D26.3 Technical Report # 3 on 3D Time-varying Scene Capture Technologies

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

Deliverable Nature: R
Number: D26.3
Contractual Date of Delivery: M41
Actual Date of Delivery: M43
Report Date: 5 March 2008
Task: WP7
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: Bodo Rosenhahn (rosenhahn@mpi-inf.mpg.de)

Editor: Bodo Rosenhahn (rosenhahn@mpi-inf.mpg.de)

5 March 2008

3D Time-varying Scene Capture Technologies
TC1 WP7 Technical Report 3

EDITOR

Bodo Rosenhahn (**MPG**)

Contributing Partners to Technical Report:

Bilkent University (**Bilkent**)
Bremer Instiut fuer angewandte Strahltechnik GmbH (**BIAS**)
Bulgarian Academy of Sciences (**CLOSPI-BAS**)
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 fuer Informatik (**MPG**)
University of Tuebingen (**UNI-TUEBINGEN**)
University of Hannover (**UHANN**)
Technische Universitaet Berlin (**TUB**)

REVIEWERS

Joern Ostermann (**UHANN**)
Elena Stoykova (**CLOSPI-BAS**)
Lukas Ahrenberg (**MPG**)
Atanas Gotchev (**TUT**)

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

TABLE OF CONTENTS

Executive Summary

This technical report summarizes the scientific results achieved by partners cooperating in work package 7 of the 3DTV project, which performs "Joint research on 3D time-varying scene capture technology". It is the third technical report published by the WP7 partners and summarizes the results obtained during the project period between month 30 and month 42.

Sixteen partners cooperate in WP7 on the technological and algorithmic foundations of an important sub-part in the 3D television production process, namely 3D scene recording and scene reconstruction. The technological concepts developed in WP7 will enable efficient and accurate reconstruction of 3D TV scene representations from captured real-world footage. The faithful reconstruction of such dynamic scene models is a prerequisite for high-quality 3D content display, and therefore the methods developed in this work package also lay the algorithmic foundations for many other work packages in the 3DTV NoE.
The partners in WP7 have defined seven high-priority subtasks in order to systematically investigate the different technological alternatives, and in order to lay the algorithmic foundations for very basic problems that are of relevance to each possible alternative. These sub-areas are *Multicamera Techniques, Single Camera Techniques, Human Face and Body Techniques, Holographic Camera Techniques, Pattern Projection Techniques, Motion Analysis and Tracking Algorithms, and Object Segmentation Approaches.* In the reported period, the project partners have obtained significant new scientific results in many of the primary research areas.
Overall, **57** papers were published in top tier journals, books and conference proceedings. This is a notable increase in the number of published papers in comparison to the previous reporting period. The high scientific output also demonstrates the high research dynamics in the work package, as well as the good progress of individual research projects.

The main part of this report is a collection of abstracts, ordered into subsections corresponding to each task. The individual subtasks are not to be seen as strictly separate fields of research since many of the attacked problems lie on the boundary between multiple subtasks. For instance, there are contributions to the Multicamera task that also deals with tracking of the human body. Thus the subtasks in this report should be considered as guidelines that the WP7 partners have given themselves in order to systematically define their joint research direction.

Task	Num. Publications TR#3	TR#2	TR#1	Joint by 3DTV partners TR#3	TR#2	TR#1
Multicamera	29	10	6	3	1	1
Single Camera	8	4	2	2	3	1
Human Face and Body	6	6	7	1	2	1
Holographic Camera Techniques	5	3	2		0	0
Pattern Projection	4	6	2	1	0	0
Motion Analysis and Tracking	2	4	2		1	1
Registration	shifted to Multi/Single Camera	2	2		0	0
Object representation and Segmentation	3	1	2	0	0	0
	Total	Total	Total	Total	Total	Total
Total, WP7	**57**	36	25	7	7	4

Total number of publications as well as joint publications (by 3DTV partners) arranged per task. The numbers for this report are given in the TR#3 columns.

As can be seen in the above table, the number of publications is again higher for this period than the previous two ones. The number of joint publications has remained the same. The general increase of the number of publications can very likely be amounted to the fact that we are now further into the NoE lifespan, and several projects (and PhD-theses) are going to be finalized within the next months leading to more activity in publications. The number of joint publications has not increased. This is probably again due to the fact, that many research partners are finalizing their theses and therefore spend less amount of time in exchange projects to other groups. The distribution of the number of publications per task is about the same as for the previous period. The somewhat uneven distribution of publications is simply due to the different number of researchers and partners working in these areas during the period. It is also worth mentioning that the Registration task was moved as subtask in Single- and Multi-camera techniques.

Conferences	PhD-theses + Reports	Journals + Book chapters	Submitted articles
30	12	10	5

Numbers for types of publications

The above table separates the 57 publications in their different types. As can be seen, most articles have been published in conferences and journals. The conference articles include **high-impact conferences** and **journals** in the fields of Computer graphics and Computer Vision, such as Eurographics, CVPR, ICIP or PAMI.

Contributors	**Number of publications**
MPI	15
Tuebingen	9
TUT	6
Koc	4
TU Berlin	5
Metu	3
ITI-Certh	5
UHANN	2
Momentum	2
BIAS	5
CLOSPI-BAS	4
UIL	2
Bilkent	1

Contributors to WP-7

The above table shows the involved research groups with their number of publications within WP-7.

Representatives for the WP7 partners have met two times during the period. Both occasions were in connection to TC1 meetings. After a meeting in September 2006, WP7 first met in Bodrum, Turkey, May 2007 (GRM-II) and then again in Aberdeen, UK, November 2007 (EU review meeting). Both meetings were successful from a WP perspective. There were discussions between partners, and a good chance to meet up and discuss the status of the work package and related issues.

During the period covered by this report, WP7 saw many collaborations of contributing partners. Seventeen research visits have been reported as fully or partially treating WP7 related research. Some amount of the joint work has also led to joint publications; seven of the reports attached to this document are direct results of the collaborations within the NoE, and several others have been influenced by discussions during meetings and research visits. Below is a table that summarizes some further data of WP7:

Number of partners	17
Number of research visits related to WP7	17
Number of publications	57
Number of joint publications	7
Software uploads related to WP7	14
Data uploads by WP7 partners	10
PhD Theses	8
MSc Theses	4

WP7 at a glance.

During the last year six PhD theses have been finished and two PhD Theses are expected to be finished within the next two years. Four MSc theses have been accomplished as well.

The report thus is a proof for successful cooperation. The variety of technological achievements summarized in the following lay the algorithmic foundations to bring us one step further towards general three-dimensional recording, capture and processing technology. Some outstanding results are given in Section 3.1.5 (3DTV – Panoramic 3D Model Acquisition and its 3D Visualization on the Interactive FogScreen), Section 3.1.15 (CVPR-paper on Marker-less Deformable Mesh Tracking for Human Shape and Motion Capture) and Section 3.1.17 (Reconstructing Human Shape, Motion and Appearance from Multi-view Video). For the final months, we recommend to accelerate the efforts on high-profile demonstrations.

1 Introduction

3D Television augments the traditional TV technology by showing the viewer not only sequences of 2D images but streams of three-dimensional scene representations. To the viewer at home this will mean a completely new media experience. He will perceive the displayed events in a more immersive way, and he may even get the chance to choose his own viewpoint to watch the displayed events. In the future, three-dimensional movies will become a standard and provide enhanced interactivity options, e.g. by allowing the user to navigate through the scenes.

The production pipeline for 2D television has developed into a mature and well-understood process over many years. Scenes are recorded with cameras from single view-points, captured image streams are post-processed, transferred to receivers, and displayed on planar screens. In contrast, the production process for 3D television requires a fundamental rethinking of the underlying technology. Scenes have to be recorded with multiple imaging devices that may be augmented with additional sensor technology to capture the three-dimensional nature of real scenes.

In addition, the data format used in 3D television is a lot more complex. Rather than normal video streams, time-varying computational models of the recorded scenes are required that comprise of descriptions of the scenes' shape, motion, and multi-view appearance. The reconstruction of these models from the multi-view sensor data is one of the major challenges that we face today. Finally, the captured scene descriptions have to be shown to the viewer in three-dimensions which requires completely new display technology.

Today, 3D Television is still in its early days. Many technological and computational problems in scene acquisition, scene reconstruction, and scene display are either unsolved or bring today's technology to its limits. Furthermore, the problems to be solved require expertise from many different areas in science and engineering, ranging from computer science, over physics, to electrical engineering. The European Union's network of Excellence on 3D television brings together a multi-disciplinary group of 21 leading European researcher centers that jointly works on the solution to the myriad of challenging technological problems. The project partners have split their research effort into seven major research work packages, each of which is focusing on one of the most intriguing technological problems that we face on the way towards 3D television.

The last twelve months has been an active and exciting period for WP7. The research projects carried out by partners, both joint and individual are continuing, and several more are being discussed. A direct result of the successful work is the texts and scientific papers presented in this report, five of them jointly authored by researchers from WP7 partners. During the reported period, the successful research in WP7 lead to **57** publications in major international conferences, journals as well as edited books and technical reports. There are currently seven tasks identified as high priority research areas within Work Package 7. These are:

Multicamera, Single Camera, Human Face and Body, Holographic Camera Techniques, Pattern Projection, Motion Analysis and Tracking and *Object Segmentation*. Most research by the partners in WP 7 is generally focused towards these tasks.

However, naturally many projects will of course touch several areas. For instance, 3D scene capture technologies, such as Single- and Multi-camera methods and applications, will naturally use techniques from the Tracking task. Thus, even though this report organizes the abstracts of the contributed texts into sub-sections corresponding to the tasks, many of the contributions contain work in several WP7 areas.

This report is arranged as follows: In the next Section, an analysis of the presented material is given. In Section 3, abstracts of selected papers and technical reports produced by WP7 partners during the project period are presented. These are organized into subsections corresponding to the identified research tasks. Each Section contains a brief conclusion summarizing the research tasks. Section 5 contains a bibliography of references cited in the abstracts. Finally, attached to the document are reprints of the submitted papers and reports.

2 Analysis of the Results Reported in the Publications

In this section, we briefly analyze the scientific results presented in this report and describe their scientific impact. Given the variety of algorithmic challenges that one faces in time-varying scene capture, the individual tasks of WP7 naturally address many different topics. In a sense, this also mirrors the very many technologies being developed for 3D recording, reconstruction and vision. Analyzing the work, we can roughly say that the research performed under the name *Scene Capture* may be thought of as recording methods together with vision and applications. This, by itself, is already a notable contribution to the scientific community, because the work of WP7 is an excellent example for a research effort in which joint expertise from different disciplines, e.g. engineering, computer vision and computer graphics, enables novel applications.

In the reported period, the project partners developed a variety of important methods that advance the field and pave the trail for mature 3D video and 3D TV capturing and reconstruction technology.

On the recording side, novel ways for camera synchronization were developed which is an important prerequisite for high-quality input video data. WP7 also proposes a new hardware architecture for holographic recording. Finally, there are several contributions dealing with Pattern Projection techniques which push that area forward by analyzing and evaluating fringe generation, phase retrieval as well as color recording.

There has also been a lot of work in vision and applications, dealing with very versatile topics such as 3D reconstruction, feature detection and tracking of body models. In the 3D reconstruction field several papers in the Single Camera task deal with advances in 2D to 3D conversion of video data, being an important technology to convert standard TV footage into 3D TV footage. However, there is also work in complete 3D reconstruction from multiple cameras, for example using Weighted Minimal Hypersurfaces. We also see a trend where not only a three dimensional model is reconstructed, but also lighting and material properties of the scene. This may be an important future step as this information could be used to render scenes on 3D-TV display systems under user-controlled environment conditions. Furthermore, compositing of 3D Video and 3DTV objects is facilitated.

Instead of a full reconstruction, model estimation and detection can be used for known objects in a scene. This report presents several approaches to tracking, motion estimation and detection. Many of the model-based results are focusing on the human face and body as they are natural candidates in many video sequences. Work dealing with body model animation and 3D face and speech recognition are representative for this area, but there is also work on tracking humans in multiple view video, as well as traffic monitoring which could prove to be important for large area scenery. In addition to the more application oriented contributions there is also work in basic vision and image registration. One paper is especially oriented towards image registration on mobile devices for instance. Finally, there have been some first steps taken towards dealing with High Dynamic Range recording, which probably will be a important part of future display systems enabling the reproduction of such higher fidelity video streams.

In the following subsections, the contributing partners analyze their results obtained in the different research areas and illustrate their impact on the scientific community.

2.1 Multicamera

Future 3D Television critically relies on mechanisms for automatically acquiring and visualizing high quality models of humans and 3D content of indoor and outdoor scenes. The envisioned goal is that a photo-realistic 3D real-time rendering from the actual and potentially arbitrary viewpoint of the beholder who is watching 3DTV becomes possible. Such scenes include movie sets in studios, e.g., for talk shows, TV series and blockbuster movies, but also outdoor scenes, e.g., buildings in a neighborhood for a car chase or cultural heritage sites for a documentary. To achieve the goal of robustly acquiring 3D content, many techniques were developed within the Multicamera task of WP7.

In a first line of research, researchers from WSI/GRIS Tübingen developed the Wägele, a mobile platform for acquisition of 3D models of indoor and outdoor environments which can be used to provide the 3D background models where potential 3D actors can be embedded, as well as for surveillance tasks. A system for semi-immersive visualization was also developed giving an impression on how a future 3D television system could be. Recent improvements are being focused on acquiring 3D models in complicated environments where small dimensions and simple usability are essential. To accomplish this task, a multi-sensor scene acquisition device (SAD) which merges information from different devices is used. Another similar approach is developed by Koc University, where a volumetric fusion technique is used for surface reconstruction. During the process of developing 3D scene acquisition devices the clear necessity of high-dynamic range (HDR) modules have been recognized. It is determined by the highly-varying illumination conditions of indoor and outdoor scenes. Responding to this necessity, the team from Tampere University of Technology (TUT) developed a new method for color HDR working in Luminance-Chrominance space. It adds an extended functionality to the SAD developed by WSI/GRIS in building improved 3D models.

Another line of research focuses on acquiring human models and their motions from real actors performing. A marker-based system is developed by Koc University which is used for analysis of dancing figures. Marker-less motion capture systems are also presented. By using as underlying model a static scan of the real subject, a number of motion capture methods were developed at the Max-Planck Institut fuer Informatik. The new motion capture methods are able to capture not only the motion but also the surface deformations of performing actors, which is a great improvement over traditional motion capture techniques. In this context a new animation paradigm is also developed simplifying the process of motion retargeting and skinning in traditional character animation.

Other research presented in this report deals with the problem on how to visualize and edit captured data. Methods are presented to convert mesh animations into traditional skeleton-based animations or to convert them to a new artistic format: an animation collage. New visualization systems for 3D Video and editing techniques for video post-processing are also presented.

In the past months two Ph.D. theses and three Master theses were successfully concluded and two more Ph.D. theses are expected to be concluded soon. These theses deal with all aspects of the 3D TV pipeline: acquisition, editing and visualization. The contributions are from MPG, TUE, TUT, Koc and Yoghurt.

2.2 Single Camera

In Technical Report #3, single camera scene extraction research is completely devoted to shape-from-motion approaches. Among 7 novel contributions, there are 2 joint research outputs with contributions from different partners. In addition, there is an MSc thesis completed during the reported period. In the contributions, there are a number of main stream research directions in single camera research, one of which aims to render multi-views of a scene without explicitly determining the dense 3D structure, but rather utilizing image-based rendering technology. These views could be utilized as inputs not only for the current auto-stereoscopic displays, but also for the next-generation high resolution 3D displays by exploiting super-resolution techniques. In a different high priority research direction, conversion of 2D broadcast video into 3D is pursued that could yield 3D scene structure explicitly, capable of being utilized in any kind of 3DTV display, even for the holographic TVs. A number of fundamental problems, such as self-calibration, moving object segmentation and dense depth estimation, are jointly approached to result with a promising an end-to-end algorithm, capable of 2D-to-3D conversion. Apart from this system, a novel outlier rejection and moving object segmentation is also proposed for obtaining robust estimates of the epipolar geometry between consecutive frames. Finally, 3D sparse scene extraction problem is reformulated, in order to consider the rate-distortion efficiency of the resulting scene representation during single camera extraction stage. Hence, 3D scene extraction is achieved in a hierarchical manner by gradually increasing the number sparse 3D reconstructed points, while considering the amount of bits to encode this information, as well as the quality of the resulting representation. An MSc thesis, dealing with efficient non-uniform to uniform resampling complements the report on single camera techniques. The involved partners are from TUB, TUT and METU.

2.3 Human Face and Body

Human beings are well trained in detecting and tracking human bodies and faces. Therefore, the cognitive capabilities of human observers to detect un-natural animated faces and bodies very quickly, requires to deal with Human Faces and bodies as a special hi-priority research area. The Human Face and Body specific techniques are presented in section 3.3. In recent years the techniques related to facial analysis and synthesis, human body motion analysis and animating 3D avatars have received increasing interest. They build a basis for new applications that are important for 3DTV.

One part of the research devotes to the algorithms related to human face and concentrates on the following topics: 3D face and facial expression recognition, detection of face and facial features in images, estimation and analysis of facial animation parameter patterns. Research results are presented in sections 3.3.1 to 3.3.4.

The part related to human body gives an overview about advances in tracking and recognition of human motion and presents a framework for joint music and video analysis for automatic human body motion synthesis. A recent joint work is given in section 3.3.5. An MSc thesis on the topic of face and facial feature detection has been jointly supervised in TUT and UHANN and is expected to be concluded in April 2008. It is given in Section 3.3.6. Partners from ITI-CERTH, KOC, Momentum, TUT and UHANN contributed to this field of research.

2.4 Holographic Camera Techniques

Digital holography as compared to traditional methods of holography is seen as the way forward in realizing practical, mass media 3D displays. The covered research advances in this NoE range from different systems, sensor calibration up to the computation of out-of-plane displacements (based on the Mellin transform). Many works are very basic and their impact will become more important during the next years. However, there are still many technological advances needed. The Holographic Camera Techniques group has been working towards this end. Future problems are addressed in the European joint project Real3D, which should start February 2008. In the consortium of this project two partners of the NoE 3DTV are participating: Bilkent and BIAS.

2.5 Pattern Projection

The long-time objective of the 3D fringe projection profilometry is to provide a real-time mode of operation which can be done by phase demodulation of a single pattern (single-shot acquisition) or by recording multiple patterns at high acquisition speed that are processed by the well-known algorithms. During the last two decades, various methods for single frame fringe pattern demodulation have been explored. The most straightforward Fourier transform phase demodulation suffers from limitation on height variation of the investigated object. The wavelet based processing overcomes this limitation but is more computationally expensive. Over the years spatial modifications of the phase-stepping approach have been developed based on assumption of slowly varying phase which hampers analysis of fringe patterns with higher frequency content. Recently, alternative spatial analysis methods for phase retrieval have been reported as regularized phase tracking by solving a set of linear equations which involves time-consuming iterative procedure; fitting-error modified spatial fringe modulation, phase demodulation based on fringe skeletonizing when an extreme map is introduced by locating the fringes minima and maxima, phase-stepping recovery of objects by numerical generation of multiple frames from a single recorded frame. The drawback of many of the spatial analysis methods is inevitable averaging over several pixels in the neighborhood of the point of interest. Single-shot measurement are also described in the literature as simultaneous projection of three colour patterns (red, green and blue) on the object at different angles and Fourier analysis of the deformed image recorded by a single CCD camera, a phase-stepping method for measuring the 3-D surface profile of a moving object by projection of a sinusoidal grating pattern and continuous intensity acquisition by three phase-shifted linear array sensors positioned along the projected stripes or a high-resolution 3D measurement of absolute coordinates using three phase-shifted fringe patterns coded with three primary colors and recorded at data acquisition speed of 90 fps.

In the technical report D26.2 a technical solution of a single-shot pattern projection profilometric system was described with simultaneous projection and recording of four phase-shifted fringe patterns which are generated at four different wavelengths. The system includes a pattern projection module with four projection elements irradiated by four near-infrared diode lasers and a registration module with four CCD cameras. As candidates for a projection element, a sinusoidal phase grating and a holographic optical element which reconstructs two point sources have been discussed and the first test experiments have been made. Technical simplicity of a set-up, easy manufacturing and reproducibility of the desired modulation and spacing, high efficiency, minimization of the phase-shifting error, and independence of the spatial period of the diffraction pattern on the wavelength advocated strongly for the choice of

the phase grating. This solution determined the aim of the work performed during the reported in D26.3 period which was to study the diffractive properties of a sinusoidal phase grating for incorporation as a pattern projection element in a multi-source and multi-camera profilometric system. Two challenges should be overcome for successful operation of such a system that are connected to inherent limitations of the phase-shifting algorithm - the requirements for a sinusoidal fringe profile and for equal background and contrast of fringes in the recorded fringe patterns. Partners from CLOSPI-BAS and ITI-CERTH contributed in this section.

2.6 Motion Analysis and Tracking

Many applications in the field of analysis-by-synthesis require algorithms for tracking and analysis of moving objects in image streams. Section 3.6 presents two papers in this field. One deals with traffic monitoring based on fast motion detection in regions of interest and the other one exploits homographies and multi-hypothesis tracking in traffic monitoring. Both papers are from partners at ITI-CERTH.

