

# 3-D Head Pose Recovery for Interactive Virtual Reality Avatars

Marius D. Cordea, Dorina C. Petriu, *Member, IEEE*, Emil M. Petriu, *Fellow, IEEE*, Nicolas D. Georganas, *Fellow, IEEE*, and Thomas E. Whalen

**Abstract**—This paper discusses a tracking method allowing real-time recovery of the three-dimensional (3-D) position and orientation of a moving head. The described method uses a wireframe model of the head, a feature-based matching algorithm, and an extended Kalman filter estimator. The resulting motion tracking system works in a realistic environment without makeup on the face, with uncalibrated camera, and unknown lighting conditions and background.

**Index Terms**—3-D head, avatar, real-time tracking, virtual reality.

## I. INTRODUCTION

MODEL-BASED video coding (MBVC) has recently emerged as a very low bit rate video compression method suitable for collaborative virtual environment (CVE) applications [1]. The MBVC increases coding efficiency by using knowledge about the scene content and describing the real-world geometry by three-dimensional (3-D) model objects. The principle of this compression is to generate a parametric model of the image seen at the emission end and to transmit only the characteristic parameters describing how the model changes in time. These differential parameters are then used to animate the model of the image recovered at the reception end.

The first step in a full automatic MBVC system is the *face detection* allowing the identification and location of the face in the first image frames. The next step is *motion estimation* encompassing global 3-D-motion recovery, local motion estimation, expression and emotion analysis, etc. The problem is not trivial, as 3-D motion parameters have to be extracted from a sequence of two-dimensional (2-D) images of the performer's head-and-shoulders.

This paper discusses a *3-D tracking method* for the real-time measurement of six head motion parameters, namely 3-D position and orientation, and the focal length of the camera. This

method uses a 3-D wireframe head model, a 2-D feature-based matching algorithm, and an extended Kalman filter (EKF) estimator. Our global motion tracking system is meant to work in a realistic CVE without makeup on the speaker's face, with uncalibrated camera, and unknown lighting conditions and background.

## II. TRACKING HEAD MOTION

The general problem of recovering 3-D position parameters from 2-D images could be solved using different 2-D views of the 3-D objects. If these images are taken at the same time, the problem is solved by *stereovision* [2], [3] or *trifocal tensor* [4]. Another approach using monocular 2-D images of moving objects is known as *structure-from-motion* (SFM) [5].

Given 2-D-object images, the SFM problem aims to recover

- i) the 3-D object coordinates;
- ii) the relative 3-D camera-object motion;
- iii) camera geometry (camera calibration).

The SFM framework (Fig. 1) consists of two main modules

- i) *Tracking module*, delivering the 2-D point measurements  $p_i(u_i, v_i)$  of the tracked features, where  $i = 1, \dots, m$ , and  $m$  is the number of measurement points.
- ii) *Estimator module* (for the estimation of 3-D geometry and motion), delivering a state vector

$$s = (t_x, t_y, t_z, \alpha, \beta, \lambda, f, X_i, Y_i, Z_i) \quad (1)$$

where  $(t_x, t_y, t_z, \alpha, \beta, \lambda)$  are the six 3-D camera/object relative motions, namely translation and rotation,  $f$  is the camera focal length, and  $P_i(X_i, Y_i, Z_i)$  is the object geometry, where  $i = 1, \dots, m$  and  $m$  is the number of tracked features.

To detect and locate a human face, the system will process the image, identifying relevant features, and then use these features to recognize and determine the location of the face. Tracking finds and locates the relevant facial features in a sequence of images. Tracking should allow estimating the motion while locating the face. There are three main tracking techniques [6], [7].

- i) *Feature-based* methods extract image features and track their movement from frame to frame. Image features are low-level image descriptors, such as "regions," "edges," and "point features." Reliable tracking of *regions* is often difficult, since minor changes between frames can lead to very different segmentation in consecutive frames. Arbitrarily curving *edges* are difficult to describe and track. Trackers based on *point features* such as nostrils, corners

Manuscript received May 29, 2001; revised May 1, 2002. This work was supported in part by Communications and Information Technology Ontario (CITO), the STENTOR New Media Fund, and the Communications Research Centre (CRC) of Canada.

