D28.3 Technical Report # 32 on 3D Time-varying Scene Representation Technologies

Project Number: 511568
Project Acronym: 3DTV
Title: Integrated Three-Dimensional Television – Capture, Transmission and Display

Deliverable Nature: R
Number: D28.3
Contractual Date of Delivery: M41
Actual Date of Delivery: M43
Report Date: 30 March 2007
Task: WP8
Dissemination level: CO
Start Date of Project: 01 September 2004
Duration: 48 months
Organisation name of lead contractor for this deliverable: METU

Name of responsible: A. Aydın Alatan (alatan@eee.metu.edu.tr)

Editor: A. Aydın Alatan (alatan@eee.metu.edu.tr)

30 March 2008

3D Time-varying Scene Representation Technologies
TC1 WP8 Technical Report 3

EDITOR

A. Aydın Alatan (**METU**)

Contributing Partners to Technical Report:

Bilkent University (**Bilkent**)
Fraunhofer Gesellschaft zur Förderung der angewandten Forschung e.V. (**FhG-HHI**)
Institute of Media Technology, Technische Universitaet Ilmenau (**UIL**)
Informatics and Telematics Institute, Centre for Research and Technology Hellas (**ITI-CERTH**)
Koç University (**KU**)
Middle East Technical University (**METU**)
Momentum Bilgisayar Yazılım, Danışmanlık, Ticaret A.Ş. (**Momentum**)
Max-Planck-Institut fur Informatik (**MPG**)
University of West Bohemia in Plzen (**Plzen**)
University of Hannover (**UHANN**)
Technical University of Berlin (**TUB**)

REVIEWERS

Ugur Güdükbay (**Bilkent**)
Christian Weigel (**UIL**)
Xenophon Zaboulis (**ITI-CERTH**)
Tanju Erdem (**Momentum**)

Project Number: 511568
Project Acronym: 3DTV
Title: Integrated Three-Dimensional Television – Capture, Transmission and Display

TC1 WP8 Technical Report #3

TABLE OF CONTENTS

Executive Summary

This report summarizes the research efforts in EC IST 3DTV Network of Excellence (NoE) under Work Package 8 (WP8), entitled *3-D Scene Representation Technologies*, during Year-III of this project. The technical report consists of 22 novel research manuscripts, out of which 8 are obtained as a result of joint research activities between WP8 contributors. The research efforts are examined under point, surface, volume, human face & body and object specific representation topics.

Point-based representation research activities under WP8 mostly focus on description of a 3D scene by using only a single dense depth map, which is defined by the reference view of the recording camera. Such a representation can be obtained not only from defocused stereo-images by a novel approach (Section 3.1.2), but it can also be determined from multi-view sequences by plane-sweeping of segmented regions with homogenous color (Section 3.1.3). This representation is further improved from a single (dense) depth map to multiple (dense) depth fields (Section 3.1.1), since such an enriched representation should yield better results during generation of intermediate views for different applications of 3DTV. This research on multi-view-plus-depth investigates for the first time the influence of compression of multi-view video and depth maps on virtual intermediate views. As a final research effort on point-based representations, it is shown that dense depth maps with their associated views could be exploited together for segmentation and representation of the scene in terms of objects (Section 3.1.4).

For surface representations, building 3D mesh representations of dynamic or static scenes from multi-view video sequences is an important task for creating 3DTV applications. 3D mesh representations might serve for realistic rendering of real scenes directly in 3D via 3D displays or can be used for increasing the efficiency of video transmission by optimizing the rate-distortion trade-off. In this context, two novel mesh reconstruction methods, have been developed, one for efficiently representing the geometry of a dynamic scene explicitly in 3D using multi-camera video data (Section 3.2.1), and one for constructing an intermediate 3D mesh structure of the geometry of a scene from video sequences with the primary goal of improving the rate-distortion efficiency in video transmission (Section 3.2.2). On the other hand, in a pioneering research effort, extraction of mesh vertices is achieved by considering rate-distortion efficiency of the resulting mesh (Section 3.2.3). Hence, new vertices are only extracted to obtain a finer mesh, if there is enough bit-budget for such a refinement (Section 3.2.4). Finally, a preliminary research effort for upgrading a static mesh into a dynamic one is also presented (Section 3.2.5).

Volumetric representations enable high quality and realistic volumetric reconstruction of imaged scenes during the production of 3D content. High content quality is crucial in the acceptability of 3DTV technology by the general audience. In accordance, computational efficiency and automatic content production is critical in the reduction of the production cost. Texture uniqueness queue is one of the most powerful tools to achieve this quality (Section 3.3.1). By dividing 3D scene via planes and modifying the orientations, texture uniqueness can be enforced in order to determine the 3D structure of small arbitrarily shaped object regions (Section 3.3.2). Plane and angle sweeping can also be improved by size-modulation so that the projection area of patches in the acquired images is invariant to distance and rotation (Section 3.3.5). Finally, applicability of such techniques to cultural heritage scenario is shown to be possible (Section 3.3.3).

Representation of human face and body realistically is a fundamental requirement for a 3DTV system. The research efforts for human face representation have emphasis on multi-modal approaches that consider audio-visual clues together. Natural looking lip animation, synchronized with incoming speech, is essential for realistic character animation and it is best achieved by tri-viseme based HMM structure (Section 3.4.1). In a different multi-modal approach, in order to generate a dancing avatar, the video is analyzed to capture the time-varying posture of a human, whereas the musical audio signal is processed to extract the beat information (Section 3.4.2). In this manner, synthesis of a dancing avatar is achieved which is capable to performing dance figures for different music genres. Human

walking behavior is further studied based on inverse kinematics techniques resulting in an interactive hierarchical motion control system (Section 3.4.4). An extensive survey on human animation is presented with an emphasis on multi-layered human body models and motion control techniques for walking behavior (Section 3.4.3).

Modeling, animation, and rendering of human figures is a very active research area in computer graphics and computer vision since almost every 3D scene contains humans. However, different degrees of realism are needed in different applications, such as entertainment (game industry, film industry), medical applications, and 3DTV applications. Motion capture-based approaches and inverse kinematics techniques are among the best alternatives for realistic animation of human models existing in 3D scenes (Section 3.5.1). Within the context of object-specific representations for 3DTV, physically-based techniques for cloth simulation are being utilized (Section 3.5.3). Physically-based techniques generate realistic results, since they simulate the physics of materials (Section 3.5.4). As the procedural techniques are simple and practical, they provide a suitable alternative for the visualization of knitwear and woven cloth for 3DTV applications (Section 3.5.2).

For pseudo-3D representations, the proposed work mainly focuses on the development of a GPU assisted image processing framework for the generation of 3D video objects (Section 3.6.1 and Section 3.6.2). On the other hand, the relationship between light fields and holograms are examined as complementary representations, both of which can be converted into each other. (Section 3.6.3).

A comprehensive survey of 3D scene representation technologies, and their applicability for 3DTV applications, is also presented in Section 3.7.1. A number of completed theses work under the scope of WP8 is also listed in Section 3.7.2.

This report presents the promising results of research collaboration between contributors of WP8 on one of the fundamental technologies, i.e., the 3D scene representation technologies, in 3DTV systems. The joint efforts will continue along the same direction in Year IV, in order to determine preferable 3-D scene descriptions for 3DTV systems.

1. Introduction

Former WP8 research activities for 3DTV NoE have been already presented in Technical Reports #1 and #2. In those reports, high priority areas have been investigated through a number of joint and solo research activities. This report is a continuation of those efforts towards obtaining various efficient and powerful 3D scene representations applicable to 3DTV systems.

Noting that 3-D scene representation can be simply defined as the description of the observed scenery in terms of geometric primitives, such as *points*, *surfaces* or *volumes*, the research activities also follow a similar categorization. Special attention is devoted to the 3-D representations of *humans*, as well as animation of specific physically-modeled *objects*, under separate categories.

Following the structure of the previous technical reports, this document also has a similar outline. The technical report starts with an analysis of the results reported in the publications and continues with the presentation of the abstracts of these manuscripts. A discussion section with future directions concludes this report. All the related publications are presented in the Appendix.