2.7 Object-Based Segmentation

In this section two region-based methods for object tracking using active contours in a dynamic programming framework are presented. Furthermore, a preprocessing algorithm for traditional chroma keying systems using a simple background illumination correction based approach for improving matting problems with uneven or poor lighting in the background using the computational power of GPU computing is presented. In the final paper a modular framework for the abovementioned keying process is described. Partners from UIL and Momentum contributed to this section.

3 Abstracts of Papers and Technical Reports

In this section, paper abstracts overview the work of the partners involved in WP7 and the results obtained within the reporting period. Each of the related sub-sections in the following includes an introduction about the respective research topic, conclusions, and future plans. The full papers and technical reports are collected in the related annex in the end of this document.

3.1 Multicamera

This section reports the research progress on projects related to acquisition, editing and visualization of data in multicamera systems. In Sect. 3.1.1, a derivation of the Euler-Lagrange equation in arbitrary dimensional space is presented which opens the possibility to solve problems involving minimal hypersurfaces in dimension higher than three. Two applications of this new framework are presented: reconstruction of temporally coherent

geometry from multiple video streams and volumetric reconstruction of refractive and transparent natural phenomena.

Efficient and comfortable acquisition of large 3D scenes is an important topic for many current and future applications like cultural heritage, web applications and 3DTV. A platform for collecting the data needed to build such models, the Wägele, is presented in Sect. 3.1.2. It is equipped with three laser range scanners, a panoramic camera, 3D-attitude and heading reference system and GPS. Further improvements include an omnidirectional stereo vision approach based on graph cut techniques (Sect. 3.1.3) which also enables the application of the system for surveillance tasks (Sect. 3.1.4). A semi-immersive 3D visualization system is also presented giving an impression on how a future 3D Television could look like, Sect. 3.1.5.

Although many approaches to 3D scene acquisition have been presented in the last years, none of them are able to acquire 3D models in complicated environments where small dimensions and simple usability are essential. In order to solve this problem, the Scene Acquisition Device (SAD) is presented (Sect. 3.1.6). The proposed system consists of three sensors: a time-of-flight range sensor combined with a standard color camera and a miniature inertia sensor. A method for enhancing the 3D data calculated from the range output of a 3D time-of-flight camera is also presented (Sect. 3.1.7). Joining all three sensors, a working scene acquisition system which allows for fast and simple acquisition of arbitrarily large 3D environments is developed (Sect. 3.1.8) and a method for self localization within such scenes presented in Sect. 3.1.9.

Inferring three-dimensional shapes of objects is important for 3DTV, virtual reality, digital preservation of cultural heritage, and computer graphics in general. In Sect. 3.1.10 a hybrid surface reconstruction scheme that combines shape from silhouette and shape from optical triangulation is presented eliminating the shortcomings of the combined methods and enabling the generation of high quality, robust, and watertight 3D models.

Sect. 3.1.11 describes an automated marker-based multicamera motion capture system to capture the body movements of a dancer from multi-view video recordings of a dance performance. Correlation between the body movement patterns and the musical audio signal are investigated towards the goal of synthesizing an audio-driven dancing avatar.

Sect. 3.1.12, 3.1.13 and 3.1.14 present two different variations of an efficient approach to turn laser-scanned human geometry into a realistically moving virtual avatar. Instead of relying on the classical skeleton-based animation pipeline, the methods use a mesh-based Laplacian editing scheme to drive the motion of the scanned model. The frameworks elegantly solve the motion retargeting problem and produces realistic non-rigid surface deformation with minimal user interaction. Realistic animations can easily be generated from a variety of input motion descriptions, which are exemplified by applying the methods to both marker-free and marker-based motion capture data.

New algorithms to jointly capture the motion and the dynamic shape of humans from multiple video streams without using optical markers are presented in Sect. 3.1.15 and 3.1.16. Instead of relying on kinematic skeletons, as traditional motion capture methods, the approaches use a deformable high-quality mesh of a human as scene representation. As opposed to many related methods, the algorithms can track people wearing wide apparel, it can straightforwardly be applied to any type of subject, e.g. animals, and it preserves the connectivity of the mesh over time. An overview of other two previous model-based approaches to capture the motion, as well as the dynamic geometry of moving humans is

presented in Sect. 3.1.17. By applying a fast dynamic multi-view texturing method to the captured time-varying geometry, we are able to render convincing free-viewpoint videos of human actors.

Recently, it has become increasingly popular to represent animations not by means of a classical skeleton-based model, but in the form of deforming mesh sequences. Unfortunately, the resulting scene representation is less compact than skeletal ones and there is not yet a rich toolbox available which enables easy post-processing and modification of mesh animations. To bridge this gap, a new method is proposed that automatically extracts a plausible kinematic skeleton, skeletal motion parameters, as well as surface skinning weights from arbitrary mesh animations (Sect. 3.1.20). Using a similar idea, an automatic method to transform mesh animations into animation collages, i.e. moving assemblies of shape primitives from a database given by an artist, is presented in Sect. 3.1.19.

A novel approach to real-time shape editing that produces physically plausible deformations using an efficient and easy-to-implement volumetric approach is presented in Sect. 3.1.18. The technique is well suited for interactive shape manipulation and also provides an elegant way to animate models with captured motion data.

Visual (digital) representation of natural scenes has reached a point where spatial resolution is no longer an issue and greater realism can be achieved by adding the third dimension and utilizing a more and more realistic gamut of light and color. The latter is achieved through a group of techniques commonly known as HDR imaging. In Sect. 3.1.21, the problem of color HDR is addressed and an effective solution working entirely in luminance-chrominance space is developed. The approach is general and for any number of cameras. It has been specifically included in the Muticamera section as it is used to enrich the 3D scene acquisition techniques described in Sect.3.1.2 – 3.1.9 with HDR capabilities. Such an HDR-equipped 3D scene acquisition system is described in Sect. 3.1.22. Sect. 3.1.23 – 3.1.29 present Ph.D. and Master theses successfully concluded (or to be concluded in 2008) by the members of the Multicamera working group. The topics deal with aspects relevant to acquisition, editing and visualization of 3D TV data. In Sect. 3.1.23, the Luminance-Chrominance approach to HDR is presented. Sect. 3.1.24 presents a depth-from-stereo approach based on combined corner and SIFT feature selection and histogram based segmentation for improved depth interpolation. In Sect. 3.1.25 a real-time hierarchical stereo matching method based on graphics hardware is introduced. Sect. 3.1.26 presents methods for reconstructing and rendering time-varying natural phenomena. Sect. 3.1.27 presents new editing techniques for video post-processing. The two remaining theses, dealing with GPU data structures for video processing and vision-based graphics (Sect. 3.1.28) and a system to capture and edit moving scanned subjects (3.1.29) are expected to be concluded in early 2008.

3.1.1 Weighted Minimal Hypersurface Reconstruction

Authors: Bastian Goldluecke, Ivo Ihrke, Christian Linz, Marcus Magnor

Institutions: MPI Informatik, TU Braunschweig

Publication: Transactions on Pattern Analysis and Machine Intelligence (TPAMI)

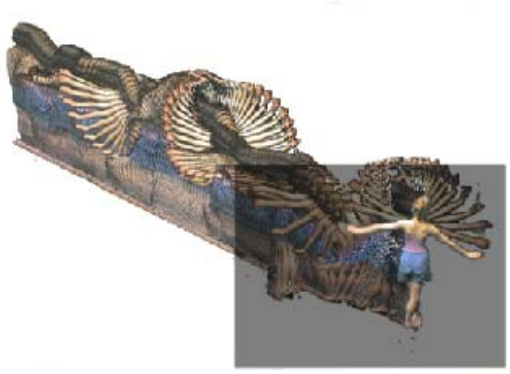


Fig. 1. *A surface evolving over time defines a hypersurface.*

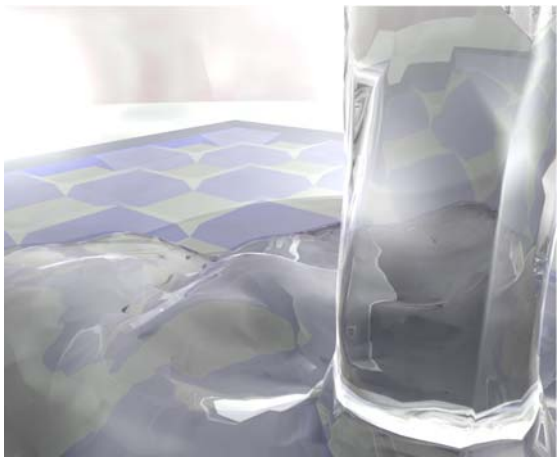


Fig. 2. *Close-up on a water surface reconstructed using our approach.*

Many problems in computer vision can be formulated as a minimization problem for an energy functional. If this functional is given as an integral of a scalar-valued weight function over an unknown hypersurface, then the sought-after minimal surface can be determined as a solution of the functional's Euler-Lagrange equation. This paper deals with a general class of weight functions that may depend on surface point coordinates as well as surface orientation. We derive the Euler-Lagrange equation in arbitrary dimensional space without the need for any surface parameterization, generalizing existing proofs. Our work opens up the possibility to solve problems involving minimal hypersurfaces in dimension higher than three, which were previously impossible to solve in practice. We also introduce two applications of our new framework: we show how to reconstruct temporally coherent geometry from multiple video streams (Fig. 1), and we use the same framework for the volumetric reconstruction of refractive and transparent natural phenomena, here bodies of flowing water (Fig. 2).

3.1.2 The Wägele: A Mobile Platform for Acquisition of 3D Models of Indoor and Outdoor Environments

Authors: Peter Biber, Sven Fleck and Wolfgang Straßer

Institutions: WSI/GRIS, University of Tübingen, Tübingen, Germany

Publication: 9th Tübingen Perception Conference (TWK 2006), 2006

Efficient and comfortable acquisition of large 3D scenes is an important topic for many current and future applications like cultural heritage, web applications and 3DTV and therefore it is a hot research topic. We have built a platform for collecting the data needed to build such models: The Wägele. It is equipped with three laser range scanners, a panoramic camera, 3D-attitude and heading reference system and GPS as shown in Fig. 3.

One of the laser scanners is mounted to record range values horizontally. This data is used to build a two dimensional map and to localize the mobile platform with respect to this map. Our techniques to tackle this problem are borrowed from robotics and in essence we have to solve the simultaneous localization and mapping (SLAM) problem. The other laser scanners are mounted perpendicularly; their data yields the geometric information for the 3D model in form of slices of the environment. Additionally, panoramic images are taken regularly.

After a recording session the collected data is assembled to create a consistent 3D model in an offline processing step. First a 2D map of the scene is built and all scans of the localization scanner are matched to this map. After this step the position and orientation of the Wägele is known at each time step. The panoramic camera has been calibrated, and as the relative positions of all sensors are known, geometry creation and texture mapping is easy with the known positions. The result of this whole process is an unstructured point cloud with color attributes per point. For outdoor scenes, we georeference these models so that they can be embedded in coarse resolution models built from aerial images and digital elevation models.

Besides 3D model-acquisition, the platform also serves as a multimodal data acquisition platform. Due to the fact that the platform is localized continuously it can provide ground truth data for computer vision and other signal processing algorithms like stereo or visual localization and tracking algorithms.

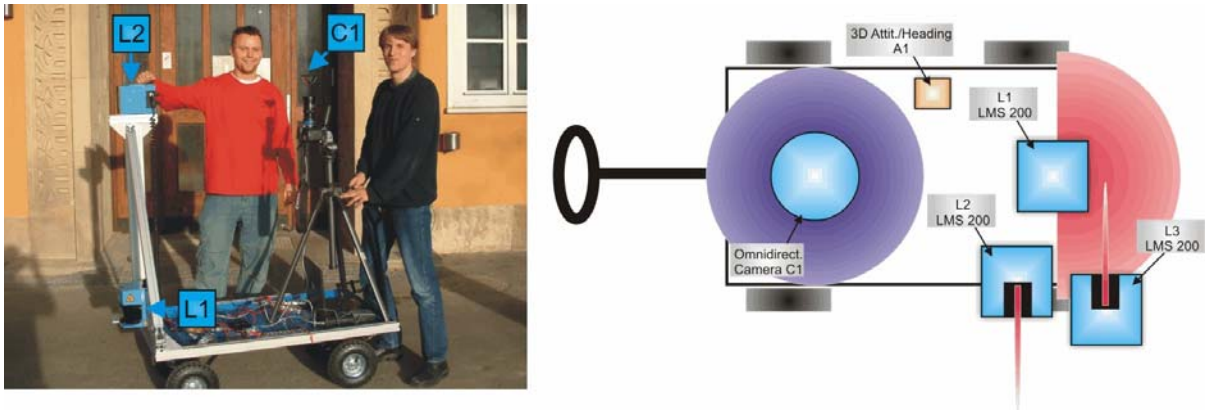


Fig. 3. *Our 3D model acquisition platform – The Wägele.*

3.1.3 Omnidirectional Stereo based 3D Model Acquisition on the Wägele

Authors: Sven Fleck, Florian Busch, Peter Biber and Wolfgang Straßer

Institutions: WSI/GRIS, University of Tübingen, Tübingen, Germany

Publication: 9th Tübingen Perception Conference (TWK 2006), 2006

We present an omnidirectional stereo vision approach based on graph cut techniques in conjunction with a mobile sensor platform ("The Wägele") where it is employed. The platform comprises an 8 Mpixel omnidirectional camera in conjunction with a laser scanner, no odometry is necessary. 3D models are acquired just by moving the platform around and recording omnidirectional images in conjunction with their poses in regular intervals. The result of this whole process is a colored point cloud. The poses and thus the according external camera parameters are determined by probabilistic matching of laser scans. However, the stereo vision algorithm does not rely on this laser range data directly, any other method of localization would do. The stereo pipeline computes dense depth maps using pairs of panoramic images taken from different positions. First, for each pixel in the first image the epipolar curve in the second image is created and a difference value for each disparity on this epipolar curve is computed. Afterwards, a graph cut algorithm as the core component is applied. The graph cut approach has become quite attractive for various vision problems targeting high quality. Our algorithm follows the work of Kolmogorov & Zabih and is extended to omnidirectional imaging. The key of graph cut is formulating the correspondence problem as an energy minimization problem. The minimization is done iteratively by transforming this problem into several minimum cut problems based on ®-expansion moves until convergence is reached. Our energy function comprises an SSD-based matching cost, an occlusion term and a smoothness term. Afterwards, several post processing steps are applied: sub-disparity refinement, epipoles removal, floor correction and filling of unknown values. Results of both indoor and outdoor scenes are presented. Besides targeting on an appealing visual quality (from a perceptual perspective) it is also important to objectively give a quantitative measure of the quality of our stereo method. This requires two components: acquiring ground truth and the design of a quality measure. Two additional calibrated laser range scanners on the Wägele deliver a well approximation of ground truth of the geometry. Texturing it leads also to point clouds that can be visually compared to our stereo results by the beholders. In terms of quantitative comparison, first results of our 3D based measure are presented.

3.1.4 3D Modeling of Indoor Environments for a Robotic Security Guard

Authors: P. Biber[1], S. Fleck[1], T. Duckett[2], and M. Wand[1]

Institutions: [1] WSI/GRIS, University of Tübingen, Tübingen, Germany and [2]AASS Research Center, Department of Technology, Örebro University, SE-70182 Örebro, Sweden

Publication: Book chapter in 3D Imaging for Safety and Security. By Koschan, Andreas Pollefeys, Marc Abidi, Mongi, ISBN: 1402061811 ISBN13: 9781402061813

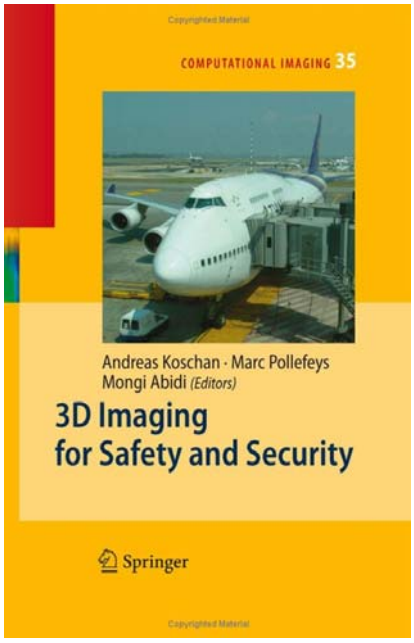


Fig. 4. *Book on 3D Imaging for Safety and Security*

Autonomous mobile robots will play a major role in future security, surveillance and 3D model acquisition tasks for large scale environments such as shopping malls, airports, hospitals and museums. Robotic security guards will autonomously survey such environments, unless a remote human operator takes over control. In this context a 3D model can convey much more useful information than the typical 2D maps used in many robotic applications today, both for visualization of information and as human machine interface for remote control. We address the challenge of building such a model of a large environment (50x60m²) using data from the robot's own sensors: a 2D laser scanner and a panoramic camera. The data are processed in a pipeline that comprises automatic, semiautomatic and manual stages. The user can interact with the reconstruction process where necessary to ensure robustness and completeness of the model. A hybrid representation, tailored to the application, has been chosen: floors and walls are represented efficiently by textured planes. Non-planar structures like stairs and tables, which are represented by point clouds, can be added if desired. Our methods to extract these structures include: simultaneous localization and mapping in 2D and wall extraction based on laser scanner range data, building textures from multiple omnidirectional images using multiresolution blending, and calculation of 3D geometry by a graph cut stereo technique. Various renderings illustrate the usability of the model for visualizing the security guard's position and environment and at the same time serve as potential application for future 3DTV visualization systems.

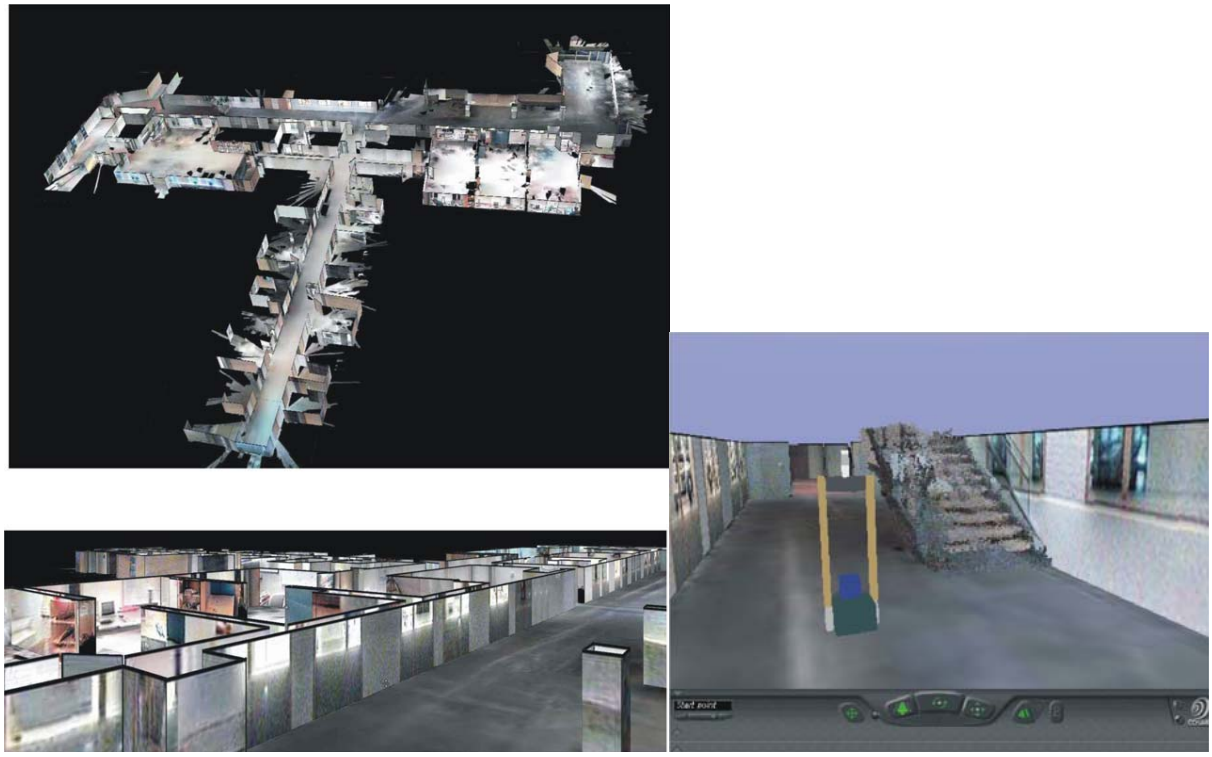


Fig. 5. *Left: Two views of the resulting VRML model. Right. 3D model enriched by results of the graph cut based stereo approach. The resulting model is augmented with virtual content, here for example the position of the robot.*

3.1.5 3DTV- Panoramic 3D Model Acquisition and its 3D Visualization on the Interactive FogScreen

Authors: Sven Fleck and Florian Busch and Peter Biber and Wolfgang Strasser and Ismo Rakkolainen and Stephen DiVerdi and Tobias Hoellerer
Institutions: WSI/GRIS, University of Tübingen, Tübingen, Germany
FogScreen Inc., Espoo, Finland
University of California, Santa Barbara, USA

Publication: IEEE International Conference on Image Processing (ICIP) -- 3DTV Special Session, 2006

Future 3D Television critically relies on mechanisms for automatically acquiring and visualizing high quality 3D content of both indoor and outdoor scenes. The envisioned goal is that a photo-realistic 3D real-time rendering from the actual and potentially arbitrary viewpoint of the beholder who is watching 3DTV becomes possible. Such scenes include movie sets in studios, e.g., for talk shows, TV series and blockbuster movies, but also outdoor scenes, e.g., buildings in a neighborhood for a car chase or cultural heritage sites for a documentary. The goal of 3D model acquisition is to provide the 3D background models where potential 3D actors can be embedded. We present both the 3D acquisition (Fig. 6) and semi-immersive 3D visualization (Fig. 7) to give an impression how a future 3D Television system could be like.

