MODELLING OF 3D SCENES FROM THE ANALYSIS OF STEREOSCOPIC IMAGE SEQUENCES

C.–E. Liedtke, R. Koch
Institut für Theoretische Nachrichtentechnik und Informationsverarbeitung, Universität Hannover
Appelstraße 9A, D–3000 Hannover 1, F. R. Germany

ABSTRACT

An approach for the modelling of complex 3D scenes like outdoor street views from a sequence of stereoscopic image pairs is presented. Starting with conventional stereoscopic correspondence analysis a 3D model scene with true 3D geometry is generated. Not only the scene geometry but also surface texture is stored within the model. 3D camera motion can be estimated directly from the image sequence to track camera motion and to fuse measurements throughout the sequence into a common 3D model scene.

Keywords: 3D Scene Analysis, Stereoscopic Sequence Analysis, Scene Reconstruction, Close Range Photogrammetry.

INTRODUCTION

There is an increasing request for obtaining computer graphics models of natural scenes in order to create realistic visual sensations by producing computer animations. Applications are

– in the movie and TV industry, for film production, synthetic studio environments,
– in future communication systems and virtual reality applications [1],
– in advertising, instruction and education,
– for the visualization part of various training simulators, like flight– and driving simulators,
– for planning purposes in landscape and city planning, architecture, etc. .

Though considerable progress has been made in the past years in computer graphics and 3D visualization techniques it is still not possible to produce with reasonable effort natural looking landscapes, persons, animals, etc.. A solution to the problem is expected from the integration of computer graphics with the analysis of natural scenes by machine vision techniques.

In the following an approach based on the analysis of sequences of stereoscopic image pairs is described, which has been applied to the modelling of complex outdoor street scenes as they are needed for driving simulators, city planning purposes etc..

CONCEPT

Conventional stereoscopic image analysis tries to reconstruct the 3D scene from pairs of camera images through triangulation of corresponding 2D image points while the relative orientation of the cameras to each other is known. Through triangulation a depth map can be constructed where the distance of each corresponding point to the camera focus point is measured. This approach is sufficient for simple scene geometry without occluded areas but will fail when analyzing complex scenes like street views.

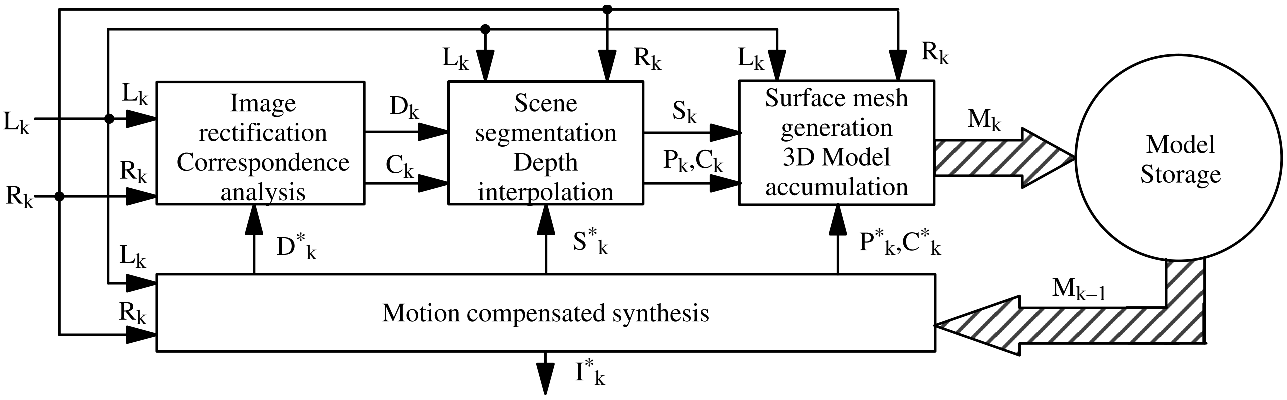
To reconstruct a complex 3D scene a series of problems need to be solved, especially the problems of 2D correspondence and 3D registration. To be able to triangulate the scene from a particular pair of images a calibration of the projection geometry and the relative orientation of the cameras to each other is needed. When using a binocular camera setup, the calibration can be performed once before the measurement and will remain constant throughout the measurement phase. Measurement of scene geometry relative to such a binocular camera system can then be obtained through triangulation of corresponding image points.

