10th Workshop "Theoretical Foundations of Computer Vision"

Multi-Image Search, Filtering, Reasoning and Visualization

March 13 – March 17, 2000

Many problems in computer vision involve multiple images. These could be images taken by multiple cameras of the same scene, video sequences or multispectral images, etc. This workshop was concerned with the theoretical, algorithmic and implementation issues in multi-image acquisition, storage, retrieval, processing, analysis, manipulation and visualization. 38 talks were presented in 10 technical sessions: active vision; motion detection, modeling and understanding; omnidirectional vision; terrain reconstruction and calibration; scene reconstruction and image separation; object and scene visualization; image and video databases; object shape recovery; pose estimation, object recognition and modeling; and image properties and reasoning.

In addition, 5 working groups (WG) were organized: learning, action and development; image and video databases; visualization; color and photometry; new camera technologies. Three talks were presented in these groups. The charge to each WG was to assess the state-of-the-art of the given topic and to identify challenging and promising research directions. The abstracts of the talks and the reports of the 5 WGs are included in this booklet.

The 48 participants of the workshop came from 15 countries: 13 from Germany, 8 from Israel, 6 from USA, 5 from New Zealand, 2 from Japan, Canada, Czech Republic, China, Taiwan, and one from UK, Singapore, Australia, The Netherlands, Denmark and Greece. In experience and age, they range from beginning graduate students to senior professors. All participants found the Workshop discussions (both inside and outside the lecture hall) most stimulating.

A refereed proceedings book is planned to be published soon by Springer in the lecture notes series.

We are grateful to the administration and the service staff of the Dagstuhl enterprise for providing an ideal environment for our Workshop. The serene surroundings of the Schloss also contributed to the success of the Workshop.

Alfred Bruckstein Thomas Huang Reinhard Klette SongDe Ma

List of Speakers

Monday

R. Klette, Y. Aloimonos, A. Bruckstein, J.L. Barron, R. Owens, J.T.-J. Yang, V. Krüger

Tuesday

K. Daniilidis, S. Peleg, G. Gimel'farb, R. Reulke, S.S. Beauchemin, S. Ma, Y. Pritch, C.-E. Liedtke, V. Hlavac, C. Perwass, Y. Schechner

Wednesday

K. Aizawa, S.-K. Wei, F. Huang, R. Koch, Y. Kenmochi, N. Kiryati

Thursday

K.-L. Chan, T. Huang, T. Tan, J. Weng, L. Hermes, S. Tsekeridou, T. Werner, C.-Y. Chen, Y. Pritch, D. Segal, M. Störring, G. Sommer, J. Pauli, J. Ostermann, M. Porat

Friday

M. Felsberg, S. Buchholz, M.R. Ahmadzadeh

In the subsequent collection of abstracts (following the order above) the speaker is listed first in case of co-authors.

Motion, Shape Analysis, and Open Problems

REINHARD KLETTE
The University of Auckland

Researchers from Sport and Exercises Sciences and from CITR Tamaki, all university of Auckland, are working towards kinematic studies of human body motion (athletes, disabled children) also incorporating 3D surface data generated by computer vision techniques. A main interest within this kinematics project is directed on calculations of moments of inertia for body parts segmented according to specified body part definitions. The project uses commercial motion analysis software and a six-camera system for tracking markers within a movement volume. There also exist several commercial products for shape analysis, e.g., whole body scanners using structured light, or phase spectrum information. Shape from boundaries and photometric stereo is also used for whole body scanners to improve the accuracy compared to phase based techniques, and to reduce the "engineering efforts" compared to structured lighting.

This project was used to discuss a more general issue: How to evaluate or compare solutions, and how to specify areas for future research? "Open problems" may be listed in the context of an application project which are highly dependent on used equipment (high-speed, strobe light camera systems etc.), the geometric and photometric set-up (illumination spectrum, geometric proportions between subject surfaces, motion volume and distance to at least two cameras etc.) etc. Diverse projects as the sketched one seem to make the definition of open problems and the evaluation of obtained results very specific.

Computer vision research benefits from precise definitions of *open problems* such that progress is "measurable", independent upon "specific" project details. As an illustrative example, the integration problem of discrete vector fields can be stated in a mathematical way and its solution is critical within surface-normal based shape recovery approaches. As in mathematics, where the solution of a well-defined problem allows to state new problems resulting from this solution, computer vision research may also benefit from such an approach: specification and study of well-defined open problems makes progress in different research groups more transparent.

Video Geometry and Statistics: Space and Action

YIANNIS ALOIMONOS
University of Maryland

Models of real-world objects and actions can be obtained only through visual information. Given video of an object or a scene captured by a moving camera, a prerequisite for model building is to recover the three-dimensional (3D) motion of the camera which consists of a rotation and a translation at each instant. It is shown here that a spherical eye (an eye or system of eyes providing panoramic vision) is superior to a camera-type

eye (an eye with restricted field of view such as a common video camera) as regards the competence of 3D motion estimation. This result is derived from a geometric/statistical analysis of all the possible computational models that can be used for estimating 3D motion from an image sequence. Regardless of the estimation procedure for a camera-type eye, the parameters of the 3D rigid motion (translation and rotation) contain errors satisfying specific geometric constraints. Thus, translation is always confused with rotation, resulting in inaccurate results. This confusion does not happen for the case of panoramic vision. Insights obtained from this study point to new ways of constructing powerful imaging devices that suit particular tasks in visualization and virtual reality better than conventional cameras, thus leading to a new camera technology. Such new eyes are constructed by putting together multiple existing video cameras in specific ways, thus obtaining eyes from eyes. For a new eye of this kind we describe an implementation for deriving models of scenes and actions from video data, while avoiding the correspondence problem in the video sequence.