M. D. Cordea, E. M. Petriu, and N. D. Georganas are with the School of Information Technology and Engineering, University of Ottawa, Ottawa, ON K1N 6N5 Canada (e-mail: mcordea@uottawa.ca; petriu@site.uottawa.ca; georganas@site.uottawa.ca).

D. C. Petriu is with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1N 6N5 Canada (e-mail: petriu@sce.carleton.ca).

T. E. Whalen is with the CRC—Communications Research Center, Ottawa, ON K1N 6N5 Canada (e-mail: thom.whelen@crc.ca).

Digital Object Identifier 10.1109/TIM.2002.802261

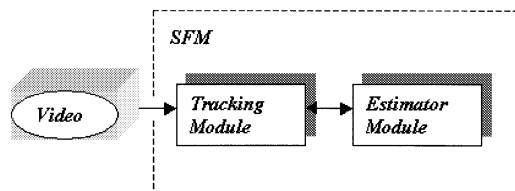


Fig. 1. The structure-from-motion (SFM) framework.

of eyes, mouth endpoints, and tips of eyebrows are increasingly used in computer vision applications [8], [9]. However, in a scene where objects move erratically, the noisy image data and spatial and temporal sub-sampling can make motion and acceleration estimation difficult.

- ii) *Optical-flow* methods use spatial and temporal partial derivatives to estimate the image flow at each location in the image. Algorithms for recovering optical flow [6] are based on a set of assumptions about the world that, by necessity, are simplifications and hence may be violated in practice, resulting in gross measurement errors. Moreover, the extraction of the optical flow from an image sequence is a highly computational task.
- iii) *Correlation-based* methods are popular for tracking objects [10], [11]. They use the sum of the absolute differences between template and search area pixel intensities as a difference measure. On the negative side, the correlation tracking methods are sensitive to changes in overall illumination changes between frames of the sequence.

We employ a *feature-based* tracking technique to obtain the 2-D observations, which SFM can use to infer the 3-D information.

The SFM problem can be formulated as a parameter estimation problem: “Given a number of noisy measurements of 2-D-tracker positions, we have to optimally recover the SFM components of (1).”

We have adapted the SFM approach of Azarbayejani and Pentland [5] to recursively recover the 3-D motion and perspective camera geometry from feature correspondences over a sequence of 2-D images. To speed up the calculations we are using a motion model that simplifies the Jacobian. The EKF is used to solve the SFM problem, resulting in an accurate, stable and real-time solution. The EKF takes into consideration the non-linear aspect of mapping. We use a perspective camera model to reflect the mapping between the 3-D world and its projection. In Section III, we present an EKF-based technique, used to recover 3-D motion parameters and camera focal length.

### III. EXTENDED KALMAN FILTER FOR 3-D TRACKING

The continuous imaging process is sampled at discrete time intervals by grabbing images at a constant time interval. These images are then sequentially analyzed using an EKF to determine the motion trajectory of the face within a determined error range.

The EKF converts the 2-D feature position measurements, using a perspective camera model, into 3-D estimates of the position and orientation of the head [5], [12], [13]. The EKF recursive approach captures both the cause-effect and the dynamic nature of the tracking, offering also a probabilistic framework for uncertainty representation.

The EKF is applied to nonlinear systems and consists of two stages: time updates (or prediction) and measurement updates (or correction). At each iteration, the filter provides an optimal estimate of the current state using the current input measurement and produces an estimate of the future state using the underlying state model. The values, which we want to smooth and predict independently, are the tracker state parameters.

The EKF state and measurement equations can be expressed as

$$s(k+1) = As(k) + \xi(k) \quad (2)$$

$$m(k) = Hs(k) + \eta(k) \quad (3)$$

where  $s$  is the state vector,  $m$  is the measurement vector,  $A$  is the state transition matrix,  $H$  is the Jacobian that relates state to measurement, and  $\xi(k)$  and  $\eta(k)$  are error terms modeled as Gaussian white noise.