2. Analysis of the Results Reported in Publications

In this part of the report, the presented research efforts will be examined separately, while indicating their novelties against the state-of-the-art.

Point-based (dense) approaches are expected be the representation technology for the pioneering 3DTV systems due to the recent and upcoming ISO MPEG standards. In MPEG-C (ISO/IEC 23002) Part-3 (Auxiliary Video Data Representation), description of the associated dense depth field of a video is already standardized. In the next few years, a similar extension will be standardized for multi-view video, where there will be an associated dense depth field for every view. For these representations to be promising, the required dense depth data should be extracted from (mono or multi-view) image sequences. In Section 3.1.2, dense depth extraction is achieved from a number of defocused stereo image pairs. In their method, the relative blurring between a number of defocused images are utilized as the 3D structure cue. However, the authors apply these principles to stereo cameras, in order to obtain two cues from defocus as well as stereo data. In a practical camera setup with optical real-apertures, disparity cannot be recovered on unfocused regions due to the increased ambiguity. This paper solves the problem uniquely by a novel anisotropic disparity estimation and pinhole image recovery embedding a stereo confocal constraint.

In Section 3.1.3, extraction of dense depth fields is achieved from multi-view video after modeling the scene by arbitrary shaped regions which are assumed to be planar. In this novel approach, plane- and angle-sweeping are applied to these regions while considering texture uniqueness clue in different views via projections. The resulting depth is further improved by applying a MRF-based formulation in order to relax the planarity constraint. Utilization of multi-view images is shown to improve the results especially in occluding regions. This method is one of the leading efforts for obtaining a piecewise planar scene for arbitrary

shaped natural regions; hence, a quite efficient representation that could be converted into meshes from dense depth field.

One of the leading research efforts on multi-view and depth representation is also presented in Section 3.1.1. This paper investigates for the first time in the literature the influence of coding of multi-view video plus depth on virtual intermediate views. Hence, it is possible to realize the required quality of encoded depth in order to obtain good quality intermediate virtual views to be utilized either in autostereoscopic displays or freeview-TV. As an important conclusion, the simulation results clearly indicate that coding artifacts on depth data strongly influence the reconstruction quality of rendered arbitrary views in a freeview-TV scenario. For depth data compressed at relatively low bit-rates, the reconstructions result with scattering artifacts, mainly around depth discontinuities. Hence, it is concluded that reasonable results can only be obtained at relative high bit-rates for dense depth compression.

Dense depth representation with its associated video can also be used for segmentation of 3D video objects, which have coherence in its texture, motion and 3D structure information, as in Section 3.1.4. Multi-view video permits all these 3 modalities to be extracted from itself to obtain semantically meaningful objects in the scene for any higher level interpretation. In the proposed method, optical flow estimation is obtained via region-based matching that has consistent parameterization with color segmentation and dense depth map estimation algorithms. This approach can be assumed to be one of the earliest methods, which aims to segment objects from multi-view data, with quite promising results.

Surface description is another promising representation for 3DTV systems with a number of available and upcoming standards. Mesh and other polygonal surface representations have been quite popular for decades, especially for the artificially generated content. In this report, two main directions on surface representation research are being examined. In one of these approaches, time-consistent dynamic meshes are generated and tracked, whereas in the other direction, the rate-distortion efficiency of the resulting meshes is pursued.

In Section 3.2.1, a novel approach for developing a mesh-based surface deformation framework for the general problem of 3D shape recovery is proposed. On the other hand, the method in Section 3.2.2 extends this deformation framework and uses it to track the surface geometry of a dynamic object from its multi-camera video acquisitions, producing an overall time-consistent dynamic mesh representation that can be encoded efficiently in terms of small-scale displacements and mesh restructuring operations. As compared to the few works that aim to construct time-consistent mesh representations of real scenes from multi-view video, the proposed method is much faster, it can track the surface geometry over longer periods, it is able to generate topologically correct and smooth mesh representations, and it can also handle objects with non-rigid motion.

In the 3D mesh compression literature, mesh representation typically initiates from a given set of vertices (and possibly their connections) and the compression algorithm aims to encode this given mesh in the most efficient manner. In Sections 3.2.3 and 3.2.4, a completely new approach is followed, where the extraction, representation and compression stages are related to each other in the sense that compression block feedbacks the bit-budget to extraction stage in case of any refinement is required by adding new vertices. In this manner, finer meshes are only resulted if there are enough available bits and more importantly, new vertices are selected in such a way that the distortion between original structure and mesh representation is minimized. There are quite interesting simulation results for this research between the

proposed mesh (surface) and dense depth (point) representations. According to these results, proposed mesh representation yield better results against dense depth maps.

Volume representations are usually obtained as a result of various 3D extraction techniques (such as *shape-from-silhouette* or *plane-sweeping*) and can be assumed as an intermediate 3D representation before the conversion to either point- or surface-based representations. In this report, the fundamental extraction technique before representations is plane- and angle-sweeping of the parallel planes which are defined in the scene (parallel to the reference camera). In Section 3.3.1, the extensive elaboration of such techniques is presented, followed with a novel method that extends state-of-the-art in increasing the accuracy, precision and efficiency of the way that texture uniqueness cue is implemented.

In Section 3.3.2, the previous approach is further improved by inclusion of angle-sweeping to plane-sweeping strategy for arbitrary shape segmented regions on different views, instead of pixels (or small rectangular regions) Novelty of this paper extends state-of-the-art in providing a segment-based stereo algorithm which can be utilized as a first step of a more sophisticated stereo algorithm. By this approach, stereo errors typically due to aperture phenomena can be reduced, which at the same time, this first step can facilitate a coarse-to-fine acceleration of traditional stereo approaches. It should be noted that this approach is also improved in Section 3.1.3, to obtain a dense depth map after relaxation of the planarity constraints.

An application of the aforementioned techniques is presented in Section 3.3.3. State of the art is extended in the following ways: First, the use of SIFT features is utilized to increase the accuracy of camera motion estimation and, consequently, of the obtained reconstruction. Second a completely automatic methodology for creating 3D VRML models of archaeological monuments in the Google Earth™ platform from 2D images is provided.

Plane-sweeping and application of texture uniqueness cue is further improved in the method at Section 3.3.4. This paper extends the state-of-the-art in stereo reconstruction by increasing the accuracy of the obtained reconstruction, due to the proposed *size-modulation* of the back-projection surface.

In two different joint research studies, representation of human face and body is approached in a multi-modal manner, considering both visual and audio information together. In Section 3.4.1, for realistic lip animation of synthetic models, audio information is being exploited. This research concludes via experimental comparisons that for realistic lip animation of 3D characters, utilization of viseme-based acoustic units is more successful compared to state-of-the-art phoneme-based acoustic unit method.

In Section 3.4.2, a novel framework for audio-driven human body motion analysis and synthesis is proposed. The problem has been addressed in the context of dance performance and it has been considered the most simplistic scenario possible, in which only a single dance figure is associated with each musical genre. In a typical dance performance, the body movements of the dancer are primarily driven by, and hence, highly correlated with, the musical audio signal; the work presented in this paper can be thought of as a first attempt to model this correlation towards the goal of automatic synthesis of a dancing avatar driven by musical audio. The experiments show that the avatar can successfully recognize the genre changes in a given audio track and synthesize the correct dance figures in a very realistic manner.

The book chapter on the human modeling and animation in Section 3.4.3 discusses some aspects of human modeling, animation, and rendering, with an emphasis on multi-layered human body models and motion control techniques for walking behavior.

A study in Section 3.4.4 presents an interactive hierarchical motion control system for the animation of human figure locomotion. The articulated figure animation system creates movements using motion control techniques at different levels, like goal directed motion and walking. Inverse Kinematics using Analytical Methods (IKAN) software, developed at the University of Pennsylvania, is utilized for controlling the motion of the articulated body.

