Mobile Face Capture for Virtual Face Videos

Chandan K. Reddy , George C. Stockman
Department of CSE
Michigan State Univ.
East Lansing, MI 48824
{reddycha, stockman}@cse.msu.edu

Jannick P. Rolland
School of Optics
Univ. of Central Florida
Orlando, FL 32816
jannick@odalab.ucf.edu

Frank A. Biocca
Department of Telecom.
Michigan State Univ.
East Lansing, MI 48824
biocca@msu.edu

***Abstract*— Minimally obtrusive face capture is essential for applications in tele-conferencing, collaborative work, and mobile applications. Acquiring and coding video for such applications is a challenging problem, particularly for a mobile user. A pilot head-mounted display system (HMD) is shown that captures two side views of the face and generates in real-time a quality frontal video of the wearer. The face is captured with little obstruction of the users field of view. The frontal views are generated by warping and blending the side views after a calibration step. In tests of a bench prototype, the generated virtual videos compared well with real video based on both objective and qualitative criteria. With success of the pilot work, we are continuing development of the mobile HMD system.**

I. Introduction

Communication of the expressive human face is important to tele-communication and distributed collaborative work. For example, how can a mobile emergency response team be networked to develop added capabilities for intelligent actions? How can a remote student experience a stronger presence of the teacher? In addition to the several sophisticated collaborative work environments [5], [9], [15], there is a strong popular trend for the merger of cell phone and video functionality at consumer prices. At both ends of the technology spectrum, there is a problem producing quality video of a persons face without interfering with that persons ability to perform some task requiring both visual and motor attention. When the person is mobile, the technology of most collaborative environments is unusable. The solution proposed here is to modify a helmet mounted display (HMD) for minimally intrusive face capture. The prototype HMD has small mirrors held above the temples and viewed by small video cameras above the ears, creating a helmet that is balanced and light and with minimal occlusion of the wearers field of view. (Figure 1). The complete HMD design includes components that display remote faces and scenes to the wearer as well as reality augmentation for the wearers environment [2], [6]. In this paper, we study only the methods that provide a virtual frontal video of the HMD wearer. This virtual video (VV) is synthesized by warping and blending the two real side view videos.

A prototype HMD facial capture system has been developed. The development of the video processing reported here was isolated from the HMD device and performed using a fixed lab bench and conventional computer. Porting and integration of the video processing with the mobile HMD hardware is ongoing work.

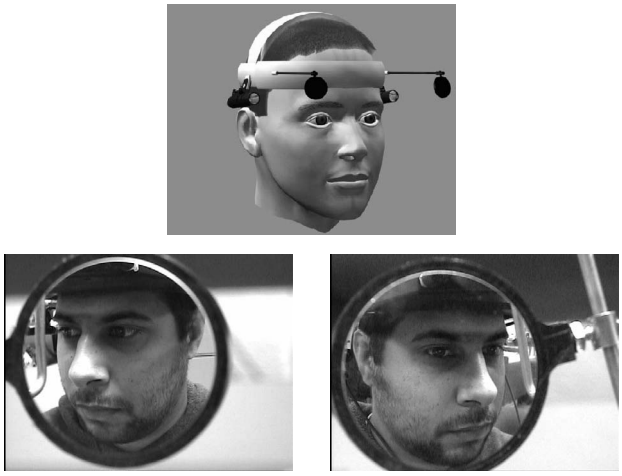

Fig. 1. Face capture concept and images from protype HMD.

Section II briefly describes some related background. Section III describes the design and analysis of the prototype system. The image processing methods are given in Section IV. Assessment of the results are discussed in Section V and conclusions in Section VI.

II. Background

Faces have been captured passively in rooms instrumented with a set of cameras [5], [8], where stereo computations can be done using selected viewpoints. Other objects can be captured using the same methods. Such hardware configurations are unavailable for mobile use in arbitrary environments, however. Other work [1], [13] has shown that faces can be captured using a single camera and processing that uses knowledge of the human face. Either the face has to move relative to the camera, or assumptions of symmetry are employed. Our approach is to use two cameras affixed to the head, which is necessary to convey non symmetrical facial expression, such as the closing of one eye and not the other, or the reflection of a fire on only one side of the face.