The observations are the 2-D feature coordinates  $(u, v)$ , which are concatenated into a measurement vector  $m(k)$  at each time step. The observation vector is the back-projection of the  $s$  state vector containing the relative 3-D camera-scene motion, and the camera internal geometry, namely the focal length. In our case, the state vector is  $s(\text{translation}, \text{rotation}, \text{velocity}, \text{focal\_length})$  that contains the relative 3-D camera-object translation, rotation and their velocities, and camera focal length.

The EKF requires a physical dynamic model of the motion and a measurement model relating image feature locations to motion parameters. Additionally, a representation of the object (user’s head) is required.

#### A. Motion Model

The dynamic model is a discrete-time Newtonian physical model of a rigid body motion, moving with constant velocity. The state vector

$$s(t_x, t_y, t_z, \omega_x, \omega_y, \omega_z, f, \dot{t}_x, \dot{t}_y, \dot{t}_z, \dot{\omega}_x, \dot{\omega}_y, \dot{\omega}_z)$$

consists of 13 elements grouped as the relative camera-object translation  $(t_x, t_y, t_z)$ , the small inter-frame rotation  $(\omega_x, \omega_y, \omega_z)$ , the camera focal length  $f$ , the translational velocity  $(\dot{t}_x, \dot{t}_y, \dot{t}_z)$ , and the rotational velocity  $(\dot{\omega}_x, \dot{\omega}_y, \dot{\omega}_z)$ .

The state (1) could be written as

$$\begin{pmatrix} t_i \\ \omega_i \\ f \\ \dot{t}_i \\ \dot{\omega}_i \end{pmatrix}_{k+1} = \begin{pmatrix} I & 0 & 0 & I\Delta\tau & 0 \\ 0 & I & 0 & 0 & I\Delta\tau \\ 0 & 0 & I & 0 & 0 \\ 0 & 0 & 0 & I & 0 \\ 0 & 0 & 0 & 0 & I \end{pmatrix} \cdot \begin{pmatrix} t_i \\ \omega_i \\ f \\ \dot{t}_i \\ \dot{\omega}_i \end{pmatrix}_k + \xi(k) \quad (4)$$

where  $i = x, y, z$  is the index of the coordinate axes of the camera reference frame,  $I$  is the identity matrix, and  $\Delta\tau$  is the inter-frame time.

#### B. Measurement Model

The measurement model relates the state vector  $s$  to the 2-D-image location  $(u_k, v_k)$  of each image feature point,  $p_k$ . The

point  $p_k(X_k, Y_k, Z_k)$  of the object reference frame becomes the point  $p_{ck}(X_{ck}, Y_{ck}, Z_{ck})$  of the camera reference frame, where

$$\begin{pmatrix} X_{ck} \\ Y_{ck} \\ Z_{ck} \end{pmatrix} = T(t_x, t_y, t_z) + R(\alpha, \beta, \gamma) \begin{pmatrix} X_k \\ Y_k \\ Z_k \end{pmatrix} \quad k=1, \dots, N \quad (5)$$

where  $T$  and  $R$  represent the object (or camera) translation and rotation matrices, and  $N$  is the number of points.

The observed perspective projection is given by

$$\begin{pmatrix} u_k \\ v_k \end{pmatrix} = \frac{f}{Z_{ck}} \begin{pmatrix} X_{ck} \\ Y_{ck} \end{pmatrix}, \quad k=1, \dots, N \quad (6)$$

where  $f$  is the camera focal length.

At each filter cycle, we have to calculate the partial derivatives of  $u$  and  $v$  with respect to each of the unknown parameters. Lowe [14] proposed a re-parameterization of the projection equations, to simplify the calculation of the  $H$  Jacobian, by expressing the translations in the camera coordinate system rather than model coordinates. In this case, the measurement equation will take the following form:

$$\begin{pmatrix} X_{ck} \\ Y_{ck} \\ Z_{ck} \end{pmatrix} = R(\alpha, \beta, \gamma) \begin{pmatrix} X_k \\ Y_k \\ Z_k \end{pmatrix} \quad (7)$$

$$\begin{pmatrix} u_k \\ v_k \end{pmatrix} = \begin{pmatrix} \frac{f}{Z_{ck}+t_z} X_{ck} + t_x \\ \frac{f}{Z_{ck}+t_z} Y_{ck} + t_y \end{pmatrix}. \quad (8)$$

When  $N$  points are tracked, there are  $2N$  measurements (coordinates of point projections) at each frame and seven parameters to be recovered (six motion parameters plus camera focal length). Both motion and focal length are over-determined at each frame when  $2N > 7$ , which happens when  $N \geq 4$ , i.e., when tracking four or more points. When camera parameters are known beforehand, we need  $N \geq 3$  points to recover the 3-D motion.