During correspondence analysis one tries to uniquely identify the projections of a scene element onto the camera targets. In a complex natural scene parts of the scene may be occluded to the camera system so that the camera has to be moved throughout the scene. The scene is then analyzed from a sequence of image pairs. To register all measurements into a common scene coordinate system the 3D motion of the camera system is tracked and all measurements are included to build a true 3D model of the scene.

The presented approach addresses the problems stated above for building a true 3D model of a complex scene from a sequence of stereoscopic image pairs. Not only the scene geometry but also the surface texture is stored within the model. It is therefore possible to synthesize realistic looking image sequences from the model using 3D computer graphics methods.

The concept is shown in Fig.1. The input at time instant k is a stereo pair L_k, R_k. In a preprocessing step the image pair is rectified to obtain an image pair where the camera axes are parallel and both cameras are displaced along horizontal image plane coordinates only. This image rectification greatly simplifies correspondence analysis and the search space is reduced to parallel horizontal epipolar lines **E**. The camera model as introduced by Yakimovsky and Cunningham is used [2].

From the rectified images a disparity map D_k is obtained by correlation matching techniques. The quality of the



L_k: Left image | D_k: Disparity map | S_k: Segmentation map | M_k: 3D Scene model
R_k: Right image | C_k: Confidence map | P_k: Depth map
k: Index indicating the present time instant
I^*_k: Virtual camera image from model
$^*_k$: = data predicted from model M_{k-1} using motion compensation

Fig.1: Concept of 3D Scene Analysis from Sequences of Stereo Images

match and therefore the quality of each displacement value is recorded in a confidence map C_k. The scene is then segmented into meaningful objects based on depth information from the disparity map D_k and texture and edge information from the input images L_k and R_k. The label of each region belonging to an individual object is stored in the segmentation map S_k. The displacement map D_k is converted into a depth map P_k. Due to noise in the input images and regions of low confidence in the confidence map C_k, the segmentaton map S_k, the displacement map D_k, and the corresponding depth map P_k are neither dense nor smooth. Interpolation is used to smooth the object surfaces and to fill the gaps. Prior knowledge about the content of the scene and its properties can be used to improve the segmentation and interpolation tasks.

In order to obtain a more efficient description of the 3D scene surface the depth map is converted into a triangular surface mesh. In the 3D scene model M_k the surface mesh, which represents the geometric properties of the scene and the surface texture, which represents the photometric information, are stored. From the geometric and photometric information realistic looking image sequences I^*_k can be synthesized.

The relative motion between camera and the scene from time instant k–1 to time instant k is estimated from the model M_{k-1} and the input data L_k and R_k. By motion compensated synthesis a displacement map D^*_k, a segmentation and confidence map S^*_k and C^*_k, and a depth map P^*_k can be predicted from the model and can be used in a feedback loop to improve the accuracy and the efficiency of the analysis process. The motion compensation of the depth map enables 3D registration and permits to accumulate the modelling results of different views of the scene.

The analysis of a stereoscopic image pair is split into an image analysis in the 2D image domain and in an incremental model building in 3D space. In the following, the individual processing steps are explained in more detail.