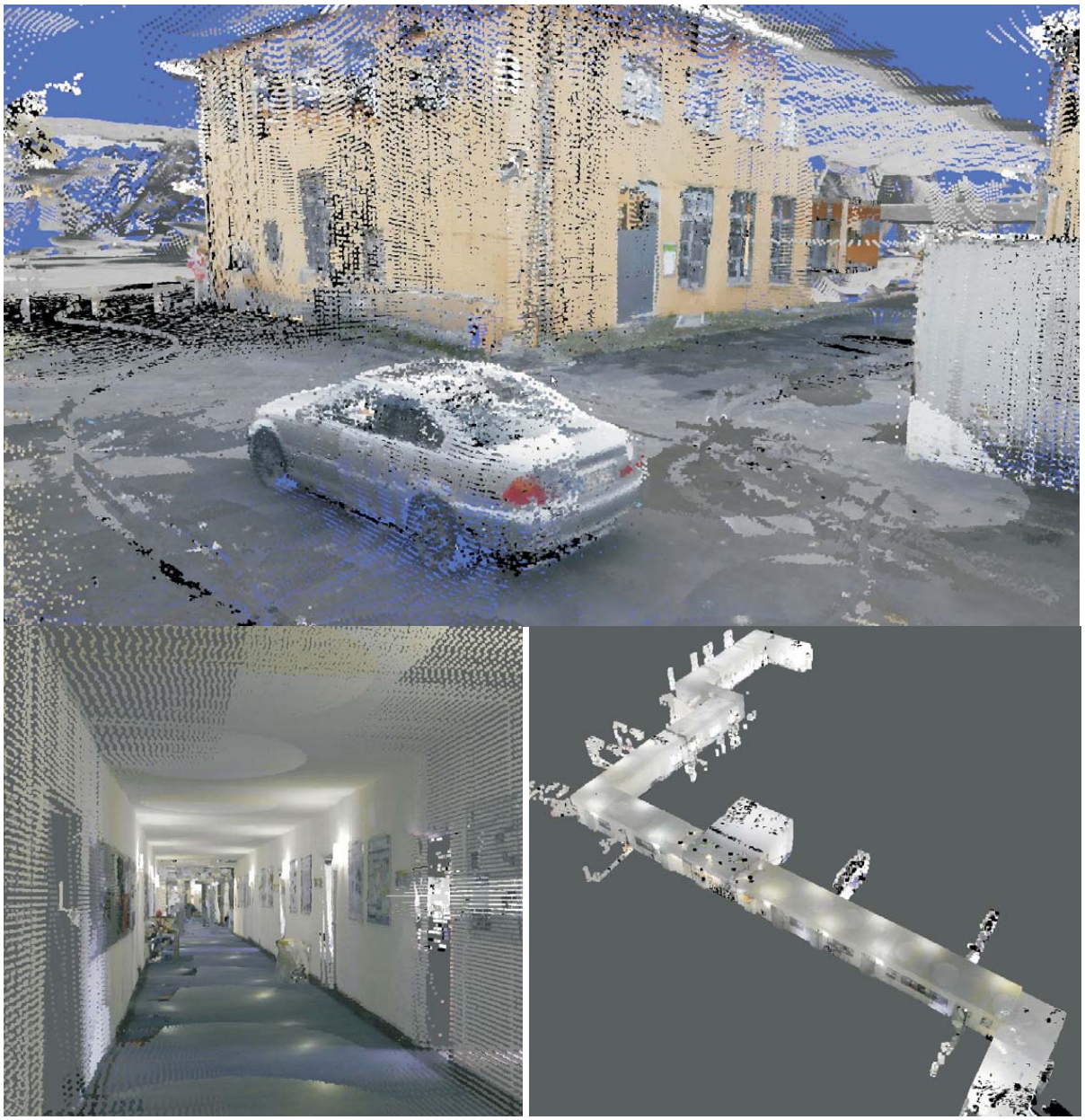


Fig. 6. *Top: Outdoor car scene. Bottom: Renderings of our whole institute's hallway*

Fig. 7. *FogScreen Visualization: Visualization of the 3D models acquired by our Wägele model acquisition platform on the Interactive FogScreen*

3.1.6 SAD – A Novel Multisensor Scene Acquisition Device

Authors: Benjamin Huhle and Philipp Jenke and Wolfgang Straßer

Institutions: WSI/GRIS, University of Tübingen, Tübingen, Germany

Publication: 10th Tübingen Perception Conference (TWK 2007), 2007

Many approaches to 3D scene acquisition with a variety of possible sensor-setups have been presented in the last years, many of them employ multiple laser scanners mounted on small carts or even real cars. In contrast, we present a Scene Acquisition Device (SAD) dedicated to the acquisition of 3D models in complicated environments where small dimensions and simple usability are essential.

The proposed system consists of three sensors: A time-of-flight range sensor (PMD Vision 19k) combined with a standard color camera (Matrix Vision BlueFox) and a miniature inertia sensor (XSens MTi) build a handy acquisition device that allows for fast interactive capturing of color and geometry in arbitrary environments.

To gain from the strengths of the different sensors we post-process the raw data employing color and depth information in an integrated manner.High-quality color data can be used to improve geometry. In contrast to usual approaches, where color images are used for texturing only, we consider both modalities when removing outliers and smoothing the noisy low-resolution depth data. Therefore, an iterative outlier removal algorithm is proposed that classifies valid depth measurements based on their local neighborhood utilizing color and depth values. The classification result is further used in the second step where a global MRF-based noise reduction is applied. The post-processing results in clean and smooth datasets. Both optimization steps are performed in little computation time such that a live visualization in 3D enables the user to obtain a preview of the snapshots from freely chosen virtual view-points.

For the assembly of large scenes, several frames of the multisensor-system have to be registered in order to create consistent large models. Therefore, an initial orientation estimate is given by the inertia sensor.

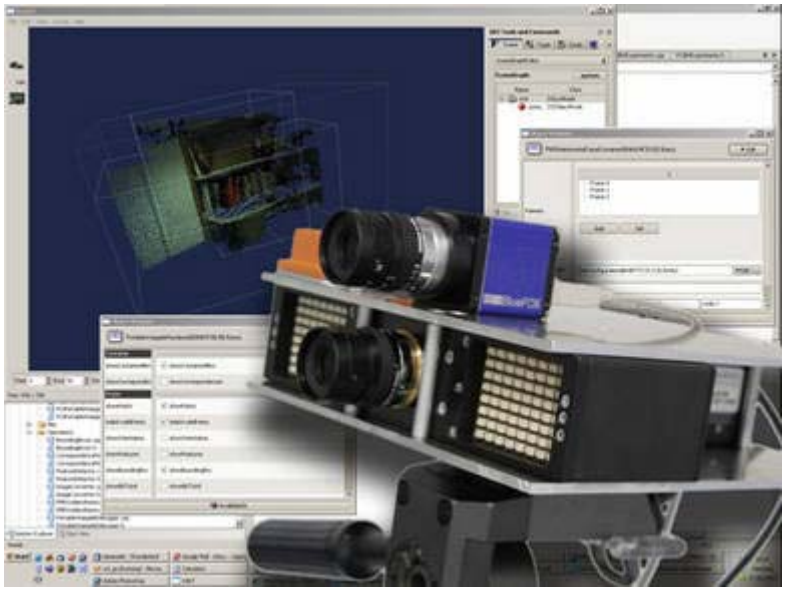


Fig. 8. *SAD Multisensor Scene Acquisition Device*

3.1.7 Integrating 3D Time-Of-Flight Camera Data And High Resolution Images For 3DTV Applications

Authors: Benjamin Huhle, Sven Fleck and Andreas Schilling

Institutions: WSI/GRIS, University of Tübingen, Tübingen, Germany

Publication: IEEE 3DTV Conference, Kos, Greece, May 7-9, 2007.

Applying the machine-learning technique of inference in Markov Random Fields we build improved 3D models by integrating two different modalities. Visual input from a standard color camera delivers high-resolution texture data but also enables us to enhance the 3D data calculated from the range output of a 3D time-of-flight camera in terms of noise and spatial resolution. The proposed method to increase the visual quality of the 3D data makes this kind of camera a promising device for various upcoming 3DTV applications. With our two-camera setup as illustrated in Fig. 9 we believe that the design of lowcost, fast and highly portable 3D scene acquisition systems will be possible in the near future.

Fig. 9. *SAD – Our scene acquisition device for 3D content – e.g., for 3DTV*

Our results (please see Fig. 10) show that a significant gain in the visual quality of the rendered 3D model derived from the different sensor data is achievable. Compared to previous work our approach does not make use of mechanical components as rotating laser scanners. The presented camera setup is applicable also to dynamical scenes whereas most laser scanner based acquisition platforms are constraint to completely static environments. In comparison with stereo systems, our setup is also able to work in lowly illuminated or textureless scenarios where stereo systems generally fail.

Fig. 10. *3D model acquired by our SAD acquisition system*

3.1.8 On-the-Fly Scene Acquisition with a Handy Multisensor-System

Authors: Benjamin Huhle and Philipp Jenke and Wolfgang Straßer

Institutions: WSI/GRIS, University of Tübingen, Tübingen, Germany

Publication: Dynamic 3D Imaging Workshop in Conjunction with DAGM 2007, 2007

We present a scene acquisition system which allows for fast and simple acquisition of arbitrarily large 3D environments (see Fig. 11). We propose a small device which acquires and processes frames consisting of depth and color information at interactive rates. This allows the operator to control the acquisition process on the fly. However, no user input or prior knowledge of the scene are required. In each step of the processing pipeline color and depth data are used in combination in order to gain from different strengths of the sensors. A novel registration method is introduced that combines geometry and color information for enhanced robustness and precision. We evaluate the performance of the system and present results from acquisition in different environments.

Fig. 11. *Left: SAD scene acquisition device. Right: NDT grid used for local registration*

The acquisition of a typical office environment is shown in Fig. 12. Such a scene usually has many color features (posters on the wall and many different items in the shelves). Fig. 13 shows the coffee corner of our institute acquired by SAD.

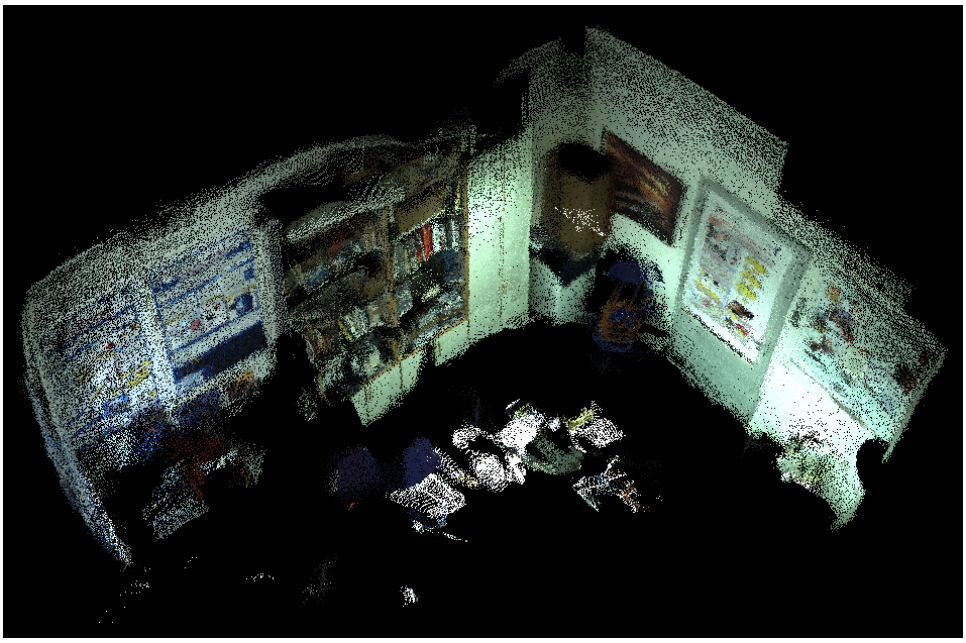


Fig. 12. *Scene acquired by SAD – office environment*

Fig. 13. *Scene acquired by SAD – coffe corner at our institute*

3.1.9 Self-Localization in Scanned 3DTV Sets

Authors: Philipp Jenke, Benjamin Huhle, Wolfgang Straßer

Institutions: WSI/GRIS, University of Tübingen, Tübingen, Germany

Publication: 3DTV CON - The True Vision, 2007

Future 3D Television applications will offer the viewer to freely choose his viewpoint during transmission. A lot of research in the field of 3DTV therefore concentrated on capturing photo-realistic 3D models of studio or movie sets. In this paper, however, we concentrate on the problem of self localization within such scenes. As input we expect a 3D model of an arbitrary environment. Therein, we are able to localize a low-cost portable sensor-system based on a 3D time-of-flight camera. Point clouds acquired with this system from arbitrary viewpoints are registered to the given model in order to estimate its position and orientation in the scene. Examples are shown in Fig. 14 and Fig. 15.

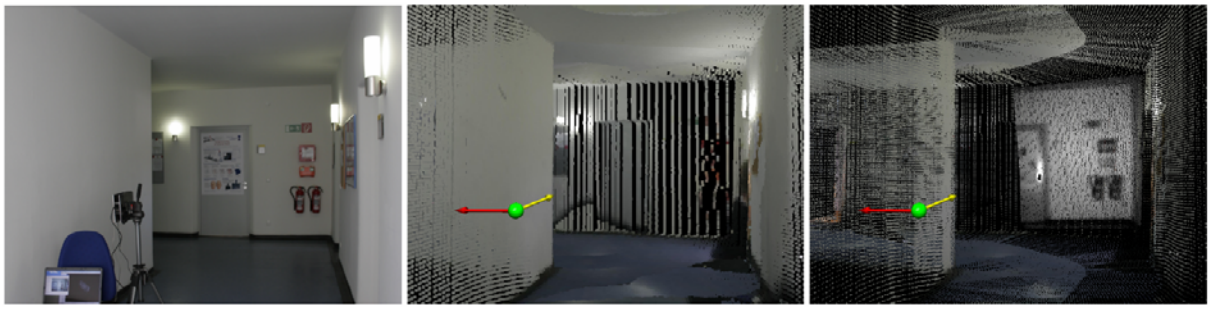


Fig. 14. *Left: Image taken with digital camera. Middle: Screenshot from within base model. Right: Registered observer data included. The yellow arrow marks the estimated view direction, the red arrow the estimated up vector and the green sphere the camera position. Rendering point size varies between middle and right image.*

Fig. 15. *Underlying model used in the self-localization process. The model shows our department main floor, acquired by our Wägele model acquisition platform. Observer camera frames (reference and up vector) are visualized.*

3.1.10 A Volumetric Fusion Technique for Surface Reconstruction from Silhouettes and Range Data

Authors: Y. Yemez and C. J. Wetherilt

Institutions: Koç University

Publication: Computer Vision and Image Understanding, Vol. 105, No. 1, pp. 30–41, 2007.

Optical triangulation, an active reconstruction technique, is known to be an accurate method but has several shortcomings due to occlusion and laser reflectance properties of the object

surface, that often lead to holes and inaccuracies on the recovered surface. Shape from silhouette, on the other hand, as a passive reconstruction technique, yields robust, hole-free reconstruction of the visual hull of the object. In this paper, a hybrid surface reconstruction method that fuses geometrical information acquired from silhouette images and optical triangulation is presented. Our motivation is to recover the geometry from silhouettes on those parts of the surface which the range data fail to capture. A volumetric octree representation is first obtained from the silhouette images and then carved by range points to amend the missing cavity information. An isolevel value on each surface cube of the carved octree structure is accumulated using local surface triangulations obtained separately from range data and silhouettes. The marching cubes algorithm is then applied for triangulation of the volumetric representation. The performance of the proposed technique is demonstrated on several real objects. In Fig. 16, we provide the reconstructed models using our method by fusing shape from silhouette and optical triangulation.

Fig. 16. *(Cup object, first row) Silhouette reconstruction, optical triangulation reconstruction, and two views from the model obtained by fusion, respectively. (Head object, second row) Silhouette reconstruction, two views from optical triangulation reconstruction, and two views from fusion. (Elephant object, third row) Silhouette reconstruction, optical triangulation reconstructions from two separate scans, and two views from fusion.*

3.1.11 Multicamera Audio-Visual Analysis of Dance Figures

Authors: F. Ofli, Y. Demir, E. Erzin, Y. Yemez, and A. M. Tekalp

Institutions: Koç University

Publication: IEEE Int. Conf. on Multimedia and Expo ICME'07, pp. 1703-1706, Beijing, China, 2007.

We present an automated system for multicamera motion capture and audio-visual analysis of dance figures. The multiview video of a dancing actor is acquired using 8 synchronized cameras. The motion capture technique is based on 3D tracking of the markers attached to the person's body in the scene, using stereo color information without need for an explicit 3D model. The resulting set of 3D points is then used to extract the body motion features as 3D displacement vectors whereas MFC coefficients serve as the audio features. In the first stage of multimodal analysis, we perform Hidden Markov Model (HMM) based unsupervised temporal segmentation of the audio and body motion features, separately, to determine the recurrent elementary audio and body motion patterns. Then in the second stage, we investigate the correlation of body motion patterns with audio patterns, that can be used for estimation and synthesis of realistic audio-driven body animation. In Fig. 17, we display an example dance scene captured by our 8-camera system.

Fig. 17. *Dance scene captured by the 8-camera system available at Koc¸ University. Markers are attached at or around the joints of the body.*

3.1.12 A Simple Framework for Natural Animation of Digitized Models

Authors: E. de Aguiar, R. Zayer, C. Theobalt, M. Magnor, H.-P. Seidel

Institutions: MPI Informatik, TU Braunschweig

Publication: Proc. of SIBGRAPI.'07. IEEE Press, 2007

We present a versatile, fast and simple framework to generate animations of scanned human characters from input optical motion capture data (Fig. 18). Our method is purely mesh-based and requires only a minimum of manual interaction. The only manual step needed to create moving virtual people is the placement of a sparse set of correspondences between the input data and the mesh to be animated. The proposed algorithm implicitly generates realistic body deformations, and can easily transfer motions between human subjects of completely different shape and proportions. We feature a working prototype system that demonstrates that our method can generate convincing lifelike character animations directly from optical motion capture data.

Fig. 18. *Subsequent frames generated by our system showing the female scan authentically performing a soccer kick.*

3.1.13 Video-driven Animation of Human Body Scans

Authors: E. de Aguiar, R. Zayer, C. Theobalt, M. Magnor and H.-P. Seidel

Institutions: MPI Informatik, TU Braunschweig

Publication: In IEEE 3DTV Conference 2007, Kos Island, Greece

We present a versatile, fast and simple framework to generate animations of scanned human characters from input multi-view video sequences. Our method is purely mesh-based and requires only a minimum of manual interaction. The proposed algorithm implicitly generates realistic body deformations and can easily transfer motions between human subjects of completely different shape and proportions. We feature a working prototype system that demonstrates that our method can generate convincing lifelike character animations from marker-less optical motion capture data (Fig. 19).

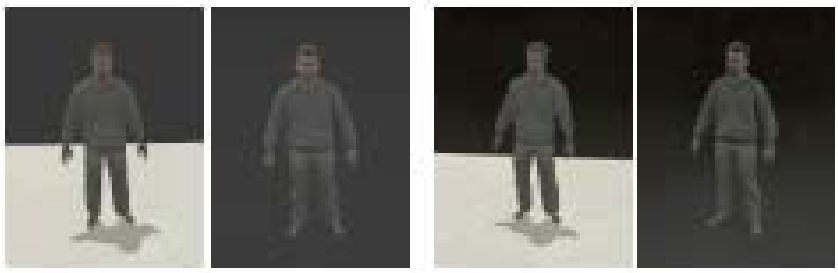


Fig. 19. *Our approach enables the creation of 3D videos with high-quality models.*

3.1.14 Rapid Animation of Laser-scanned Humans

Authors: E. de Aguiar, C. Theobalt, C. Stoll and H.-P. Seidel

Institutions: MPI Informatik

Publication: Proc. of IEEE Virtual Reality 2007, pp. 223-226, Charlotte, USA

We present a simple and efficient approach to turn laser-scanned human geometry into a realistically moving virtual avatar (Fig. 20). Instead of relying on the classical skeleton-based animation pipeline, our method uses a mesh-based Laplacian editing scheme to drive the motion of the scanned model. Our framework elegantly solves the motion retargeting problem and produces realistic non-rigid surface deformation with minimal user interaction. Realistic animations can easily be generated from a variety of input motion descriptions, which we exemplify by applying our method to both marker-free and marker-based motion capture data.

Fig. 20. *Several frames where the avatar performs soccer moves. The input motion was captured by means of a marker-based optical motion capture system and the animation is efficiently generated by our method.*

3.1.15 Marker-less Deformable Mesh Tracking for Human Shape and Motion Capture

Authors: E. de Aguiar, C. Theobalt, C. Stoll and H.-P. Seidel

Institutions: MPI Informatik

Publication: Proc. of IEEE CVPR 2007, Minneapolis, USA.