Modeling and animation of different types of objects that are represented by physically-based techniques are examined as a separate topic. In Section 3.5.1, a book chapter on the scene representation techniques for different types of objects existing in 3D scenes compares the techniques for modeling, animation, and rendering of different types of objects, both in terms of the realism they offer and the computational cost as 3DTV applications require both realism and efficiency. The techniques that allow hardware implementations seem to be the promising approaches for 3DTV.

A simple and efficient procedural method is proposed in Section 3.5.2 for the visualization of knitted and woven textiles. For this purpose, rectangular mass-spring meshes are utilized and regular garment patterns composed of quadrilaterals are cut-out from the meshes. Regularity both preserves the general cloth behavior, such as shearing and bending; thus, preserving physical accuracy and enables the definitions of complex knit and weave patterns. In most of the existing systems, woven fabrics are not simulated physically; they are simplified as 2D structures. Hence, the regular structure of cloth models is exploited and the repetitious structure of woven and knitted fabric is parametrically defined. The proposed method does all the calculations on-the-fly; it does not require preprocessing for texture generation.

A 3D graphics environment for virtual garment design and simulation is proposed in Section 3.5.3, which enables the three dimensional construction of a garment from its cloth panels, for which the underlying structure is a mass-spring model. The garment construction process is performed through automatic pattern generation, posterior correction, and seaming.

Cloth is much stiffer than a damped spring and it should not behave like a rubber. However, when it is modeled with mass-spring systems, it may behave like a rubber. In order to prevent over-elongation of cloth materials, an elongation limit is set between rest length and elongation up to the rest length of the spring, as in Section 3.5.4.

Regarding Pseudo-3D representations, currently there is not so much interest in this kind of representation within WP8, compared for example to MV and depth representations. Nonetheless, the framework explained in Section 3.6.1 and 3.6.2. and some of its algorithms could also be used by partners for other representation technologies. In terms of quality evaluation conducting tests are planned in order to find optimal parameters for optimal visual quality.

Apart from all these published work under different research topics, an important joint contribution to the related literature is also achieved by the survey paper, which is obtained as a result of the document, *WP8 Survey Report on 3D Scene Representation Technologies*. The resulting paper in Section 3.7.1 might be the sole effort that summarizes and discusses the representation technologies in the literature.

3. Abstracts of Publications for Year-III

3.1. Point representations

3.1.1. Multi-view Video plus Depth Representation and Coding

Authors: P. Merkle, A. Smolic, K. Müller, and T. Wiegand

Institutions: Heinrich Hertz Institut

Publication: Proc ICIP 2007, IEEE International Conference on Image Processing, San Antonio, TX, USA, September 2007.

A study on the video plus depth representation for multi-view video sequences is presented. Such a 3D representation enables functionalities like 3D television and free viewpoint video. Compression is based on algorithms for multi-view video coding, which exploit statistical dependencies from both temporal and inter-view reference pictures for prediction of both color and depth data. Coding efficiency of prediction structures with and without inter-view reference pictures is analyzed for multi-view video plus depth data, reporting gains in luma PSNR of up to 0.5 dB for depth and 0.3 dB for color. The main benefit from using a multi-view video plus depth representation is that intermediate views can be easily rendered. Therefore the impact on image quality of rendered arbitrary intermediate views is investigated and analyzed in a second part, comparing compressed multi-view video plus depth data at different bit rates with the uncompressed original.

3.1.2. Confocal Disparity Estimation and Recovery of Pinhole Image in Real-aperture 3-D Camera system

Authors: Jang-Heon Kim, Thomas Sikora

Institutions: Technical University of Berlin

Publication: Proc ICIP 2007, IEEE International Conference on Image Processing, San Antonio, TX, USA, September 2007.

A single dense depth estimation using stereo or defocus cannot produce a reliable result due to the ambiguity problem. In this paper, we propose a novel anisotropic disparity estimation embedding a stereo confocal constraint for real-aperture stereo camera systems. If the focal length of a real-aperture stereo camera is just changed, the depth range is localized in a focused object which can be discriminated from defocused blurring. The focal depth plane is estimated by the displacement of tensors which are derived from generalized 2-D Gaussian, since the point spread functions (PSF) in defocused blurring can be approximated by a shift-invariant Gaussian function. We localize the isotropic propagation in blurring over invariance by a sparse Laplacian kernel in Poisson solution. The matching of real-aperture stereo images is performed by observing the focal consistency. However, the isotropic propagation cannot exactly hold a non-parallel surface to the lens plane i.e. unequifocal surface. An anisotropic

regularization term is employed to suppress the isotropic propagation near the non-parallel surface boundary. Our method achieves an accurate dense disparity map by sampling the disparities in focal points from multiple defocus stereo images. The pels in focal points are utilized to recover the pinhole image (i.e. an ideally focused image for all different depths).

3.1.3. Region-Based Dense Depth Extraction from Multi-view Video

Authors: Cevahir Çığla, Xenophon Zabulis and A. Aydın Alatan

Institutions: Middle East Technical University - Informatics and Telematics Institute

Publication: Proc ICIP 2007, IEEE International Conference on Image Processing, San Antonio, TX, USA, September 2007.

A novel multi-view region-based dense depth map estimation problem is presented, based on a modified plane-sweeping strategy. In this approach, the whole scene is assumed to be region-wise planar. These planar regions are defined by back-projections of the over-segmented homogenous color regions on the images and the plane parameters are determined by angle-sweeping at different depth levels. The position and rotation of the plane patches are estimated robustly by minimizing a segment-based cost function, which considers occlusions, as well. The quality of depth map estimates is measured via reconstruction quality of the conjugate views, after warping segments into these views by the resulting homographies. Finally, a greedy-search algorithm is applied to refine the reconstruction quality and update the plane equations with visibility constraint. Based on the simulation results, it is observed that the proposed algorithm handles large un-textured regions, depth discontinuities at object boundaries, slanted surfaces, as well as occlusions.

3.1.4. Depth Assisted Object Segmentation in Multi-view Video

Authors: Cevahir Çığla and A. Aydın Alatan

Institutions: Middle East Technical University

Publication: *submitted to 3DTV-CON 2008*

In this work, a novel and unified approach for multi-view video (MVV) object segmentation is presented. In the first stage, a region-based graph-theoretic color segmentation algorithm is proposed, in which the popular Normalized Cuts segmentation method is improved with some modifications on the graph structure. Segmentation is obtained by the recursive partitioning of the weighted graph. The proposed region-based approach is also utilized during the dense depth map estimation step, based on a novel modified plane- and angle-sweeping strategy. In the proposed dense depth estimation technique, the whole scene is assumed to be region-wise planar and 3D models of these plane patches are estimated by a greedy-search algorithm that also considers visibility constraint. Finally, the image segmentation algorithm is extended to object segmentation in MVV with the additional depth and optical flow information. Optical flow estimation is obtained via region-based matching that has consistent parameterization with color segmentation and dense depth map estimation algorithms. The experiment results

indicate that the proposed approach performs successfully and segments the meaningful objects in MVV with high precision.

3.1.5. Summary, conclusion, plans

Dense depth representation research activities in WP8 mostly focus on description of a 3D scene by using only a single dense depth map, which is defined by the reference view of the recording camera. In Section 3.1.2, this representation is obtained from defocused mono-images, whereas in Section 3.1.3, 3D scene description is determined by the help of multi-view sequences. This representation is further extended to multiple-views and their associated multiple (dense) depth fields in Section 3.1.1, since such a representation should yield better results while generating intermediate views. Finally, it is shown in Section 3.1.4 that dense depth map with its associated views could be exploited together for the representation of the scene in terms of objects.

Among different 3D scene representation techniques, point-based (dense) representations are currently the leading approach for 3DTV systems due to the recent standardization efforts under ISO MPEG. Single view with its associated dense depth information (*2D-plus-depth*) has been already standardized as a representation of 3D scene. However, the current trend within these bodies is to standardize multiple-views and their associated multiple (dense) depth fields, namely *multi-view-plus-depth*. Such a representation not only handles occlusions better, but it also generates more pleasant virtual views that can be utilized in freeview-TV or 3DTV systems. During the last year of this project, the groups are expected to conduct more research on this new dense point-based representation *multi-view-plus-depth*, mainly focusing on its extraction techniques and virtual view generation by the help of this representation.

