SYNTHESIS OF MULTI VIEWPOINT IMAGES AT NON-INTERMEDIATE POSITIONS

André Redert, Emile Hendriks, Jan Biemond

Information Theory Group, Department of Electrical Engineering
Delft University of Technology
Mekelweg 4, 2628 CD Delft, The Netherlands
phone +31 15 278 6269, fax +31 15 278 1843
email {andre, emile, biemond}@it.et.tudelft.nl

ABSTRACT

In this paper we present an algorithm for the synthesis of multi viewpoint images at non-intermediate positions, based on stereoscopic images. We consider the synthesis of images from virtual camera positions and the synthesis of images for scene reconstruction using stereo displays. The algorithm provides scene reconstruction without geometric distortion and without any restriction to the position of the viewer.

All synthesized images are based on extrapolation of a single source image and a single disparity field. This provides low use of bandwidth and compatibility with mono video systems.

With teleconferencing images, the generated views were subjectively evaluated as good for viewing positions not more than one half camera baseline from the centre position. Objectively, reconstructed left and right images have PSNR values of 41 dB.

1. INTRODUCTION

In video communications, 3D imaging can greatly enhance the feeling of telepresence. This requires the acquisition, transmission and presentation of 3D scene data. Existing stereo video systems use two cameras for recording the scene and a stereo display to present it to the viewer. These systems give the viewer a sensation of depth, but a major drawback is the restriction to a very specific viewing position [4]. Any movement by the viewer will not result in the expected motion parallax, but in geometric distortion of the reconstructed scene.

A solution to this is provided by multi viewpoint systems [1,8,9], shown in Figure 1. These provide motion parallax, based on the position and motion of the viewer. Current systems use stereo cameras with large baselines at the acquisition side. Their images are analysed by a disparity estimator resulting in 3D scene data. At the presentation side new intermediate images are synthesized, based on the actual position of the viewer. The viewer position is determined by a headtracker.

With current multi viewpoint systems, the viewer has a restricted freedom of movement, corresponding to viewing positions in between the cameras. Any deviation from those positions will result in geometric distortion of the scene.

In this paper we will investigate image synthesis algorithms that provide also non-intermediate viewpoints. The aim is the geometrically correct reconstruction of that part of the original scene that is present in the 3D scene data.

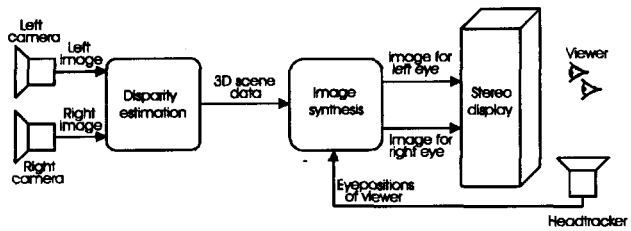


Figure 1: The multi viewpoint 3D system

With video communications as major application, the feasibility of real time hardware implementation is important[3]. Therefore we will investigate special cases of the synthesis algorithm having lower complexity.

In section 2 we will discuss image acquisition, disparity estimation and 3D scene format in detail because of its large influence on the complexity of the image synthesis algorithms. In section 3 we examine the synthesis of images from virtual cameras at any position and zoom factor, and the same orientation as the original cameras. Section 4 gives a general scene reconstruction algorithm for multi viewpoint systems and two special cases with lower complexity. Section 5 gives results with real teleconferencing images and synthetic data. Conclusions are given in section 6.

2. ACQUISITION OF 3D SCENE DATA

In this section we describe the acquisition of 3D scene data in detail, because of its large influence on the image synthesis algorithms. The camera setup, disparity estimator and 3D scene data will be explained.

2.1 Camera setup

We assume the camera setup as depicted in Figure 2. In the acquisition reference frame O_{acq}, the cameras optical centres lie on the x-axis at position $-B$ (left) and $+B$ (right). The optical axes are parallel and point in the +z direction. The focal distance of each camera is f. The left and right image planes lie at $z=f$ and are centred at the corresponding optical axes. The image planes are spanned by two perpendicular vectors, one in the +x direction with size H_c and one in the +y direction with size V_c . These are the horizontal and vertical size of the pixels on the CCD chip in the cameras.