There is little overlap in the images taken from outside the users central field of view, so the frontal view synthesized is a novel view. In previous work, novel views have been synthesized by a panoramic system[16], [12] and/or by interpolating between a set of views [3], [11]. Producing novel views in a dynamic scenario was successfully shown for a highly rigid motion [7]. This work extended interpolation

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04)
1063-6919/04 $ 20.00 IEEE

techniques to the temporal domain from the spatial domain. A novel view at a new time instant was generated by interpolating views at nearby time intervals using spatio-temporal view interpolation[14], where a dynamic 3-D scene is modelled and novel views are generated at intermediate time intervals.

We now show how to generate in real time a synthetic frontal view of a human face from two real side views.

III. System Design

The prototye system was configured with off-the-shelf hardware and software components. We built a lab bench, shown in Figure 2, on which to develop our image processing methods. The bench was built to accomodate human subjects so they could keep their heads fixed relative to two cameras and a structured light projector. The two cameras were placed so that their images would be similar to those to be obtained from the HMD optics. The light projector is used to orient the head precisely and to obtain calibration data used in image warping. In addition to the equipment shown in Figure 2, a video camera placed on top of the projector recorded the subjects face during each experiment for comparison purposes.

A. Equipment

Our prototype uses an Intel Pentium III processor running at 746 MHz with 384 MB RAM with two Matrox Meteor II standard cards. The cards are connected to the control units of lipstick cameras; Sony DXC LS1 NTSC cameras with 12 mm focal length lenses. We use Matrox Meteor II Standard that supports both multiple composite and s-video inputs. The video is digitized by a Matrox Meteor II standard capture card, yielding interlaced 320 X 240 video fields at 60 Hz. During the off-line calibration stage, the system also uses an Infocus LP350 projector to project a grid onto the user's face. Voice is recorded in the same system using a microphone.

B. Software

The API for programming and controlling this hardware is MIL-LITE 7.0. The standard Windows based sound recording software is used to record the voice of the user during the conversation. The sound file is appended to the .avi file using Adobe Premiere 6.0. Two videos are captured simultaneously at the rate of 30 frames per second.

C. Experimental Procedure

Several videos were taken of several volunteers so that the synthetic video could be compared to real video. About half of these volunteers were recruited from a class and given a small amount of course credit for their participation; the other half were research colleagues from the laboratory. The major research question was whether or not the synthetic frontal video would be of sufficient quality to support the applications intended for the HMD. The bench was set up for a general user and adjustments were made for individuals only when needed. Video and audio were recorded for each subject for offline processing. The experiment proceeded as follows. (1) The user was asked to sit in a chair and the chair was adjusted according to the convenience of the user and workspace of the cameras. (2) The projector was switched on and a grid projected onto the face in such a way that a vertical line passed through the center of the face, bisecting the face into two halves. (3) A *start command* was issued and the three cameras started recording the user's face (two side cameras and the front video recorder). (4) A microphone was switched on to record the voice. (5) After 1-2 seconds, the projector was switched off and the grid no longer projected onto the face. (6) The human subject repeated *The quick brown fox jumped over the lazy dog* continuously for 10 seconds.

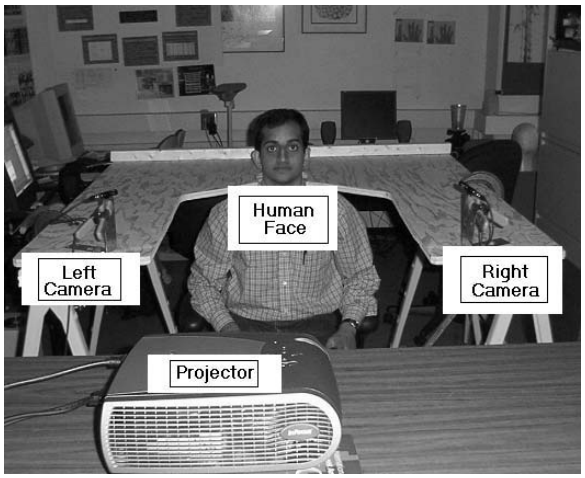

Fig. 2. Experimental prototype of the face capture system.