3.2. Mesh representations

3.2.1. A Surface Deformation Framework for 3D Shape Recovery

Authors: Y. Sahillioğlu and Y. Yemez

Institutions: Koç University

Publication: Lecture Notes in Computer Science (MCRS'06), Vol. 4105, pp. 570-577, 2006.

We present a surface deformation framework for the problem of 3D shape recovery. A spatially smooth and topologically plausible surface mesh representation is constructed via a surface evolution based technique, starting from an initial model. The initial mesh, representing the bounding surface, is refined or simplified where necessary during surface evolution using a set of local mesh transform operations so as to adapt local properties of the object surface. The final mesh obtained at convergence can adequately represent the complex surface details such as bifurcations, protrusions and large visible concavities. The performance of the proposed framework which is in fact very general and applicable to any kind of raw surface data, is demonstrated on the problem of shape reconstruction from silhouettes. Moreover, since the approach we take for surface deformation is Lagrangian, that

can track changes in connectivity and geometry of the deformable mesh during surface evolution, the proposed framework can be used to build efficient time-varying representations of dynamic scenes.

3.2.2. Time-varying Surface Reconstruction from Multi-view Video

Authors: Y. Yemez and C. Bilir

Institutions: Koç University

Publication: To be submitted to Shape Modeling International (SMI'08)

We present a fast deformation-based method for building time-varying surface models of dynamic objects from multi-view video streams. Starting from an initial mesh representation, the surface of a dynamic object is tracked over time, both in geometry and connectivity, via a mesh-based deformation technique. The mesh representation of each frame is obtained by deforming the mesh representation of the previous frame towards the optimal isosurface defined by the time-varying multi-view silhouette information, using mesh restructuring operations and vertex displacements. The whole time-varying surface is then represented as a mesh sequence that can efficiently be encoded in terms of these restructuring operations and small-scale vertex displacements along with the initial model. Our reconstruction method hence yields a compact time-varying mesh representation of the dynamic object, which is smooth both in time and space. Another advantage of the proposed method is the ability to deal with dynamic objects that may undergo a non-rigid transformation. The time-varying mesh structure of such non-rigid surfaces, which is not necessarily of fixed connectivity, can also successfully be tracked thanks to the restructuring operations employed in our deformation scheme. We demonstrate the performance of the presented method on a synthetic human body model sequence.

3.2.3. Rate-distortion based Piecewise Planar 3D Scene Geometry Representation

Authors: Evren İmre, A. Aydın Alatan, and Uğur Güdükbay

Institutions: Middle East Technical University – Bilkent University

Publication: Proc ICIP 2007, IEEE International Conference on Image Processing, San Antonio, TX, USA, September 2007.

This paper proposes a novel 3D piecewise planar reconstruction algorithm, to build a 3D scene representation that minimizes the intensity error between a particular frame and its prediction. 3D scene geometry is exploited to remove the visual redundancy between frame pairs for any predictive coding scheme. This approach associates the rate increase with the quality of representation, and is shown to be rate-distortion efficient by the experiments.

3.2.4. Oran Bozulum Güdümlü Parçalı-Düzlemsel 3D Sahne Gösterimi (Rate-Distortion Guided Piecewise Planar 3D Scene Representation)

Authors: Evren İmre, A. Aydın Alatan, and Uğur Güdükbay

Institutions: Middle East Technical University – Bilkent University

Publication: IEEE Sinyal İşleme ve Uygulamaları Kurultayı (SIU'07), Eskişehir, Turkey, June 2007. (In Turkish)

This paper proposes a novel 3D piecewise planar reconstruction algorithm, which utilizes the statistical error between a particular frame and its prediction to refine a coarse 3D piecewise planar representation. The algorithm aims utilization of 3D scene geometry to remove the visual redundancy between frame pairs in any predictive coding scheme. This approach associates the rate increase with the quality of representation for determining an efficient description for a given budget. The preliminary experiments on synthetic and real data indicate the validity of the rate distortion based approach.

3.2.5. Iso-surface extraction from time-varying data

Authors: Slavomir Petrik

Institutions: University of West Bohemia, Plzen

Publication: Technical Report of University of West Bohemia for State of the Art and Future Research

This work is focused on the iso-surfaces extraction from scalar data sets with dynamic simulation mesh. Such data sets usually originate from Computational Fluid Dynamic (CFD) simulations, where moving boundaries of a simulation domain force a simulation mesh to change itself with each discrete time step. Up to now, each time step has been treated as a stand-alone entity, because of lack of the methods capturing spatial and temporal coherency in data sets of such nature.

The work covers the existing techniques for iso-surfaces extraction from time-varying scalar data sets with static mesh and provides an introduction to the problematic of dynamic meshes and an overview of the initial research done. Finally our concept of handling data sets with time-varying meshes is described.

3.2.6. Summary, conclusion, plans

In this report, two main directions on surface representation research are being examined. In one of these approaches, time-consistent dynamic meshes are generated and tracked, whereas in the other direction, the rate-distortion efficiency of the resulting meshes is pursued.

For time-consistent dynamic meshes, two methods for building 3D mesh representations of real scenes from multi-view video sequences are proposed. The first method aims to represent the surface

geometry of a dynamic object explicitly in 3D by using multi-camera video data. The method is computationally very efficient and can successfully track the time-varying geometry of a moving object from its multi-view silhouettes, using mesh deformation. The resulting representation is both spatially and temporally smooth and space efficient. Since both the connectivity and the geometry of the object can be tracked, the method is applicable also to objects with non-rigid motion. The current drawback of the proposed algorithm is that it may fail to track the surface in case the local motion is too fast on small shape details that can be confused by the silhouette information. For the algorithm to work successfully, there is generally a compromise between the frame-rate, the speed of the motion and the size of the shape details. We plan to overcome this problem by incorporating stereo information to the current algorithm, which would not only increase its robustness, but would also increase its efficiency since in this way a faster convergence would be possible with less number of iterations.

The second method constructs an intermediate 3D mesh structure of the geometry of a scene from video sequences with the primary goal of improving the rate-distortion efficiency. The proposed 3D piecewise planar reconstruction algorithm builds a 3D scene representation that minimizes the intensity error between a particular frame and its prediction. 3D scene geometry is exploited to remove the visual redundancy between frame pairs for any predictive coding scheme. The algorithm seeks a favorable point on rate-distortion curve by refining an initial mesh through the addition of new vertices, whose locations are determined by the prediction error. The experiments indicate that the proposed algorithm can yield efficient representations, thus it is an important step towards rate-distortion optimal 3D reconstruction for multi-view compression. However, in applications in which camera, structure or both should be estimated from the sequence, the algorithm requires accurate estimates of these parameters, in order to achieve satisfactory results.

On the other hand, a completely new approach is followed for surface representations, where the extraction, representation and compression stages are associated with each other in the sense that compression block feedbacks the bit-budget to extraction stage when refinement is required by adding new vertices. In this manner, finer meshes can only be obtained if there are enough available bits, and more importantly, new vertices are selected in such a way that the distortion between original structure and mesh representation is minimized.

3.3. Volume representations

3.3.1. Utilization of the texture uniqueness cue in stereo

Authors: Xenophon Zabulis

Institutions: Informatics and Telematics Institute

Publication: In Haldun M. Ozaktas, Levent Onural (Eds.), Three-Dimensional Television: Capture, Transmission, and Display (ch. 4), Springer Verlag, 2007

The cue to depth due to the assumption of texture uniqueness has been widely utilized in approaches to shape-from-stereo. Despite the recent growth of methods that utilize spectral information (color) or silhouettes to three dimensionally reconstruct surfaces from images, the depth cue due to the texture uniqueness constraint remains relevant, as being utilized by a significant number of contemporary stereo systems. Certainly, combination with other cues is necessary for maximizing the quality of the reconstruction, since they provide of additional information and since the texture-uniqueness cue exhibits well-known weaknesses; e.g. at cases where texture is absent or at the so-called “depth discontinuities”. The goal of this work is to provide of a prolific, in terms of accuracy, precision and

efficiency, approach to the utilization of the texture uniqueness constraint which can be, thereafter, combined with other cues to depth.