We present a novel algorithm to jointly capture the motion and the dynamic shape of humans from multiple video streams without using optical markers (Fig. 21). Instead of relying on kinematic skeletons, as traditional motion capture methods, our approach uses a deformable high-quality mesh of a human as scene representation. It jointly uses an image-based 3D correspondence estimation algorithm and a fast Laplacian mesh deformation scheme to capture both motion and surface deformation of the actor from the input video footage. As opposed to many related methods, our algorithm can track people wearing wide apparel, it can straightforwardly be applied to any type of subject, e.g. animals, and it preserves the connectivity of the mesh over time. We demonstrate the performance of our approach using synthetic and captured real-world video sequences and validate its accuracy by comparison to the ground truth.

Fig. 21. *Our method realistically captures the motion and the dynamic shape of a woman wearing a Japanese kimono from only eight video streams.*

.

3.1.16 Marker-less 3D Feature Tracking for Mesh-based Motion Capture

Authors: E. de Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel

Institutions: MPI Informatik

Publication: 2nd Workshop on Human Motion. ICCV'07, Rio de Janeiro, Brazil

We present a novel algorithm that robustly tracks 3D trajectories of features on a moving human who has been recorded with multiple video cameras. Our method does so without special markers in the scene and can be used to track subjects wearing everyday apparel. By using the paths of the 3D points as constraints in a fast mesh deformation approach, we can directly animate a static human body scan such that it performs the same motion as the captured subject (Fig. 22). Our method can therefore be used to directly animate high quality geometry models from unaltered video data which opens the door to new applications in motion capture, 3D Video and computer animation. Since our method does not require a kinematic skeleton and only employs a handful of feature trajectories to generate lifelike animations with realistic surface deformations, it can also be used to track subjects wearing wide apparel, and even animals. We demonstrate the performance of our approach using several captured real-world sequences, and also validate its accuracy.

Fig. 22. *Our framework correctly tracks 3D trajectories of features over time. By combining the 3D point trajectories with our mesh deformation method, our algorithm is able to directly animate a human body scan.*

3.1.17 Reconstructing Human Shape, Motion and Appearance from Multi-view Video

Authors: C. Theobalt, E. de Aguiar, M. A. Magnor, and H.-P. Seidel

Institutions: MPI Informatik, TU Braunschweig

Publication: Chapter in book Three-Dimensional Television: Capture, Transmission, and Display. Springer, Heidelberg, 2007. In press.

In recent years, an increasing research interest in the field of 3D video processing has been observed. The goal of 3D video processing is the extraction of spatio-temporal models of dynamic scenes from multiple 2D video streams. These scene models comprise of descriptions of the shape and motion of the scene as well as its appearance. Having these dynamic representations at hand, one can display the captured real world events from novel synthetic camera perspectives. In order to put this idea into reality, algorithmic solutions to three major problems have to be found: the problem of multi-view acquisition, the problem of scene reconstruction from image data, and the problem of scene display from novel viewpoints.

Human actors are presumably the most important elements of many real-world scenes. Unfortunately, it is well-known to researchers in computer graphics and computer vision that both the analysis of shape and motion of humans from video, as well as their convincing graphical renditions are very challenging problems. To tackle the difficulties of the involved problems, we propose in this chapter three model-based approaches to capture the motion, as well as the dynamic geometry of moving humans. By applying a fast dynamic multi-view texturing method to the captured time-varying geometry, we are able to render convincing free-viewpoint videos of human actors. This chapter is a roundup of several algorithms that we have recently developed. It shall serve as an overview and make the reader aware of the most important research questions by illustrating them on state-of-the art research prototypes. Furthermore, a detailed list of pointers to related work shall enable the interested reader to explore the field in greater depth on its own.

For all the proposed methods, human performances are recorded with only eight synchronized video cameras. In the first algorithmic variant, a template model is deformed to match the shape and proportions of the captured human actor and it is made to follow the motion of the person by means of a marker-free optical motion capture approach. The second variant extends the first one, and enables the estimation not only of shape and motion parameters of the recorded subject, but also the reconstruction of dynamic surface geometry details that vary over time. The last variant shows how we can incorporate high-quality laser-scanned shapes into the overall work-flow. Using any of the presented method variants, the human performances can be rendered in real-time from arbitrary synthetic viewpoints. Time-varying surface appearance is generated by means of a dynamic multi-view texturing from the input video streams.

3.1.18 A Volumetric Approach to Interactive Shape Editing

Authors: C. Stoll, E. de Aguiar, C. Theobalt and H.-P. Seidel

Institutions: MPI Informatik

Publication: Technical Report MPI-I-2007-4-004, MPII, 2007.

We present a novel approach to real-time shape editing that produces physically plausible deformations using an efficient and easy-to-implement volumetric approach. Our algorithm alternates between a linear tetrahedral Laplacian deformation step and a differential update in which rotational transformations are approximated. By means of this iterative process we can achieve non-linear deformation results while having to solve only linear equation systems. The differential update step relies on estimating the rotational component of the deformation relative to the rest pose. This makes the method very stable as the shape can be reverted to its rest pose even after extreme deformations. Only a few point handles or area handles imposing an orientation are needed to achieve high quality deformations, which makes the approach intuitive to use. We show that our technique is well suited for interactive shape manipulation and also provides an elegant way to animate models with captured motion data.

3.1.19 Animation Collage

Authors: C. Theobalt, C. Roessl, E. de Aguiar and H.-P. Seidel

Institutions: Stanford University, INRIA Sophia-Antipolis and MPI Informatik

Publication: Proc. ACM Symposium on Computer Animation (SCA 2007), San Diego, USA

We propose a method to automatically transform mesh animations into animation collages, i.e. moving assemblies of shape primitives from a database given by an artist (Fig. 23). An animation collage is a complete reassembly of the original animation in a new abstract visual style that imitates the spatio-temporal shape and deformation of the input. Our algorithm automatically decomposes input animations into plausible approximately rigid segments and fits to each segment one shape from the database by means of a spatio-temporal matching procedure. The collage is then animated in compliance with the original's shape and motion. Apart from proposing solutions to a number of spatio-temporal alignment problems, this work is an interesting add-on to the graphics artist's toolbox with many applications in arts, non-photorealistic rendering, and animated movie productions. We exemplify the beauty of animation collages by showing results created with our software prototype.

Fig. 23. *Our method automatically generates moving 3D collages out of mesh animations by rebuilding them as moving assemblies of shape primitives. In the example above the galloping horse has been transformed into a galloping sets of fruits.*

3.1.20 Automatic Conversion of Mesh Animations into Skeleton-based Animations

Authors: E. de Aguiar, C. Theobalt, S. Thrun, and H.-P. Seidel

Institutions: MPI Informatik and Stanford University

Publication: Proc. of EUROGRAPHICS 2008 (Computer Graphics Forum, vol. 27 issue 2), Crete, Greece.

Recently, it has become increasingly popular to represent animations not by means of a classical skeleton-based model, but in the form of deforming mesh sequences. The reason for this new trend is that novel mesh deformation methods as well as new surface based scene capture techniques offer a great level of flexibility during animation creation. Unfortunately, the resulting scene representation is less compact than skeletal ones and there is not yet a rich toolbox available which enables easy post-processing and modification of mesh animations. To bridge this gap between the mesh-based and the skeletal paradigm, we propose a new method that automatically extracts a plausible kinematic skeleton, skeletal motion parameters, as well as surface skinning weights from arbitrary mesh animations (Fig. 24). By this means, deforming mesh sequences can be fully-automatically transformed into fully-rigged virtual subjects. The original input can then be quickly rendered based on the new compact bone and skin representation, and it can be easily modified using the full repertoire of already existing animation tools.

Fig. 24. *From left to right: Input animation, color-coded distribution of blending weights, and two poses of the input poses regenerated by our method.*

3.1.21 Color High Dynamic Range (HDR) Imaging: The Luminance-Chrominance Approach

Authors: Ossi Pirinen, Alessandro Foi, Atanas Gotchev

Institution: Department of Signal Processing, Tampere University of Technology, Finland

Publication: International Journal of Imaging Systems and Technology, vol. 17, No.3, pp. 152-162, 2007

This paper presents a novel and efficient approach to color in high dynamic range (HDR) imaging. In contrast to state-of-the-art methods, we propose to move the complete HDR imaging process from RGB to a luminance-chrominance color space. Our aim is to get a computationally efficient technique and to avoid any possible color distortions originating from the three RGB color channels processed separately. To achieve this, we build a camera response function for the luminance channel only and weight and compose the HDR luminance accordingly, while for the chrominance channels we apply weighting in relation with the saturation level. We demonstrate that our technique yields natural and pleasant to perceive tone-mapped images and is also more robust to noise. The technique is illustrated in Fig. 25.

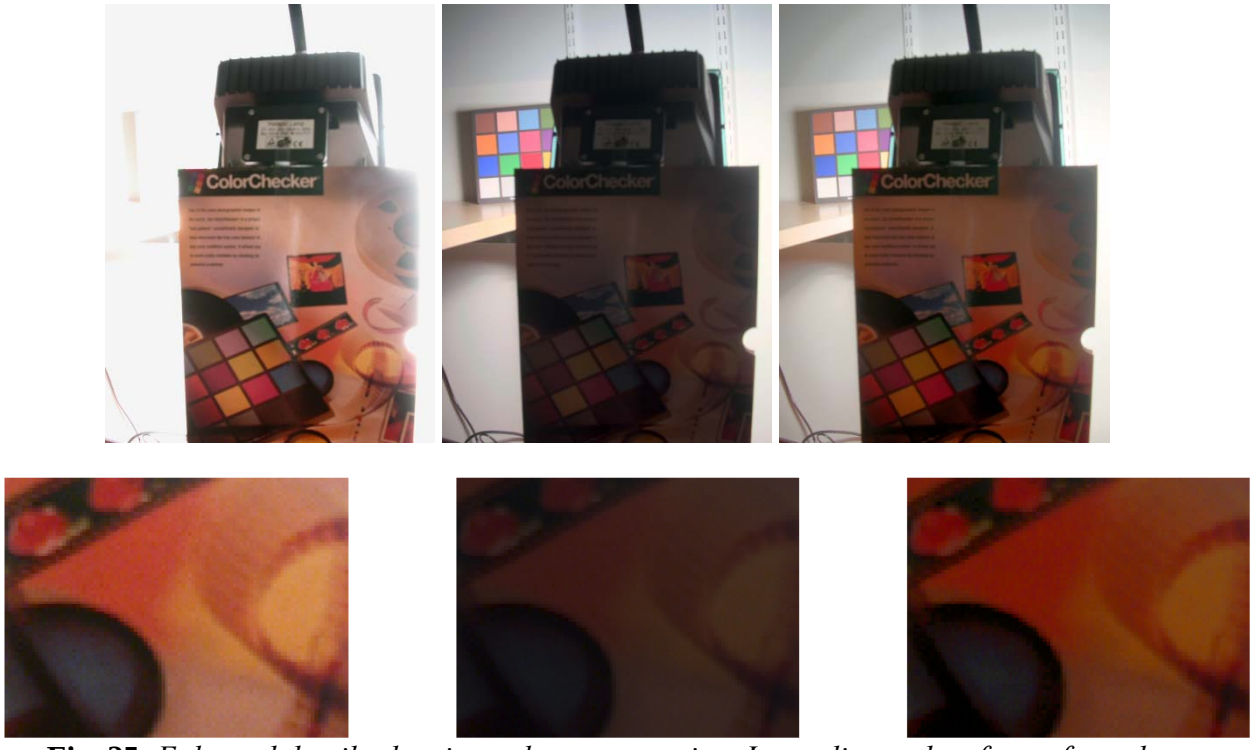


Fig. 25. *Enlarged details showing color preservation. In reading order: frame from the original sequence, tone mapped RGB HDR and tone mapped L-Cr HDR.*

3.1.22 Why HDR is Important for 3DTV Model Acquisition

Authors: Benjamin Huhle[1], Ossi Pirinen[2], Sven Fleck[1], Atanas Gotchev[2], and Wolfgang Strasser[1]

Institutions: [1]WSI/GRIS, University of Tübingen, Germany, [2]Department of Signal Processing, Tampere University of Technology, Finland

Publication: submitted to 3DTV-CON 2008, Istanbul, Turkey

Mechanisms for automatically acquiring high quality 3D content of both indoor and outdoor scenes are essential research topics within 3D Television (3DTV). The goal of 3D model acquisition is to provide the 3D background models where potential 3D actors can be embedded. Although model acquisition platforms have become available, the dynamic range of the used cameras is too limited to fully cover real world environments. We present two multi-sensor systems for 3D model acquisition, demonstrate the need for high dynamic range (HDR) capable acquisition platforms and propose a system that performs HDR computation to produce 3D models with tonemapped HDR information. Combining depth and HDR color information, a Markov Random Field based technique is used to get best out of both worlds – the high spatial resolution of the tonemapped HDR image with the depth field of a time-of-flight camera. The results are illustrated in Fig. 26.

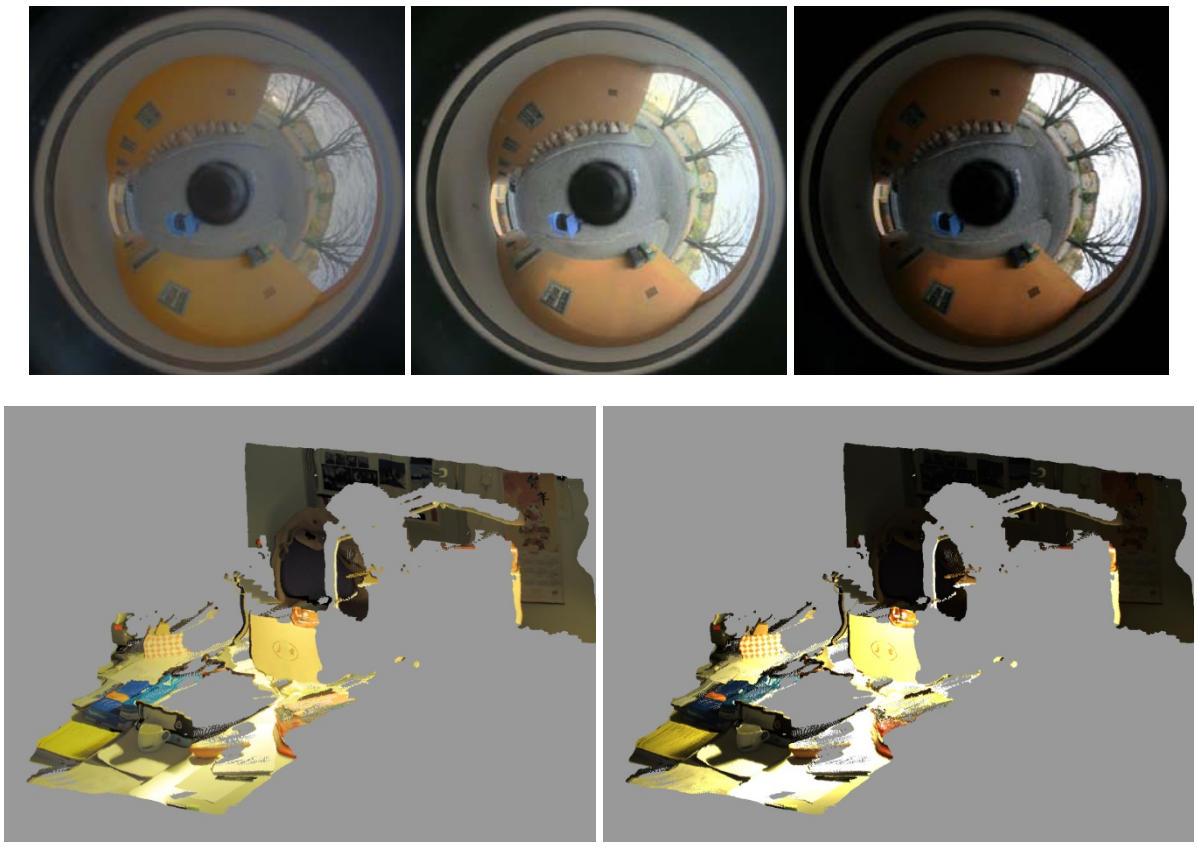


Fig. 26. *Top row: LCr-HDR tone mapped spherical image, RGB-HDR tone mapped spherical image, LDR spherical image. Bottom row: time-of-flight depthmap texturized with a tone mapped LCr-HDR and with an LDR image.*

3.1.23 High Dynamic Range Imaging in Luminance-Chrominance Space

Authors: Ossi Pirinen

Institution: Department of Signal Processing, Tampere University of Technology, Finland

Publication: MSc thesis

High dynamic range (HDR) imaging is a modern solution to the problem of sensor capability limitations in digital imaging. The issue has been addressed for gray scale as well as RGB (red, green, blue) images. Color is however not taken into account adequately profoundly in neither. This thesis presents a novel and efficient approach to color in high dynamic range (HDR) imaging.

In contrast to state-of-the-art methods, the complete HDR imaging process is moved from RGB to a luminance-chrominance color space. The result is a computationally efficient technique which avoids possible color distortions originating from the three RGB color channels processed separately. To achieve this, a camera response function is built for the luminance channel only and the HDR luminance is weighted and composed accordingly, while for the chrominance channels a saturation level dependent weighting is applied. It is demonstrated that the described technique yields natural and pleasant to perceive tone-mapped images and is also more robust to noise.

3.1.24 Stereopsis based on image segmentation

Authors: Ruixing Yang

Institution: Department of Signal Processing, Tampere University of Technology, Finland

Publication: MSc thesis

In this thesis, a hybrid algorithm for stereo map generation is investigated. It combines the properties of the feature-based and area-based approaches to achieve better performance. In the algorithm, feature correspondence searching is utilized in a precise manner. First, a corner detector is used to detect the edges and corners in the sensed image and the reference image. After removal of outliers, the remaining features are used in an area-based approach for searching correspondences. This is complemented by a Scale Invariant Feature Transform (SIFT) detector to derive a complementary set of SIFT keys. They provide an additional robustness to the corner-based features. After finding a good set of correspondences and generating a disparity map, it is required to interpolate it to generate the complete disparity map. In the algorithm, segmentation based on histogram is proposed for improving the interpolation result. The proposed algorithm benefits from both classical feature-based and area-based approaches. The selected SIFT feature correspondence searching is very robust and efficient. For edges and corners, some points are dropped to avoid the error that is easily generated by the area-based approach. For each step of the proposed algorithm, the robustness and efficiency are carefully considered, leading to an improved final result.

3.1.25 Real-time Hierarchical Stereo Matching on Graphics Hardware

Authors: Lukas Heidenreich

Institutions: MPI Informatik

Publication: MSc thesis

Stereo matching is a well-known problem in computer vision, and many researchers deal with finding good algorithms for solving it. The disadvantage of most approaches is the long time they need for detecting correspondences on common computer systems. Thus, stereo matching in real-time systems is often not possible. One possibility for speeding up the algorithms is the use of modern graphics hardware. This thesis describes a stereo matching algorithm that runs completely on graphics hardware, and calculates dense depth maps of a stereo pair of arbitrarily placed calibrated cameras in real-time, so it can be used for processing image sequences or video streams. Starting with the computation of small-sized depth maps from box-filtered images, we use a hierarchical approach and propagate repeatedly computed depths to calculate the final correct depth. Through the filtering of outliers after every step, we can further improve the calculation of the final depth map. This map can then be used for reconstructing the captured scene. Finally, the algorithm is applied with different settings on several stereo sets to measure its quality and speed.

