reshaped) eigenvectors corresponding to the l largest eigenvalues. As with the shape, any particular shape-normalized image is generated using a linear combination of the basis images added to $\mathbf{A}_0(\mathbf{x})$:

$$\mathbf{A}_0(\mathbf{x}) = \mathbf{A}_0(\mathbf{x}) + \sum_{i=1}^{l} \lambda_i A_i(\mathbf{x}) \qquad \forall\, \mathbf{x} \in \mathbf{s}_0, \tag{2}$$

where the coefficients λ_i are the appearance parameters. See the bottom row of Figure 1 for an example of an appearance component of an AAM.

To render a near-photorealistic image of a face from a set of AAM parameters, first the shape parameters, $\mathbf{p} = (p_1, \ldots, p_m)^{\mathrm{T}}$, are used to generate the shape, $\mathbf{s}$, of the AAM using Eq. (1). Next the appearance parameters $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_l)^{\mathrm{T}}$ are used to generate the AAM appearance image, $\mathbf{A}(\mathbf{x})$, using Eq. (2). Finally $\mathbf{A}(\mathbf{x})$ is warped from $\mathbf{s}_0$ to the generated shape $\mathbf{s}$.

To label an existing image automatically with an existing AAM, an analysis-by-synthesis approach is used. First, an initial estimate (e.g. the parameters representing the previous frame) of the parameters that represent the shape and appearance in the image of interest is generated. Next an image is synthesized by applying the parameters to the model, and finally a gradient-descent error minimization is used update the parameters to minimize the residual between the image being fitted to and the model-synthesized image. There are a wealth of algorithms proposed for performing this minimization [4]. The trade-off is typically speed versus accuracy.

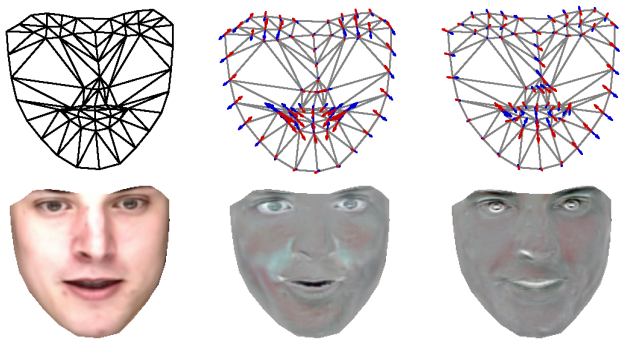

Fig. 1. The shape (top) and appearance (bottom) components of an AAM. For each row the left column illustrates the mean shape/appearance, and the center and right columns the first two modes of variation of the shape/appearance components.

2.1. Multi-segment 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 is generally not the case when considering faces and consequently some important information is considered noise and discarded, which results in blurring in the rendered images. This is most striking in the eyes and inner mouth which tend to be the more important areas of the face. An extension to improve the quality of rendering is to construct a piece-wise PCA by building independent appearance models for the different regions of the face (skin, eyes, 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 the individual segments can then be regenerated and copied into the appearance vector $A(\mathbf{x})$ before warping to the shape $\mathbf{s}$. This allows different model segments to be encoded with more/less resolution, allocating more resources to regions of the face on which a viewer's attention is likely to be focused, such as the eyes and mouth.

3. MOTION PARALLAX

Motion parallax is a visual effect by which humans gauge the distance to objects [5]. It is caused by the apparent rotation of an object as the viewer moves around that object. That is, as the viewer moves to her right, she is able to see more of the left side of the object.

Apparent motion parallax is achieved by estimating the user's viewing angle with respect to the screen, and rotating the generated view appropriately. We assume for the purposes of this paper that the user should see an unrotated view of the co-conversant when the user's face is centered at $(0, 0)$. In most cases, the apparent horizontal and vertical viewing angles (θ and ϕ, respectively) can be calculated from the overall displacement of the face along the horizontal and vertical axes (x and y, respectively), and the estimated distance from the camera d.

$$\theta = \arcsin \frac{x}{d} \qquad \phi = \arcsin \frac{y}{d}$$

While the distance d to the camera can be precisely calculated if the actual size of the user's face and the focal length of the camera are known, we have found it more expedient to simply provide the user with a tool to adjust the distance manually, requesting that they adjust it until the parallax effect is achieved.