IV. Creating the virtual video

The problem is *to generate a virtual frontal view from two side views*. The projected light grid provides a basis for mapping pixels from the side images into a virtual image with the projector's viewpoint. The grid is projected onto the face for only a few frames so that mapping tables can be built, and then is switched off for regular operation.

There are three 2D coordinate systems involved in creation of the virtual video. For discussion only, we denote a global 3D coordinate system; however, it must be emphasized that 3D coordinates are not needed for the task of the current paper.

1) World Coordinate System (WCS): *for discussion only.*
2) Left Camera Coordinate System (LCS): $I_L[s,t]$ is the left image with s,t coordinates.
3) Right Camera Coordinate System (RCS): $I_R[u,v]$ is the right image with u,v coordinates.
4) Projector Coordinate System (PCS): $V[x,y]$ is the output virtual video image with coordinates defined by the projected grid.

A. Calibration for Virtual Video Synthesis

During the calibration phase, the transformation tables are generated using the grid pattern coordinates. A rectangular grid is projected onto the face and the two side views are captured

Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW'04)
1063-6919/04 $ 20.00 IEEE

as shown in Figures 3 and 4. The location of the grid regions in the side images define where real pixel data is to be accessed for placement in the virtual video. Coordinate transformation is done between PCS and LCS and between PCS and RCS. Using transformation tables that store the locations of grid points, an algorithm can map every pixel in the front view to the appropriate side view. By centering the grid on the face, the grid also supports the correspondence between LCS and RCS and the blending of their pixels.

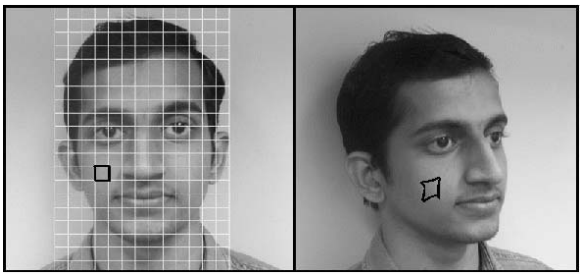

Fig. 3. Demonstration of the behaviour of the grid pattern

The behavior of a single gridded cell in the original side view and the virtual frontal view is demonstrated in Figure 3. A grid cell in the frontal image will map to a *quadrilateral* with curved edges in the side image. Bilinear interpolation is used to reconstruct the original frontal grid pattern by warping a quadrilateral into a square or a rectangle.

$$s = f_l(x, y) \text{ and } t = g_l(x, y) \tag{1}$$

$$u = f_r(x, y) \text{ and } v = g_r(x, y) \tag{2}$$

Equations 1 and 2 give the four functions determined during the calibration stage and implemented via the transformation tables. These transformation tables are then used in the operational stage immediately after the grid is switched off. During operation, it is known for each pixel $V[x, y]$ in which grid cell of LCS or RCS it lies. Bilinear interpolation is then used on the grid cell corners to access an actual pixel value to be output to the VV.

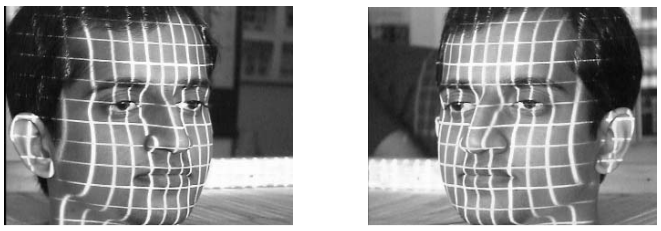

Fig. 4. Face images captured during the calibration stage.