We employ a three-parameter incremental rotation  $(\omega_x, \omega_y, \omega_z)$ , similar to that used in [5] to estimate inter-frame rotation. The incremental rotation computed at each frame step is combined into a global quaternion vector  $(q_0, q_1, q_2, q_3)$  used in the *EKF* linearization process and rotation of the 3-D-model [15].

## IV. IMPLEMENTATION AND EXPERIMENTAL RESULTS

### A. *EKF* Initialization

The 3-D model provides the initial structure parameters  $(X_i, Y_i, Z_i)$  of the Kalman filter. Each 2-D-feature point  $(u_i, v_i)$  corresponds to a structure point  $p_i(X_i, Y_i, Z_i)$ . As shown in Fig. 2, these  $(u_i, v_i)$  points are obtained by intersecting the 2-D image plane with a ray rooted in the camera's center of projection COP and aiming to the 3-D structure point on the head model.

The typical point identification problem of the 3-D pose recovery from 2-D images is solved in our case by identifying corresponding points in both the 2-D live image of the subject and the 3-D model of the subject's head. In order to aid the point

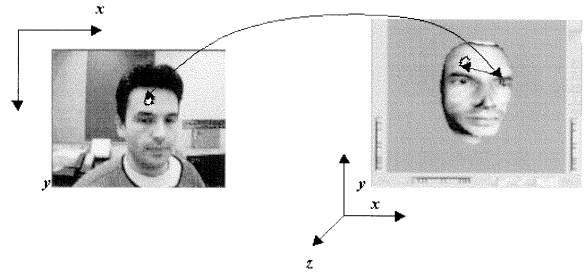


Fig. 2. Identical point selection process on Marius' image and the corresponding 3-D model projection.

identification process, we are using an augmented reality technique by projecting in the 2-D live image the 3-D mesh used to model the head.

At this development stage, it is still up to the user to arrange the scale matching between the live face image and the projected mesh. The steps of the *EKF* initialization algorithm for the multiple "point identification" procedure using this augmented reality technique are as follows.

- Step 1) The user positions his/her face at the center of the screen. Adjust the matching of the live image and the projected mesh, so that the projected mesh covers the entire facial region.
- Step 2) Left mouse click on every rigid feature point of interest on the live image. An automatic program function takes care to properly align the selected live feature point to a vertex of the projected mesh.
- Step 3) Right mouse click anywhere on the active Windows "live" image. This triggers the tracking process (by "booting" the *EKF* module).

### B. *EKF* Update

The *EKF* update stage is illustrated in Fig. 3. At each iteration, the *EKF* computes an estimate of the rigid 3-D motion that must probably correspond to the motion of the 2-D live image. We employ the Kanade–Lucas–Tomasi (KLT) [16] 2-D-gradient feature tracking method, which robustly performs the tracking reinforced by the *EKF* estimation output. An estimate of motion and camera focal length is found at each step. After the 3-D-motion and focal length are recovered, a perspective transformation will project feature points back onto the image to determine an estimated position of the 2-D feature trackers. At the next frame in the sequence, a 2-D tracking is performed starting at this 2-D estimated position. The current matching coordinates of tracked features are fed back into the Kalman filter as the observation vector, and the loop continues. The feedback from the *EKF* is used to update the 3-D-model pose parameters, i.e., it provides the 3-D head tracking information.

The recovered 3-D position and orientation are propagated to the *head modeling* block of the *CVE* system, which renders a new posture of the 3-D-model as illustrated in Fig. 4.