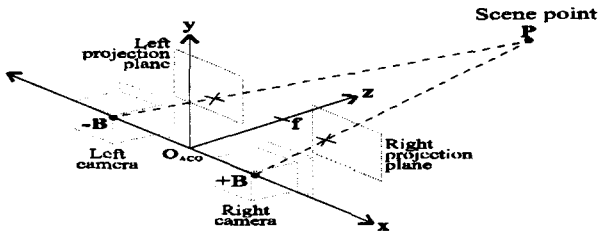


Figure 2: The camera setup

Every scene point P with coordinates P_x, P_y, P_z in the acquisition reference frame is projected onto the left and right image planes. The coordinates of the projections (in the projection reference frame or in image pixels) are:

$$\begin{bmatrix} X_L \\ X_R \\ Y \end{bmatrix} = \frac{f}{P_z} \begin{bmatrix} (P_x + B)/H_c \\ (P_x - B)/H_c \\ P_y / V_c \end{bmatrix} \quad (1)$$

For this camera setup we always have $P_z > 0$, $Y=Y_L=Y_R$ and $X_L \geq X_R$.

2.2 Disparity estimation

The task of disparity estimation is to find out which left and right pixels correspond to the same scene point, based on the luminance information in the left and right images. The P coordinates of the scene point are then given by (1). The luminance of the scene point can be estimated based on the left and right pixel luminances.

Objects that are half-occluded (visible only in one of the images) give rise to pixels that can not be paired to a pixel in the other image. In this case we can not use (1), so the P coordinates of the half-occluded objects can be found only in some other way, e.g. interpolation of coordinates of neighbouring objects. In the next section we will handle this.

For disparity estimation we use a dynamic programming algorithm similar to that of Cox et al. [2, p. 547], working on blocks of 4*4 pixels rather than single pixels.

2.3 Format of 3D scene data

Usually the 3D scene data consists of both left and right images, accompanied by a left-to-right disparity field $D_{LR}(X_L) = X_R - X_L$ and/or a right-to-left field $D_{RL}(X_R) = X_L - X_R$.

As 3D scene data we use a single interpolated centre image with luminance $I(X,Y)$ and the right-to-centre or centre-to-left disparity $D(X,Y)$. This has the following consequences. First, the reconstructed scene will have diffuse reflection properties, since to each scene point a single luminance value is assigned (luminance is not a function of viewing position). Next, occluded areas in left and right image will both be present in the centre image, horizontally compressed by a factor two, resulting in half resolution. Finally the disparity D is single valued in the centre view position, thereby excluding scene objects that are exactly in front of each other.

We chose this format because it has several advantages. First, in general the centre view is the most interesting view so it is best to generate it at the transmitter using original camera images, undistorted by a coding system. Secondly, the complexity of the synthesis algorithms is lowered as all new viewpoints will be generated in the same way using extrapolation (even for viewpoints in between the original cameras) and there is no need for weighting luminances of left and right images (only one image I). Thirdly, this format is compatible with mono video systems and provides eye contact in teleconferencing applications [5]. Finally the use of a single image saves transmission bandwidth.

To obtain the centre image and the disparity field, we first make a list of pixel pair coordinates X_L, X_R, Y and generate the corresponding centre X coordinates and the disparities D:

$$\begin{bmatrix} X \\ D \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} X_L \\ X_R \end{bmatrix} \Rightarrow \begin{bmatrix} X \\ Y \\ D \end{bmatrix} = \frac{f}{P_z} \begin{bmatrix} P_x / H_c \\ P_y / V_c \\ B / H_c \end{bmatrix} \quad (2)$$

The disparity field is the dense array $D(X,Y)$. Holes in the field due to occlusions are filled by linear interpolation of defined D in a horizontal neighbourhood.

Now we generate the centre image luminance:

$$I(X,Y) = (\tfrac{1}{2} + \tfrac{1}{2}\Delta) \cdot I_L(X_L,Y) + (\tfrac{1}{2} - \tfrac{1}{2}\Delta) \cdot I_R(X_R,Y) \quad (3)$$

With Δ equal to the derivative of the disparity field, smoothed by a uniform filter of length L [7]:

$$\Delta = \frac{D(X + \tfrac{1}{2}L) - D(X - \tfrac{1}{2}L)}{L} \quad (4)$$

If an object is visible in only one of the original images, corresponding to Δ=1 (left) or Δ=-1 (right), data is taken only from that image for the centre image. If an object is visible in both original images, corresponding to Δ=0, the original left/right image data is averaged.