3.3.2. Segment-Based Stereo Matching via Plane and Angle-Sweeping

Authors: Cevahir Çığla, Xenophon Zabulis and A. Aydın Alatan

Institutions: Middle East Technical University - Informatics and Telematics Institute

Publication: Proceedings of the IEEE 3DTV-CONFERENCE: Capture, Transmission and Display of 3D Video, Kos, Greece, May 2007.

A novel approach for segment-based stereo matching problem is presented, based on a modified plane-sweeping strategy. The space is initially divided into planes that are located at different depth levels via plane sweeping by the help of region-wise planarity assumption for the scene. Over-segmented homogenous color regions are utilized for defining planar segment boundaries and plane equations are determined by angle sweeping at different planes. The robustness of depth map estimates is improved by warping segments into the other image via the resulting homographies. In order to refine the reconstruction quality and update segment depths, as well as plane normals, with smoothness and visibility constraints, a greedy search algorithm is applied. Based on the simulation results, the proposed algorithm handles large un-textured regions, depth discontinuities at object boundaries and slanted surfaces. Moreover, the algorithm could be easily upgraded from stereo to multi-view case, since 3D plane equations are already determined.

3.3.3. 3D Reconstruction for a Cultural Heritage Virtual Tour System

Authors: Yalin Bastanlar, Erdal Yilmaz, Yasemin Yardimci Cetin, Nikos Grammalidis, Xenophon Zabulis, Georgios Triantafyllidis

Institutions: Middle East Technical University - Informatics and Telematics Institute

Publication: Submitted to the XXI Congress, of the International Society for Photogrammetry and Remote Sensing, 3-11 July 2008, Beijing, China.

The aim of this study is to build a Web-based virtual tour system, focused at the presentation of archaeological sites. The proposed approach is comprised of powerful techniques such as multi-view stereo reconstruction; omni-directional viewing based on panoramic images, as well as, integration of the above with GIS technologies. In the proposed method, the scene is captured from multiple viewpoints and its 3D geometry is extracted from the acquired images based on stereoscopic techniques. Colour information is added to this 3D reconstruction of the scene and the result is provided to a 3D visualization tool for rendering.

The 3D scene could be artificially synthesized, e.g. by a 3-D modeler, by performing surface modeling and then adding texture information. Current applications are usually Web-based and are composed of elementary and graphical textures which are displayed via a VRML

plug-in. The problem with such synthesized 3D models for cultural applications is that the feeling of reality to the end-user is lost and that the procedure to generate them is tedious and requires highly-experienced personnel.

It is argued that utilization of real-image data in texture mapping enhances the feeling of reality for the end-user. The fully automatic multi-view reconstruction of a scene from real-images is not straight forward and, thus, a complete work-path for a reconstruction and presentation of archaeological sites is proposed. In short, this path is:

1) Acquisition of multiple images (preferably of high resolution) or video-recording and subsequent selection of key frames.
2) Computation of internal camera calibration parameters.
3) Estimation of lens distortion and image rectification.
4) Extrinsic calibration of the acquired images, based on robust feature extraction, tracking and camera motion estimation techniques,
5) Multi-view stereo reconstruction of the scene using the acquired images and intrinsic and extrinsic calibration parameters.
6) Conversion of the reconstruction output to textured VRML format, which includes triangulation of points into a mesh, combination of textures from different images
7) Generation of KML/KMZ file from VRML format.
8) Display of the reconstructed portion of the archeological site, on the Google Earth™ system or other GIS tools that support KML/KMZ format.

The method is utilized for the 3D modeling of the archaeological scenes. The technique stereoscopically processes images of the scene that are acquired from multiple viewpoints to produce its 3D reconstruction. As is the case for most multi-view stereo reconstruction techniques, the accuracy of the final results depends on the quality of both intrinsic and extrinsic camera calibration. To efficiently tackle the problem of fully-automatic calibration, the proposed approach is based on state-of-the-art algorithms for this problem, as well as custom modifications of these techniques that target the accuracy-improvement of our calibration results (e.g. robust feature point detection and matching using SIFT and bundle adjustment).

After successful 3D structure computation and reconstruction, output is generated in VRML format and converted into KML/KMZ formats to be integrated to Google Earth™ platform. In this way, the reconstructed sites can easily become a part of a large geographical information system (GIS) in the near future. We have developed a prototype system which uses excavation site plans as detailed raster overlays and interactive place marks as hot-spots. Omni-directional (or 360°) viewing based on panoramic images is a popular technique among Web-based virtual-tour applications. Using a map of the archaeological site increases the comprehension of the tour and enhances the user's sense of orientation. By enhancing our system with these tools, more information can be effectively communicated to the virtual tour users in an ergonomic and educational fashion.

3.3.4. Modulating the size of back-projection surface patches, in volumetric stereo, for increasing reconstruction accuracy and robustness

Authors: Xenophon Zabulis and Georgios D. Floro

Institutions: Informatics and Telematics Institute

Publication: Proceedings of the IEEE 3DTV-CONFERENCE: Capture, Transmission and Display of 3D Video, Kos, Greece, May 2007.

This paper concerns volumetric stereo methods, which compare the back-projections of the acquired images onto a hypothetical surface patch in order to reconstruct the imaged surfaces.

In particular, it introduces a size-modulation of this patch so that its projection area in the acquired images is invariant to distance and rotation. It is shown, and explained why, that performing this modulation results in superior accuracy of the volumetric reconstruction than retaining the patch size constant, as it has been to date practiced. The proposed extension to the hypothetical patch operator is compatible with the existing volumetric approaches to stereo.

3.3.5. Summary, Conclusion and Plans

A central problem in the provision of 3D content is the high cost of its production. The reason is the lack of automatic reconstruction techniques that produce high quality results. This results in requiring highly elaborate indoor studios for capturing the motion of a single person or requiring user-intervention to provide robustness against reconstruction errors. Investing on reconstruction techniques that are automatic and produce high quality results, reduces the production cost of 3D content for 3DTV. In this context, this research is focused on improving the quality of automatic reconstruction algorithms. The latest experiments mostly involve large-scale & outdoors reconstruction of archaeological monuments. During the last year, research effort was extended in applying our high reconstruction accuracy technology in large spatial volumes. In the future, this effort will be further extended for temporally long video recordings, in order to facilitate the 3D recording of dynamic events in high quality and resolution. Future plans concern the 3D acquisition of events and be able to facilitate the recording of cultural and sport events.

3.4. Human Face and Body Specific Techniques

3.4.1. Comparison of Phoneme and Viseme Based Acoustic Units for Speech Driven Realistic Lip Animation

Authors: Elif Bozkurt, Çiğdem Eroğlu Erdem, Engin Erzin, Tanju Erdem, Mehmet Özkan

Institutions: Momentum – Koc

Publication: Proceedings of the IEEE 3DTV-CONFERENCE: Capture, Transmission and Display of 3D Video, Kos, Greece, May 2007.

Natural looking lip animation, synchronized with incoming speech, is essential for realistic character animation. In this work, we evaluate the performance of phone and viseme based acoustic units, with and without context information, for generating realistic lip synchronization using HMM based recognition systems. We conclude via objective evaluations that utilization of viseme based units with context information outperforms the other methods.