3.1.26 Reconstruction and Rendering of Time-Varying Natural Phenomena

Authors: Ivo Ihrke

Institutions: MPI Informatik

Publication: Ph.D. thesis

While computer performance increases and computer generated images get ever more realistic, the need for modeling computer graphics content is becoming stronger. To achieve photo-realism detailed scenes have to be modeled often with a significant amount of manual labour. Interdisciplinary research combining the fields of Computer Graphics, Computer Vision and Scientific Computing has led to the development of (semi-)automatic modeling tools freeing the user of labour-intensive modeling tasks. The modeling of animated content is especially challenging. Realistic motion is necessary to convince the audience of computer games, movies with mixed reality content and augmented reality applications. The goal of this thesis is to investigate automated modeling techniques for time-varying natural phenomena. The results of the presented methods are animated, three-dimensional computer models of fire, smoke and fluid flows (Fig. 25)

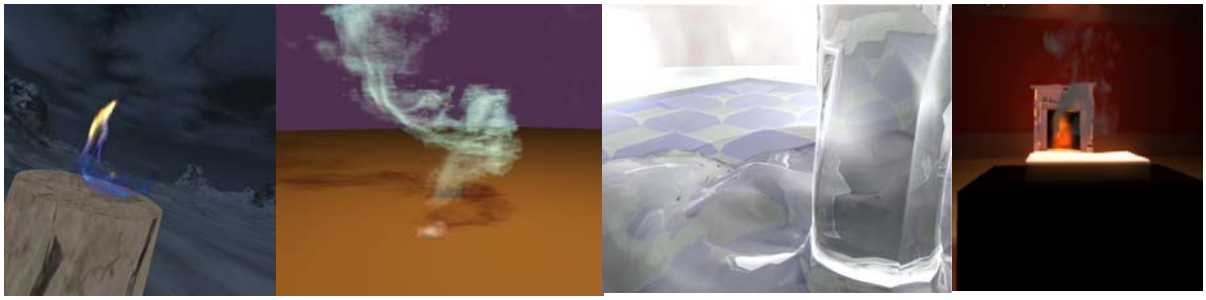


Fig. 27. *Animated three-dimensional computer models of fire, smoke and fluid flows generated by the approaches described in the thesis*

3.1.27 New Editing Techniques for Video Post-Processing

Authors: Volker Scholz

Institutions: MPI Informatik

Publication: Ph.D. thesis

Today's digital image processing tools have greatly advanced movie editing capabilities. However, considerable, time-consuming manual interaction is still necessary for post-production tasks like rotoscoping, segmentation etc. Replacement of non-rigid objects such as cloth is almost infeasible without automation, due to the high number of degrees of freedom of the surface. For general shape and motion editing of video objects, an easy-to-use interactive system which only requires a moderate amount of user interaction is desirable. This dissertation contributes to capturing cloth shape, editing cloth texture and altering object shape and motion in multi-camera and monocular video recordings. We propose a technique to capture cloth shape from a 3D scene flow by determining optical flow in several camera views. Together with a silhouette matching constraint we can track and reconstruct cloth surfaces in long video sequences. Reconstructing the surface is a prerequisite for further editing operations such as texture replacement. In the area of garment motion capture, we present a system to reconstruct time-coherent triangle meshes from multi-view video recordings. It makes use of a specially designed color pattern which allows a unique identification of color features on the garment across different camera viewpoints. Texture mapping of the acquired triangle meshes is used to replace the recorded texture with new cloth patterns. We extend this work to the more challenging single camera view case. Simultaneously extracting texture deformation and shading effects enables us to achieve texture replacement effects which are close to reality. We use the same color pattern as in the multi-camera approach. This method enables us to exchange fabric pattern designs worn by actors as a video post-processing step. Finally, we propose a system for keyframe editing of video objects. A color-based segmentation algorithm together with automatic video in-painting for filling in missing background texture allows us to edit shape and motion of 2D video objects. We present examples for altering object trajectories, applying non-rigid deformation and simulating camera motion. Our vision is that a powerful video post-processing framework gives artists additional freedom to tell the visual story of a film during editing.

3.1.28 GPU Data Structures for Video Processing and Vision-based Graphics

Authors: Gernot Ziegler

Institutions: MPI Informatik

Publication: Ph.D. thesis (expected by 2008)

Graphics hardware has in recent years become increasingly programmable. The problem in porting CPU algorithms for video and volume processing to graphics hardware lies often in the inherent restrictions of the stream processor model that the GPU is using to maintain its high performance. The serial data dependencies that are often employed to accelerate CPU processing are directly counterproductive in a parallel processing model.

This PhD thesis shows new ways of handling well-known problems in large scale video/volume analysis. In some occasions, we adapt to the restricted hardware model by re-introducing algorithms from early computer graphics research. In other occasions, we build hierarchical data structures to circumvent the random-access read/fixed write restriction that previously kept sophisticated analysis algorithms from running solely on graphics hardware. We apply known graphics concepts such as mip-maps, projective texturing, or dependent texture lookups to general purpose computing on the GPU (GPGPU), and show how video/volume processing can benefit algorithmically from being implemented in a graphics API.

In summary: We demonstrate how these novel GPU data structures rapidly increase processing speed, and many times lift processing-heavy operations into the real-time domain, making way for new, and interactive vision/graphics applications.

3.1.29 Capturing and Editing Moving Scanned Subjects

Authors: Edilson de Aguiar

Institutions: MPI Informatik

Publication: Ph.D. thesis (expected by 2008)

Recently, it has become increasingly popular to represent animations not by means of a classical skeleton-based model, but in the form of deforming mesh sequences. New marker-free skeleton-less approaches have been developed which makes the capture process easier. Unfortunately, the resulting scene representation is less compact than skeletal ones and there is not yet a rich toolbox available which enables easy post-processing and modification of mesh animations. In this thesis we present a variety of methods for capturing motion and surface deformations without the use of optical markers. The output is a mesh animation which can later be automatically converted to an skeleton-based animation or a artistic moving collage. By combining different methods, the resulting system can be directly used to create virtual humans for 3D Video, Virtual Reality and Computer Games.

3.1.30 Conclusions and future plans

The research progress on projects related to acquisition, editing and visualization of data in multicamera systems were presented in this report. In Sect. 3.1.1, two applications of a new formulation of the Lagrange equation in arbitrary dimensional space were presented enabling the reconstruction of temporally coherent geometry from multiple video streams and volumetric reconstruction of refractive and transparent natural phenomena.

In Sect. 3.1.2 - 3.1.5, 3D models are acquired simply by moving the platform around and recording images in regular intervals. This approach is similar to a classical multicamera approach for static scenes, however requires only one physical camera. Two approaches are available – a stereo vision approach based on graph cuts and a laser scanner based approach. Sect. 3.1.6 - 3.1.9 introduces a Scene Acquisition Device (SAD) and applications. SAD is a platform that combines cameras of different types – a conventional higher resolution industrial camera and PMD time of flight camera – to perform 3D model acquisition. By visualizing the 3D models acquired by the Interactive FogScreen larger parts of the 3DTV pipeline are covered.

Three paths for improvement of the acquired 3D models are envisaged: first, model acquisition with HDR imaging will overcome the limited dynamic range of the utilized camera. Developments are already ongoing – an automated HDR acquisition flow is available where the Wägele platform performs the LDR image series automatically according to the actual light conditions while minimizing the time for the platform's operator. The Luminance-Chrominance HDR image composition described in Sect. 3.1.21 - 3.1.23 has demonstrated its superiority in terms of color preservation and noise immunity, compared with state-of-the-art methods. The generation of HDR 3D models will be further improved for both Wägele and SAD platforms. Second, the simplification and filling of missing data using machine learning techniques is future work. Third, sensor fusion like within the SAD platform based on time-of-flight cameras is a novel field and a promising approach for a next generation 3DTV related (static) scene acquisition.

Sect. 3.1.10 described a novel surface reconstruction scheme that fuses the geometry information obtained separately from shape from silhouette and shape from optical triangulation techniques. The aim was to compensate for the problems associated with each method by the benefits of the other. The most prominent property of the presented method is the ability to build cavity-sensitive and hole-free models of complicated objects containing severe occlusions and sharp hollows on their surfaces. The experiments show that it is possible to produce robust and satisfactory surface reconstructions. The test objects of our experiments are examples to such complicated objects which are very difficult to reconstruct using only shape from optical triangulation, even with sophisticated range scanners. The foremost restraining factor in the overall system performance was found to be resolution related since the proposed technique is based on volumetric carving. As future work, we plan to develop a fusion scheme that is based on smooth surface deformation to overcome this problem.

3.1.11 presented an automated system for multicamera motion capture and audio-visual analysis of dance figures. The results of our joint audio-video analysis indicate that certain motion patterns are highly correlated with the musical audio channel during a dance performance. Our future work will involve modeling the correlation between temporal

patterns of visual motion and audio towards the goal of developing avatars that can learn how to dance from multicamera recordings and perform a realistic dancing act when driven by musical audio.

Sect. 3.1.12 - 3.1.14 presented two simple and efficient approaches to turn laser-scanned human geometry into a realistically moving virtual humans. The framework elegantly solves the motion retargeting problem and produces realistic non-rigid surface deformation with minimal user interaction. Realistic animations can easily be generated from a variety of input motion descriptions (marker-free and marker-based motion capture data).

Sect. 3.1.15 - 3.1.16 presented new algorithms to jointly capture the motion and the dynamic shape of humans from multiple video streams. As opposed to many related methods, the algorithms can track people wearing wide apparel, they can straightforwardly be applied to any type of subject, e.g. animals, and they preserve the connectivity of the mesh over time. As a direction of future work, we plan to integrate our method into a larger approach to reconstruct high-quality 3D videos of humans wearing arbitrary apparel. Sect. 3.1.17 described two previous model-based approaches for motion capture and 3D Video.

Sect. 3.1.18 presented a novel approach to real-time shape editing that is well suited for interactive shape manipulation providing an elegant way to animate models with captured motion data. Sect. 3.1.19 proposed a method to automatically transform mesh animations into animation collages and Sect. 3.1.20 proposed a new method to automatically extract a plausible kinematic skeleton, skeletal motion parameters, as well as surface skinning weights from arbitrary mesh animations.

Sect. 3.1.23- 3.1.29 presented Master and Ph.D. theses successfully concluded (or to be concluded in 2008) as a result of the research in WP7.

3.2 Single Camera

The upcoming sections briefly explain the outputs on single camera scene extraction technologies, whereas the resulting publications can be found in the Appendix of this technical report.

3.2.1 From 2D- to Stereo- to Multi-View Video

Authors: Sebastian Knorr, Aljoscha Smolic, and Thomas Sikora

Institutions: Technische Universaet Berlin, Fraunhofer-HHI

Publication: 3DTV-Conference 2007, Kos Island, Greece, May 07 - 09 2007

We present a new approach for generation of multi-view video from monocular video. Such multi-view video is used for instance with multi-user 3D displays or auto-stereoscopic displays with head-tracking to create a depth impression of the observed scenery. The intention of this work is not a real-time conversion of existing video material with a deduction in stereo perception, but rather a more realistic off-line conversion with high accuracy. Our approach is based on structure from motion techniques and uses image-based rendering to generate the desired multiple views for each point in time. Figure 26 illustrates the whole approach.

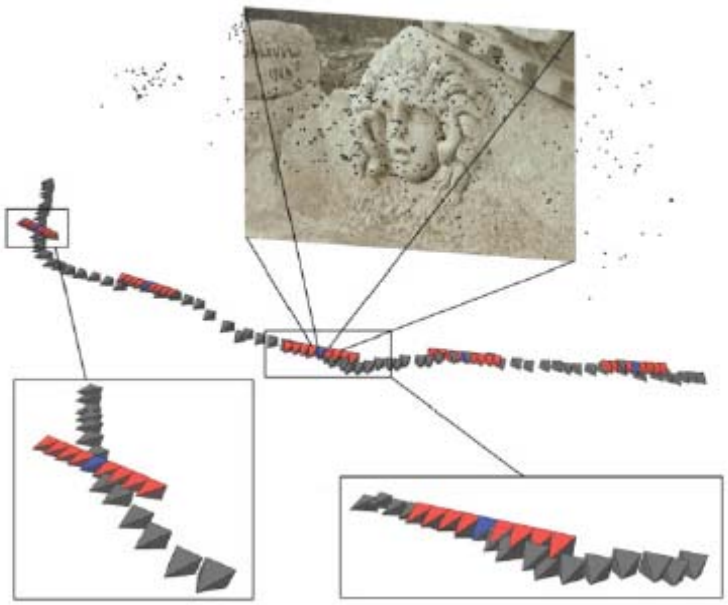


Fig. 28. *Multi-view synthesis using SfM and IBR; gray: original camera path, red: virtual stereo cameras, blue: original camera of a multi-view camera setup*

The algorithm is tested on several TV broadcast videos, as well as on sequences captured with a single handheld camera. Finally, some simulation results will show the remarkable performance of this approach.

3.2.2 Super-Resolution Stereo- and Multi-View Synthesis from Monocular Video Sequences

Authors: Sebastian Knorr, Matthias Kunter, and Thomas Sikora

Institutions: Technische Universitaet Berlin

Publication: 3-D Digital Imaging and Modeling (3DIM 2007), Montréal, Québec, Canada, August 21-23, 2007

Extending visual communication to the third dimension by providing the user with a realistic depth perception of the observed scenery instead of flat 2D images has been investigated over decades. 3DTV is in the focus of many researchers worldwide. Recent progress in related research areas may enable various 3D applications and systems in the near future.

Fig. 29. *Example of an auto-stereoscopic display*

Especially the innovations regarding the 3D display technology are tremendous. 3D displays are entering professional and consumer markets. However, the film industry still adheres to traditional capture techniques with a single camera, i.e. the conversion of existing 2D content into super-resolution 3D is highly interesting for instance for content owners. Movies may be reissued in 3D in the future (see Figure 27), e.g. the *Star Wars* episodes are currently being converted entirely into 3D.

We present a new approach for generation of super-resolution stereoscopic and multi-view video from monocular video. Such multi-view video is used for instance with multi-user 3D displays or auto-stereoscopic displays with head-tracking to create a depth impression of the observed scenery. Our approach is an extension of the realistic stereo view synthesis (RSVS) approach which is based on structure from motion techniques and image-based rendering to generate the desired stereoscopic views for each point in time. The extension relies on an additional super-resolution mode which utilizes a number of frames of the original video sequence to generate a virtual stereo frame with higher resolution. The algorithm is tested on several TV broadcast videos, as well as on sequences captured with a single handheld camera and sequences from the well known BBC documentation “Planet Earth”. Finally, some simulation results will show that RSVS is quite suitable for super-resolution 2D-3D conversion.

3.2.3 An Image-Based Rendering (IBR) Approach for Realistic Stereo View Synthesis of TV Broadcast Based on Structure from Motion

Authors: Sebastian Knorr and Thomas Sikora

Institutions: Technische Universitaet Berlin

Publication: IEEE 14th International Conference on Image Processing (ICIP'07),
San Antonio, Texas, USA, September 16 -19, 2007

In the past years, the 3D display technology has become a booming branch of research with fast technical progress. Hence, the 3D conversion of already existing 2D video material increases more and more in popularity. In this paper, a new approach for realistic stereo view synthesis (RSVS) of existing 2D video material is presented. The intention of our work is not a real-time conversion of existing video material with a deduction in stereo perception, but rather a more realistic off-line conversion with high accuracy. Our approach is based on structure from motion techniques and uses image-based rendering to reconstruct the desired stereo views for each video frame. The algorithm is tested on several TV broadcast videos, as well as on sequences captured with a single handheld camera. Finally, some simulation results will show the remarkable performance of this approach.

3.2.4 Window-Based Image Registration Using Variable Window Sizes

Authors: Andreas Krutz, Michael Frater*, and Thomas Sikora

Institutions: Technische Universitaet Berlin, *University of New South Wales, Australia

Publication: IEEE 14th International Conference on Image Processing (ICIP'07), San Antonio, Texas, USA, September 16-19, 2007

We present an unsupervised image registration algorithm to estimate the background object motion in a real video sequence. The algorithm is based on a Gaussian minimization technique. It has been shown earlier that initialization of such an approach is very important to achieve the motion parameters of the background object precisely, and that the use of a windowing technique can give better background object motion estimation results, even with large background occlusions. In some cases, however, the fixed window size initializes the gradient descent algorithm in a sub-optimal way. Here, another window size would bring the desired estimation direction. In this paper, we present a technique where variable window sizes are used to prevent these outliers. Experimental results show that the technique works very well with the considered test sequences.

3.2.5 Fast Outlier Rejection by Using Parallax-Based Rigidity Constraint for Epipolar Geometry Estimation

Authors: Engin Tola[1] and A.Aydın Alatan[2]

Institutions: [1] Computer Vision Laboratory, Ecole Polytechnique Federal de Lausanne (EPFL), Lausanne, Switzerland, [2] Dept. of Electrical & Electronics Eng. Middle East Technical University (METU), Ankara, Turkey

Publication: Multimedia Content Representation, Classification and Security, MRCS 2006, İstanbul, Turkey, September 11-13, 2006

A novel approach is presented in order to reject correspondence outliers between frames using the parallax-based rigidity constraint for epipolar geometry estimation. In this approach, the invariance of 3-D relative projective structure of a stationary scene over different views is exploited to eliminate outliers, mostly due to independently moving objects of a typical scene. The proposed approach is compared against a well-known RANSAC-based algorithm by the help of a test-bed. The results showed that the speed-up, gained by utilization of the proposed technique as a preprocessing step before RANSAC-based approach, decreases the execution time of the overall outlier rejection, significantly.

3.2.6 Towards 3-D Scene Reconstruction from Broadcast Video

Authors: Evren İmre[1], Sebastian Knorr[2], Burak Özkalaycı[1], Uğur Topay[1], A.Aydın Alatan[1], Thomas Sikora[2]

Institutions: [1] Department of EEE, Middle East Technical University, Ankara, Turkey, [2] Communication Systems Group, Technische Universitat Berlin, Berlin, Germany

Publication: Signal Processing: Image Communication 22 (2007), 108–126

Three-dimensional (3-D) scene reconstruction from broadcast video is a challenging problem with many potential applications, such as 3-D TV, free-view TV, augmented reality or three-dimensionalization of two-dimensional (2-D) media archives. In this paper, a flexible and effective system capable of efficiently reconstructing 3-D scenes from broadcast video is proposed, with the assumption that there is relative motion between camera and scene/objects. The system requires no a priori information and input, other than the video sequence itself, and capable of estimating the internal and external camera parameters and performing a 3-D motion-based segmentation, as well as computing a dense depth field. The system also serves as a showcase to present some novel approaches for moving object segmentation, sparse and dense reconstruction problems. According to the simulations for both synthetic and real data, the system achieves a promising performance for typical TV content, indicating that it is a significant step towards the 3-D reconstruction of scenes from broadcast video.

3.2.7 Rate-Distortion Based Piecewise Planar 3D Scene Geometry Representation

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

Institutions: Department of Electrical & Electronics Engineering, METU, Ankara, TURKEY, * Department of Computer Engineering, Bilkent University, Ankara, TURKEY

Publication: IEEE 14th International Conference on Image Processing (ICIP'07), San Antonio, Texas, USA, 16.-19. 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.

3.2.8 Super-Resolution Reconstruction using Non-uniform to Uniform Resampling in Spline Spaces

Author: Harish Essaky Sankaran

Institutions: Department of Signal Processing, Tampere University of Technology, Finland

Publication: MSc thesis

Reconstructing a signal from its non-uniformly sampled values is an often encountered problem in applications such as image super-resolution, virtual view generation in multi-view imaging and structure-from-motion based 2D-to-3D video conversion. A general setting assumes a set of (potentially blurred and noisy) low resolution images which are sub-pixel shifted from each other. These sub-pixel shifted images, put together, form a non-uniform grid with respect to the desired (super-resolution) grid. In this thesis, the problem of reconstructing a signal from its non-uniform samples in shift-invariant spaces is studied and two reconstruction techniques are proposed: a near least squares separable reconstruction technique and a near least squares two-dimensional reconstruction technique, both not previously applied to the super-resolution problem. The separable reconstruction technique is suitable for the special case of interlaced sampling which corresponds to translational motion between the low resolution images. The second technique is an extension of the first one to two dimensions and handles arbitrary motion between the low resolution images. Efficient realizations of both techniques are achieved by digital filtering using the transposed Farrow structure. Both techniques are very computationally efficient and demonstrate good performance compared with other recent resampling methods.

3.2.9 Conclusions and future plans

The novel methods in Sections 3.2.1, 3.2.2 and 3.2.3 yield quite attractive multi-view sequences of a scene without completely determining 3D information. In this manner, the requirement of explicitly determining 3D structure at all scene points become obsolete, although the proposed method requires limited depth range for better visualization.

On the other extreme, 3D scene structure could still be estimated, even from uncalibrated broadcast video, by successively applying self-calibration, object segmentation and dense depth estimation algorithms to the input, as well as 3D orientation estimation between views. However, there are still cases in which the observed motion is degenerate or scene has insufficient number of interest points to determine reliable epipolar geometry. These problems are still under consideration to yield much robust estimates.

3.3 Human Face and Body

Since the human face and body is an integral part of almost any 3DTV applications, capturing a 3D structure of human face and body is one of the most important research activity. The apriori knowledge about 3D structure and motion of human face and body can be used to make the algorithms more robust and efficient. Facial expression analysis and synthesis techniques have received increasing interest in recent years. Numerous new applications can be found, for instance in the motion picture/broadcast industry for animating 3D characters and low bit-rate communications. Detection and tracking of observer's view point is also needed to render the correct view according to the observer position in 3D display systems. In this section the human face and body specific techniques are presented. The section is organized as follows. In the subsection 3.3.1 an overview about advances in tracking and recognition of human motion is given. Subsection 3.3.2 describes bilinear models for 3D face and facial expression recognition. In the subsection 3.3.3 an algorithm for detection of faces and facial features in images is presented. Subsection 3.3.4 devotes to estimation and analysis of facial animation parameter patterns. In the subsection 3.3.5 a framework for joint music and video analysis for automatic human body motion synthesis is presented. Section 3.3.6 summarizes an MSc thesis work on face and facial feature detection.

3.3.1 Advances in Tracking and Recognition of Human Motion

Authors: Niki Aifanti, Angel D. Sappa, Nikos Grammalidis,and Sotiris Malassiotis

Institutions: ITI-CERTH

Publication: paper submitted to “Encyclopedia of Information Science and Technology, Second Edition” book.

Tracking and recognition of human motion has become an important research area in computer vision. In real world conditions it constitutes a complicated problem, considering cluttered backgrounds, gross illumination variations, occlusions, self-occlusions, different clothing and multiple moving objects. These ill-posed problems are usually tackled by making simplifying assumptions regarding the scene or by imposing constraints on the motion. Constraints such as that the contrast between the moving people and the background should be high and that everything in the scene should be static except for the target person, are quite often introduced in order to achieve accurate segmentation. Moreover, the motion of the target person is often confined to simple movements with limited occlusions. In addition, assumptions such as known initial position and posture of the person are usually imposed in tracking processes. In this report, to be submitted later as a “review paper” of the field, recent techniques are reviewed, with emphasis in tracking techniques based on monocular or multiple camera image sequences and pose recovery applications. Furthermore, advances in the area of Human Motion recognition techniques are also reviewed, either based on the extracted 3D pose parameters, or on low-level features (e.g. silhouettes). Techniques extensively used for recognizing human actions include Template matching algorithms, State-space approaches (e.g., HMMs) and techniques based on semantic or other higher level descriptions (e.g., natural language). Finally, a few of the recent applications of vision-based human motion tracking and recognition techniques are also discussed.