Space Fiducials for Robotic Self-Location

ALFRED M. BRUCKSTEIN
Technion, Haifa

What we see tells us a lot about where we are. This is a well-known fact and was indeed exploited to design various signs to be placed on walls, along roads etc. to enable us to visually locate ourselves in space. In the field of robotics, accurate self-location is an important task, and hence there were many efforts to address this problem by a combination of sign design and image analysis. In this talk we shall discuss some problems concerning the design of 3D objects that can serve as fiducials, in the sense that they enable us to easily locate ourselves with respect to their location. If we shall then place such objects around us, it will become an easy task (easy in the sense that no sophisticated image processing will be needed to analyze the images of these objects) to determine our absolute location in space, given that we know the location of these fiducial objects. We shall claim that in the field of robotics, the full potential of designing 3D objects for self-location was not yet fully exploited, and we shall suggest several possibilities. Some of them have been patented and are currently tested in practice.

References:

Bruckstein, A.M., Holt, R.J., Huang T.S. and Netravali A.N., "New devices for 3D pose estimation: mantis eyes, agam paintings, sundials and other space fiducials", 14th ICPR, Brisbane, 1998, & to appear: IJCV Special Issue, 2000.

Bruckstein, A.M., Holt, R.J., Huang, T.S and Netravali, A.N., "Optimum fiducials under weak perspective projections" ICCV, Corfu, 1999, & to appear: in IJCV, December 1999.

The Fusion of Image and Range Flow

JOHN L. BARRON
University of Western Ontario

We examine quantitatively range flow fields computed using local least squares and global regularization methods on image/range sequences made by a Birus range sensor moving relative to a scene. We first review the computation of local least squares range flow [DAGM1999] and then show how its computation can be cast in a global Horn and Schunck like regularization framework [ECCV2000]. This is done using range data only and using a combination (fusion) of image and range data [VI2000, Submitted 2000]. We present quantitative results for three regularization algorithms and a least squares image-range algorithm for one synthetic range sequence and one real range sequence, where the correct 3D motions are known a priori. We show that the fusion of image and range data can lead to a more accurate and dense range flow calculation than if the range flow is computed solely from range data.

Understanding Australian Sign Language

ROBYN OWENS
University of Western Australia, Nedlands

AUSLAN is the language of the Australian deaf community. It is a sign language involving the movement of both hands relative to the upper body, and it includes facial expression as well. To facilitate communication between the deaf and non-aurally affected communities, we are building a sign language understanding system that uses computer vision to translate images of signing into English. In this talk I describe the Human Motion Understanding (HMU) system, which incorporates a visual hand tracker with an adaptive fuzzy expert system to decode signs into words. The hand tracker uses a 3D kinematic model of the hand and an iterative state estimation scheme that updates the 21 degrees of freedom of this model, based on detected image features of the signing hand. The rule-based expert system decodes the sign after the kinematic parameters have been fuzzified into a small number of categories. The system outputs an English translation of the sign with a computed confidence level, and it achieves over 95% correct recognition.

Optical Tracking for Facial Animation

TZONG-JER YANG*, CHIEN-FENG HUANG† and MING OUHYOUNG†
*Digimax Production Center, Taiwan †National Taiwan University

The making of 3D facial animation is labor-intensive work, and it is very difficult to generate a realistic result. Animators face 2 major problems in the production of 3D

facial animation. The first one is how to reproduce a real person's motion, including global head motion and local facial expression. The second one is the creation of a 3D head model suitable for animation. We solve the first problem by introducing a commercial motion capture system. 20 optical markers are attached on one performer's face to calculate the local motion of facial expression while 3 additional markers on the performer's forehead are used to calculate global head motion. By eliminating the global head motion data, we get the processed local motion data which is then applied to a 3D head model with texture mapping, where the hair area is separately modeled with a 2D mesh rendered by using texture mapping, since the details of hair are difficult to be modeled. In this way, we can produce a photo-realistic facial animation sequence. Our next step is to generate a customized 3D head model by deforming a well-defined mesh model from several photographs. In this way, the process of model creation could be simplified, and the result of the head model is ready for animation purpose, since mesh design and locations of facial control points have been carefully pre-defined.

Gabor Wavelet Networks for Object Representation in Active Vision Systems

VOLKER KRÜGER
University of Kiel

The choice of the object representation is crucial for an effective and efficient performance of cognitive tasks in active vision systems such as object recognition, fixation, etc. The object representation should allow a dynamical adaption of the representation to the task and it should be possible to relate the representation to the action that is taken. For example for approximate information only an approximate representation makes sense while for precise information a precise representation should be used. In this article we want to introduce the Gabor wavelet network as a model based approach for an effective and efficient object representation. The Gabor wavelet network has several advantages such as invariance to some degree with respect to translation, rotation and dilation. Furthermore, the use of Gabor filters ensured that geometrical and textural object features are encoded. The feasibility of the Gabor filters as a model for local object features ensures a considerable data reduction while at the same time allowing *any* desired precision of the object representation ranging from a sparse to a photo-realistic representation. The feasibility of the object representation is verified by various experiments, including face recognition, face tracking and pose estimation.

A Unifying Theory on Omnidirectional Viewing

KONSTANTINOS DANIILIDIS and C. GEYER
University of Pennsylvania

Omnidirectional vision systems can provide panoramic alertness in surveillance, improve navigational capabilities, and produce panoramic images for multimedia. Catadioptric realizations of omnidirectional vision combine reflective surfaces and lenses. A particular class of them, the central panoramic systems, invented by S. Nayar, preserve the uniqueness of the projection viewpoint. In fact, every central projection system including the well known perspective projection on a plane falls into this category.