Some implementation details are as follows. A rectangular grid of dimension 400 x 400 is projected onto the face. The grid is made by repeating three colored lines. We used white, green and cyan colors because of their bright appearance over the skin color. The first few frames have the grid projected onto the face before the grid is turned off. One of the frames with the grid is taken and the transformation tables are generated. The size of the grid pattern that is projected in the calibration stage plays a significant role in the quality of the video. This size was decided based on the trade-off between the quality of the video and execution time. An appropriate grid size was chosen based on trial and error. We started by projecting a sparse grid pattern onto the face and then increasing the density of the grid pattern. At one point, the increase in the density did not significantly improve the quality of the face image but consumed too much time. At that point, the grid was finalized with a grid cell size of row-width 24 pixels and column-width 18 pixels. Figure 4 shows the frames that are captured during the calibration stage of the experiment. (This calibration step is feasible for use in collaborative rooms; however, we are currently working on removing it to have one procedure applicable to mobile users as well.)

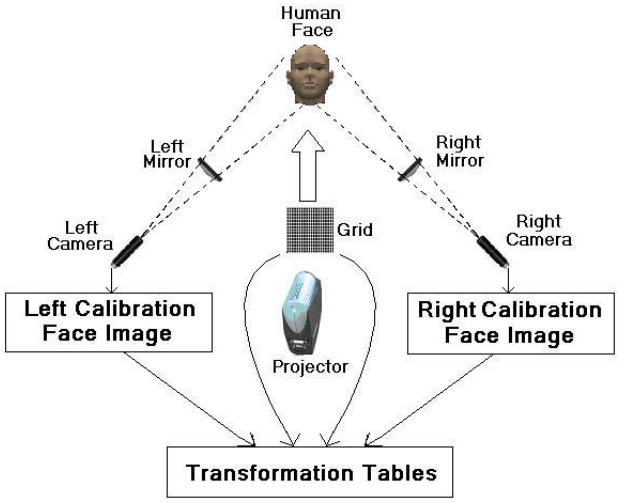

Fig. 5. The off-line calibration stage during the synthesis of the virtual frontal view.

B. Virtual Video Synthesis

Using the transformation tables generated in the calibration phase each virtual frontal frame is generated. The algorithm reconstructs each (x,y) coordinate in the virtual view by accessing the corresponding location in the transformation table and retrieving the pixel in I_L (or I_R) using interpolation. Then, a 1D linear smoothing filter is used to smooth the intensity across the vertical midline of the face. Without this, a human viewer usually perceives a slight intensity edge at the midline of the face.

Figure 6 shows the complete block diagram of the operational phase. Since the transformation is based on the bilinear interpolation technique, each pixel can be generated only when it is inside four grid coordinate points. Because the grid is not defined well at the periphery of the face, our algorithm is unable to generate the ears and hair portion of the face.

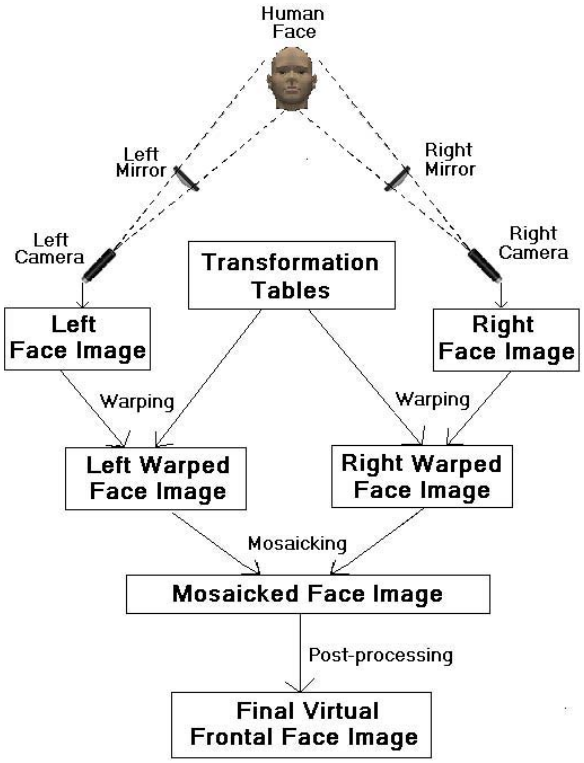

Fig. 6. Operational stage during the synthesis of the virtual frontal view

The results [1] of the warping during the calibration and the operation stage is shown in Figures 7 to 9.