3.4.2. An Audio-Driven Dancing Avatar

Authors: F. Ofli, Y. Demir, E. Bozkurt, E. Erzin, Y. Yemez, M. Tekalp, T. Erdem, K. Balcı, I. Kizoğlu, L. Akarun, C. Canton-Ferrer, J. Tilmanne

Institutions: Koc-Momentum

Publication: Journal on Multimodal User Interfaces, in review

We present a framework for training and synthesis of an audio-driven dancing avatar. The avatar is trained for a given musical genre using the multi-camera video recordings of a dance performance. The video is analyzed to capture the time-varying posture of the dancer's body whereas the musical audio signal is processed to extract the beat information. We consider two different marker-based schemes for the motion capture problem. The first scheme uses 3D joint positions to represent the body motion whereas the second uses joint angles. Body movements of the dancer are characterized by a set of recurring semantic motion patterns, i.e., dance figures. Each dance figure is modeled in a supervised manner with a set of HMM (Hidden Markov Model) structures and the associated beat frequency. In the synthesis phase, an audio signal of unknown musical type is first classified, within a time interval, into one of the genres that have been learnt in the analysis phase, based on mel frequency cepstral coefficients (MFCC). The motion parameters of the corresponding dance figures are then synthesized via the trained HMM structures in synchrony with the audio signal based on the estimated tempo information. Finally, the generated motion parameters, either the joint angles or the 3D joint positions of the body, are animated along with the musical audio using two different animation tools that we have developed. Experimental results demonstrate the effectiveness of the proposed framework.

3.4.3. Modeling, Animation, and Rendering of Human Figures

Authors: Uğur Güdükbay, Bülent Özgüç, Aydemir Memişoğlu, and M. Şahin Yeşil

Institution: Bilkent University

Publication: Three-Dimensional Television - Capture, Transmission, Display, (Chapter 7), Edited by H. M. Özaktaş and L. Onural, Springer, 2007.

Human body modeling and animation has long been an important and challenging area in computer graphics. The reason for this is two-fold. First, the human body is so complex that no current model comes even close to its true nature. The second reason is that our eyes are so sensitive to human figures that we can easily identify unrealistic body shapes (or body motions).

Today many fields use 3D virtual humans in action: video games, films, television, virtual reality, ergonomics, medicine, biomechanics, etc. We can classify all these applications into three categories: film production for arts and entertainment, real-time applications, such as robotics, video games and virtual environments, and simulations, such as computer-aided ergonomics for the automotive industry, virtual actors, biomedical research, and military simulations. The type of application determines the complexity of the models. For example video games or virtual reality applications require the lowest possible ratio between the computation cost and capabilities of the model. However, for biomedical research, realism is essential and the animated model should obey physical laws. Hence, the models are designed and animated according to the specific area in which they are applied.

Humans are an indispensable part of dynamic 3D scenes. Therefore, human face and body specific representations and animation techniques should be heavily used in a 3DTV framework to achieve the goals of real-time implementation and realism. Techniques of 3D motion data collection, such as motion capture, can be incorporated in human model animation. Continuous video and motion recording at high sampling rates produce huge amounts of data. Key-frame transmission that can be regenerated into continuous motion using interpolation techniques will reduce the size of the transmitted data significantly.

To study human modeling and animation, many techniques based on kinematics, dynamics, biomechanics, and robotics have been developed by researchers. In order to produce realistic animations, rendering is also an inseparable part. Furthermore, hair, garment, interaction of multiple avatars, expression of feelings, behavior under extreme conditions (such as accidents, deep sea diving, etc.) and many more real life experiences make the problem as complicated as one's imagination.

The human body has a rigid skeleton. This is not the case with some other living, artificial or imaginary objects. If the animation aims at a particular instance of bone fracture, maybe for an orthopedical simulation, then the rules all of a sudden change. As long as the subject excludes these non-articulated body behaviors, there is a reasonable starting point, a skeleton that is an articulated object with joints and rigid elements. It is natural, then, to assume that if a proper motion is given to the skeleton, one can build up the rest of the body on top of this. Layers include muscles, skin, hair and garments that can somehow be realistically rendered based on skeleton motion, plus some external forces, such as wind and gravity, to add more realism, at least to hair and garment. This obviously is a reverse way of looking at things; it is the muscles that expand or contract to give motion to the skeleton, but if the ultimate aim is to generate a realistic animation visually, and if the muscles can be accurately modeled, the order in which the forces are originated can be reversed. This makes the skeletal motion to be the starting source of animation.

It is very difficult to fit all the aspects of human modeling and animation into a limited scope of a book chapter. Thus, this chapter discusses some aspects of human modeling, animation,

and rendering, with an emphasis on multi-layered human body models and motion control techniques for walking behavior.

3.4.4. Motion Control for Realistic Walking Behavior Using Inverse Kinematics

Author: Aydemir Memişoğlu, Uğur Güdükbay and Bülent Özgüç

Institution: Bilkent University

Publication: Proceedings of the IEEE 3DTV-CONFERENCE: Capture, Transmission and Display of 3D Video, Kos, Greece, May 2007.

This study presents an interactive hierarchical motion control system for the animation of human figure locomotion. The articulated figure animation system creates movements using motion control techniques at different levels, like goal directed motion and walking. Inverse Kinematics using Analytical Methods (IKAN) software, developed at the University of Pennsylvania, is utilized for controlling the motion of the articulated body.

3.4.5. Summary, conclusion, plans

Four different Hidden Markov Model structures for realistic lip animation have been tested given a speech file as the only input. The performances of the phone, tri-phone, viseme and tri-viseme acoustic units are considered for HMM based viseme recognition. Based on the objective viseme recognition rates, one can conclude that the tri-viseme based HMM structure outperforms the other structures.

We have developed a framework for audio-driven human body motion analysis and synthesis. We have addressed the problem in the context of dance performance and considered the most simplistic scenario possible in which only a single dance figure is associated with each musical genre. The experiments show that the avatar can successfully recognize the genre changes in a given audio track and synthesize the correct dance figures in a very realistic manner. The avatar can also keep track of the changing beat information and adjust the speed of the dance movements accordingly. A crucial task during avatar training is to capture the motion of the dancer in an accurate manner. For this, we have developed a marker-based algorithm based on annealing particle filtering that can automatically extract the human posture from multi view video without any human intervention. Our future work will involve unsupervised training of our dancing avatar for different musical genres in more complicated scenarios in which the dance figures are more sophisticated in structure, having certain syntactic rules and hierarchies of figures. In order to achieve this, we will also need to consider various musical audio features other than beat and tempo, such as tonality, harmony and melody.

On the other hand, multi-layered modeling of human motion based on anatomical approach yields realistic and real time results for generating avatars in motion for three dimensional computer animations. Films produced by these techniques have become very popular and wide spread, many with major box-office success. The motion, up until recently, was specified by the animators. The current developments in motion capture can also provide data

for animation thus making human model animation a readily integral part of three dimensional television systems.

Building up motion databases and using various motion definitions from such databases, as well as modifying parts of this data whenever needed, provides a very powerful tool for the entertainment industry for generating three dimensional realistic humans in motion. Many other research applications ranging from medical to flight simulators also need correct human models.

3.5 Object Specific Representations: Modeling, Rendering and Animation Techniques

3.5.1. Three-Dimensional Scene Representations - Modeling, Animation, and Rendering Techniques

Author: Uğur Güdükbay, Funda Durupınar

Institution: Bilkent University

Publication: In Three-Dimensional Television – Capture, Transmission, Display, (Chapter 6), Edited by H. M. Özaktaş and L. Onural, Springer, 2007.