In this paper, we provide a unifying theory for all central catadioptric systems. We show that all of them are isomorphic to projective mappings from the sphere to a plane with a projection center on the perpendicular to the plane. Subcases are the stereographic projection equivalent to parabolic projection and the central planar projection equivalent to every conventional camera. We define a duality among projections of points and lines as well as among different mappings.

This unification is novel and has a a significant impact on the 3D interpretation of images. We present new invariances inherent in parabolic projections and a unifying calibration scheme from one view. We describe the implied advantages of catadioptric systems and explain why images arising in central catadioptric systems contain more information than images from conventional cameras. One example is that intrinsic calibration from a single view is possible for parabolic catadioptric systems given only three lines. Another example is metric rectification using only affine information about the scene.

OmniStereo: Panoramic Stereo in All Directions

SHMUEL PELEG
The Hebrew University of Jerusalem

A representation of stereo views in all directions using only two panoramic images is developed. In this compact representation, one panorama is generated for the left eye, and another panorama is generated for the right eye. A panoramic stereo pair provides a stereo sensation up to a full 360 degrees, and is generated with a new scene-to-image multiple viewpoint projection, where each viewing direction is viewed from a different viewpoint.

Stereo panoramas unifies, for the first time in a compact and efficient way, the two important elements for immersive environments: full-view panoramas and stereo. They can be created from 3D models or from real images. Once pre-recorded, playback of stereo panoramas involves only simple warping, with no depth or correspondence information. With stereo panorama viewers have the ability to freely view, in stereo, all directions.

Initial Image Orientation for Multiple View Stereo

GEORGY L. GIMEL'FARB and JIAN QUAN ZHANG
The University of Auckland

Initial image orientation allows to start an iterative process of uncalibrated 3D reconstruction which alternately calibrates the cameras and estimates the desired terrain model. We consider a practical problem of matching images under relatively large geometric distortions and use the RADIUS data set "M" as an example of an unordered set of multiple views of the same scene. To match such images, we use the least square technique with affine approximation of geometric distortions. Experiments show that the exhaustive search over a sparse grid of possible relative shifts, combined with a gradient-based search for the affine parameters around each grid point, allows in many cases to find the close matches and roughly estimate the relative orientation of one image with respect to other images of the same scene. Drawbacks of such a search and problems to be solved for organizing more efficient matching are discussed.

Recent Progress in Digital Photogrammetric Stereo Cameras and Data Evaluation

RALF REULKE
DLR Berlin

German Aerospace Center (DLR) was involved in line scanner experiments in the last 15 years. An example is the satellite sensor MOMS-02, which was successfully flown in the space shuttle D2 mission and later on PRIRODA, a module of the Russian space station MIR. In the last two years the Institute for Space Sensor Technology and Planetary Exploration is also involved in a commercial camera development of a digital three line stereo sensor of the LH-Systems company. The paper presents actual results in the developments of high resolution airborne and spaceborne scanner. The detector of such a sensor can be a large format CCD-matrix or a CCD-line. Stereo with a CCD-line camera was proposed first by Derenyi (University of New Brunswick, Canada 1970) and independently from Hofmann (about 1985). It was shown by Derenyi and Hofmann, that CCD-line stereo needs at least 3 lines. It is possible to arrange all CCD-lines and additional RGB and multispectral channels on one focal plane with one optics. One stereo line looks forward, the second one backward and the third in nadir direction. Each object on ground is covered three times and with exact attitude knowledge a stereo reconstruction is possible. The exact knowledge about flight pass must be measured with additional INS/GPS system. Airborne and spaceborne sensors are working alternatively. Spaceborne imagery cannot replace airborne imagery because of limitation in ground resolution and flexibility of spaceborne sensors. But with spaceborne sensors it is possible to map regions which are not reachable with airplanes.

Modelling and Removing Radial and Tangential Distortions in Spherical Lenses

STEVEN S. BEAUCHEMIN, R. BAJCSY and G. GIVATY
University of Pennsylvania

Spherical cameras are variable-resolution imaging systems and promising devices for autonomous navigation purposes, mainly because of their wide viewing angle which increases the capabilities of vision-based obstacle avoidance schemes. In addition, spherical lenses resemble the primate eye in their projective models and are biologically relevant. However, the calibration of spherical lenses for Computer Vision is a recent research topic and current procedures for pinhole camera calibration are inadequate when applied to spherical lenses. We present a novel method for spherical-lens camera calibration which models the lens radial and tangential distortions and determines the optical center and the angular deviations of the CCD sensor array within a unified numerical procedure. Contrary to other methods, there is no need for special equipment such as low-power laser beams or non-standard numerical procedures for finding the optical center. Numerical experiments, convergence and robustness analyses are presented.

Two Challenging Problems in Computer Vision: from Local to Global Approaches, and from Top-down to Bottom-up Approaches

SONG DE MA
Chinese Academy of Sciences

The traditional computer vision system deals with local features in the first step and uses bottom-up approaches. Time complexity, robustness and difficulties to link the local features to the model's knowledge in recognition step are the main difficulties in such systems. Since many years, people have proposed many ideas to solve these problems such as model based vision, or active vision. The main ideas are to link the local features to the global features and to deal with the interaction between bottom-up and top-down approaches. We argue that in doing this, to find the proper intermediate features by perception organization is the key point. We present the work on conic and quadric surface based stereo, motion, and pose determination. It is shown that the reconstruction can be done more robust and the result is more related to the object model. We present also the energy-based perception organization method to obtain the intermediate features and to measure the similarity in matching which is a necessary step in using top-down approaches.