(a) (b) (c)

Fig. 7. Frontal view generation during the calibration stage and reconstruction of the frontal image from the side view using the grid: (a) left image captured during the calibration stage. (b) operational left image warped into virtual image plus calibration stripes. (c) operational left image without stripes. The result of the reconstructed frontal view from the transformation tables and the right image is shown in Figure 8 below.

Some other post-processing is needed.The frames with the gridded pattern are deleted from the final output: these can be identified by a large shift in intensity when the projected grid is switched off. The microphone recording of the voice of the user, stored in a separate .wav file, is appended to the video file and the final output is obtained.

Finally, we need to mention necessary color balancing of the cameras. Even though software based approaches for color balancing can be taken, the color balancing in our work is done at the hardware level. Before the cameras are used for calibration, they are balanced using the white balancing technique. A single white paper is shown to both cameras and cameras are white balanced instantly. Ideally, the synthetic video will have minimal lighting variation due to the different cameras. However, it is important to note that there may be significant lighting variation in the user's environment that should be represented in the output as information to be communicated. A fire fighter, for example, may have more light on one side of her face than on the other – or a surgeon, or driver of a car. Software color balancing is likely to defeat such desirable color imbalance.

[1]Our videos are in color and are on our web site. Even if this paper appears without color, the authors feel that the concept is still properly shown.

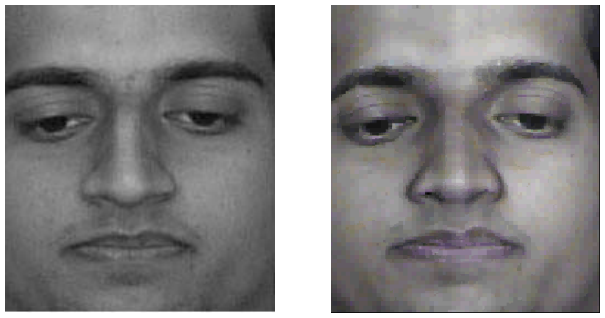

Fig. 8. (a) Frontal view obtained from the camcorder and (b) virtual frontal view generated by our algorithm.

V. Assessment of Results

The virtual video of the face must be adequate to support the communication of identity, mental state, gesture, and gaze direction. We report some objective comparisons between the synthesized and real videos and our own qualitative assessment.

A. Objective Evaluation

The real video frames from the camcorder and the virtual video frames were normalized to the same size of 200 x 200 and compared using cross correlation and interpoint distances between salient face features. Five images that were considered for evaluation are shown in Figure 9. Important items considered were the smoothness and accuracy of lips and eyes and their movements, the quality of the intensities, and the synchronization of the audio and video. In particular, we were looking for breaks at the centerline of the face due to blending and for other distortions that may have been caused by the sensing and warping process.

1) Normalized Cross-correlation: The cross correlation between regions of the virtual image and real image was computed for rectangular regions containing the eyes and mouth (Figure 10). As Table I shows, there was high correlation between the real and the virtual images taken at the same instant of time. Frames 2 and 3 shown in Figure 9 contain facial expressions (eye and lip movements) that were quite different from the expression used during the calibration stage and the generated view gave a slightly lower correlation value when compared with the other frames. Also, the facial

Fig. 9. Images considered for objective evaluation (a)Top row: real video frames (b) Bottom row: virtual video frames

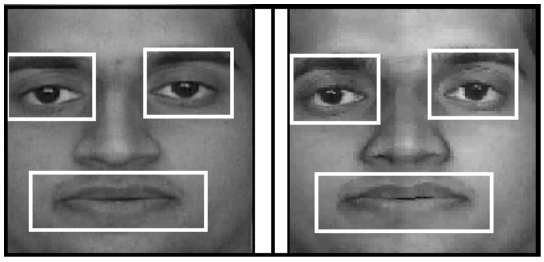

Fig. 10. (a) Facial regions compared using normalized cross-correlation (Left: real view and Right: virtual view.)

expressions in the frames 1 and 4 were similar to that of the expression in the calibration frame. Hence, these frames have a higher correlation value compared to the rest. The eye and lip regions were considered for evaluating the system because during any facial movement, these regions change significantly and are more important in communication.