3. SYNTHESIS OF IMAGES FROM VIRTUAL CAMERAS

As an extension to intermediate view generation we consider the synthesis of images from virtual cameras at any position and any zoom factor, with the same orientation. First we will give the geometric relation between the centre image and the image from the virtual camera. Next we will discuss our rendering method.

3.1 Geometric relation virtual and centre camera

The virtual camera is positioned at $(S_x, S_y, S_z)B$, and has a focal length $S_{zoom} f$. Every scene point P is projected onto the virtual camera image according to:

$$\begin{bmatrix} X_V \\ Y_V \end{bmatrix} = \frac{S_{zoom} \cdot f}{P_z - S_z \cdot B} \begin{bmatrix} (P_x - S_x \cdot B)/H_c \\ (P_y - S_y \cdot B)/V_c \end{bmatrix} \quad (5)$$

Using (2) and (5), the virtual camera image can be acquired by translating all points of the centre image according to:

$$\begin{bmatrix} X_V \\ Y_V \end{bmatrix} = \frac{S_{zoom}}{1 - S_z \frac{H_c}{f} \cdot D} \begin{bmatrix} X - S_x \cdot D \\ Y - S_y \cdot D \cdot R \end{bmatrix} \quad (6)$$

With $R = H_c / V_c$, the pixel ratio of the camera.

Although the disparity D was calculated using a horizontal displacement between the original cameras, its function is the same for both the horizontal and vertical pixel displacement (apart from the pixel ratio).

For $-1 \le S_x \le 1$, $S_y = 0$, $S_z = 0$ and $S_{zoom} = 1$, (6) is equal to normal image interpolation [1,8,9].

3.2 Image rendering

Two problems can be encountered when rendering images using (6). Pixels in the virtual image may become overdefined (two or more assignments) or remain undefined (no assignments).

In the overdefined case multiple pixels of the centre image are translated to the same pixel in the virtual camera image. Then we select the one that is closest to the virtual camera, corresponding to maximal D.

Pixels with undefined luminance in the virtual image are caused by insufficient information in the 3D scene data. For the generation of non-intermediate views, this is very likely to happen. Correct reconstruction of the scene is therefore restricted to the part present in the 3D scene data. To avoid clear visibility of the 'holes' in the reconstructed scene, we use linear interpolation of the luminance based on neighbouring pixels.

4. SYNTHESIS OF IMAGES FOR SCENE RECONSTRUCTION

In this section we will present an algorithm for the synthesis of images for a multi viewpoint 3D video system. The aim is the geometrically correct reconstruction of the recorded scene on a stereo display.

Figure 3 shows the presentation side of the system. The stereo display is centered in the presentation reference frame O_{pres}. The display image plane is spanned by two vectors with sizes H_d and V_d, the horizontal and vertical pixel size. We assume that the display pixel ratio is R, the same as the ratio of the camera.

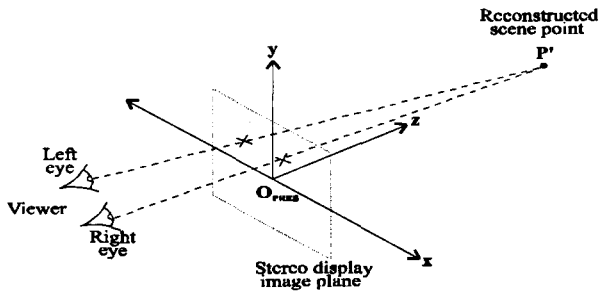


Figure 3: The presentation side

We select the reconstructed point $P' = P+Q$ with Q_z a constant and $Q_x = Q_y = 0$. The depth shift Q_z is introduced here for three reasons. First, in teleconferencing it is desired to have a small person to person distance. Secondly we would like to shift the reconstructed scene to be centered in the display. This minimizes visual strain due to conflicts between convergence and accommodation of the eye [6]. Finally, the complexity of the synthesis algorithm drops for two specific choices of Q_z.

From the reconstructed scene point P', two light rays enter the viewer eyes via the centres of the irises. The left and right light ray intersect the display in T_L and T_R, measured in pixels. The determination of the left and right intersection points is independent, so we will restrict ourselves to the synthesis of multi viewpoint images for a single eye at position E_x, E_y, E_z (with $E_z < 0$). Rendering of images is done as described in section 3.