## V. CALIBRATION

In order to validate the accuracy of our 3-D-head tracking system, we developed a rapid calibration technique. A previ-

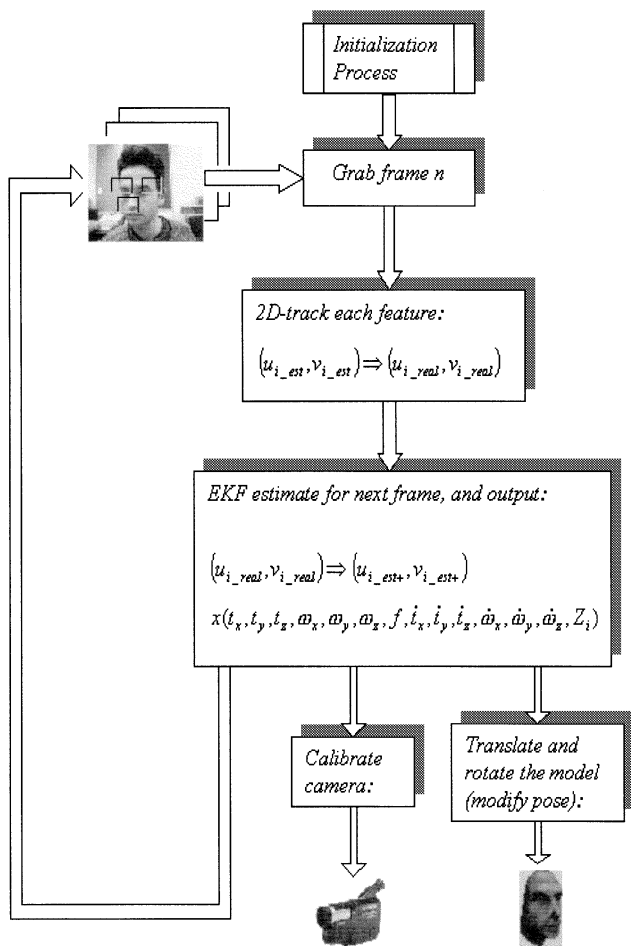


Fig. 3. Continuous 3-D pose recovery using EKF.



Fig. 4. Tracking the head motion.

ously recorded sequence of 2-D images representing 3-D head model poses is played as the “live” image and tracked with our EKF framework.

The estimated motion values  $(t_x^e, t_y^e, t_z^e, \theta_x^e, \theta_y^e, \theta_z^e)$  are compared with the measured motion values  $(t_x^m, t_y^m, t_z^m, \theta_x^m, \theta_y^m, \theta_z^m)$  of the synthetic image sequence. In the above representation  $(t_x, t_y, t_z)$  is the 3-D position, and  $(\theta_x, \theta_y, \theta_z)$  is the 3-D orientation of the head. The resulting errors show the effect of both human-aided 3-D/2-D point-identification and 3-D tracking.

We minimized the errors by fine-tuning the initialization process of the EKF.

Fig. 5 shows the *recovered* versus *real* 3-D-orientation for a calibrated sequence.

We have found experimentally in one case that the RMS difference between true and recovered rotation angles is  $3.035^\circ$

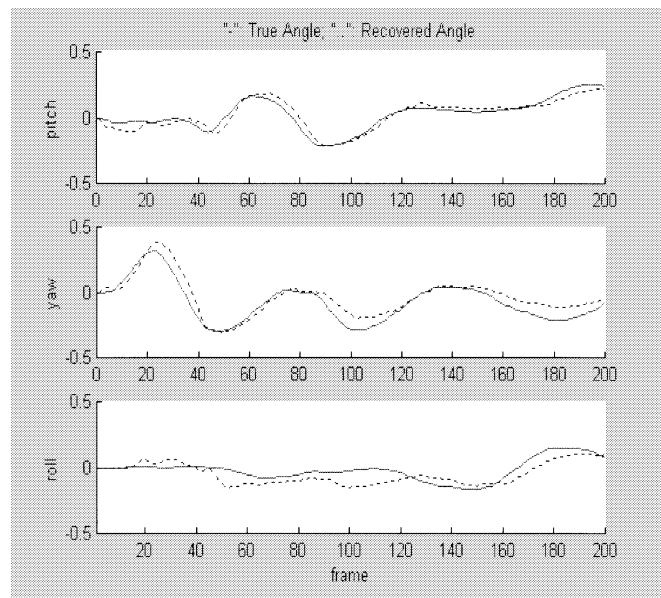


Fig. 5. True and recovered rotation angles: EKF-4 points.

when tracking four points. These statistics are comparable to the Polhemus sensor accuracy [5] indicating that the vision estimate is at least as accurate as the Polhemus sensor.