4. SYSTEM ARCHITECTURE

The system presented consists of two computers with commodity digital video cameras and standard displays connected by a network. As each frame of video captured by a camera is processed by the local computer, the system fits the user's AAM to that frame and extracts the appropriate model parameters to capture the image and location of the local user's face. The model parameters are then sent across the network to the remote machine, which then reconstructs the image of the local user's face. The camera-relative location of the local user's face is then used to perform the appropriate rotations to achieve motion parallax three-dimensionality. In this way, the system can provide head tracking for three-dimensional display, while encoding the image of the user's face as a small number of parameters for low-bandwidth transfer. The overall system architecture is shown in Figure 2.

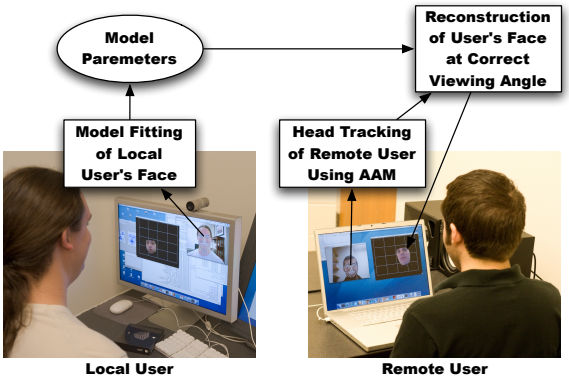

Fig. 2. This diagram displays the system architecture. After fitting the local user's AAM to a frame of video, model parameters are sent across the network to the remote user's computer, where they are used to reconstruct the image of the local user.

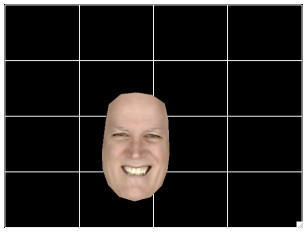

Fig. 3. An example of a reconstructed facial image, as it is shown to the remote user.

It is important to note that each computer must have access to pre-constructed AAMs of both the local and remote user. While it is possible to construct multi-person models, in general each model is tailored to a single individual. Full models are large (between 5 and 15 MB, depending on the resolution of the model images and the complexity of the model) but need only be shared once. The bulk of the shared model size is due to the appearance eigenvectors, each of which is a mapping from a parameter to each pixel, essentially a full-resolution image in itself. The size of the models can therefore be reduced by subsampling the mean image and the appearance eigenvectors to reduce the number of pixels represented by the appearance model. A full-face example model, shown in Figure 3, requires just over 13 MB to perform full-quality reconstruction. As seen in Figure 4, however, it is quite possible to construct a reasonable image after transmitting less than a single megabyte. Although the resolution of each reconstructed image would be quite low, it might be sufficient for display on, for example, a two-inch-wide cellular phone screen. A multi-segment model could be used to ensure that there is sufficient resolution in the parts of the face that are deemed more important, for example the eyes and mouth.

It would also be possible to send a small but working model before the conversation begins (taking a few tens of seconds to send) and to send improvements to the model as bandwidth allows during the conversation. The addition of a multi-segment model would allow the most important segments to be updated first, while details such as skin texture might follow later.

Because the downsampled model will still use the same parameters as the full-pixel model, it is even possible to reconstruct a higher-resolution image from a lower-resolution camera, provided the original model was built at the higher resolution. While the individual pixels of this higher-resolution model are not guaranteed to be accurate, the resulting images should still appear to be high-resolution images of the co-conversant and the model will ensure that the resulting face is always valid, in the sense that it will be similar to the images originally labelled for the model.

