TELECONFERENCING EYE CONTACT USING A VIRTUAL CAMERA

Maximilian Ott [*] *John P. Lewis* [*] *Ingemar Cox* [+]

[*] C&C Research Laboratories, NEC USA [+] NEC Research Institute
4 Independence Way
Princeton, NJ 08540, USA
`max@ccrl.nj.nec.com`

ABSTRACT

To preserve eye contact in teleconferencing both the camera and the monitor need to be positioned on the same optical axis which, in practice, is usually not possible. We propose a method to construct the view from a virtual coaxial centered camera given two cameras mounted on either side of the monitor. Stereoscopic analysis of the two camera views provides a partial three-dimensional description of the scene. With this information it is possible to "rotate" one of the views to obtain a centered coaxial view that preserves eye contact.

KEYWORDS: Teleconferencing, eye contact, stereo matching, camera calibration.

INTRODUCTION

Dissatisfaction with teleconferencing facilities can often be traced to missing eye contact. This is an inherent problem of the physical setup. To preserve eye contact a user would need to look simultaneously at the monitor and into the camera. However, the camera and monitor cannot be physically in the same location. One solution is to place a half-mirror oriented 45 degree to the gaze direction. The monitor can be still observed while the reflection of the face is picked up by a camera. However, this arrangement is bulky and only half the monitor's intensity reaches the user's eyes.

Our proposed method is based on the fact that if we can obtain a three-dimensional description of the scene it is possible to construct an image seen from any desired view point. This 3D information is available by using two (or more) cameras and solving the stereo correspondence problem.

Our procedure consists of four steps, calibration, stereo matching, reconstruction, and interpolation (Figure 1). The calibration step converts the view from two tilted cameras into two parallel views. The second step matches pixels between the two views to obtain a displacement map. The last two steps construct the desired coaxial virtual view from between the two cameras, thereby recovering eye contact.

CALIBRATION

Stereo algorithms match features along epipolar lines [3]. Epipolar lines coincide with image scanlines if the camera axis are parallel. Under these circumstances finding the matching points becomes a traditional one-dimensional stereo correspondence problem.

In a realistic setting however, it is necessary to tilt the cameras inwards considerably to obtain a complete view of the face. The images obtained from the inward-oriented cameras are reprojected onto parallel aligned virtual imaging planes. In an initial step the projected coordinates of a known test grid are measured which allows us to calculate the position of both cameras in a common coordinate system. We then obtain a correction matrix which would turn the cameras into a parallel orientation. This correction matrix can be applied to all subsequent video images provided the camera positions are not changed.

STEREO MATCHING

The stereo algorithm determines the pixel correspondences in the two images and produces a disparity map which indicates the relative offset between a pixel in the left image and its corresponding pixel in the right image. The disparity map contains the critical information needed to generate the virtual view.

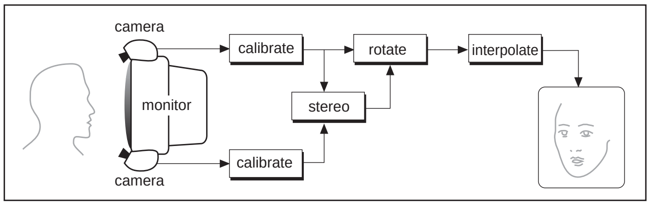
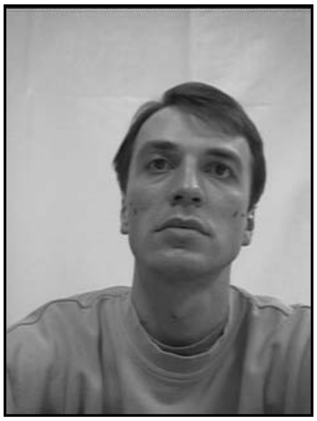
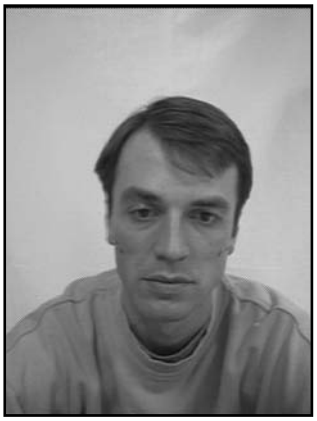