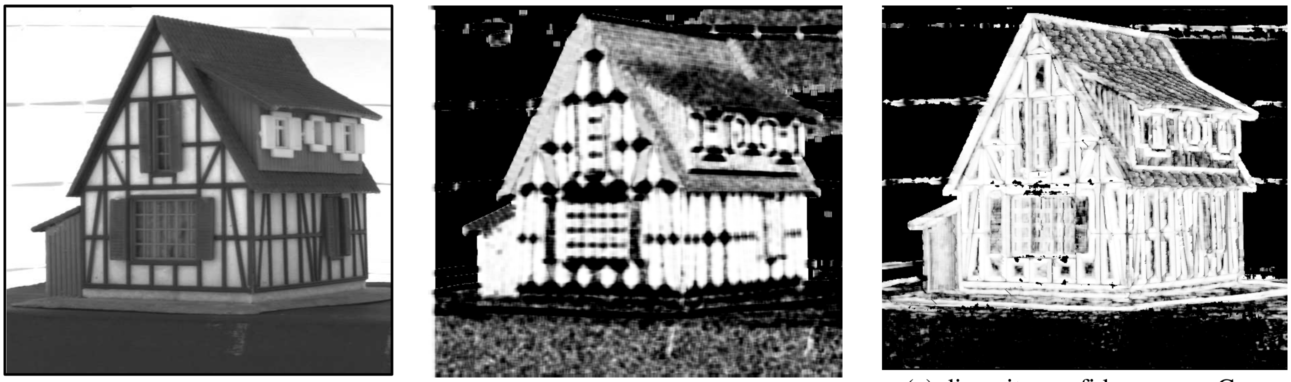
ANALYSIS IN THE IMAGE DOMAIN

The 2D analysis of a stereoscopic image pair consists of correspondence analysis and object generation. Correspondence analysis tries to locally estimate image plane correspondences while during object generation areas in the image that belong to physically connected regions are merged by similarity measures through segmentation. Each region is interpolated to yield a dense depth map.

Correspondence Analysis

The correspondence problem is solved using an area based matching process. Candidates for corresponding points are selected from the evaluation of the gradient **G**. Only areas exceeding a minimum image gradient value $|\mathbf{G}| > G_{min}$ can be candidates for correspondences. The quality of a candidate can further be judged by the direction of the gradient compared with the direction of the epipolar line **E**. The quality is maximum, when **G** and **E** are parallel and is minimum when they are both perpendicular. The quality measure C_1 in Eq.(1) is calculated for each image pixel and pixels with $C_1 > 0$ are selected as candidates for correspondence analysis.

The search space in the other image is reduced to a one–dimensional search along the epipolar line between minimum and maximum disparity values derived from the known minimum and maximum scene distance.The most probable candidate pair is the pair with maximum normalized cross correlation value (NCC–value). In complex scenes there may be multiple maxima or false maxima in the search space due to occlusions, repeated structures or image noise. This ambiguity can be reduced when uniqueness and ordering constraints are exploited. These constraints are employed in an optimum search procedure using dynamic programming that matches all correspondences between left and right image on the same epipolar line. The dynamic programming algorithm was adapted from the work of Cox [3]. The disparity value obtained for each candidate is re-



(a) left original image

(b) epipolar gradient confidence map C_1

(c) disparity confidence map C_2 (light = high confidence, dark = low confidence)

Fig. 2: Confidence Measures for Correspondence Analysis

corded in a disparity map. The quality C_2 of the disparity measurement can be judged on the basis of the NCC–value according to Eq.(1). Since low NCC–values (below 0.5) are often related to false matches C_2 is therefore defined to be zero below a value NCC_{min}.

$$C_1 = \left\{ \begin{array}{ll} 0 & \text{for } |\mathbf{G}| < G_{min} \\ \dfrac{\mathbf{G} \cdot \mathbf{E}}{|\mathbf{G}|} & \text{else} \end{array} \right\} \qquad (1)$$

$$C_2 = \left\{ \begin{array}{ll} 0 & \text{for } NCC < NCC_{min} \\ NCC & \text{else} \end{array} \right\}$$

Both quality measures can be merged to one disparity confidence measure $C_c = C_1 \cdot C_2$ with $\{0 \leq C_c < 1\}$ that contains the combined quality measure for each candidate. C_c is recorded for each candidate pixel in a confidence map C_k. Fig. 2 gives an example for the gradient confidence map C_1 and the disparity confidence map C_2. The resulting disparity map is displayed in Fig.3a, where dark gray values indicate small disparities (background) and light gray values large disparities (foreground). Black regions are regions where the confidence measure is zero.

Object Generation

The correspondence analysis yields a disparity map based on local depth measurement only. As the disparity map shows, the measurements are corrupted by noise. Regions that describe physical object surfaces must be found and gaps must be closed. Based on similarity measures the segmentation divides the viewed scene into contiguous surfaces. As similarity measure the estimated disparities as well as grey level statistics are used to group pixels into object regions. The region boundaries are then corrected from the grey level image with a contour approximation by assuming that physical object boundaries most often create grey level edges in the image. The result of image segmentation is shown in Fig. 3b. The segmentation areas corresponding to the background and the foreground areas have been merged and are displayed in white, whereas the single object area is labeled black. During image segmentation knowledge about the objects can be included interactively by marking edges in the image.