Cameras for Stereo Panoramic Imaging

SHMUEL PELEG, YAEL PRITCH, and MOSHE BEN-EZRA
The Hebrew University of Jerusalem

A panorama for visual stereo consists of a pair of panoramic images, where one panorama is for the left eye, and another panorama is for the right eye. A panoramic stereo pair provides a stereo sensation up to a full 360 degrees. A stereo panorama cannot be photographed by two omnidirectional cameras from two viewpoints. It is normally constructed by mosaicing together images from a rotating stereo pair, or from a single moving camera. Capturing stereo panoramic images by a rotating camera makes it impossible to capture dynamic scenes at video rates, and limits stereo panoramic imaging to stationary scenes.

This paper presents two possibilities for capturing stereo panoramic images using optics, without any moving parts. A special mirror is introduced such that viewing the scene through this mirror creates the same rays as those used with the rotating cameras. Such a mirror enables the capture of stereo panoramic movies with a regular video camera. A lens for stereo panorama is also introduced. The designs of the mirror and of the lens are based on curves whose caustic is a circle.

Knowledge-Based Concepts for the Fusion of Multisensor and Multitemporal Aerial Images

CLAUS-E. LIEDTKE
University of Hannover

The increasing amount of remotely sensed imagery from multiple platforms requires efficient analysis techniques. The leading idea of the presented work is to automate the interpretation of multisensor and multitemporal remote sensing images by the use of common prior knowledge about landscape scenes. In addition the system can use specific map knowledge of a GIS, information about sensor projections, and temporal changes of scene objects. Prior expert knowledge about the scene content is represented explicitly by a semantic net. A common concept has been developed to distinguish between the semantics of objects and their visual appearance in the different sensors considering the physical principle of the sensor and the material and surface properties of the objects. A flexible control system is used for the automated analysis, which employs mixtures of bottom up and top down strategies for image analysis dependent on the respective state of interpretation. The control strategy employs rule based systems and is independent of the application. The system permits the fusion of several sensors like multispectral camera images, SAR-images, laser-scans etc. and it can be used for the fusion of images taken at different instances of time. Sensor fusion can be achieved on a pixel level, which requires prior rectification of the images, on feature level, which means that the same

object may show up differently in different sensors, and on object level, which means that different parts of an object can more accurately be recognized in different sensors. Results are shown for the extraction of roads from multisensor images. The approach for a multitemporal image analysis is illustrated for the monitoring of moorland areas and the recognition and extraction of an industrial fairground from an industrial area in an urban scene.

Scene Reconstruction from Uncalibrated Images

VACLAV HLAVAC, MARTIN URBAN, TOMAS PAJDLA, and TOMAS WERNER
Czech Technical University

Projective reconstruction recovers 3D points in projective space from their several projections in 2D images. The method allowing projective reconstruction based on concatenation of trifocal constraints around a reference view is presented. This "cake" configuration simplifies computations significantly. The algorithm relies on linear estimates only and the estimates are "closed" to image data. The method requires correspondences only across triplets of views. The method is not symmetrical with respect to views. The reference view plays a special role. The method can be viewed as a generalization of Hartley's algorithm or as a particular application of Triggs' closure relations. An accuracy and stability of the proposed algorithm with respect to pixel errors were tested. Experiments on real data are presented too.

3D-Reconstruction from Vanishing Points

CHRISTIAN PERWASS* and J. LASENBY†
*University of Kiel †Cambridge University

We present a 3D-reconstruction algorithm which reconstructs a scene from two static images. The two images are taken with unknown cameras from unknown positions. We assume that apart from point matches we also know the projections of a number of sets of parallel world lines. The latter are used to find vanishing points but also to constrain the reconstruction. Furthermore, we take the 3D-coordinate frame of the first camera as the basis we reconstruct in, and find the rotation, translation and the internal parameters of the second camera *relative* to the first. That is, we do not find the internal calibration of the first camera. Using modern cameras, 3D-Reconstructions are still of a good quality, though, and we are not restricted to using vanishing points that define an orthogonal world frame. However, it is also not difficult to extend the algorithm to find the internal calibration of the first camera, if a set of orthogonal vanishing points is given.

Note that relative translation, rotation and internal parameters are not found explicitly. This is a disadvantage if these values have to be known, but an advantage if we

are only interested in a 3D-reconstruction. Another advantage of our algorithm is that it is fast. On a Pentium II/233MHz under Windows 98 it took on average 160ms for a calibration (10000 trials). This time includes updating of dialog boxes and OpenGL windows. In an optimized program this time could probably be reduced to less than half. The algorithm is also robust, where the robustness depends mostly on the set of vanishing points used. The more similar the directions the vanishing points describe, the less robust the algorithm is.

We derived this algorithm using Geometric Algebra (GA). The advantage of using GA is that we are working directly with the basis frames of the cameras and not only with image point coordinates. This enables us, for example, to derive the relation between the collineation of the plane at infinity and a camera matrix in a straight forward manner. This relation forms the basis of our reconstruction algorithm.

Multi-Valued Images and Their Separation

YOAV Y. SCHECHNER*, NAHUM KIRYATI† and JOSEPH SHAMIR*
*Technion - Haifa †Tel Aviv University

Consider scenes deteriorated by reflections off a semi-reflecting medium (e.g., a glass window) that lies between the observer and an object. We present two approaches to recover the superimposed scenes. The first one is based on a focus cue, and can be generalized to volumetric imaging with multiple layers. The second method, based on a polarization cue, can automatically label the reconstructed scenes as reflected/transmitted. It is also demonstrated how to blindly determine the imaging PSF or the orientation of the invisible (semi-reflecting) surface in space in such situations.