TABLE I

RESULTS OF NORMALIZED CROSS-CORRELATION BETWEEN THE REAL AND THE VIRTUAL FRONTAL VIEWS APPLIED IN REGIONS AROUND THE EYES AND MOUTH.

video	left eye	right eye	mouth	eyes + mouth	complete
Frame1	0.988	0.987	0.993	0.989	0.989
Frame2	0.969	0.972	0.985	0.978	0.985
Frame3	0.969	0.967	0.992	0.978	0.986
Frame4	0.991	0.989	0.993	0.990	0.990
Frame5	0.985	0.986	0.992	0.988	0.989

2) Euclidean distance measure: We computed the difference in the normalized Euclidean distances between some of the most prominent feature points. The feature points are chosen in such a way that one of them is relatively static with respect to the other. For some prominent feature points, such as corners of the eyes, nose tip, corners of the mouth, the corners of the eyes are relatively static when compared with the corners of the mouth. Figure 11 shows the most prominent facial feature points and the distances between those points. Let R_{ij} represent the Euclidean distance between two feature points i and j in the real frontal image and V_{ij} represent the Euclidean distance between two feature points in the virtual frontal image. The difference in the Euclidean distance is $D_{ij} = |R_{ij} - V_{ij}|$. The average error ϵ for comparing the face images is defined by $\epsilon = \frac{1}{6}[D_{af} + D_{bf} + D_{cf} + D_{cg} + D_{dg} + D_{eg}]$.

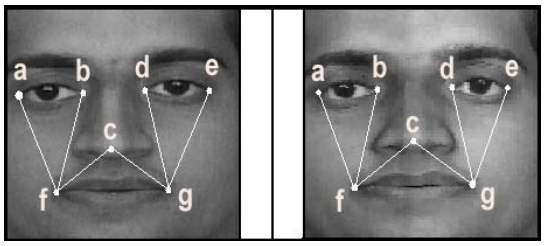

Fig. 11. Facial feature points and the distances that are considered for evaluation using Euclidean distance measure (Left: real view. Right: virtual view.)

TABLE II

EUCLIDEAN DISTANCE MEASUREMENT OF THE PROMINENT FACIAL DISTANCES IN THE REAL IMAGE AND VIRTUAL IMAGE AND THE DEFINED AVERAGE ERROR. ALL DIMENSIONS ARE IN PIXELS.

Frames	D_{af}	D_{bf}	D_{cf}	D_{cg}	D_{dg}	D_{eg}	Error(ϵ)
Frame1	2.00	0.80	4.15	3.49	2.95	3.46	2.80
Frame2	0.59	3.00	0.79	4.91	0.63	0.80	1.79
Frame3	1.88	3.84	4.29	4.34	2.68	1.83	3.14
Frame4	1.09	2.97	2.10	6.33	3.01	4.08	3.36
Frame5	1.62	2.21	5.57	4.99	1.24	1.90	2.92

The results in Table II indicate small errors in the Euclidean distance measurements of the order of 3 pixels in an image of size 200 X 200. The facial feature points in the five frames were selected manually and hence the errors might have also been caused by the instability of manual selection. One can note that the error values of D_{cf} and D_{cg} are larger than the others. This is probably because the nose tip is not as robustly located compared to eye corners.

B. Subjective Evaluation

A preliminary subjective study was done by the authors. In general, the quality of the videos was assessed as adequate to support the variety of intended applications. The two halves of all the videos are well synchronized and color balanced. The quality of the audio is good and it has been synchronized well with the lip movements. Some observed problems were distortion in the eyes and teeth and in some cases a cross-eyed appearance. The face appears slightly bulged compared with the real videos, which is probably due to the combined radial distortions of the camera and projector lenses.