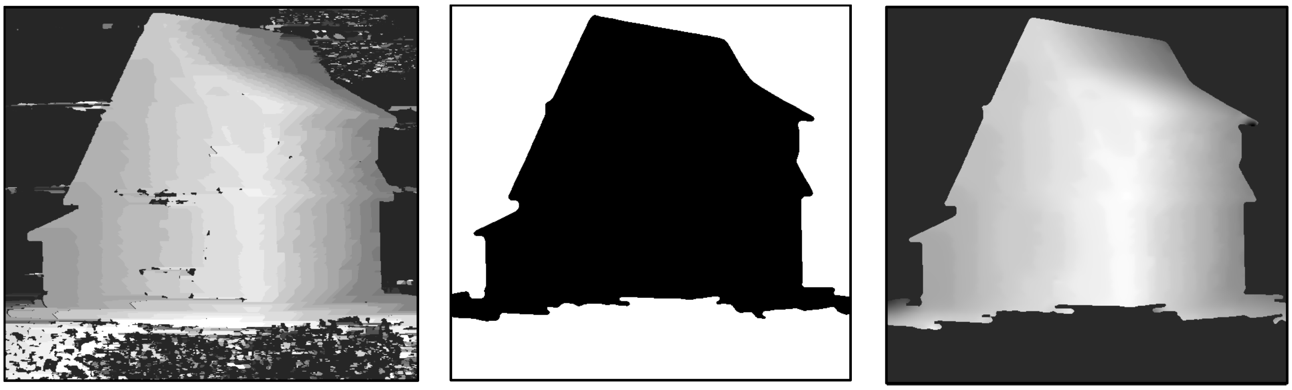
Once the disparity map is segmented into object regions all measurements of one region are interpolated by a thin plate surface model that calculates the best quadratic surface approximation of the disparity map based on the uncertain depth measures. A multi grid surface reconstruction algorithm described by Terzopoulos [4] was chosen to calculate the interpolation with a finite element approximation. The interpolation fills out gaps in areas where no disparity calculation is possible and creates a dense disparity map. The result of the smooth disparity interpolation for the object area is shown in Fig.3c. Dark gray values indicate small disparities (background) and light gray values large disparities (foreground).

INCREMENTAL MODEL BUILDING

The interpolated depth map contains the visible scene geometry measured from a single camera view point. Whenever the scene contains occluded surfaces then the camera must be moved around the objects and the measurements from multiple view points must be included. For that purpose the 2D depth map is converted into a 3D surface description that can be modified to include hidden surfaces.

Surface mesh generation

The depth map is converted into a piecewise continuous, parametric 3D surface description by spanning a wireframe in space for each contiguous object surface. For each object region the depth map is approximated by triangular, planar surface patches. The triangular mesh was chosen because it is capable to approximate arbitrary surface geometries without singularities. On the surface of each triangular patch the object surface texture is stored in a texture map from which a naturally looking view of the original objects can be synthesized with texture mapping. In Fig.4a the wireframe for the object "house" is shown. The image segmentation for this object was performed interactively to mark straight edges at the house corners. For each triangular patch the corresponding image texture is stored and used to synthesize a computer generated view which is shown in Fig.4c. The surface geometry was calculated from the interpolated disparity map while the surface texture was taken from the left original image.



(a) Disparity map (b) Segmentation mask (c) Interpolated disparity map (light values = near to camera, dark values = far from camera)

Fig.3: Segmentation and Interpolation

Motion estimation

The first 3D model which has been obtained from a single camera view point may be erroneous due to incorrect correspondence analysis and occluding parts. When complex scenes with occluding objects are analyzed then measurements from multiple view points have to be integrated into the 3D surface model. Therefore it is necessary to estimate the 3D motion of the camera and possible object motions in the scene from the image sequence and to fuse the multiple depth measurements into a consistent 3D scene model.

In this section an algorithm to directly estimate 3D scene motion from a monocular or stereoscopic image sequence is described briefly. A complete discussion of the algorithm can be found in [5].

An object is defined as a rigid 3D–surface in space that is spanned by a set of N control points. A set of six motion parameters is associated with each object. The object center **G** is the mean position vector of all N object control points. Each object control point $\mathbf{P}_{\mathbf{i}(k-1)}$ at frame k–1 is transformed to its new position $\mathbf{P}_{\mathbf{i}(k)}$ in frame k according to the general motion Eq.(2) between frame k–1 and k.

$$\mathbf{P}_{\mathbf{i}(k)} = [\mathbf{R}_G] \cdot (\mathbf{P}_{\mathbf{i}(k-1)} - \mathbf{G}) + \mathbf{G} + \mathbf{T} \quad \text{with} \qquad (2)$$

T=Translation vector and $[\mathbf{R_G}]$=matrix of rotation vector **R**.