Based on a viewers eye at any position E and a reconstructed scene point P' given by the shift Q_z and (2), the intersection X_T, Y_T of the light ray EP' and the display is given by:

$$\begin{bmatrix} X_T \\ Y_T \end{bmatrix} = \frac{-\frac{B}{G}\cdot E_z \cdot \begin{bmatrix} X \\ Y \end{bmatrix} + \left(Bf + Q_z H_c \cdot D\right)\begin{bmatrix} E_x/H_d \\ E_y/V_d \end{bmatrix}}{Bf + (Q_z - E_z)\cdot H_c \cdot D} \tag{7}$$

with $G = H_d / H_c = V_d / V_c$.

For two specific Q_z, the general reconstruction formula (7) reduces to a lower complexity version. First, for $Q_z = 0$ the disparity term in the enumerator disappears. This results in:

$$\begin{bmatrix} X_T \\ Y_T \end{bmatrix} = \frac{\frac{-E_z}{Gf}\begin{bmatrix} X \\ Y \end{bmatrix} + \begin{bmatrix} E_x/H_d \\ E_y/V_d \end{bmatrix}}{1 - \frac{E_z}{B}\cdot\frac{H_c}{f}\cdot D} \tag{8}$$

Secondly, for $Q_z = E_z$, the denominator in (7) becomes constant. The reconstructed scene becomes 'attached' to the viewer, which is very unnatural. Now we introduce $Z_{sys} = Gf$. If $Q_z = E_z = -Z_{sys}$, (7) reduces to:

$$\begin{bmatrix} X_T \\ Y_T \end{bmatrix} = \begin{bmatrix} X \\ Y \end{bmatrix} - \frac{(D - D_{offset})}{B}\begin{bmatrix} E_x \\ E_y \cdot R \end{bmatrix} \tag{9}$$

With $D_{offset} = B / H_d$. This is equal to (6) with $S_{zoom} = 1$, $S_z = 0$, $S_x = E_x/B$ and $S_y = E_y/B$ plus a shift. This shift is equal to the shift encountered in stereo video systems [4].

Using (9) in stead of (7), the viewer has freedom of movement only in the horizontal and vertical direction, but the system complexity is substantially lower.

For teleconferencing applications, the camera baseline will be a little larger than the physical size of the display. In that case the shift will be so large that the original camera images do not overlap at all, making disparity estimation impossible. Solutions for this are the use of shifted lenses [4] or a slightly converging camera setup. If this is not possible, options are to use (7) or to accept reconstruction errors.

5. EXPERIMENTAL RESULTS

Figure 4 shows the left and right image from a typical teleconferencing application. Figure 5 shows the disparity field and interpolated centre image, according to the algorithm in section 2 with $L = 8$.

Figure 6 shows synthesized virtual camera images generated by (6) and by (9) with zero shift. Camera positions are $S_x \in \{-1,0,1\}$, $S_y \in \{-½,0,½\}$, $S_z = 0$ and $S_{zoom} = 1$. Objectively, the PSNR values of the reconstructed left and right images are 41 dB. We evaluated these images subjectively on a mono

display. The images look very good for $|S_x| < 1.5$ and $|S_y| < 0.5$. With larger S values, the image quality degrades smoothly.