3.3.2 Bilinear Models for 3D Face and Facial Expression Recognition

Authors: Iordanis Mpiperis, Sotiris Malassiotis and Michael G. Strintzis

Institutions: ITI-CERTH

Publication: paper submitted to "IEEE Transactions On Information Forensics And Security*" journal.*

In this work, we explore bilinear models for jointly addressing 3D face and facial expression recognition. An elastic deformable model algorithm that establishes correspondence among a set of faces is proposed first. Then bilinear models that decouple the identity and facial expression factors are constructed. Fitting these models to unknown faces enables us to perform face recognition invariant to facial expressions and facial expression recognition with unknown identity. Experimental results on a public database demonstrate the superiority of this approach in comparison with other state-of-the-art techniques.

The main limitation of the proposed technique is the need of a large bootstrap set which should also be annotated with respect to facial expressions. The more different expressions are present in the bootstrap set, the better is the fit to the novel faces. Training with a few expressions leads to unbalanced generalization ability in favour of identity which leads to better surface approximation but poorer expression control. However, building and annotating in practice a bootstrap set may be difficult considering that a great number of possibly ambiguous expressions have to be classified into a finite number of expression classes.

Another factor that affects performance is the accuracy of point correspondence between faces. In our experiments we observed that poor correspondence affects substantially bilinear model training and eventually recognition performance. The problem is twofold: During training, the bilinear model cannot learn the true identity-expression manifold implying errors

in bilinear parameters estimation. During testing, expression manipulation is actually applied on a slightly (or quite) different face. This error is further amplified by inaccurate bilinear parameters leading to a distorted facial surface.

3.3.3 Automatic Detection of Face and Facial Features Using Scalable Filters

Authors: S. Piekh

Institutions: UHANN (Leibniz University of Hannover)

Publication: internal report

A method for detection of face and facial features in images is presented. The proposed method avoids the application of luminance templates and the problems associated with the generation and properly rotation and scaling of templates. Special scalable filters are used to calculate the energy function, that indicates the probability for eye and mouth regions. Then the candidates for eye and mouth areas are extracted. Finally the proper triplet of areas for both eyes and mouth is estimated on basis of human face geometry. The eyes and mouth regions indicates the face. The algorithm shows a detection rate 88% on used test images.

3.3.4 Estimation and Analysis of Facial Animation Parameter Patterns

Authors: Ferda Ofli, Engin Erzin, Yucel Yemez, and A. Murat Tekalp

Institutions: Koç University

Publication: IEEE Int. Conf. on Image Processing (ICIP'07), San Antonio, 2007

State of the art visual speaker animation methods are capable of generating synchronized lip movements automatically from speech content; however, they lack automatic synthesis of speaker emotions and gestures from speech. Head gestures and face expressions are usually added manually by artists, which is costly and may look unrealistic.

In this work we propose a framework for estimation and analysis of temporal facial expression patterns of a speaker. The proposed system aims to learn personalized elementary dynamic facial expression patterns for a particular speaker. These facial expression patterns coded by FAP sequences can be used for personalized emotion estimation or realistic synthesis of a speaker. We use head-and-shoulder stereo video sequences to track lip, eye, eyebrow, and eyelid motion of a speaker in 3D (see Fig. 30). MPEG-4 Facial Definition Parameters (FDPs) are used as the feature set, and temporal facial expression patterns are represented by the MPEG-4 Facial Animation Parameters (FAPs) (see Fig. 31). We perform Hidden Markov Model (HMM) based unsupervised temporal segmentation of upper and lower facial expression features separately to determine recurrent elementary facial expression patterns for a particular speaker. The block diagram of the proposed scheme is displayed in Fig. 32. These facial expression patterns coded by FAP sequences, which may not be tied with prespecified emotions, can be used for personalized emotion estimation and synthesis of a speaker.

We have conducted experiments using the MVGL-MASAL database. The database includes four recordings of a single person telling stories in Turkish. Experimental results show that the semantic face patterns can successfully be clustered using the proposed system. Video sequences for the temporal facial expression patterns are available online at http://mvgl.ku.edu.tr/faceanalysis.

Fig. 30. *Example image sequence that demonstrates the performance of tracking by Active Appearance Models.*

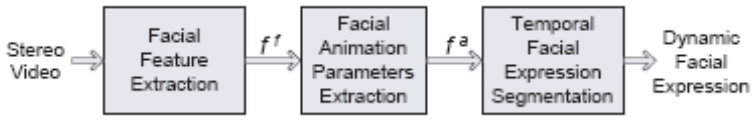


Fig. 31. *Block diagram of the proposed analysis system.*

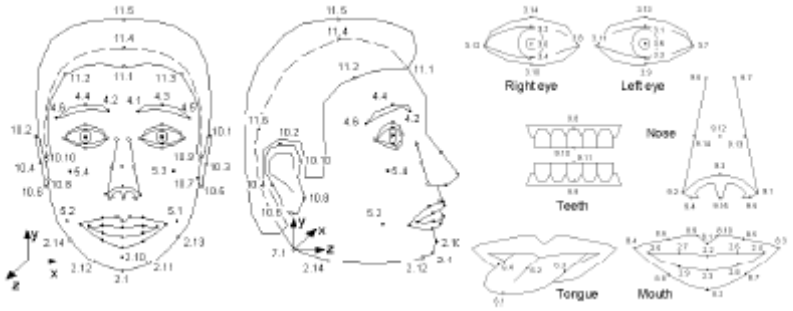


Fig. 32. *The set of MPEG-4 Facial Definition Parameters. There are 84 feature points on morphological places of the neutral head model.*

3.3.5 Music and Video Analysis for Automatic Human Body Animation

Authors: F.Ofli, E.Bozkurt,, Y.Yemez, E.Erzin, A. M.Tekalp, Ç.E.Erdem, A.T.Erdem, M.K.Özkan

Institutions: Koç - Momentum

Publication: internal report

This report presents a framework for joint music and video analysis for automatic human body motion synthesis. We address the problem in the context of a dance performance, where gestures and movements of the dancer are mainly driven by a musical piece and characterized by the repetition of a set of dance figures. The system is trained in a supervised manner using the multi view video recordings of the dancer. Music is analyzed to extract beat and tempo information. The joint analysis of music and motion features provides a correlation model that is then used to animate a dancing avatar when driven with any musical piece of the same genre. Experimental results demonstrate the effectiveness of the proposed algorithm.

3.3.6 Algorithms for Face and Facial Feature Detection

Author: Xinghan Luo

Institutions: Department of Signal Processing, Tampere University of Technology, Finland; TNT, Leibniz University of Hannover, Germany

Publication: MSc thesis (jointly supervised at TUT and UHANN)

Automatic face detection and facial feature localization are important computer vision problems, which have challenged a number of researchers to develop fast and accurate

algorithms for their solutions. This MSc thesis, structured in two parts, addresses two modern approaches for face and facial feature detection, namely the cascade Adaptive Boosting (AdaBoost) face detection and Active Appearance Model (AAM) based facial feature detection. In the first part, the classical cascade AdaBoost method and its extensions are overviewed and a Matlab based test platform is described. To develop it, training and test face databases have been collected and generated. Existing Matlab packages for face detection have been modified to include a multiple detection elimination module and a graphical user interface for easy manipulation and demonstration. The demo software has been also extended with an OpenCV-based implementation. In the second part, facial feature detection as a first step in a facial animation application is studied. An AAM based facial feature detection system, aimed at lip motion and face shape detection/tracking has been targeted. Its development has included proper landmark definition for lip/face shape models, training/test set selection and marking, and modification of an existing AAM module built on the OpenCV library. Specifically, the detection accuracy has been improved by landmark optimization based on training set self-rebuilt convergent iterations. Detection/tracking results presented demonstrate a significant improvement compared with previous implementations. As the main contribution of this thesis project, an accurate facial feature gemotric database has been set up based on automatic accurate facial feature detection. It is expected to improve the performance of facial animation techniques being developed for 3DTV-related applications.

3.3.7 Conclusions and future plans

In this section the human face and body related techniques were presented.

Subsection 3.3.1 presents an overview about advances in tracking and recognition of human motion. In subsection 3.3.2 a technique for simultaneous 3D face and facial expression recognition is proposed. A novel model-based approach for establishing point correspondence among faces is presented which involves the solution of a simple linear system. It was proposed using directed distances both from the mesh to the cloud of facial points and inversely which leads to a smoother force field and thus more plausible anatomical correspondence. Another advantage of this approach is that correspondence may be performed fully automatically after training the system with a number of facial surfaces annotated with anatomical salient points. It was provided a solution for the problem of open mouth in this case whose configuration might cause problems during the establishment of correspondence. Then, it was proposed bilinear models for face and expression recognition and the general solution of the error minimization during the training of the symmetric bilinear model. The algorithm was finally evaluated on a public database under tough conditions in case of face recognition and proved its superiority over other relevant techniques. An algorithm for detection of faces and facial features in images was presented in subsection 3.3.3. The method uses the special scalable filters for detection and avoids the application of luminance templates and the problems associated with this technique. The algorithm shows a detection rate 88%. In subsection 3.3.4 a dynamic facial expression analysis system is presented, that extracts personalized recurrent elementary expression patterns for a specific user. The system successfully tracks a user's 3D head position and orientation, as well as 3D dynamic facial expressions by using a stereo camera setup. In the analysis, the elementary facial expression patterns for a specific speaker are defined. Experimental results indicate that facial expressions vary from person to person, and even in time for the same person. Thus, the proposed analysis proves extremely valuable for synthesis of personalized facial expressions rather than using a generic set of facial expressions for all people. The future research direction will be towards the goal of synthesizing a realistic

audio-driven facial animation scheme that also includes facial expressions. Subsection 3.3.5 presents a novel framework for audio-driven human body motion analysis and synthesis. The problem is addressed in the context of dance performance and considered a simple scenario possible in which only a single dance figure is associated with each musical genre. The dancing avatar has been trained for salsa and belly. 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. Future research within this topic involve unsupervised training of the 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. To achieve this, various musical audio features other than beat and tempo, such as tonality, harmony and melody has to be considered. The facial feature detection system described in Subsection 3.3.6 will be utilized in a face animation system being developed at TNT, UHANN.

3.4 Holographic Camera Techniques

A widely spread way to capture 3D data is based on triangulation. In its easiest way a point is projected onto the surface under investigation from an oblique direction relative to the observation direction. From the lateral displacement and with the help of the known angle between the two directions the distance of the illuminated point from a calibrated point in the observing system can be calculated. The single point can be replaced by a full line in the so-called light intersection method. From the recorded form of the projected line the contour of the object along this line is determined. The method can be further improved by projecting several lines and by introducing gray-values. Furthermore we can modify these fringes which not necessarily must be straight lines.

In 3.4.1 a 3D-camera based on these principles is developed at BIAS in cooperation with an industrial partner. The special design of this camera enables measuring even in the interior of tubes and is especially suited for augmented reality applications. 3.4.2 describes in detail a new method for calibrating a 3D-camera based on the fringe projection principle.

The fringe projection techniques for 3D capturing work with white light. They would work also with coherent light but the inevitable speckles would disturb severely. On the other hand coherent light enables interferometric methods for capturing 3D data, the most prominent of these being holography. In holography the wave field is encoded in an interference pattern using a mutually coherent reference wave. While in holography this reference wave can be nearly arbitrary, there is a special way where a shifted version of the original wave field is used as reference. In this method called shearography we speak about self-reference. The special advantages of shearography is the robustness against lateral rigid body motions of the object relative to the capturing optics.

In 3.4.3 possible approaches to digital holographic 3D-TV is outlined. The proposed systems range from partial application of holography in capture or reconstruction and the transmission of holographic data or already reconstructed 3D data to the all holographic approach with digital holographic capture by CCD, transmission of the digital holograms and holographic reconstruction using a spatial light modulator. The special problems which have to be solved are indicated.

In 3.4.4 methods for determining and compensating out-of-plane displacements which may disturb the applications are described. These procedures are based on the numerical Mellin transform. The digital holographic method used here is the lensless Fourier transform holography, where the (virtual) source of a divergent reference beam is placed in the plane of the object with the result of an easy reconstruction by the Fourier transform instead of a Fresnel transform.

3.4.5 describes a compact shearing interferometer where the shearing is accomplished by a reflective liquid crystal spatial light modulator. The special advantages of this approach are demonstrated.

3.4.1 3D-camera for Scene Capturing and Augmented Reality Applications

Authors: T. Bothe, A. Gesierich, W. Li, C. v. Kopylow, N. Köpp, W. Jüptner

Institutions: BIAS – Bremer Institut für Angewandte Strahltechnik; VEW – Vereinigte Elektronikwerkstätten Bremen

Publication:Proceedings of 3DTV Conference, Kos, 2007

Coordinate based 3D multimedia applications benefit from cost effective, compact and easy to use profilers like the miniaturized 3D-camera that works on basis of the fringe projection technique. The system uses a compact housing and is usable like a video camera with minimum stabilization like a tripod. Camera and projector are assembled with parallel optical axes having coplanar projection and imaging planes. Their axes distance is comparable to the human eyes' distance, giving a compact system of shoebox-size and allow measuring high gradient objects like the interior of tubes and delivering captured scenes with minimum loss by shadowing.

Additionally, the 3D-camera can be used for the inverse projection technique, allowing single-frame video rate capture and to virtually place information like virtual labels or defect maps onto the surface of objects, thus, allowing augmented reality applications.

In the paper, the concept and realization for the 3D-camera is described and an overview of possible applications is given.

3.4.2 Beam Based Calibration for Optical Imaging Device

Authors: W. Li, M. Schulte, T. Bothe, C. v. Kopylow, N. Köpp, W. Jüptner

Institutions: VEW – Vereinigte Elektronikwerkstätten Bremen; BIAS – Bremer Institut für Angewandte Strahltechnik

Publication: Proceedings of 3DTV Conference, Kos, 2007

Well calibrated optical imaging devices are needed in 3DTV. An optical imaging device, for example camera and projector, maps beams in object space into 2D points in image space. The mapping function, or lens distortion, must be calibrated in order to build the correct object-image relation. Current calibration methods try to model the distortion by analytical

functions. This will easily fail when the distortion is irregular, especially near the edge of the view field.

In this paper, we propose a new flexible technique to easily and accurately calibrate single or combined optical imaging devices. We directly calculate the beam's vector in a device-based coordinate system for each sensor pixel element without analytically modeling the lens distortion. A TFT monitor, which displays fringe patterns, is used as calibration object. The imaging device is mounted on a computer controlled shift/rotation setup and observes the monitor at few (>2) different positions.

The calibration for a fringe projection system has been implemented and the results will be demonstrated and compared to classical techniques to show the improvement of the new technique.

3.4.3 Digital Holography Methods in 3D-TV

Authors: Thomas Kreis

Institutions: BIAS – Bremer Institut für Angewandte Strahltechnik

Publication: Proceedings of 3DTV Conference, Kos, 2007

A promising approach to 3D-TV is the concept of digital holography. One way is to record the interference pattern which is generated by superposition of the wave field reflected from the scene and a mutually coherent reference wave by a CCD- or CMOS-array, transmitting the data, and displaying the wave field by feeding the data to a spatial light modulator which is illuminated with coherent light. Another way is to perform a numerical reconstruction and to transmit and display the reconstructed 3D-data by a non-holographic method. Both these ways require an analysis of their limits and of the reconstruction algorithms before a practical implementation of digital holographic methods for 3D-TV is started. Conservative estimates of the limits rest on the sampling theorem while recent results have shown that reliable reconstructions are also possible even with sub-sampling for moderate boundary conditions. Here the preliminaries of an all-digital-holographic approach to 3D-TV are given and some implications are discussed.

3.4.4 Determination of Large-Scale Out-of-Plane Displacements in Digital Fourier Holography

Authors: E.Kolenovic, Th. Kreis, C. v. Kopylow, W. Jüptner

Institutions: BIAS – Bremer Institut für Angewandte Strahltechnik

Publication: Applied Optics, vol. 46, no 16, 3118 – 3125, 2007

A novel approach for the determination of large-scale out-of-plane displacements from digital Fourier holograms is presented. The proposed method is invariant to lateral object shifts. It is based on the determination of the scaling of the reconstructed image that occurs when the recording distance is changed. For a precise determination of the scaling factor, we utilize the Mellin transform. After the discussion of mathematical and computational issues,

experimental results are presented to verify the theoretical considerations. The results show that displacements of at least up to 8.4% from the initial recording distance can be detected with this approach. The displacements could be determined with a deviation of typically less than 1.0% from the set values.

3.4.5 Compact Lateral Shearing Interferometer to Determine Continuous Wave Fronts

Authors: C. Falldorf, C. v. Kopylow, W. Jüptner

Institutions: BIAS – Bremer Institut für Angewandte Strahltechnik

Publication: Proceedings of 3DTV Conference, Kos, 2007

In this work we present a very compact implementation of a lateral shearing interferometer which makes use of the birefringent properties of a reflective liquid crystal spatial light modulator (SLM). The main advantage of this approach is its considerable tolerance against environmental disturbances and its high degree of flexibility. A great variety of lateral and/or radial shears can be realized without the requirement of moving parts.

The setup can be used to determine continuous wave fronts reflected by coherently illuminated objects. Thus it is capable of acquiring 3-dimensional interferometric data in a very flexible and robust way. Experimental results are presented as an example of application. The lateral phase distribution of a disturbed wave front is recorded with the proposed interferometer.

3.4.6 Conclusions and future plans

The coherent optical methods offer good prospects in 3D-TV. The partners were able to contribute to several very basic problems (e.g. in calibration) and are happy to report, that many of the existing problems are addressed in the European joint project Real3D, which should start February 2008. In the consortium of this project two partners of the NoE 3DTV are participating: Bilkent and BIAS.

3.5 Pattern Projection

Information about the 3D coordinates of an object in pattern projection systems is encoded in the phase of a periodic fringe pattern which has been projected onto the object and deformed by its surface. Hence accuracy of 3D object capture is directly related to accuracy of phase retrieval from the recorded fringe patterns. A survey of different phase-retrieval methods was made in the invited paper reported in Section 3.5.1. The aim of the work performed during the reported in D26.3 period was to prove applicability of a sinusoidal phase grating as a pattern projection element in a multi-wavelength and multi-camera 3D profilometric system based on a phase-shifting algorithm. Satisfactory performance of such a system requires a sinusoidal

profile and equal contrast of the projected fringes as well as an equal background in the recorded by different cameras fringe patterns. To fulfill the research goal we evaluated the spectral content of the diffractive pattern created by the sinusoidal phase grating in the Fresnel zone for collimated and divergent beam illumination and simulated performance of the multiwavelength system for different test objects (see Sections 3.5.3-3.5.4). The accuracy of phase retrieval and shape restoration evaluated from the simulations gave the acceptable regions of variation of grating parameters. The experiments in a static mode of operation of the system were also performed.

3.5.1 Pattern Projection Approach and Time-of-Flight Range Imaging

Institutions: CLOSPI-BAS, ITI-CERTH

Publication: **Section V** and **Section VI** in the invited paper E. Stoykova, A.Alatan, P. Benzie, N. Grammalidis, S. Malassiotis, J. Ostermann, S. Piekh, V. Sainov, C. Theobalt, T. Thevar, X. Zabulis, "3-D Time-Varying Scene Capture Technologies – A Survey", **IEEE TCSVT**, vol.17, November, 1568-1586 (2007).

Section V of the invited paper presents a survey of pattern projection techniques, in which information of the object shape and color is encoded in a 2D pattern which is projected onto and reflected from the object, for the case of structured light and sinusoidal fringe projection. Pattern projection techniques such as coded light or fringe projection for real-time extraction of 3D objects positions and color information could manifest themselves as a low cost alternative to traditional camera-based methods. The extraction of point clouds and color coordinates of the recorded objects out of 2D images by using a one-shot system relies on development of recording algorithms from a single image or on design of special methods for simultaneous acquisition of several images. There are also some active imaging devices capable of 3D extraction such as the 3D time-of-flight camera, which provides 3D image data of its environments by means of a modulated infrared light source. Section VI of the invited paper outlines the principle and some basic issues of time-of-flight range imaging systems. Resolution achieved by the modern time-of-flight range-imaging systems which are capable of determining the distance map, as well as the local brightness in the scene in real time makes them a useful tool in middle accuracy applications of computer vision.

3.5.2 Pattern Projection with a Sinusoidal Phase Grating

Authors: Elena Stoykova, Jana Harizanova, Ventseslav Sainov

Institutions: CLOSPI-BAS

Publication: presented as an oral presentation at ***3DTV conference, 7-9 May 2007***, Kos island, Greece and published in IEEE proceedings of the conference.