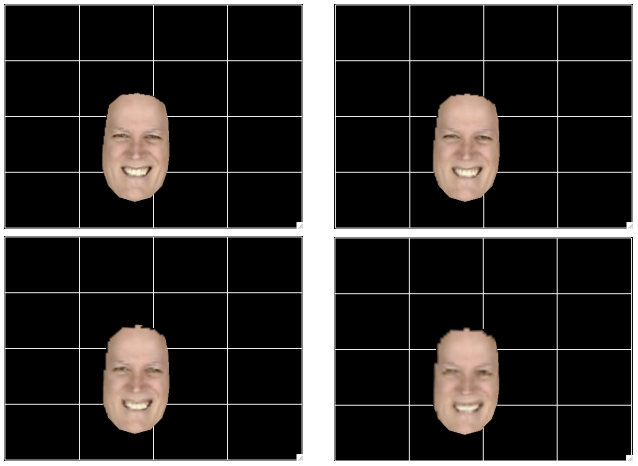

Fig. 4. These images were the result of downsampling the model; in order from top-left to bottom-right, the model generating each image is of size 3 MB, 1.6 MB, 925 KB, and 400 KB. The fully rendered image, shown in Figure 3, uses a model of size 13 MB.

5. RESULTS

While the exact network performance of the system depends on the fidelity of the model and the exact choice of labelled frames, fully labelled models generally require about thirty parameters (fifteen for shape and fifteen for appearance) for video-realistic reproduction. If each parameter is transmitted at floating-point precision, a thirty-parameter model results in 240 bytes of data transmitted between the machines per full frame. At these speeds, a 30 fps video stream requires a total (two-way) transfer of 7200 Bytes per second, i.e., 57.6Kbps. This number reflects raw parameter values, and could almost certainly be reduced further by keyframing or other forms of compression. Theobald, et. al. [7] were able to reduce the transmission size to 3.6 Kbps with minimal perceptual impact.

Two fully labelled pre-shared models were used for performance evaluation. A shape model including the forehead, eyes, eyebrows, nose, outer and inner mouth, and chin was designed and built. Each model extracted roughly 95% of facial variance within the chosen region and the resulting real-time rendering of the face was near-photo-realistic, as shown in Figure 3. In an experiment using same-quality facial renderings, none of the 28 participants guessed that the reconstructed video was not a cropped live video stream[8]. A video demonstrating the system can be downloaded from:
`http://people.virginia.edu/~trb6e/wiamis09/demo.mov`

One side of the system ran on a Macbook Pro with a Dual-core 2.5 GHz Xeon processor, 4 GB of RAM, and a built-in iSight web camera. The other side ran on a Mac Pro with Dual Quad-core 3.8 GHz Xeon processors, 8GB of RAM, and an external iSight camera. To perform fitting, transfer and reproduction, the Macbook Pro required on average 33.2% of the processing power of a single core

at any given time, and maintained a maximum memory usage of 75 megabytes of memory. Because the amount of memory usage and processing power scale down with the number of pixels in the fit and reconstructed models, each model could easily be downsampled to match the video and processing capabilities of the receiving device.

6. DISCUSSION AND FUTURE WORK

As a result of the statistical representation provided by the AAMs, the bandwidth consumed by this system is meager by today's standards. Because the system requires only a single camera per user, it could be incorporated into low-bandwidth consumer devices, such as laptops, cell phones, and PDAs, many of which are already manufactured with a single digital camera.

Tracking the location of the local user's face with the AAM and displaying an image of the remote user's face rotated appropriately provides apparent three-dimensionality via motion parallax for increased co-presence. It is expected that this increase in co-presence will be associated with a more comfortable interactive experience. A user study to test this hypothesis empirically is currently in progress.

Because the display of each user is reconstructed from a digital model, it is possible to perform manipulations on the transmitted video by altering the model. For example, Theobald, et. al. [9] have mapped the appearance of one face onto the motions of a second person by manipulating the AAM. This manipulation allows the generation of near-photo-realistic, real-time video where the user's perceived identity, including gender, have been changed. Other possibilities for manipulation include expression and the dynamics of conversations. This provides numerous possibilities for research in the social sciences, as well as media applications, such as animation and online games. For example, an online game could capture the expression of a player, transmit it to a central server using negligible bandwidth, and map that expression to the face of the player's character for display to other players. The possibility of undetectable manipulation also raises questions of trust in video-conferencing, but these issues are beyond the scope of this paper.