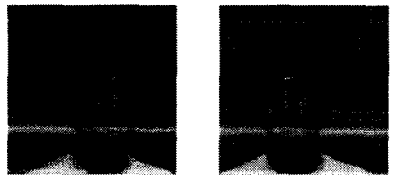


Figure 4: Original left and right camera images

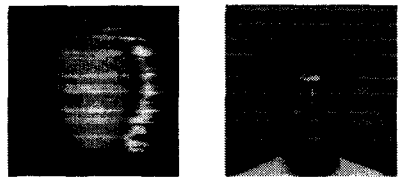


Figure 5: The disparity field and the interpolated centre image

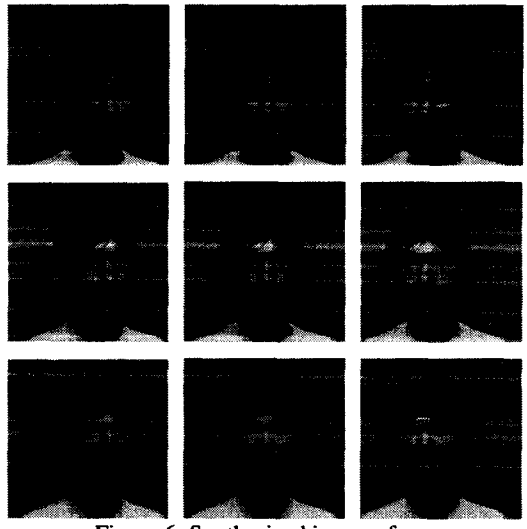
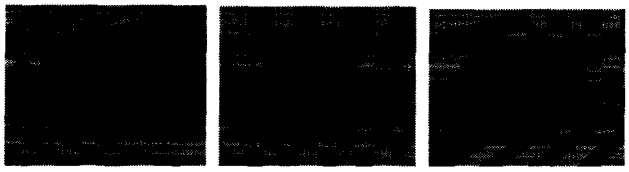


Figure 6: Synthesized images from virtual cameras.

We evaluated the reconstruction algorithm given by (8) using synthetic scene data, avoiding the incompleteness of the 3D scene data described in section 2. The synthetic scene is a box of 64*48*40 cm, equal to the physical size of our display device, with a cube of 24*24*24 cm centered in the display image plane.



a) b) c)

Figure 6: Synthesized images for geometrically correct reconstruction at viewer position E_x, E_y, E_z in meters =
a) -0.5 , 0, -0.5 b) 0, -0.5, -1.5 c) 1, 0.5, -1.5

Currently we do not have a headtracker available. We generated stereo pairs according to several viewing positions. As expected, the scene looks very good when viewed from the appropriate position and becomes distorted quite heavily when seen from other viewpoints.

6. CONCLUSIONS

We presented a new image synthesis algorithm for multi viewpoint 3D video systems. The algorithm provides scene reconstruction without geometric distortion and without any restriction to the position of the viewer.

All synthesized images are based on extrapolation of a single source image and a single disparity field. This provides low use of bandwidth and compatibility with mono video systems.

With teleconferencing images, the generated views were subjectively evaluated as good, for viewing positions not more than one half camera baseline from the centre position. Objectively, reconstructed left and right images have PSNR values of 41 dB.

7. ACKNOWLEDGEMENT

This work was done in the framework of the European ACTS project PANORAMA. One of the major goals is the realization of a real time multi viewpoint system in hardware.

REFERENCES

[1] B. Chupeau and P. Salmon, "Synthesis of intermediate pictures for autostereoscopic multiview displays", in Proceedings Workshop on HDTV '94, Turin, Italy

[2] I.J. Cox, S.L. Hingorani and S.B. Rao, "A maximum likelihood stereo algorithm", *Computer vision and image understanding*, Vol. 63, No. 3, 1996, pp. 542-567

[3] E.A. Hendriks and Gy. Marosi, "Recursive disparity estimation algorithm for real time stereoscopic video applications", Proceedings of the International Conference on Image Processing (ICIP) '96, pp. 891-894

[4] R. Kutka, "Reconstruction of correct 3-D perception on screens viewed at different distances", *IEEE Transactions on communications*, Vol. 42, No. 1, 1994, pp. 29-33

[5] J. Liu, I.P. Beldie and M. Wöpking, "A computational approach to establish eye-contact in videocommunication", in Proceedings of the International Workshop on Three Dimensional Imaging (IWS3DI), Santorini, Greece, 1995, pp. 229-234

[6] S. Pastoor, "3D-television: A survey of recent research results on subjective requirements", *Signal processing: Image Communication* 4, 1991, pp. 21-32

[7] P.A. Redert and E.A. Hendriks, "Disparity map coding for 3D teleconferencing applications", to appear in proceedings of IS&T/SPIE VCIP, San Jose, USA, 1997

[8] B.L. Tseng and D. Anastassiou, "A theoretical study on an accurate reconstruction of multiview images based on the Viterbi algorithm", Proceedings of the International Conference on Image Processing (ICIP) '95, pp. 378-381

[9] J.S. McVeigh, M.W. Siegel and A.G. Jordan, "Intermediate view synthesis considering occluded and ambiguously referenced image regions", *Signal Processing: Image Communication* 9, 1996, pp. 21-28