Implicit 3D Approach to Image Generation: Object-Based Visual Effects by Linear Processing of Multiple Differently Focused Images

KIYOHARU AIZAWA
University of Tokyo

A new image generation scheme is introduced. The scheme linearly fuses multiple images, which are differently focused, into a new image in which objects in the scene is applied arbitrary linear processing such as focus(blurring), enhancement, extraction, shifting etc,. THe novelty of the work is that it does not require any segmentation to produce visual effects on objects in the scene. It typically uses two images for the scene: in one of them, the foreground is in focus and the background is out of focus, in the other image, vice versa. A linear imaging model is introduced, based on which an identity equation is derived between the original images and the target image in

which the object in the scene is selectively visually manipulated, and the target image is directly produced from the original images. It is shown that it can effectively produce visual effects for synthetic and real images. Various visual effects are examined such as focus manipulation, motion blur, enhancement, extraction, shifting etc,. A special camera is also introduced, by which synchronized three differently focused video can be captured, and dynamic scene can also handled by the scheme.

Classification of Image Acquisitions for 3D Scenes - Visualization and Reconstruction

SHOU-KANG WEI
The University of Auckland

Many image acquisitions have been used or proposed for 3D scene visualization and reconstruction. Different image acquisitions require different analysis approaches for the understanding of geometric and photometric situations relevant to these different approaches or methods to reconstruct or visualize a 3D scene. We are interested to characterize a broad range of image acquisitions such that their limitations and preferable applications can be understood. In this paper, we propose a classification scheme and specify what considerations are necessary to the classification. Four binary classifiers are used and studied. The resulting classification tree is constructed and presented with various examples from existing approaches such as QuickTimeVR, multi-viewpoint panoramic images, binocular stereo etc.

Epipolar Geometry in Concentric Panoramas

FAY HUANG
The University of Auckland

This work involves a study of the epipolar geometry for concentric panoramic images. An epipolar curve equation is derived, which enables us to plot the epipolar curve in one of the concentric panoramic images by specifying an image point in the other image. We also discuss conclusions based on this equation and justify our observations by plotting curves in examples of the synthetic images using various values of acquisition parameters. In both, the synthetic images and the real concentric panoramic images that we acquired, we show that the corresponding points in the symmetrical image pair lie on the same image rows.

Combining Structure from Motion and Image Based Rendering

REINHARD KOCH*, BENNO HEIGL†, and MARC POLLEFEYS**
*University of Kiel, †University of Erlangen-Nuremberg, and **Katholic University, Belgium

Image based rendering (lightfield rendering) allows fast visualization of complex scenes by view interpolation from images of densely spaced camera viewpoints. The lightfield data structure requires calibrated viewpoints, and rendering quality can be improved substantially when local scene depth is known for each viewpoint. Structure from motion allows to reconstruct camera calibration and 3D scene geometry from an image sequence.

In this contribution we propose to combine lightfield rendering with structure from motion. The advantage of the combined approach w.r.t. a pure geometric structure recovery is that the estimated geometry need not be globally consistent but is updated locally depending on the rendering viewpoint. In addition we can visualize view-dependent surface characteristics like specular reflections, and only a coarse geometric approximation of scene geometry is needed for the view interpolation. Thus geometric detail can be scaled according to the requested visualization quality.

Criteria for Surface Representation Techniques

YUKIKO KENMOCHI* and REINHARD KLETTE†
*Japan Advanced Institute of Science and Technology
†The University of Auckland

There are various surface representation techniques which have been developed and used in the fields of computer vision, three-dimensional image analysis and computer graphics, such as representations of local/global polyhedral surfaces, parametric surfaces, or geometric surfaces. For the achievement of a given task in one of these fields, we expect to choose an effective and relevant technique if it exists. Otherwise, there would be a need to design or provide a (new) matching technique for the given task. In order to make such a choice/provision of a surface representation technique operational, we suggest and discuss three criteria for the classification of surface representation techniques:

(a) suitability for topological characterizations (i.e. unique definition of topological features),

(b) support for surface deformations (that is, for algorithmic descriptions of deformations), and

(c) soundness with respect to geometric feature extractions such as surface area estimation ("convergence to the true value").

We illustrate these criteria via a discussion of a specific surface representation scheme, and suggested ways for performance analysis. We show

(1) how to solve the problem of topological ambiguity by "transforming" the classical marching cubes method into a simplicial surface representation model, see [Y. Kenmochi and A. Imiya: "Marching Cubes Method with Connectivity". in: Proc. ICIP'1999, Vol. 4, 361–365],

(2) how to describe surface deformations with/without topological changes using this simplicial model, and

(3) how to specify multigrid convergence for surface area estimations for the experimental evaluation of given surface representation techniques.

Two of the above criteria can be ensured by the simplicial surface representation technique. The criterion of multigrid surface area estimation might be not satisfied by any existing surface representation scheme. Further criteria may be defined and discussed within the context of multi-image applications.

Towards Segmentation from Multiple Cues: Symmetry and Color

NAHUM KIRYATI and ROY SHOR
Tel Aviv University

Segmentation is a notoriously difficult problem. It has been recognized that cue integration might be the key to better segmentation algorithms. This paper demonstrates the combined use of color and symmetry for detecting regions of interest (ROI), using face detection and wooden artifact detection as working examples. A functional that unifies color fitness and color-symmetry within elliptic supports is defined. Using this functional, the ROI detection problem becomes a five-dimensional global optimization problem. Exhaustive-search is inapplicable due its prohibitive computational cost. An adaptive random search method rapidly converges to the correct solution. The added value obtained by combining color and symmetry is demonstrated.