## VI. CONCLUSION

Tracking 3-D pose parameters of a moving target (head) from a sequence of 2-D-images motion is not a trivial task. The effects of head motion and facial expressions are combined in these images, so it is crucial to successfully separate the rigid from the nonrigid motion of the head (“pose/expression separation”). The head pose has to be accurately computed before attempting to recover the expressions.

The *3-D tracking* model-based algorithm discussed in this paper allows automatic recovery of six head-parameters: the 3-D position and orientation.

Our system was implemented on the 800 MHz PC-platform processing images provided by a commercial USB web camera at 30 fps. Experimental results show that this tracking system works well in a realistic videoconferencing environment, without makeup highlighting the speaker’s facial features, unknown lighting conditions, and unknown scene background.

From a computational point of view, this tracking system is based on the inversion of a  $2N$ -by- $2N$  matrix, where  $N$  is the number of points that are tracked. Both motion and focal length are over-determined at each frame for  $2N > 7$ , which happens when  $N \geq 4$ . Good results are routinely obtained even when tracking fewer than ten points. The computational effort needed for such a small size matrix inversion is easily handled in real-time even by a mid-level PC-platform. Our software-implemented SFM recovery system has successfully tracked a human face for minutes. If the head motions were relatively slow, our system was also able to deal effectively with occasional out-of-plane rotations and short occlusions. It is reasonable to expect that this limitation will be less restrictive as the speed of the PC-platform is increasing.

The measurement errors affecting our system are essentially due to image quantization and sampling. These errors are modeled as Gaussian distribution noise that is included in the iterative EKF model. This explains the remarkable robustness of the tracking system.

#### REFERENCES

- [1] K. Aizawa and T. S. Huang, "Model based image coding: Advanced video coding techniques for very low bit-rate applications," *Proc. IEEE*, vol. 3, Feb. 1995.
- [2] O. Faugeras, "What can be seen in three dimensions from an uncalibrated stereo rig?," in *Proc. 2nd Eur. Conf. Computer Vision*, Santa Margherita Ligure, Italy, 1992, pp. 563–578.
- [3] R. Hartley, R. Gupta, and T. Chang, "Stereo from uncalibrated cameras," in *Proc. Conf. Comput. Vision Pattern Recognit.*, Urbana-Champaign, IL, 1992, pp. 761–764.
- [4] R. Hartley, "Lines and points in three views—An integrated approach," in *Proc. ARPA IU Workshop. DARPA*, 1994.
- [5] A. Azarbayejani and A. Pentland, "Recursive estimation of motion, structure, and focal length," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, no. June, 1995.
- [6] B. K. P. Horn, *Robot Vision*. Cambridge, MA: MIT, 1986.
- [7] L. S. Shapiro, "Affine Analysis of Image Sequences," Ph.D. dissertation, Sharp Lab. of Europe, Oxford, U.K., 1995.
- [8] V. S. S. Hwang, "Tracking feature points in time-varying images using an opportunistic selection approach," *PR*, vol. 22, no. 3, pp. 247–256, 1989.
- [9] J. Shi and C. Tomasi, "Good features to track," in *IEEE Conf. Comput. Vision Pattern Recognit.*, Seattle, WA, June 1994.
- [10] G. D. Hager and P. N. Buelhumeur, "Real-time tracking of image regions with changes in geometry and illumination," in *IEEE Conf. Computer Vision Pattern Recognit.*, 1996, pp. 403–410.
- [11] T. Jebara and A. Pentland, *Parameterized Structure From Motion for 3-D Adaptive Feedback Tracking of Faces*. Cambridge, MA: Media Laboratory, 1996.
- [12] T. J. Broida and R. Chellappa, "Estimation of object motion parameters from noisy images," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 8, pp. 90–99, January 1986.
- [13] D. B. Gennery, "Visual tracking of known 3-dimensional object," *Int. J. Comput. Vis.*, vol. 7, no. 3, pp. 243–270, 1992.
- [14] D. G. Lowe, "Three-dimensional object recognition from single two-dimensional images," *Artif. Intell.*, vol. 31, no. 3, pp. 355–395, Mar. 1987.
- [15] K. Shoemake, *Quaternions*. Philadelphia, PA: Dept. Computer and Information Science, Univ. Pennsylvania.
- [16] J. Shi and C. Tomasi, "Good features to track," in *IEEE Conf. Comput. Vision Pattern Recognit.*, Seattle, WA, June 1994.