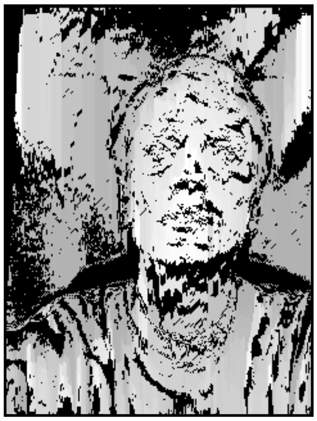
Figure 1: The *virtual* camera system.



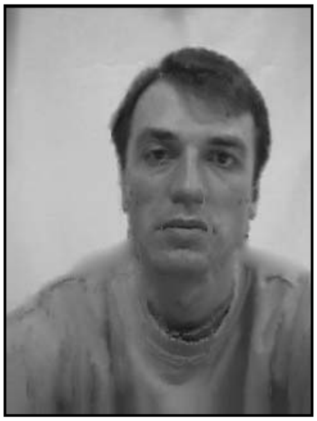
Bottom Camera



Top Camera



Disparity Map



Virtual View

We use a stereo algorithm described in [1] which matches individual pixel intensities using a maximum likelihood cost function and several cohesivity constraints. There are several advantages to using this algorithm. First, it provides a dense disparity map. Second, feature extraction and adaptive windowing techniques common to other stereo algorithms are avoided. Third, the stereo algorithm is fast and highly parallel, offering the potential for real-time implementation. For the purposes of this paper, the stereo algorithm uses dynamic programming to very efficiently find the solution – a 512x512 images takes approximately 50 seconds on a MIPS R3000 processor at 35MHz. A high percentage of correct matches and very little smearing of depth discontinuities are obtained.

CONSTRUCTING THE CENTERED VIEW

Because depth can be obtained from disparity, the disparity map effectively provides a partial three-dimensional scene description. A three-dimensional scene can be rotated, translated, and projected to an arbitrary viewpoint. Our scene information is limited to the form of a disparity field $z = f(x, y)$. For a virtual view exactly midway between parallel cameras the pixels from one view are simply *shifted* by half of the disparity at each pixel.

INTERPOLATION

Several practical problems are encountered in the construction of the virtual view. While the human head is fairly convex, so most points are visible to both cameras, some occlusions will exist. Points present in the left image but occluded in the right image (and vice versa) have no associated disparity values.

Although holes in the disparity map could be filled in by interpolating neighboring areas, we found that a simpler implementation and superior visual effects resulted from ignoring image areas whose disparity is unknown, thereby causing the construction problem to be manifested as holes in the final image. These holes are then handled by a suitable scattered data interpolation method. Thin-plate spline interpolation[2] was found to provide high quality interpolation in our implementation, but a simpler method would probably be adopted in a real-time implementation.

RESULTS

At the top of the page we show the results of the proposed method. While no eye contact is present in either the bottom camera or top camera images, the virtual image compensates for this, providing good eye contact with the observer.

The black areas in the disparity map denote regions of the left image where no match was obtained. Areas of large occlusion, such as around the neck area demonstrate the limitations of the smoothing step. Errors in the disparity map caused by incorrect pixel correspondences can cause artifacts in the constructed image, as for example, in the right corner of the mouth. In a motion sequence, these artifacts are expected to appear as noise and some form of temporal averaging should reduce their visibility.

We plan to apply this method to a motion sequence to test its robustness as well as to evaluate the observed distortions.

REFERENCES

[1] Cox, I. J., Hingorani, S., Maggs, B. M., and Rao, S. B. Stereo without disparity gradient smoothing: a bayesian sensor fusion solution. In *British Machine Vision Conference* (1992).

[2] Grimson, W. E. L. *From Images to Surfaces*. MIT Press, Cambridge, USA, 1981.

[3] Zisserman, A. *Notes on Geometric Invariance in Vision*. Tutorial, British Machine Vision Conference (1992).