Region-Based Image Retrieval and Image Classification

KAP-LUK CHAN, XUEJIAN XIONG and FAN LIU
Nanyang Technological University - Singapore

When retrieving images containing multiple objects or regions with very different perceptual properties, a global content description of these images will not be able to represent the content accurately. A region-based image retrieval approach is advocated. Images can be partitioned into regions according the homogeneity various image attributes. Regions are represented by their own local properties. Global image content description can be constructed from the regional representations. Image retrieval can be done by entire image content or by region content. Methods have been developed for both types of retrieval. Image classification is achieved by clustering of database which aims to reduce the search space and speed up the process. The clustering results

also reveals the power of the features, be it color, texture or shape. The clustering of database and the image segmentation use the same unsupervised optimal fuzzy clustering algorithm. Experimental results show that image retrieval by global representation based region feature sets or by regions give better retrieval rates than by global content descriptors.

Video Databases: Some Image Processing and Computer Vision Issues

THOMAS HUANG
University of Illinois at Urbana-Champaign

Driving force for research: Push and pull applications for video (archives and real-time) on the Internet.

Required functionalities for video tools: Browsing, searching (retrieval), filtering, summarizing/listing.

Modes of query: Key words and phrases, sketch/draw, by example and content similarity, intelligent dialog between user and system.

Challenging research issues:

bridging the gap between low-level features and high-level concepts (multimedia understanding);

how to take advantage of "human-in-the-loop";

learning issues (incremental learning, training with a small set of labeled data and a large set of unlabeled data, unsupervised search for "interesting" spatial-temporal patterns, etc.);

fusion of multimodal - image sequences, audio, closed-captions, etc. - cues;

clustering and structuring of large video data for ease of retrieval.

A Generic Model for Image Watermarking

TIENIU TAN
Chinese Academy of Sciences

Digital watermarking has attracted a great deal of attention from both the academic community, industry and governmental organizations since the middle 1990s. Invisible digital watermarking embeds watermark information such as a company logo or an individual's signature into host data while without causing perceptually noticeable differences between the original and the watermarked data. As such there are numerous potential applications for watermarking, ranging from copyright protection to secrete communication. A large number of algorithms have been proposed, many of which watermark the host data in the transform domain such as DCT, FFT and wavelet

transforms. In this talk, a generic digital image watermarking model is described which can unify and explain most existing transform domain based methods. Fundamental issues of the generic model which are common to these methods are also addressed. These include the selection of image features (or transform coefficients)for carrying watermark information, the design of robust watermarks, the estimation of the maximum watermark energy and maximum watermark capacity under a given image quality measure, and the determination of thresholds in statistical watermark detection.

Discriminant Regression Tree for Image-Based Search and Automatic Skill Development

JUYANG (JOHN) WENG and WEY S. HWANG
Michigan State University

Context-based image or video search faces a challenging problem: defining features. It depends very much on the query type. Even if the query type is fixed, the features are difficult to design by hand. So far, feature selection is very much an art instead of science. In this talk, we aim to address this feature definition problem (more general than just selecting from given set of features) systematically. More generally, image-based search is a learning problem. Given training images with labels, the task of the trained system is to search and classify new images. We address this image-based search problem in a more general context — enabling machines to automatically develop skills. Further, we want to do this incrementally, instead of in a batch fashion. This is motivated by human mental development from infancy to adulthood. Our developmental algorithm runs in a single mode, where is no separate training and learning modes. The system has input ports called sensors and output ports called effectors. For image based search, at least two input ports are needed: video and label inputs. The output port gives the label of the current image frame (or video segment). The system runs in real time with real-time video stream. If the human trainer needs to train the system, he feeds labels through the label input for the image frame that is currently fed into the system. If the class label input is absent, the output port from the system produces the predicted class label automatically for the current frame. Humans can teach the system at any time. If the class label produced from the system is almost always correct, the training is completed. Otherwise, sample video for the problematic scenes can be fed into the system with the desired label, to improve the system's performance for these problem scenes. In this talk, we present our SAIL-2 developmental algorithm which automatically derive features from the two input source: image stream and label, and it incrementally builds a hierarchical regression tree for fast, real-time image-based search and classification. Our SAIL-2 algorithm has been also tested for real-time robots to enable robots to develop mental skill online, in real-time, incrementally. This represents a new approach to making man-made device, called the developmental approach.

Support Vector Machines for Land Usage Classification in Remote Sensing Imagery

LOTHAR HERMES, JOACHIM M. BUHMANN, and JAN PUZICHA
University of Bonn

Land usage classification is an essential part of many remote sensing applications for mapping, inventory, and yield estimation. We evaluate the potential of the recently introduced support vector machines for remote sensing applications. Moreover, we expand this discriminative technique by a novel Bayesian approach to estimate the confidence of each classification. These estimates are combined with a priori knowledge about topological relations of class labels. To do this, a contextual classification step based on the iterative conditional mode algorithm is used. As shown for Landsat TM imagery, this strategy is highly competitive and outperforms several commonly used classification schemes like gaussian mixture models, maximum likelihood classifiers, and k-nearest neighbor algorithms. Preliminary results are also shown for synthetic aperture radar data.