**Marius D. Cordea** received the engineering degree from the Polytechnic Institute of Cluj, Romania, and the M.S. degree in electrical and computer engineering from the University of Ottawa, Ottawa, ON, Canada. He is currently pursuing the Ph.D. degree at the School of Information Technology and Engineering, University of Ottawa.

His research interests include interactive virtual environments, pattern recognition, and animation languages.

**Dorina C. Petriu** (M'90) received the Dipl. Eng. degree in computer engineering from the Polytechnic University of Timisoara, Romania, and the Ph.D. degree in electrical engineering from Carleton University, Ottawa, ON, Canada.

She is currently an Associate Professor with the Department of Systems and Computer Engineering, Carleton University, Ottawa. Her research interests are in the areas of performance modeling and software engineering, with emphasis on integrating performance engineering into the software development process. She was a contributor to the UML Performance Profile standardized recently by OMG. Current research projects include automatic derivation of software performance models from UML design specifications, scalability analysis of virtual private networks, and software development and modeling for virtual environments and robotics.

Dr. Petriu is a member of ACM.

**Emil M. Petriu** (M'86–SM'88–F'01) received the Dipl. Eng. and Dr. Eng. degrees from the Polytechnic Institute of Timisoara, Romania, in 1969 and 1978, respectively.

He is currently a Professor with the School of Information Technology and Engineering, University of Ottawa, ON, Canada, where he has been since 1985. His research interests include test and measurement systems, interactive virtual environments, intelligent sensors, robot sensing and perception, neural-networks, and fuzzy control.

Dr. Petriu is a Registered Professional Engineer in the province of Ontario, Canada. He is a Fellow of the Engineering Institute of Canada and a Fellow of the Canadian Academy of Engineering. He is currently serving as Vice-President (Conferences), member of the AdCom, and Co-Chair of TC-15 of the IEEE Instrumentation and Measurement Society. He is an Associate Editor of the IEEE TRANSACTIONS ON INSTRUMENTATION AND MEASUREMENT and member of the Editorial Board of the IEEE I&M Magazine.

**Nicolas D. Georganas** (F'90) received the Dipl. Ing. degree in electrical engineering from the National Technical University of Athens, Greece, in 1966 and the Ph.D. degree in electrical engineering (summa cum laude) from the University of Ottawa, ON, Canada, in 1970.

He is a Professor with the School of Information Technology and Engineering, University of Ottawa. His research interests are in multimedia communications and collaborative virtual environments.

Dr. Georganas is a Fellow of the Engineering Institute of Canada, a Fellow of the Canadian Academy of Engineering, and a Fellow of the Royal Society of Canada. In 1998, he was selected as the University of Ottawa Researcher of the Year and also received the University 150th Anniversary Gold Medal.

**Thomas E. Whalen** received the Ph.D. degree in experimental psychology from Dalhousie University, Halifax, NS, Canada, in 1979.

Since 1979, he has been conducting research in human–computer interactions at the Communications Research Centre (CRC), a research institute of the Government of Canada. His current research interests include access to information through natural language queries, retrieval of images using visual features, Web-based training, and the control of avatars in shared virtual environments. He is also an Adjunct Professor with the Psychology Department of Carleton University, Ottawa, ON, Canada, and with the Department of Management Science and Finance at St. Mary's University.