The image displacement vector $\mathbf{d} = \mathbf{p}_2 - \mathbf{p}_1$ describes the projection of the observation point displacement $\mathbf{P}_{(k)} - \mathbf{P}_{(k-1)}$ with corresponding grey level values I_1 and I_2 onto the image plane. I_2 is measured at position of $\mathbf{p}_1$ at frame k, where I_1 is taken from image position $\mathbf{p}_1$ at frame k–1. When assuming a linear dependency of the surface texture between I_1 and I_2 and a brightness constancy constraint between frame k–1 and k it is possible to predict I_2 from I_1 and its corresponding image intensity gradients and hence to estimate **d** from the measurable difference $I_2 - I_1$. When approximating the spatial derivatives as finite differences the optical flow vector $\mathbf{d} = (d_x, d_y)^T$ can be predicted from the image gradients $\mathbf{g} = (g_x, g_y)^T$ and the temporal image intensity difference $\Delta I_{\mathbf{p1}} = I_2 - I_1$ between frame k and k–1 at $\mathbf{p}_1$ in Eq.(3) :

$$\Delta I_{\mathbf{p1}} = \mathbf{g} \cdot \mathbf{d} = g_x \cdot (p_{2x} - p_{1x}) + g_y \cdot (p_{2y} - p_{1y}) \qquad (3)$$

In Eq.(3) **d** is related to intensity differences. Substituting the perspective projection of $\mathbf{P}_{(k-1)}$ and $\mathbf{P}_{(k)}$ for $\mathbf{p}_1$ and $\mathbf{p}_2$ in Eq.(3) yields a direct geometric to photometric transform that relates the spatial movement of **P** between frame k–1 and k to temporal intensity changes in the image sequence at $\mathbf{p}_1$.

With this approach, rigid 3D object motion can be estimated directly from the image sequence when the object shape $\mathbf{P}_{(k-1)}$ is known. Assuming that rotation between successive images is small, $[\mathbf{R_G}]$ can be linearized and ΔI_{p1} is expressed as a function of the parameters **R** and **T**:

$$\begin{aligned}
\Delta I_{p1} \cdot P_z^2 = {} & f \cdot g_x \cdot P_z \cdot \mathbf{T_x} + f \cdot g_y \cdot P_z \cdot \mathbf{T_y} \qquad (4)\\
& - [\, \Delta I_{p1} \cdot P_z + f \cdot P_x g_x + f \cdot P_y g_y \,] \cdot \mathbf{T_z} \\
& - [\, \Delta I_{p1} \cdot P_z \cdot (P_y - G_y) + f \cdot P_x \cdot g_x \cdot (P_y - G_y) \\
& + f \cdot P_y \cdot g_y \cdot (P_y - G_y) + f \cdot P_z \cdot g_y \cdot (P_z - G_z) \,] \cdot \mathbf{R_x} \\
& + [\, \Delta I_{p1} \cdot P_z \cdot (P_x - G_x) + f \cdot P_x \cdot g_x \cdot (P_x - G_x) \\
& + f \cdot P_y \cdot g_y \cdot (P_x - G_x) + f \cdot P_z \cdot g_x \cdot (P_z - G_z) \,] \cdot \mathbf{R_y} \\
& + [\, f \cdot P_x \cdot g_y \cdot (P_x - G_x) - f \cdot P_z \cdot g_x \cdot (P_y - G_y)] \cdot \mathbf{R_z}
\end{aligned}$$

Significantly more than the minimum number of six linear independent equations are evaluated to solve for the six motion parameters **R** and **T** by using a linear regression. To account for the linearizing of $[\mathbf{R_G}]$ and nonlinear grey level distribution, the estimation is iterated.

For 3D motion estimation the object shape is assumed to be known. An initial estimate of the scene shape was generated from stereoscopic image analysis. When the initial estimate fails this dependency may affect the analysis and will sometimes lead to estimation errors. As long as the initial shape approximation is reliable, however, this dependency can be neglected.