Audio-Visual Content Analysis and Interaction for Content-based Video Indexing

SOFIA TSEKERIDOU
Aristotle University of Thessaloniki

The content-based video parsing approach presented, analyzes both information sources (auditory and visual) and accounts for their inter-relations and synergy to extract high-level semantic information. Both shot-based and object-based access to the visual information is employed. Due to the temporal nature of video, time has to be accounted for. Thus, time-constrained labelling functions are generated. Audio source parsing leads to the extraction of a speaker identity mapping function over time. Visual source parsing results in the extraction of a talking face shot mapping over time. Integration of the audio and visual mappings constrained by interaction rules leads to more detailed video content descriptions and even partial detection of its context. Furthermore, it results in the refinement of content findings from single information sources.

Selection of Optimal Set of Reference Images for Image-based Rendering

TOMAS WERNER, VACLAV HLAVAC, MARTIN MATOUSEK, ALES LEONARDIS, and TOMAS PAJDLA
Czech Technical University

Image-based rendering is an alternative technique to traditional rendering of an explicit Euclidean 3-D model. While much research has been done in geometry of image-based rendering, almost nobody considered the problem of finding a set of reference images that is small but represents the scene sufficiently well. This is a topic of our work.

Our approach to selection proposes to select a small subset of reference (representing) images out of a large set of captured primary images of a scene. The selection is an optimization task in which a function of reference images set and the synthesis error is minimized. The synthesis error is an error introduced by replacing a real image with the corresponding synthesized image, rendered from a pair of reference images.

We make a restriction to (1) 1-DOF camera motion (e.g., a video sequence) and (2) image interpolation from two reference images (generally, images can be also extrapolated and more than two reference images can be used for rendering). Under this assumption, we propose three algorithms for selection based on: (1) sequential growing, (2) dynamic programming, and (3) growing and selection of plausible domains.

As image interpolation is not a strong enough model even if the scene is a convex polyhedron, there is need to use image extrapolation. The third algorithm is, at least theoretically, extensible for image extrapolation and also for 2-DOF camera motion.

Integration of Shape from Occluding Contours and Photometric Stereo by Fusing Depth and Orientation Data

CHIA-YEN CHEN* and RADIM SARA†
*The University of Auckland †Czech Technical University

In this work, we investigated the surface recovered by fusion of depth and orientation data. By fusing these two types of complementary data, we aim to provide a more accurate surface recovery method. The surface obtained by fusion should also have no observable artifacts along the boundary where two different types of data are fused.

For the fusion process, we determine the weighting function for measured data by examining the compatibility of the data. The dot products of the surface orientations are used to determine the data compatibility function. In this work, the compatibility is given as a binary function, having 1 where the data are compatible and 0 where the data are incompatible. The fusion algorithm is implemented according to the work of D. Terzopoulis, which minimized an error function of the unknown surface.

We used synthesized surfaces and data for the purpose of evaluation. It has been found that the surfaces produced by the fusion of data do yield higher accuracy than surfaces produced by either SFC or PSM alone. There were also no artifacts in the concavity boundaries.

Our next step is to modify the fusion algorithm and evaluate its application on real data.

Automatic Disparity Control in Stereo Panoramas (OmniStereo)

YAEL PRITCH, MOSHE BEN-EZRA, and SHMUEL PELEG
The Hebrew University of Jerusalem

An omnistereo panorama consists of a pair of panoramic images, where one panorama is for the left eye, and another panorama is for the right eye. An omnistereo pair provides a stereo sensation up to a full 360 degrees. OmniStereo panoramas can be created by mosaicing images from a rotating video camera, or by specially designed cameras.

The stereo sensation is a function of the disparity between the left and right images. This disparity is a function of the ratio of the distance between the cameras (the baseline) and the distance to the object: disparity is larger with longer baseline and close objects. Since our eyes are a fixed distance apart, we loose stereo sensation for far away objects.

It is possible to control the disparity in omnistereo panoramas which are generated by mosaicing images from a rotating camera. The baseline can be made larger for far away scenes, and smaller for nearer scenes. A method is described for the construction of omnistereo panoramas having larger baselines for faraway scenes, and smaller baseline for closer scenes. the baseline can change within the panorama from directions with closer objects to directions with further objects.

3D Reconstruction from Tangent-of-Sight Measurements of a Moving Object Seen from a Moving Camera

DANA SEGAL
The Hebrew University of Jerusalem

Consider the situation of a monocular image sequence with known ego-motion observing a 3D point moving simultaneously but along a path of up to second order, i.e. it can trace a line in 3D or a conic shaped path. We wish to reconstruct the 3D path from the projection of the tangent to the path at each time instance. This problem is analogue to the "trajectory triangulation" of lines and conic sections recently introduced in [Avidan&Shashua-cvpr99,Shashua&Avidan&Werman-iccv99], but instead of observing a point projection we observe a tangent projection and thus obtain a far simpler solution to the problem.

We show that the 3D path can be solved for in a natural manner, and linearly, using degenerate quadric envelopes - specifically the disk quadric. Our approach works seamlessly with both linear and second order paths, thus there is no need to know in advance the shape of the path as with the previous approaches for which lines and conics were treated as distinct. Our approach is linear in both straight line and conic paths, unlike the non-linear solution associated with point trajectory.

We provide experiments that show that our method behaves extremely well on a wide variety of scenarios, including those with multiple moving objects along lines and conic shaped paths.