In this work we derived an expression for the complex amplitude of coherent light distribution in the Fresnel diffraction zone behind a thin phase grating with purely sinusoidal modulation at illumination with a unit-amplitude normally incident plane wave. The expression was used to analyze the frequency content of the projected fringes (Fig. 31) at different grating parameters and wavelengths and to show applicability of the phase-stepping algorithm. The

intensity distribution exhibits periodicity in lateral and longitudinal directions. In longitudinal direction the grating reproduces itself at Talbot planes. The results of test measurements of relative 3D coordinates performed with interferometrically recorded on holographic plates sinusoidal phase gratings were also presented for the case of a single wavelength illumination.

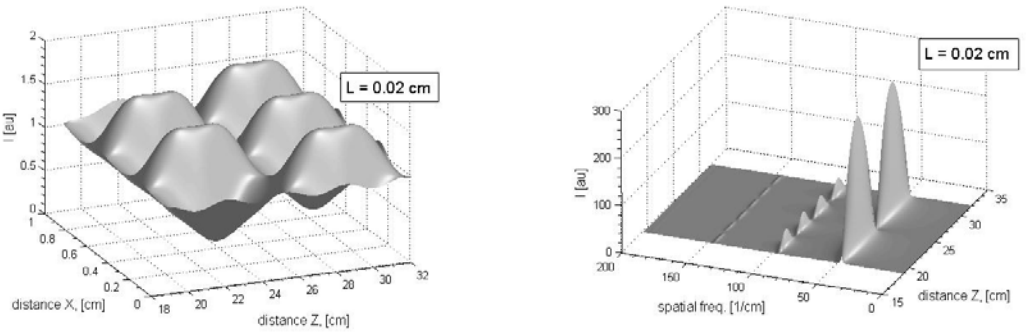


Fig. 33. *Intensity distribution (left) and spatial frequency spectrum (right) of the light transmitted by a sinusoidal phase grating with spacing L = 0.02 cm in the Fresnel zone as a function of the distance behind the grating at illumination with a collimated beam. For convenience, the part of the spectrum at the fundamental frequency is omitted. The wavelength is λ = 633 nm.*

The fact that the spacing of fringes created by the sinusoidal phase grating does not depend on the wavelength (Fig. 32) can be used to illuminate the object by spatially similar fringes at different wavelengths in order to record simultaneously the deformed fringe patterns by separate CCD cameras and to overcome the main drawback of the temporal phase-shifting profilometry in which patterns acquisition is made successively in time. As a second task, we evaluated the depth of the spatial zone in which the multi-wavelength illumination yielded fringe patterns of equal contrast.

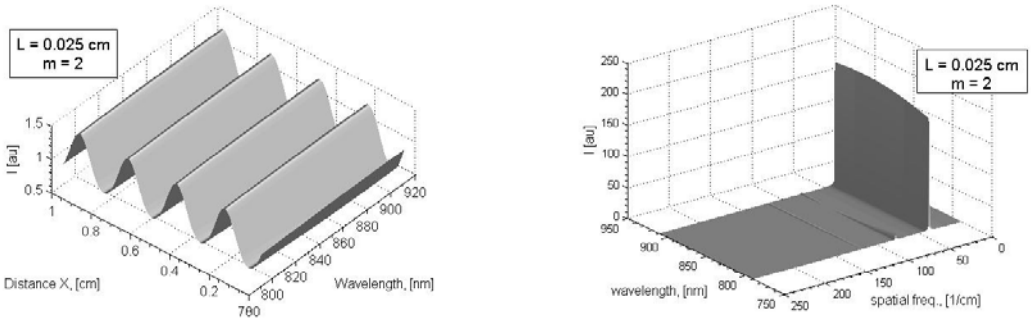


Fig. 34. *Intensity distribution (left) and the spatial frequency spectrum (right) of the light transmitted by a sinusoidal phase grating with spacing L = 0.025 cm in the Fresnel zone as a function of the wavelength at 10.5 cm behind the grating at illumination with a collimated beam.*

3.5.3 Real-time Multi-Camera System for Measurement of 3D Coordinates by Pattern Projection

Authors: Ventseslav Sainov, Elena Stoykova, Jana Harizanova

Institutions: CLOSPI-BAS

Publication: oral presentation at ***Optical Metrology'2007***, 17-22 June 2007, Munich, Germany; published in Proc. SPIE, vol 6616, 2007, 6616OA (6616-08)

The work continues the study of a sinusoidal phase grating as a projection element in pattern projection profilometry. The goal was to obtain values of the parameters which ensured a sinusoidal profile of the projected fringes in order to capture 3D objects without systematic errors. If the grating is over modulated, it creates fringe patterns with increased amplitudes of the second, third and higher harmonics. Analysis of the frequency content of the fringes gave the maximal modulation of the phase grating that was acceptable for accurate phase-shifting measurement.

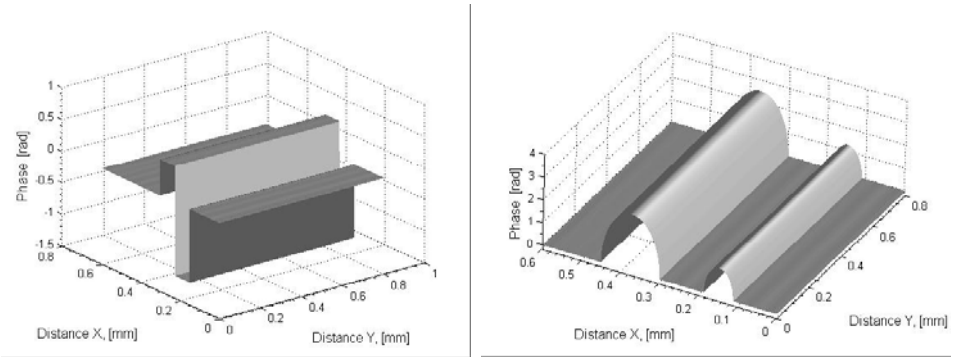


Fig. 35. *Reconstruction of 3D objects with a four-step phase-stepping algorithm from four fringe patterns generated at □1 = 790 nm, □2 = 810 nm, □3= 850 nm and □4 = 910 nm and shifted at π/2 by using four identical phase gratings with a spatial period L = 0.025 cm and low modulation parameter.*

Operation of a four-wavelength profilometric system with four spatially phase-shifted at π/2 sinusoidal phase gratings illuminated with four diode lasers at wavelengths 790 nm, 810 nm, 850 nm and 910 nm was simulated (Fig.33). Systematic phase error in reconstruction of a plane surface from four fringe patterns projected at 790 nm, 810 nm, 850 nm and 910 nm respectively using four identical sinusoidal phase gratings was studied at increasing grating modulation and at the presence of a Gaussian white noise.

3.5.4 Pattern Projection with a Sinusoidal Phase Grating

Authors: Elena Stoykova, Jana Harizanova, Ventseslav Sainov

Institutions: CLOSPI-BAS

Publication: submitted to EURASIP Journal on Advances in Signal Processing

This work studies the fringes created by a sinusoidal phase grating for the more important case of divergent light illumination. With increase of the angle of divergence the number of Talbot planes diminishes and the spatial region where the frequency content of the fringes is practically constant becomes substantially larger than in the case of collimated illumination. Thus coordinates of larger objects can be measured. Since the distance between the light source and the grating became a crucial parameter for the quality of fringes, the fringe profiles were calculated for a spherical wave in paraxial approximation and were recorded with a CCD camera at different values of this parameter in order to find the optimal geometry.

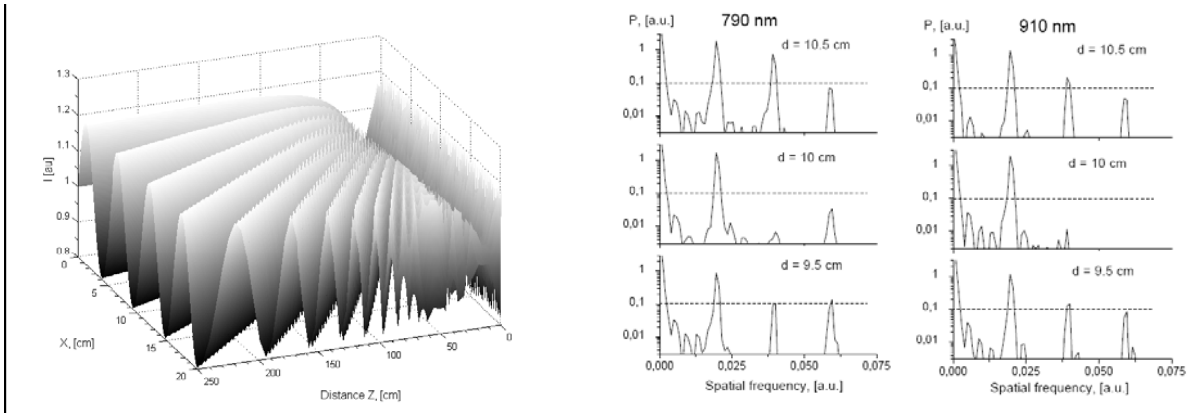


Fig. 36. *Intensity distribution (left, calculation) and frequency content (right, experiment) of patterns created by a sinusoidal phase grating at divergent light illumination. The parameter d is the distance between the light source and the grating.*

The systematic errors in reconstruction of a plane surface were compared for:

i) single wavelength and multi wavelength illumination;

ii) collimated and divergent light illumination.

It was shown that for a collimated single wavelength illumination the error, which in this case is due only to higher harmonics, showed periodicity along the distance from the grating. In the case of collimated multi-wavelength illumination the error was also due to the different contrast of the projected patterns being acceptable in a very narrow region. In the case of divergent illumination the error could be kept quite small in a large spatial region which was a promising result for capture of large objects. The first experimental results obtained for a multi-wavelength plane surface reconstruction as well as simulation of a multi-wavelength 3D reconstruction of a dome (Fig. 35) proved the conclusion that the sinusoidal phase grating could be a suitable projection element in the pattern projection profilometry.

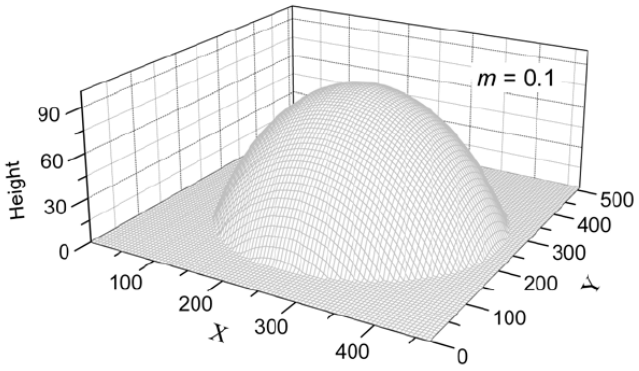


Fig. 37. *Reconstruction of a 3D surface by a phase-shifting technique from four fringe patterns generated with four-wavelength divergent illumination. The projection distance is 1.3 m; the grating spacing is L = 0.025 cm.*

3.5.5 Conclusions and future plans

The research efforts within the "Pattern projection" section for the reported period were oriented towards accurate retrieval of 3D absolute coordinates with a fringe projection profilometric system which combines single-shot acquisition with phase shifting processing algorithm for "real time" surface measurement. The system relies on a sinusoidal phase grating which creates periodic fringe patterns with the same spacing at different wavelengths. This advantage of the sinusoidal phase grating as a projection element permits to develop a simple portable device using four single mode diode lasers emitting at four different wavelengths λ_1 = 790 nm, λ_2 = 810 nm, λ_3 = 850 nm and λ_4 = 910 nm and four CCD cameras for simultaneous recording of the fringe patterns. The main advantages of coherent illumination by NIR diode lasers are connected with higher efficiency, larger focal depth and fringes contrast, that lead to higher sensitivity and accuracy of measurement. The detailed technical description of the system is given in D26.2. The problem of higher harmonics influence has been thoroughly analyzed. Both cases of collimated and divergent beam illumination have been studied. Careful maintaining of grating modulation could ensure vanishing contribution of higher diffraction orders. The additional source of error is the difference in the contrast of projected patterns at multi-wavelength illumination. The performance of the proposed four-wavelength system was checked by simulation of reconstruction of a plane and a 3D object (dome) for four identical gratings with L = 0.025 cm at different values of the grating modulation parameter. The results were experimentally verified. The made analysis confirms that the sinusoidal phase grating can serve as a projection element in a profilometric system as well as to be used at multi-wavelength illumination. An appropriate software for fringe pattern processing – phase retrieval, denoising, filtration, unwrapping as well as for 3D visualization of reconstructed surfaces has been developed.

The future research in this area will concentrate on experiments with static and moving objects for a multi-wavelength illumination. We plan to perform 3D capture of micro-objects by using a collimated illuminating beam. In the case of a scene with large objects the error introduced by the divergence of light and hence the different spacing of fringes in depth should be analyzed. On the basis of the results reported in D26.2 and D26.3 we envisage to develop a more sophisticated system which includes two multi-wavelength projection modules with time multiplexing for a double symmetrical illumination of the scene to avoid shadowing. Each module has four diode lasers emitting in the near infra-red and four identical phase gratings as projection elements. The modules differ only by the spacing of the used sinusoidal phase gratings to achieve absolute coordinate measurement without a reference plane. In addition, for capture of color coordinates we plan to use a white-light illumination (only visible part of the spectrum) of the scene and RGB recording by a color CCD camera calibrated to yield the same point cloud as provided by the four monochrome CCD cameras which are used for 3D coordinates measurement. Different tasks connected with calibration of the system and transforming of the data in a global coordinate system should be solved.

3.6 Motion Analysis and Tracking

First, a new video sensor for real-time motion detection for a specific interest area in scenes is proposed which are used such in traffic monitoring systems. The goal is the extension of the auto-scope vehicle detection system by improving the background subtraction method of the polygon-shaped detection. The detection and tracking of multiple objects techniques can be integrated in smart cameras which are used in broad field of multiple cameras installed. A

network of these intelligent cameras are usable such for traffic control of the aircraft parking area (APRON) at airports. These techniques are consecutively described in both subsections.

3.6.1 An Efficient Sensor for Traffic Monitoring and Tracking Applications Based on Fast Motion Detection at the Areas of Interest

Authors: Nikolaos Zournis-Karouzos, Alexandra Koutsia, Kosmas Dimitropoulos and Nikos Grammalidis

Institutions: Aristotle University of Thessaloniki, ITI-CERTH (Informatics and Telematics Institute – Centre of Research and Technology Hellas)

Publication: To be presented at VISAPP International Conference on Computer Vision Theory and Applications, January 2008, Funchal, Madeira – Portugal

A novel video sensor for real-time motion detection at specific user-defined regions of interest is proposed, designed primarily for traffic monitoring, surveillance and tracking applications. The ultimate goal is to extend the capabilities and to alleviate shortcomings of embedded motion detection video sensors (like Autoscope®) for target tracking and surveillance applications, including road traffic monitoring or Advanced Surface Movement, Guidance and Control Systems (A-SMGCS) at airports. Specifically, the new sensor a) supports virtual detectors with a generalized (polygonal) shape, thus providing additional flexibility in the design of detector configurations, b) is based on fast implementations of recent state-of-the art background extraction and update techniques and c) constitutes a generic, inexpensive software solution, which can be used with any video camera. Figure 36 illustrates an example how it will look like

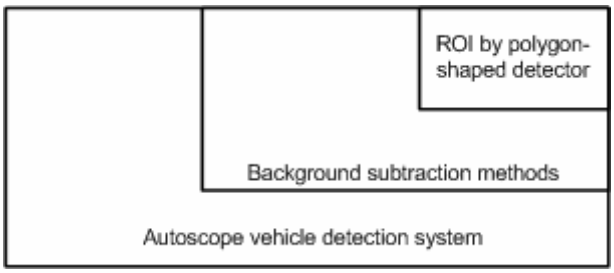


Fig. 38. *Background subtraction methods as a component in an Autoscope system.*

Four state-of-the-art background modelling, subtraction and update techniques were extended so that they are applied only within specific regions of interest, defined by a set of polygon-shaped detectors. Experimental results have shown that the extension results to a very significant reduction of the complexity and execution times, therefore even techniques with increased computational complexity, like the Bayes-based or the Non-parametric Model approaches can be considered suitable for integration in real-time systems using the proposed technique. The new video sensor meets the expectations in terms of real-time performance and demonstrates the additional functionalities, according to which it was designed. The final goal is to use this new sensor as an alternative, improved version of the Autoscope video sensors for the targeted applications.

3.6.2 Traffic Monitoring Using Multiple Cameras, Homographies and Multi-Hypothesis Tracking

Authors: A. Koutsia, T. Semertzidis, K.Dimitropoulos, N. Grammalidis, A. Kantidakis, K. Georgouleas and P. Violakis

Institutions: ITI-CERTH (Informatics and Telematics Institute – Centre of Research and Technology Hellas), MARAC Electronics

Publication: An article submitted to EURASIP JASP 3DTV special issue, poster submitted to ISPRS 2008

Traffic control and monitoring using video sensors has drawn increasing attention recently due to the significant advances in the field of computer vision. However, robust and accurate detection and tracking of moving objects still remains a difficult problem for the majority of computer vision applications. Especially in case of outdoor video surveillance systems, the visual tracking problem is particularly challenging due to illumination or background changes, occlusions problems etc. In this paper, our aim is to present a novel multi-camera video surveillance system, which supports functionalities such as detection, tracking and classification of objects moving within the supervised area.

The system is based on a network of intelligent autonomous tracking units, which capture and process images from a network of pre-calibrated visual sensors. The results of the image processing are transmitted to a Sensor Data Fusion (SDF) server located in a traffic control centre through wired or wireless transmission (Figure 37). Subsequent steps of data fusion and target tracking are performed at the SDF server as well as ground situation display in order to facilitate human operators at the surveillance centre.

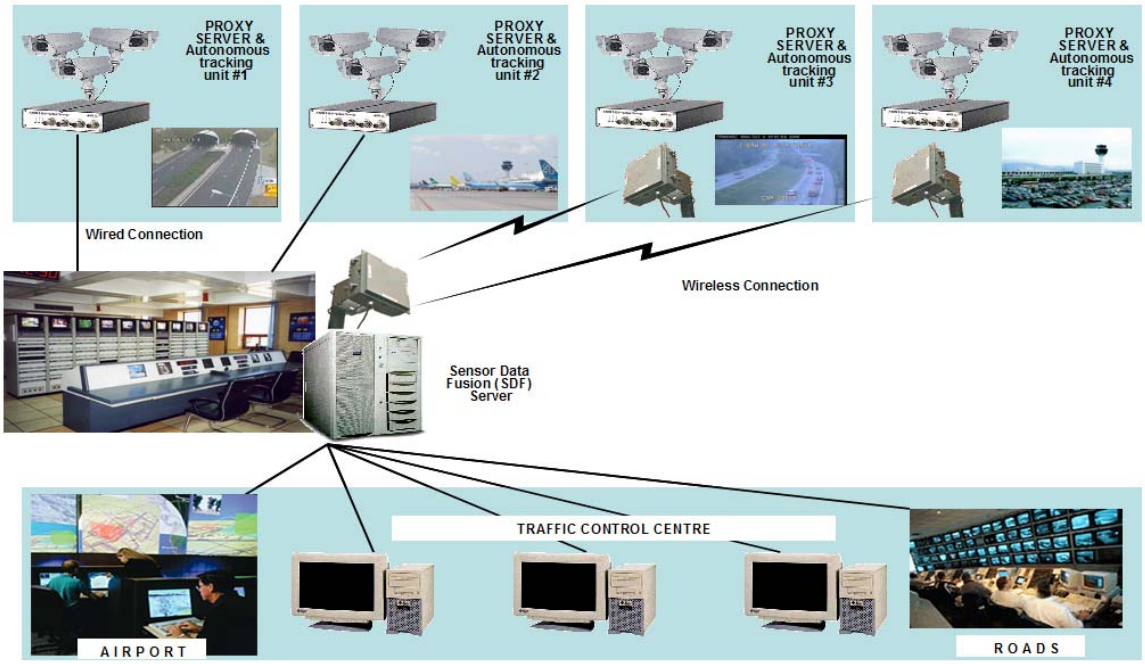


Fig. 39. *An overview of the TRAVIS system*

The fundamental goal is the development of a moving target tracking system which is fully scalable and parameterized and can be applied in a broad field of applications in the future. Two prototypes have been developed, each one focusing on a different aspect of traffic monitoring:

• Traffic control of tunnels at highways: This application aims to investigate the ability of the proposed system to monitor and control highway tunnels traffic in order to trace events that can lead to accidents. The tracing of such events leads to instant warning of drivers through special traffic lights placed at the entrance of the tunnel. The system is also able to provide the traffic control centre with statistical information about the traffic inside the tunnel (traffic volume, average vehicle speed etc) as well as warnings in case of accidents using a user friendly graphical interface.

• Traffic control of the aircraft parking area (APRON) at airports: This prototype is addressed to the APRON controllers, who are responsible for the control of movements (airplanes, cars, buses, humans, etc.) occurring at the aircraft parking area. The objective of this application is to provide airport supervision authorities with a ground situation display in order to facilitate the traffic management at the APRON. The system also provides alarms for the avoidance of accidents, thus increasing the safety levels at airports.

3.6.3 Conclusions and future plans

Systems for special applications such traffic monitoring and surveillance systems have been described in which detection, tracking of objects techniques are virtually integrated in smart cameras. Using of multiple cameras covering a broad field of view need a sophisticated tracking network system which has been proposed in this work.

Beside the improvements of the proposed systems a creation of a 3D synthetic representation of the scene under surveillance which could also be rendered at any 3D display is planned as a possible future work.