Synchronization in the two videos is crucial in our application. Since, two views of a face with lip movements are merged together, any small changes in the synchronization will have high impact on the misalignment of the lips. This synchronization was evaluated based on sensitive movements such as eyeball movements and blinking eyelids. Similarly, mouth movements were examined in the virtual videos. Figures 12 to 13 show some of these effects.

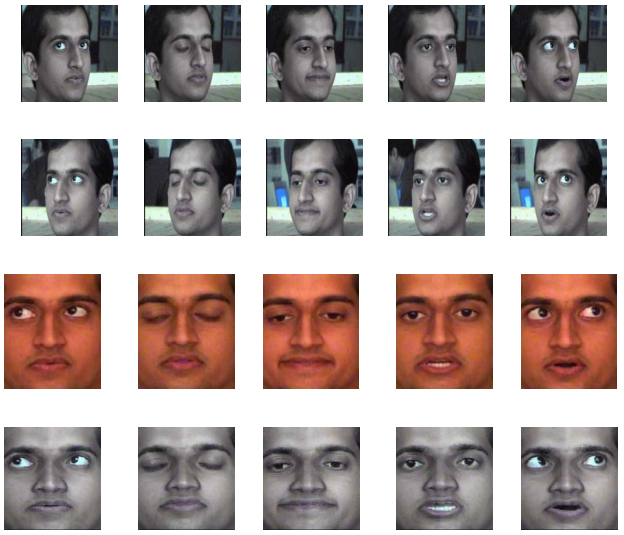

Fig. 12. (a) Top row: images captured from the left camera. (b) Second row: images captured using the right camera. (c) Third row: images captured using camcorder that is placed in front of the face. (d) Final row: virtual frontal views generated from the images in the first two rows

Fig. 13. Synchronization of the eyelids during blinking: real video is in the top row and the virtual video is in the bottom row.

C. Execution time

Our analysis indicates that a real-time mobile system can be built. The total computation time consists of (1) transfering the images into buffers, (2) warping by interpolating each of the grid blocks, and (3) linearly smoothing each output image. The average time is about 60 ms per frame using a 746 MHz computer. Less than 30 ms would be considered to be real-time: this can be achieved with a current computer with clock rate of 2.6 GHz. Future implementations will require more power to mosaick training data into the video to account for features occluded from the cameras.

D. Current Hardware

At the time of writing, virtual video synthesis is being developed on a tethered prototype. Figure 1 shows two frames acquired from this prototype HMD. The side videos are stable even when the wearer speaks and moves. The depth of field, which had been a concern in the optical design, is good as is the field of view of the face. Our future virtual frames should be better than those shown in this paper, which have been cropped due to constraints in working with the frame buffers. We need to show more hair and ears, in particular, something that is difficult to do in other systems that create a model from one frontal camera view. While we develop our algorithms on the prototype HMD, a new version on the HMD is also being developed that should improve the compactness and lightness of the HMD.

VI. Concluding Discussion

We have presented methods to synthesize a real time video of the frontal view of the face by merging two real side videos. The algorithm being used can be made to work in real-time. The working prototype has been tested on a diverse set of 7 individuals. From comparisons of the virtual videos with real videos, we expect that important facial expressions will be represented adequately not distorted by more than 2%. We conclude that our HMD can support the intended tele-communication applications. More detail of this work can be found in the thesis of Reddy [10]. Feasibility being shown, there are significant implementation details to work out that are in progress.

Calibration using a projected grid is critical to the current algorithms. Using it, we have also created 3D texture-mapped face models by calibrating the cameras and projector in the WCS, which we have not discussed in this paper. 3D models present the opportunity for greater compression of the signal and for arbitrary frontal viewpoints, which are needed for virtual face-to-face collaboration and especially when more than two persons are collaborating. Although technically feasible, structured light projection is an obtrusive step in the process and would be cumbersome in the field. We are working on eliminating the use of structured light by making use of a generic mesh model of the face. Other projects have shown some success in tracking faces in video using adaptive meshes: our problem should be easier, since the two cameras are fixed to the head.