The primary limitation of the system is related to the creation of the models. Hand-labeling facial images for an accurate, robust AAM takes approximately two hours, though it need only be done once. This labeling, if done incorrectly, can add error-variance to the models, resulting in poor fit and unrealistic synthesized video. Models are also susceptible to poor fitting in lighting conditions drastically different from the lighting in the original environment. Further work is therefore needed to better automate this process and reduce the amount of hand-labeling needed by the model.

Some work is already in progress on robust generic statistical models, such as CLMs [10]. While these models are not yet viable for real-time tracking and redisplay, they may in the future be easy ways to automatically label, or possibly even replace, the use of AAMs for video-conferencing. Because these models differ only in the way that the fit of the model is achieved, the present system would still be able to interface with them normally.

Even with the current implementation that requires labeling time to create models, it seems likely that users would be willing to devote such time in order to be able to perform high-presence video-conferencing on their mobile devices. The online gaming and chat communities have already demonstrated their willingness to spend long hours tuning avatars to precise specifications, and it seems likely that they would similarly be willing to spend the time required to label images for an accurate AAM.

As previously mentioned, the size of a complete AAM model file (5-15 megabytes) makes it difficult to share. For those devices such as cellular phones with tight bandwidth restrictions or low processing power, lower resolution models could easily be shared at the beginning of the conversation. To optimize this transfer, devices intending to engage in video-conference might first send specifications about display and camera resolution and bandwidth so that each side could quickly downsample the appearance components of a high-quality model for transmission. Again, multi-segment models could ensure that important areas of the face are rendered at higher resolution than less important areas.

Alternately, a model could "trickle" down to the user over a long period of time before the video-conference. For example, a model could be transmitted to a user's cell phone once the co-conversant's contact information was entered. For any video-conference planned in advance, appropriate models with could be easily shared this way.

7. CONCLUSIONS

This paper demonstrates a real-time, full frame-rate, low-bandwidth video-conferencing system that provides an experience of co-presence via motion parallax using a single commodity camera. It is believed that this system may increase the acceptance and usage of video-conferencing technologies as communications tools by bringing these qualities to consumer devices.

8. REFERENCES

[1] W. Barfield, C. Hendrix, and K. Bystrom. Visualizing the structure of virtual objects using head tracked stereoscopic displays. In *Proceedings of the 1997 Virtual Reality Annual International Symposium*, 1997.

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

[3] B. Lei and E. Hendriks. Middle view stereo representation: An efficient architecture for teleconference with handling occlusions. In *IEEE International Conference on Image Processing*, 2001.

[4] I. Matthews and S. Baker. Active appearance models revisited. *International Journal of Computer Vision*, 60(20), 2004.

[5] B. O'Conaill, S. Whittaker, and S. Wilbur. Conversations over videoconferences: An evaluation of the spoken aspects of video-mediated communication. *Human-Computer Interaction*, 8(4):389–428, 1993.

[6] B. Theobald, J. Bangham, I. Matthews, and G. Cawley. Near-videorealistic synthetic talking faces: Implementation and evaluation. *Speech Communication*, 44:127–140, 2004.

[7] B. Theobald, S. Kruse, G. Cawley, and J. Bangham. Towards a low bandwidth talking head using appearance models. *Journal of Image and Vision Computing (IVC)*, 21:1077–1205, 2003.

[8] B.-J. Theobald, I. Matthews, M. Mangini, J. R. Spies, T. Brick, J. F. Cohn, and S. Boker. Mapping and manipulating visual prosody. (In Review).

[9] B.-J. Theobald, I. A. Matthews, J. F. Cohn, and S. M. Boker. Real-time expression cloning using appearance models. In *International Conference on Multimodal Interaction*, 2007.

[10] Y. Wang, S. Lucey, and J. Cohn. Enforcing convexity for improved alignment with constrained local models. In *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2008.