Accumulation of depth maps

For each image pair of the sequence L_k, R_k a depth map D_k was calculated by stereoscopic analysis together with its associated confidence map C_k. 3D camera motion between successive frames was estimated which allows to register the image pairs relative to another. The goal of sequence accumulation is to fuse the depth measurements from the

(a) 3D surface model from depth map with interaction during segmentation (b) reconstructed image from textured surface model for second view point (c) original image of house from second view point

Fig.4: Results of image analysis

image sequence into a consistent 3D scene model to improve estimation quality. Consistency is achieved by compensating the camera motion from frame k–1 to k. The scene model M_{k-1} is transformed into the model M^*_k with the estimated motion parameters. From the model geometry in this position a prediction of the disparity map D^*_k can be computed and compared with the measured disparity map D_k to detect geometric errors.

Two types of geometric errors can be identified: Gross errors where the disparity estimation failed due to occlusions, repeated structures, or noise; and quantization errors due to the limited resolution of the disparity d. Gross errors can be detected and excluded from the analysis by comparing the predicted disparity map D^*_k with the measured disparity map D_k at frame k. In areas of large disparity difference only the measurement with higher confidence C_c should be chosen. The quantization of d leads to a quantization in depth estimation which can be severe for a small base line between the cameras and coarse resolution of d. The depth measurements are improved by weighted depth accumulation from the motion compensated sequence of depth maps. For each observation point $\mathbf{P}$ of the surface there exist a confidence value C_c from Eq.(1) that expresses the measurement accuracy. The confidence value C_c is converted into the weight S according to Eq.(5) that can easily be accumulated throughout the sequence. Each observation point holds not only its position $\mathbf{P}_{k-1}$ in space but also its corresponding confidence weight S_{k-1}. $\mathbf{P}_{k-1}$ is transformed to $\mathbf{P}^*_k$ according to Eq.(2) and its projection (x,y) in the image is computed. The disparity d_k, measured in frame k at image position (x,y) with corresponding weight S_k, is converted to a depth measurement $\mathbf{P}_k$ and fused with $\mathbf{P}^*_k$ in a weighted accumulation to compute the improved depth estimate $\mathbf{P}_k$' and weight S_k':

$$\mathbf{P}_k' = \frac{\mathbf{P}_k^* \cdot S_{k-1} + \mathbf{P}_k \cdot S_k}{S_{k-1} + S_k}, \quad S_k' = S_k + S_{k-1} \quad \text{with} \quad S = \frac{C_c}{1 - C_c} \qquad (5)$$

The information fusing process described above can only be applied to an existing surface. When new objects and prior unseen object surfaces appear, the surface mesh must be extended from the new measurement. Once the surface is built, the fusing process can continue.

Some results are presented in the Fig.4 above. A stereoscopic image sequence of a camera moving around a toy house was analyzed. From each camera position a stereoscopic depth map was generated and a 3D model was built from the depth maps. Fig.4a shows the model shape represented by a mesh of triangles. Fig.4b represents a reconstructed projection of the house from a different camera position. The quality of the modelling can be judged by comparing the reconstruction from the model in Fig.4b with Fig.4c which displays the original input image of the house from the same view point as in Fig.4b.

REFERENCES

[1] H. Harashima and F. Kishino: Intelligent Image Coding and Communications with Realistic Sensations – Recent Trends. IEICE Transactions, Vol.E 74, No.6, June 1991.

[2] Yakimovski, Y., R. Cunningham, A System for Extracting 3D Measurements from a Stereo Pair of TV Cameras, Computer Graphics and Image Processing Vol 7, pp. 195–210, 1978.

[3] Cox, I., Hingorani, S., Maggs, B., Rao, S., 1992. Stereo without Regularisation. British Machine Vision Conference, Leeds, UK, pp. 337–346, David Hogg & Roger Boyle (ed.), Springer Verlag.

[4] Terzopoulos, D., 1988. The computation of visible–surface representations, IEEE Trans. Patt. Anal. Mach. Intell., Vol 10, pp.417–438.

[5] R. Koch: Model–Based 3D Scene Analysis from Stereoscopic Image Sequences. Proceedings of the International Society of Photogrammetry and Remote Sensing, ISPRS–Conference, Washington, D.C., USA, Aug. 1992.