3.7 Object-based representation and segmentation

Object- based representation and segmentation is an important and challenging research area, which has many important applications including object-based video coding, video postproduction, content-based indexing and retrieval, surveillance, and 3D scene reconstruction for 3D TV.

The next section addresses two new region-based methods for object tracking using active contours in a dynamic programming framework. Next, a preprocessing algorithm for traditional chroma keying systems using a simple background illumination correction based approach for improving matting problems with uneven or poor lighting in the background using the computational power of GPU computing is presented. In the final abstract a modular framework for the above mentioned keying process is described.

3.7.1 Video Object Segmentation and Tracking Using Region Based Statistics

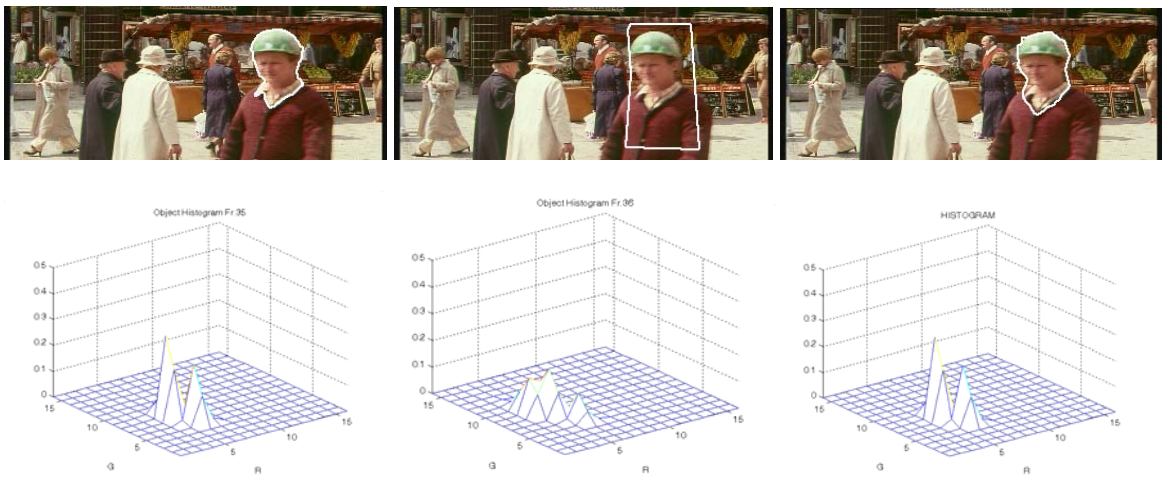
Authors: Çiğdem Eroğlu Erdem

Institution: Momentum

Publication: Signal Processing: Image Communication, Vol.22, No.10, pp.891-905, 200.712

Two new region-based methods for video object tracking using active contours are presented. The first method is based on the assumption that the color histogram of the tracked object is nearly stationary from frame to frame. The proposed method is based on minimizing the color histogram difference between the estimated objects at a reference frame and the current frame using a dynamic programming framework. The second method is defined for scenes where there is an out-of-focus blur difference between the object of interest and the background. In such scenes, the proposed "defocus energy" can be utilized for automatic segmentation of the object boundary, and it can be combined with the histogram method to track the object more efficiently. Experiments demonstrate that the proposed methods are successful in difficult scenes with significant background clutter.

A demonstration of color histogram matching using snakes is given in the following Figure 38.



(a) (b) (c)

Fig. 40. *A demonstration of color histogram matching using snakes applied on 'Market sequence'. (a) The user-initialized boundary in frame 35 (top) and the corresponding reference histogram (bottom). (b) The initial location of the boundary in frame 36 (top) and the corresponding histogram. (c) The final boundary after 30 iterations (top), and the object histogram (bottom).*

3.7.2 GPU-Based Background Illumination Correction for Blue Screen Matching

Authors: Nicolas Ley, Christian Weigel

Institution: Technische Universität Ilmenau (UIL)

Publication: Proc. of European Signal Processing Conference, Poznan (Poland) September 2007

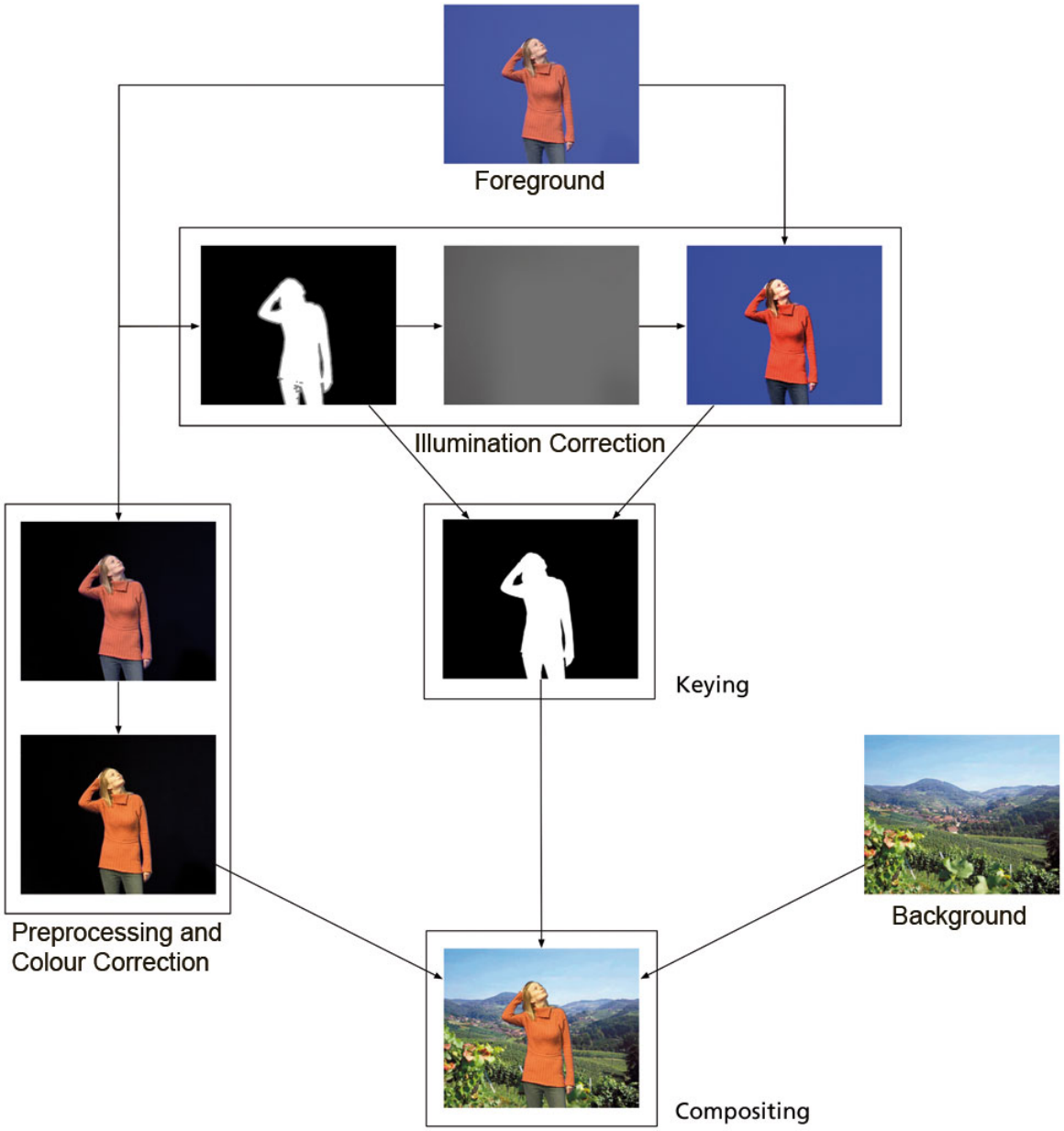


Fig. 41. *Background illumination correction*

Separation of foreground objects from an almost constant backing color for video applications is still a common problem. For non-realtime situations there is a wide variety of different powerful mathematical approaches that can deal with most of the matting problems. For SD/HD studio realtime keyers most solutions are not applicable due to their algorithm complexity or high effort in user interaction. Excellent hardware keyers, such as UltimatteTM

work on most occasions, but even under controlled lighting in a blue-/greenscreen matting problems may occur, or creativity is limited by necessary lighting conditions. As a preprocessing algorithm for traditional chroma keying systems, we present a simple background illumination correction based approach for improving matting problems with uneven or poor lit blue-/greenscreens (Figure 39). Using the computational power of GPU computing (GPGPU) the presented algorithm is realtime capable and offers an improvement for achievable mattes quality.

3.7.3 Segmentation in video sequences for compositing - applications in television production.

Authors: Michael Kirchner, Jan Röder, Lars Hörchens

Institution: Technische Universität Ilmenau (UIL)

Publication: Presentation of '12. Dortmunder Fernsehseminar', 20 - 21 March 2007, Dortmund

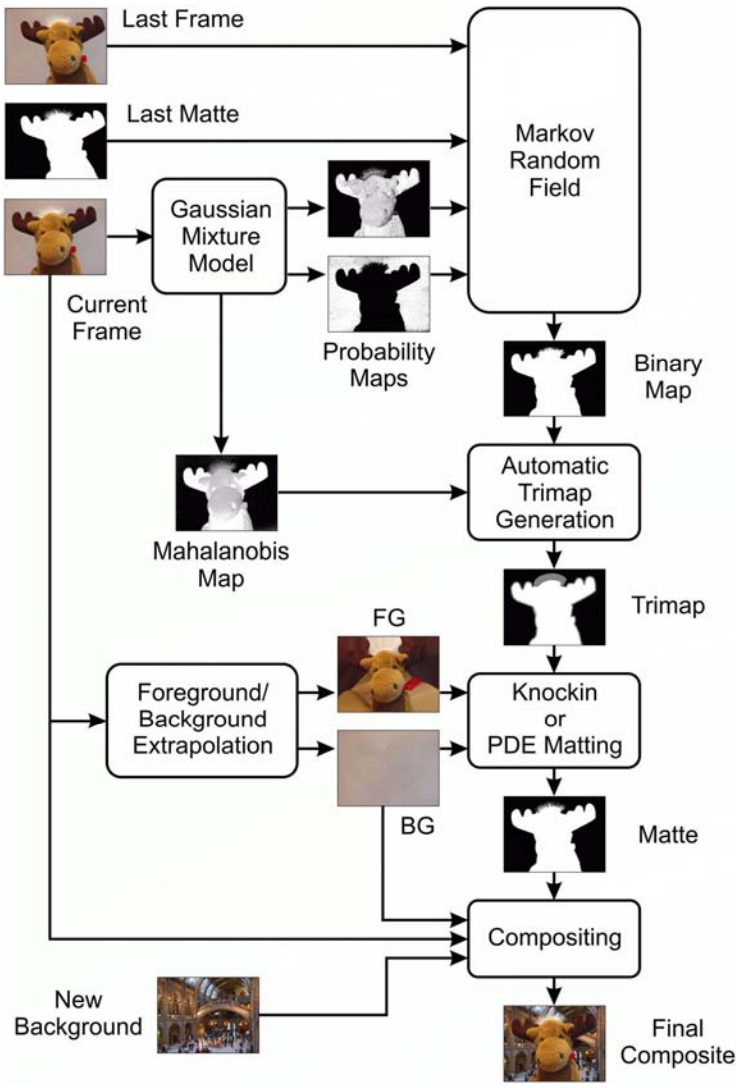


Fig. 42. *Digital Matting Framework*

After an introduction to the principal problems of chroma keyers this article describes a modular framework for the keying process (Figure 40). This framework, developed at UIL, allows including temporal and local constraints to weaken the above mentioned problems. Further the results of some test sequences are explained.

3.7.4 Conclusions and future plans

At the first abstract, two new region-based methods for object tracking using active contours in a dynamic programming framework are presented. The color histogram based method assumes that the color histogram of the tracked object is stationary from frame to frame. The "defocus energy" based method is useful when there are out-of-focus regions/objects in the scene. Experimental results on various video sequences demonstrate that the proposed methods are successful in difficult scenes with significant background clutter or scenes containing out-of-focus regions.

Regarding the second and the third abstract, in the field of image processing a next step is the implementation of the functionality of further video studio components, such as the video mixer, using the GPU.

At the actual segmentation step the integration of other image features to the segmentation process is being evaluated and algorithms for finding the optimal feature set are being developed

Currently, there are not much contributors to this research topic within the NoE. Nonetheless, the step of segmentation is always important for the whole chain of 3DTV production when it comes to deal with single objects. Some inter-workpackage collaborations could have been defined and future work will evaluate if researchers contributing to other work packages can benefit from this research area.

4 Conclusions

This report presented the projects and research performed by the partners of WP7 during the project months 30 to 42. During this time, several important contributions to the area 3D scene capture have been made by 3DTV partners.

The tasks within this project are concerned with recording, processing and analysis of 3D scenes. The huge number of different approaches to tackle this problem and their variety makes WP7 a most varying, diverse and interesting work package. Within this WP, recording oriented tasks, such as Multi- and Single-Camera are treated, as well as Holographic and Pattern projection methods. Contributions in processing and computer vision such as Object-based segmentation are presented as well as specific work on Human Face and Body techniques.

The overall knowledge about 3D time varying scene capture gathered in our NoE is a unique pool of expertise that is shared by all 3D partners. All in all, WP7 is a very successful research effort. In future we will make our knowledge available to the contributors to the other work packages. Also several general overview papers have been published. In the following, the partners are summarizing their results presented in the different subsections and give an outlook to their future activities.

Now, we summarize the main results, the project partners presented in the different subsections.

4.1 Multicamera

The research topic on multicamera deals with mechanisms for automatically acquiring and visualization of high quality 3D scenes. The envisioned goal is photo-realism during 3D real-time rendering from an arbitrary viewpoint. Possible scenes are movie sets in studios or outdoor scenes. To make a step towards this ultimate vision, many different techniques were developed in the multicamera research at WP7. The high number of publications and articles at high impact conferences (Eurographics, TPAMI, CVPR or ICME) clearly show the excellent contributions of the NoE to the international community. Currently photo-realism is not fully reached, but some of the contributions (e.g. Section 3.1.13 - 3.1.16) are on the edge of this goal. It is additionally supported by an effective HDR technique (Section 3.1.21 - 3.1.23). Scene reconstruction using different sensor methods, as done with PMD-sensors look most promising at this time.

4.2 Single Camera

In the presented contributions are a number of main stream research directions. One aims to render multi-views of a scene without explicitly determining the dense 3D structure, but rather utilizing image-based rendering technology. These views could be utilized as inputs not only for the current auto-stereoscopic displays, but also for the next-generation high resolution 3D displays by exploiting super-resolution techniques. In a different high priority research direction, conversion of 2D broadcast video into 3D is pursued that could yield 3D scene structure explicitly, capable of being utilized in any kind of 3DTV display, even for the holographic TVs. A number of fundamental problems, such as self-calibration, moving object segmentation and dense depth estimation, are jointly approached to result with a promising an

end-to-end algorithm, capable of 2D-to-3D conversion. Apart from this system, a novel outlier rejection and moving object segmentation is also proposed for obtaining robust estimates of the epipolar geometry between consecutive frames. Finally, 3D sparse scene extraction problem is reformulated, in order to consider the rate-distortion efficiency of the resulting scene representation during single camera extraction stage. Hence, 3D scene extraction is achieved in a hierarchical manner by gradually increasing the number sparse 3D reconstructed points, while considering the amount of bits to encode this information, as well as the quality of the resulting representation. Currently, a startup commercializing 2D-3D conversion technology in case of camera pans is setup.

4.3 Human Face and Body

Human faces and bodies are an integral part of many 3D-TV applications. Especially a-priori knowledge about 3D structure and models can be used to make existing algorithms more stable and robust. This is especially important, since human observers are very picky in the identification of un-natural human faces or movements. The presented works deal with synthesis, but also action recognition and multi-media fusion which mark current state-of-the art approaches. The current technology still requires improvements in order to achieve photo-realism.

4.4 Holographic Camera Techniques

Digital holography as compared to traditional methods of holography is seen as the way forward in realizing practical, mass media 3D displays. However, there are still many technological advances needed to achieve this goal. The Holographic Camera Techniques group has been working towards this end. Future problems are addressed in the European joint project Real3D, which should start February 2008. In the consortium of this project two partners of the NoE 3DTV are participating: Bilkent and BIAS.

4.5 Pattern Projection

The research on pattern projection within the reported period of the 3DTV project focused on development of low-cost portable devices for 3D coordinates measurement in real time by using of color encoded structured light or fringe projection. In the last case information about the coordinates is encoded in the phase of the reflected from the object pattern. A survey on the existing techniques for phase retrieval indicated that the most suitable for real-time capture in a large dynamic range are the phase-stepping algorithms. The survey was performed within WP7 and WP11. The presented works included theoretical and experimental analysis of the frequency content of projected fringes for collimated and divergent light illumination, estimation of the systematic errors due to higher harmonics and differences in the contrast of fringes projected at the different wavelengths in reconstruction of a plane reference surface and a dome for both types of illumination, simulation of the four-wavelength system operation for a dome surface and objects with sharp edges, as well as profilometric measurements of test objects. Proper software was developed to solve the mentioned tasks. The obtained results show that there exist optimal grating parameters and optical geometry which ensure accurate measurement of 3D coordinates.

For future activity we plan to use the four-wavelength system for capture of static and moving objects in a large scale of dimensions as well as to develop a more sophisticated system which includes two multi-wavelength projection modules with time multiplexing for a double symmetrical illumination of the scene to avoid shadowing. In addition, we plan to use a white-light illumination for capture of color coordinates (only visible part of the spectrum) of the scene by RGB recording with a color CCD camera calibrated to yield the same point cloud as provided by the four monochrome CCD cameras which are used for 3D coordinates measurement. Different tasks connected with calibration of the system and transforming of the data in a global coordinate system should be solved.

4.6 Motion Analysis and Tracking

Tracking of objects in video streams is till now on of the most crucial source of information many scene reconstruction algorithms rely on. Since the research topics are unsolved till now, it is important to treat this as a hi-priority research topic with impact in surveillance or scene analysis. Especially in the context of networks of cameras (e.g. in commercial TV studios) the topic is highly interesting and recognized in the community.

4.7 Object-Based Segmentation

Similar to tracking, also object-based representation and segmentation is an important and challenging research area, which has applications in object-based video coding, video post production or scene reconstruction of multiple moving objects in dynamic scenes recorded with moving cameras. Especially the GPU-implementations presented in this section are of high interest for the research community. Currently, we achieve object-segmentation for use in post production in case of still cameras.

5 References

[1] P.Viola and M. Jones, "Robust Real-Time Face Detection", *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137-154, 2004.

[2] Y. Freund and R. Schapire, "A Decision-Theoretic Generalization of Online Learning and an Application to Boosting", Journal of Computer and System Sciences, vol. 55, no. 1, pp. 119-139, 1997.

[3] S. Mallat and Z. Zhang, "Matching Pursuit with Time-Frequency Dictionaries", *IEEE Transactions on Signal Processing*, vol. 41, pp. 3397-3415, 1993.

[4] S. Chen, D. Donoho and M. Saunders, *Atomic Decomposition by Basis Pursuit*, Technical Report, Department of Statistics, Stanford University, 1995.

[5] I. F. Gorodnitsky, and B. D. Rao, "Sparse Signal Reconstruction from Limited Data Using FOCUSS: A Re-weighted Minimum Norm Algorithm", *IEEE Trans. on Signal Processing*, vol. 45, no. 3, March 1997.

[6] M. Aharon, M. Elad and A. M. Bruckstein, "The K-SVD: An Algorithm for Designing of Overcomplete Dictionaries for Sparse Representation", to appear in *IEEE Trans. On Signal Processing*

[7] N.G.Kingsbury, "Complex wavelets for shift invariant analysis and filtering of signals", *Journal of Applied and Computational Harmonic Analysis*, vol 10, no 3, pp. 234-253, May 2001.

6 Annex

6.1 Multicamera

6.1.1 Weighted Minimal Hypersurface Reconstruction

6.1.2 The Wägele: A mobile platform for Acquisition of 3D Models of Indoor and Outdoor Environments

6.1.3 Omnidirectional Stereo Based 3D Model Acquisition on the Wägele

6.1.4 3D Modeling of Inoor Environments for a Robotics Security Guard

6.1.5 3DTV-Panoramic 3D Model Acquisition and its 3D Visualization on the Interactive FogScreen

6.1.6 SAD A novel Multisensor Scene Acquisition Device

6.1.7 Integrating 3D Time-Of-Flight Camera Data and High resolution images for 3DTV Applications

6.1.8 On-the-fly Scene Acquisition with a Handy Multi-Sensor System

6.1.9 Self-Localization in Scanned 3DTV Sets

6.1.10 A Volumetric Fusion Technique for Surface Reconstruction from Silhouettes and Range Data

6.1.11 Multicamera Audio-Visual Analysis of Dance Figures

6.1.12 A Simple Framework for Natural Animation of Digitized Models

6.1.13 Video-driven Animation of Human Body Scans

6.1.14 Rapid Animation of Laser-scanned Humans

6.1.15 Markerless Deformable Mesh Tracking for Human Shape and Motion Capture

6.1.16 Markerless 3D Feature Tracking for Mesh-based Motion Capture

6.1.17 Reconstructing Human Shape, Motion and Appearance from Multi-view Video

6.1.18 A Volumetric Approach to Interactive Shape editing

6.1.19 Animation Collage