Modeling the behavior and appearance of captured three-dimensional (3D) objects is a fundamental requirement for scene representation in a three-dimensional television (3DTV) framework. By using the data acquired from multiple cameras, it is possible to model a scene with high quality visual results. In fact, 3D scene capturing and representation phases are highly correlated. Information acquired from the capturing phase can be employed in the representation phase by using computer graphics and image processing techniques. The resultant model then allows the users to interact with the scene, not just remain observers but be participants themselves. Thus, the main considerations for the quality of a scene representation technique are basically the accuracy of the technique about how the results correspond to the original scene and the efficiency of the technique as real-time performance is required. 3D shape modeling is an essential component of scene representation for 3DTV. Time-varying mesh representations provide a suitable way of representing 3D shapes. With these methods, the static components of a scene are constructed only once and the other objects are modeled as dynamic components, thus the computational time to represent 3D scenes is reduced. Polygonal meshes are efficiently used in shape modeling due to their builtin representation in hardware. Thus, they are suitable for applications such as 3DTV where real-time performance is required. Alternatively, volumetric representations can be used in shape modeling. The basic volume elements, voxels, of a 3D space correspond to the 2D pixels of an image. Volumetric techniques require large amounts of data in order to represent a scene or object accurately. Images acquired from multiple calibrated cameras provide the necessary information for volumetric models. Thus, these methods are intuitive for 3DTV. However, recent research shows that point-based approaches are the most suitable shape modeling techniques for 3DTV. The reason is that results of 3D data acquisition

methods such as laser scans already represent the scene in a point-based manner. 3D scene representation has two components: *geometry* and *texture*. Geometry representation is handled by modeling the shape of an object or a scene. Since the scenes mostly contain dynamic objects that move and deform in different ways, modeling the motion becomes important. Animation techniques that have potential for real-time hardware implementations are promising approaches to be used in a 3DTV framework. Texture representation is handled by the underlying rendering technique. Scan-line rendering techniques are suitable for 3DTV as they are hardware-supported and efficient. In addition, image-based rendering is a very successful and promising rendering scheme for 3DTV as it directly makes use of the captured images. This chapter provides introductory knowledge for the modeling, animation, and rendering techniques used in computer graphics. It is not an exhaustive survey of these topics and includes only representatives of each, focusing on techniques relevant to 3DTV. The interested reader is referred to the references for an in-depth discussion of the topics covered. The chapter is organized as follows. First, different 3D scene representation techniques, namely mesh-based representations, volumetric methods, and point-based techniques, will be discussed. Then, we will explain animation techniques for modeling object behavior. Finally, we will discuss illumination models and rendering techniques for 3D scenes containing different types of objects and lighting conditions.

3.5.2. Procedural Visualization of Knitwear and Woven Cloth

Author: Funda Durupınar and Uğur Güdükbay

Institution: Bilkent University

Publication: Computers & Graphics, Vol. 31, No. 5, pp. 778-783, October 2007.

In this paper, a procedural method for the visualization of knitted and woven fabrics is presented. The proposed method is compatible with a mass-spring model and makes use of the regular warp–weft structure of the cloth. The visualization parameters for the loops and threads are easily mapped to the animated mass-spring model. The simulation idea underlying both knitted and woven fabrics is similar as we represent both structures in 3D. As the proposed method is simple and practical, we can achieve near real-time rendering performance with good visual quality.

3.5.3. A Virtual Garment Design and Simulation System

Author: Funda Durupınar and Uğur Güdükbay

Institution: Bilkent University

Publication: In Proceedings of Information Visualization (IV'07), E. Banissi et al. (Eds.), pp. 862-870, Zurich, Switzerland, IEEE Computer Society, 2007.

In this paper, a 3D graphics environment for virtual garment design and simulation is presented. The proposed system enables the three dimensional construction of a garment from

its cloth panels, for which the underlying structure is a mass-spring model. The garment construction process is performed through automatic pattern generation, posterior correction, and seaming. Afterwards, it is possible to do fitting on virtual mannequins as if in a real life tailor's workshop. The system provides the users with the flexibility to design their own garment patterns and make changes on the garment even after the dressing of the model. Furthermore, rendering alternatives for the visualization of knitted and woven fabric are presented.

3.5.4. Practical and Realistic Animation of Cloth

Author: Serkan Bayraktar, Uğur Güdükbay and Bülent Özgüç

Institution: Bilkent University

Publication: Proceedings of the IEEE 3DTV-CONFERENCE: Capture, Transmission and Display of 3D Video, Kos, Greece, May 2007.

In this paper, we propose a system for the practical animation of cloth materials. A mass spring based cloth model is used. Explicit time integration methods are used to solve the equations of motion. We update the spring constants dynamically according to the net force acting on them. In this way, spring constants do not grow arbitrarily to introduce numerical instability and realistic cloth appearance without over elongation is obtained.

3.5.5. Summary, conclusion, plans

One promising approach for realistic animation of human models is using motion capture-based approaches since very realistic animations can be generated using these techniques. We make research on *filtering* and *keyframe reduction* of motion capture data. Filtering is necessary to eliminate any jitter introduced by a motion capture system. Key-frame reduction allows animators to easily edit motion data by representing animation curves with a significantly smaller number of key frames. Since the curve simplification-based keyframe reduction technique produce promising results, we plan to investigate these techniques further.

Inverse kinematics-based approaches still provide one of the best alternatives for realistic motion control of human figures in 3D scenes. We develop high level and low-level inverse kinematics based motion control techniques for realistic humans walking behavior. We plan to extend this research for other complex human behaviors.

Within the context of object-specific representations for 3DTV, physically-based techniques generate realistic results for cloth simulation since they simulate the physics of materials. Since the procedural techniques are simple and practical, they provide an alternative for the visualization of knitwear and woven cloth for 3DTV applications. In the future, we will further elaborate on physically-based cloth simulation and procedural visualization of cloth materials.

3.6. Pseudo-3D Representations

3.6.1. Framework for real time 3D video object generation

Authors: Christian Weigel, Stefan Werner, Peter Schübell

Institutions:UIL

Publication: A real-time image-based rendering framework (unpublished)

We present a software framework designed for real-time image-based rendering applications. Due to its modular structure the framework suits for a number of tasks mainly in the computer vision area. First, we introduce the state of the art software design using parallel processing paradigms. Then, we show examples of 3DTV applications based on the system namely the pre-processing and synthesis of so called 3D video objects based on the theory of the trifocal transfer. We point out that due to the possibility of freely configuring the system the calculation of pre-processing and synthesis algorithms are accelerated significantly. Finally, we present possible future applications.

3.6.2. Evaluation of different 3D video object synthesis methods

Author: Christian Weigel

Institutions:UIL, HHI

Publication: (UIL solo) Trifocal Transfer on Commodity Graphics Hardware (unpublished)

We present an algorithm and its implementation for the interactive view synthesis of 3D video objects. The algorithm is based on the theory of trifocal transfer by using dense depth information of two cameras. In order to implement the algorithm and its post processing stages at interactive rates we employ commodity graphics hardware. We present issues that arise when using the pipeline architecture of the GPU for non-computer graphics algorithms and introduce approaches how to solve these problems with focus on our application. Finally, we show that the employment of a GPU can speed up the synthesis significantly.

3.6.3. A Bidirectional Light Field - Hologram Transform

TC1 WP8 Technical Report #3

Author: Remo Ziegler, Simon Bucheli, Lukas Ahrenberg*, Marcus Magnor, Markus Gross*

Institutions: ETH Zurich, MPI, TU Braunschweig

Publication: EUROGRAPHICS 2007

In this paper, we propose a novel framework to represent visual information. Extending the notion of conventional image-based rendering, our framework makes joint use of both light fields and holograms as complementary representations. We demonstrate how light fields can be transformed into holograms, and vice versa. By exploiting the advantages of either representation, our proposed dual representation and processing pipeline is able to overcome the limitations inherent to light fields and holograms alone. We show various examples from synthetic and real light fields to digital holograms demonstrating advantages of either representation, such as speckle-free images, ghosting-free images, aliasing-free recording, natural light recording, aperture-dependent effects and real-time rendering which can all be achieved using the same framework. Capturing holograms under white light illumination is one promising application for future work.

3.6.4. Summary, conclusion, plans

The development of a test and demonstration framework for interactive video object generation is still one of the main tasks within this subgroup. Still, this work is mostly done solo by UIL. Beside the informative exchanges of knowledge no interfaces between participants within this WP could be identified. Nonetheless, this work is still useful for all partners of the NoE which need to combine image processing algorithms in a workflow and evaluate their performance. The novelty of this frame compared to similar image processing frameworks is the ability of researchers to use CPU and GPU accelerated 3D related image processing algorithms simply by combining modules in an XML file. In future version this process will be assisted visually.