There is a problem due to occlusion in the blending of the two side images. Some facial surface points that should be displayed in the frontal image are not visible in the side images. For example, the two cameras cannot see the back of the mouth. A solution that we are developing is to take training data from the user and to patch it into the synthetic video. The user will have to generate a basis for all possible future output material and the system will have to contain methods to index to the right material and blend it with the regular warped output. Ezzat and Poggio have shown this to be possible [4]. A related problem is that facial deformations that make significant alterations to the face surface will not be rendered well by the static warp. Examples are tongue thrusts and severe facial distortions. We will also address this problem in future work. The current system is good for moderate facial distortion: it will not crash when severe cases are encountered, but the virtual video will show a discontinuity in important facial features.

URL FOR ASSOCIATED MATERIALS

We have samples of data on the webpages at www.cse.msu.edu/∼stockman/FCHMD/. Included are four separate videos, one from the left camera, one from the right camera, the frontal virtual video synthesized from those, and the real frontal video for comparison. This is one of the better data sets, and was acquired from one of the authors. We do not have permission from other subjects to show their data sets. Other documents concerning this work are also included.

ACKNOWLEDGEMENTS

This research was supported in part by the following grants: NSF IIS-0222831, an MSU Foundation Strategic Partnership grant, NSF IIS 00-82016 ITR and EIA-99-86051

REFERENCES

[1] S. Basu, I.A. Essa, and A.P. Pentland. Motion regularization for model-based head tracking. In *Proceedings of 13th international conference on Pattern Recognitio n*, 1996.

[2] F. Biocca and J. Rolland. Teleportal face-to-face system: Tele-work augmented reality system. In *U.S. Patent Application 6550-00048; MSU 99-029, Michigan State University and Central Florida University, pending*, 2000.

[3] S. E. Chen and L. Williams. View interpolation for image synthesis. In *Proceedings of SIGGRPAH93*, pages 279–288, 1993.

[4] T. Ezzat and T. Poggio. Visual speech synthesis by morphing visemes. *Int. Journal of Computer Vision*, 38:45–57, 2000.

[5] 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,*, 1994.

[6] H. Hua, C. Gao, F. Biocca, and J. Rolland. An ultralight and compact design and implementation of head-mounted projective displays. In *Proceedings of IEEE Virtual Reality 2001*, pages 175–182, 2001.

[7] R.A. Manning and C.R. Dyer. Interpolating view and scene motion by dynamic view morphing. In *Proceedings of International Conference on Computer Vision and Pattern Recognition*, pages 388–394, 1999.

[8] Y. Onoe, K. Yamazawa, H. Takemura, and N. Yokoya. Telepresence by real-time view-dependent image generation from omnidirectional video streams. *Computer Vision and Image Understanding*, 71(2):154–165, 1998.

[9] 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*, pages 179–188, 1998.

[10] C. Reddy. A non-obtrusive head mounted face capture system. Master's thesis, Computer Science and Engineering Dept., Michigan State University, August 2003.

[11] S. M. Seitz and C. R. Dyer. View morphing. In *Proceedings of SIGGRAPH96*, pages 21–30, 1996.

[12] S.M. Seitz. The space of all stereo images. In *Proceedings of International Conference on Computer Vision*, pages 26–33, 2001.

[13] Jacob Strom, Tony Jebara, Sumit Basu, and Alex Pentland. Real time tracking and modeling of faces: An ekf-based analysis by synthesis approach. In *Proceedings of the IEEE Modeling People Workshop at ICCV'99*, september 1999.

[14] S. Vedula, S. Baker, and T. Kanade. Spatio-temporal view interpolation. In *Proceedings of the 13th ACM Eurographics Workshop on Rendering*, June 2002.

[15] L. Q. Xu, B. Lei, and E. Hendriks. Computer vision for a 3-d visualization and telepresence collaborative working environment. *BT Technology Journal*, 20(1):64–74, 2002.

[16] J.Y. Zheng and S. Tsuji. Panoramic representation for route recognition by a mobile robot. *International Jounal of Computer Vision,*, 9:55–76, 1992.