The interrupted collaboration between UIL and HHI from the first year is still not continued in this period which is clearly dissatisfactory. The reason is still the huge effort in software development at UIL. Within this scope it was also found, that the capturing methods for both types of analysis and representations differ a lot and the creation of commonly used test material is very difficulty. Furthermore, the focus at UIL shifted to the sole evaluation of pixel based methods. Therefore the pixel based synthesis of virtual views was developed further as shown in the paper. Most of GPU assisted algorithms focus on real time disparity estimation. The contribution is one of the first implementing the trifocal transfer algorithm by means of general purpose GPU methods.

We plan to process more and more parts of the whole processing chain for Pseudo-3D objects in real time in order to allow for interactive display. Currently there is not so much interest in this kind of representation within the WP compared for example to MV and depth representations. Therefore most of the work is a solo contribution by UIL. Nonetheless the framework and some of its algorithm could also be used by partners for other representation technologies. In terms of quality evaluation we plan to conduct tests in order to find optimal parameters for optimal visual quality. In collaboration with partner from WP12 subjective tests with quantitative and also qualitative approaches will from the basis for this plan.

3.7. Other Research Outputs for 3D Representation Technologies

3.7.1. Scene Representation Technologies for 3DTV – A Survey

Author: A. Alatan, Y. Yemez, U. Gudukbay, X. Zabulis, K. Muller, C. Erdem, C. Weigel, and A. Smolic

Institution: Middle East Technical University, Koç University, Bilkent University, ITI-CERTH, Fraunhofer Institute for Telecommunications-Heinrich-Hertz-Institut, Momentum A.S, TÜBITAK-MAM-TEKSEB, Technical University of Ilmenau

Publication: IEEE Transactions on Circuits and Systems for Video Technology, Special Issue on Multi-view Video Coding and 3DTV, November 2007.

3-D scene representation is utilized during scene extraction, modeling, transmission and display stages of a 3DTV framework. To this end, different representation technologies are proposed to fulfill the requirements of 3DTV paradigm. Dense point-based methods are appropriate for free-view 3DTV applications, since they can generate novel views easily. As surface representations, polygonal meshes are quite popular due to their generality and current hardware support. Unfortunately, there is no inherent smoothness in their description and the resulting renderings may contain unrealistic artifacts. NURBS surfaces have embedded smoothness and efficient tools for editing and animation, but they are more suitable for synthetic content. Smooth subdivision surfaces, which offer a good compromise between polygonal meshes and NURBS surfaces, require sophisticated geometry modeling tools and are usually difficult to obtain. One recent trend in surface representation is point-based modeling which can meet most of the requirements of 3DTV; however, the relevant state-of-the-art is not yet mature enough. On the other hand, volumetric representations encapsulate neighborhood information that is useful for the reconstruction of surfaces with their parallel implementations for multi-view stereo algorithms. Apart from the representation of 3-D structure by different primitives, texturing of scenes is also essential for a realistic scene rendering. Image-based rendering techniques directly render novel views of a scene from the acquired images, since they do not require any explicit geometry or texture representation. 3-D human face and body modeling facilitate the realistic animation and rendering of human figures that is quite crucial for 3DTV that might demand real-time animation of human bodies. Physically based modeling and animation techniques produce impressive results, thus have potential for use in a 3DTV framework for modeling and animating dynamic scenes. As a concluding remark, it can be argued that 3-D scene and texture representation techniques are mature enough to serve and fulfill the requirements of 3-D extraction, transmission and display sides in a 3DTV scenario, X3D, 3DTV.

3.7.2. Completed Theses on 3D Scene Representation Technologies

The following theses have been completed within NoE between 2006-2007:

1) Onur Önder, Combined Filtering and Keyframe Reduction for Motion Capture Data, M.S. Thesis, Bilkent University, Department of Computer Engineering, July 2007.

2) Evren İmre, Prioritized 3D scene reconstruction and rate-distortion efficient representation for video sequences, PhD Thesis, METU, Department of Electrical-Electronics Engineering, September 2007

3) Cevahir Çığla, Dense depth map estimation for object segmentation in multi-view video, MS Thesis, METU, Department of Electrical-Electronics Engineering, July 2007

4) Yusuf Sahillioğlu, Fusion of Shape from Stereo, Silhouette and Optical Triangulation Techniques for 3D Reconstruction, MS Thesis, Koç University, Department of Computer Engineering, July 2006.

5) M. Franc, Methods for Polygonal Mesh Simplification, University of West Bohemia, PhD Thesis, 2007

6) Lukas Ahrenberg, Methods for transform, analysis and rendering of complete light representations, PhD Thesis, Max-Planck-Institut fur Informatik, July 2007.

4. Conclusions and Future Directions

This technical report presents the outputs of 25 research activities, whilst 8 of these papers are result of joint efforts. It should be noted that WP8 Technical Reports #1 and #2 had 19 and 20 research outputs, respectively, both with 6 joint publications. In Year-III, research efforts continue in the high-priority joint research areas, based on updated DoW of 3DTV NoE .

For point-based 3D scene representation, at the end of Year-III, the research efforts have been focusing on multi-view-plus-depth representation, which will be standardized by ISO MPEG. In the last year of NoE, both METU and HHI are expected to focus on this new representation with new extraction techniques and virtual view generation algorithms.

KOC and PLZEN continue to work on time-consistent dynamic mesh representations. KOC has more focus on the initial extraction of such meshes and tracking them in time. As compared to the few works that aim to construct time-consistent mesh representations of real scenes from multi-view video, the proposed method by KOC is much faster, it can track the surface geometry over longer periods, it is able to generate topologically correct and smooth mesh representations, and it can also handle objects with non-rigid motion. PLZEN has presented some initial research output in dynamic time-varying mesh representations, where mesh connectivity and parameters might also vary in time. METU and BILKENT have proposed a novel mesh reconstruction algorithm that considers the compression efficiency of every vertex included into this mesh. In Year-IV, the efforts will continue jointly to apply this algorithm for more general cases.

For volumetric representations, during Year-III, research effort was extended in applying high reconstruction accuracy technology in large spatial volumes. In the future, this effort will be further extended for temporally long video recordings, in order to facilitate the 3D recording of dynamic events in high quality and resolution by ITI. Future plans between ITI and METU concern the 3D acquisition of events and be able to facilitate the recording of cultural and sport events.

Modeling, animation, and rendering of human figures is a very active research area in computer graphics and computer vision since almost every 3D scene contains humans. These are very challenging research topics since different degrees of realism are needed in different applications, such as entertainment (game industry, film industry), medical applications, and 3DTV applications. One promising approach for realistic animation of human models is using motion capture-based approaches since very realistic animations can be generated using these techniques. BILKENT makes research on *filtering* and *keyframe reduction* of motion capture data. Filtering is necessary to eliminate any jitter introduced by a motion capture system. Key-frame reduction allows animators to easily edit motion data by representing animation curves with a significantly smaller number of key frames. Since the curve simplification-based keyframe reduction technique produce promising results, BILKENT plans to investigate these techniques further.

Human face and body specific research has shifted its focus to multi-modal analysis and synthesis in Year III. KOC and MOMENTUM has jointly developed more realistic lip

animation that considers both visual and audio analysis together. In another multi-modal approach, a dancing object is analyzed together with the associated music in order to be able to synthesize a dancing avatar for any unknown musical content by again KOC and MOMENTUM. BILKENT has produced some outputs for human motion modeling, especially via inverse kinematics.

Inverse kinematics-based approaches still provide one of the best alternatives for realistic motion control of human figures in 3D scenes. We develop high level and low-level inverse kinematics based motion control techniques for realistic humans walking behavior. BILKENT plans to extend this research for other complex human behaviors.

Within the context of object-specific representations for 3DTV, physically-based techniques generate realistic results for cloth simulation since they simulate the physics of materials. Since the procedural techniques are simple and practical, they provide an alternative for the visualization of knitwear and woven cloth for 3DTV applications. In the future, BILKENT will further elaborate on physically-based cloth simulation and procedural visualization of cloth materials.

For Pseudo-3D representation only solo work is accomplished by UIL. The work mainly focuses on the development of a GPU assisted image processing framework for the generation of 3D video objects which could also be useful for other partners within the consortium.

5. Annex