Estimation of the Illuminant Color from Human Skin Color

MORITZ STÖRRING
Lab. of Computer Vision & Media Technology, University of Aalborg

Color is an important and useful feature for object tracking and recognition in computer vision. However, it has the difficulty that the color of the object changes if the illuminant color changes. But under known illuminant color it becomes a robust feature. There are more and more computer vision applications tracking humans, for example in interfaces for human computer interaction or automatic camera men, where skin color is an often used feature. Hence, it would be of significant importance to know the illuminant color in such applications. A novel method is proposed to estimate the current illuminant color from skin color observations. The method is based on a physical model of reflections, the assumption that illuminant colors are located close to the Planckian locus, and the knowledge about the camera parameters. The method is empirically tested using real images. The average estimation error of the correlated color temperature is as small as 180K. Applications are for example in color based tracking to adapt to changes in lighting and in visualization to re-render image colors to their appearance under canonical viewing conditions.

Pose Estimation Using Geometric Constraints

GERALD SOMMER, BODO ROSENHAHN, and YIWEN ZHAN
University of Kiel

The paper concerns 2D-3D pose estimation in the algebraic language of kinematics. The pose estimation problem is modelled on the base of several geometric constraint equations. In that way the projective geometric aspect of the topic is only implicitly represented and thus, pose estimation is a pure kinematic problem. The dynamic measurements of these constraints are either points or lines. The optimal estimation of line movements or line based constraints in space requires a linear algebraic embedding. The authors propose the use of motor algebra to model screw displacements. This is

a degenerate geometric algebra in which line transformations are linear ones. Instead of using matrix based LMS optimization, the development of special extended Kalman filters is proposed. In this paper an extended Kalman filter for estimating rotation and translation of a 2D-3D point-plane constraint in terms of rotors will be presented. The experiments aim to compare the use of different constraints and different methods of optimal estimating the pose parameters. The paper presents some preliminary results of that comparison which so far confirm the assumptions with respect to the behavior of the extended Kalman filter estimation.

Neural Network Learning for Visual Object Recognition

JOSEF PAULI
University of Kiel

We propose a learning approach for object recognition, which does not require a priori knowledge of three-dimensional geometric shapes. Instead, the task-relevant knowledge about objects is grounded on patterns of photometric appearance or filter responses. In a Robot Vision application significant variations may originate from changes in the spatial relation among robot effectors, cameras, and environmental objects. For the task of recognition, the variation manifold must be learned on the basis of visual demonstration. Algebraically, the manifold is specified by an implicit function for which certain deviations from ideal value 0 are accepted. Functions like these will serve as operators for the recognition of objects under varying view angle, view distance, illumination, or background, and also serve as operators for the recognition of scored situations. For function approximation, mixtures of Gaussian basis function networks are used in combination with Karhunen-Loeve expansion and support vector principles. The learning system can be biased by controlled camera motions and continual perception-action cycles. This enables the application of specific image transformations and the exploitation of correlations between close consecutive view patterns, and thus leads to constrained, simplified manifolds. The aim is to find a compromise between efficiency, invariance, and discriminability of the recognition function. The talk presented the theoretical background and real applications.

Interactive Services using Talking Faces

JÖRN OSTERMANN
AT&T Labs - Research, New Jersey

Facial animation has been combined with text-to-speech synthesis to create innovative multimodal interfaces. In this paper, we present an architecture for this multimodal interface. A face model is downloaded from a server into a client. The client uses an MPEG-4 compliant speech synthesizer that animates the head. The server sends text

and animation data to the client in addition to regular content to be displayed in a web browser.

We present 2 technologies for creating and rendering face models. For 3D face models, a computer graphics model is created using modeling software or adapting a mesh to a 3D head scan. This model is then animated by moving individual vertices of the mesh of the face model. Alternatively, we use sample-based face models. A set of samples of the mouth area is extracted from a video of a real person talking. The samples are classified according to mouth shape and related sounds. This database of mouth shapes is used to generate appropriate mouth movements according to the speech pronounced by the TTS system.

We believe that a talking face can support electronic commerce by providing a more friendly, helpful and intuitive user interface when compared to a regular web browser. In order to substantiate these claims, we undertook experiments to understand user reaction to interactive services designed with synthetic characters. In one experiment, participants played the 'Social Dilemma' game with the computer as a partner. Results indicate that users cooperate more with a computer when an animated face is representing the computer during the game. A simulated commercial application was evaluated, also comparing the performance of facial animation, text-to-speech and text only conditions. According to the results, the use of facial animation in the design of interactive services was favorably rated for most of the attributes in these experiments.

Image Reconstruction from Partial Information

MOSHE PORAT
Technion - Haifa

Spectral transforms like Fourier are central in image processing and computer vision. Due to their complex nature, it is of interest to investigate the separate roles of the Phase (angle) and the Magnitude (absolute value) of such transforms in image representation.

A few years ago it was found that localized phase is significantly more informative in image representation than global-phase, especially when iterative reconstruction algorithms are used (Behar, Porat & Zeevi, IEEE-SP, 1992). The reason for this major difference was discovered only very recently. It is shown that the structure of the signal plays a major role in the reconstruction process (Urieli, Porat and Cohen, IEEE-IP, 1998). Two types of signals, **geometric** and **symmetric** are situated in the two extremes of the best and worst reconstruction cases respectively. It is demonstrated that localized segments of a signal tend to be monotonic and thus closed to the geometric case, resulting in faster convergence of the algorithms used.

Having investigated the role of Fourier Magnitude in image representation as well, it has been found and proven that geometric signals are also the optimal case for image reconstruction from Fourier magnitude (Shapiro and Porat, CC Publication 278, 1999). Our conclusion is that the role of geometric signals in image representation and pro-

cessing deserves further attention and could be instrumental in image representation by partial information.

Structure Multivector for Local Analysis of Images

MICHAEL FELSBERG
University of Kiel

The structure multivector is a new approach for analyzing the local properties of a two-dimensional signal (e.g. images). It combines the classical concepts of the structure tensor and the analytic signal in a new way. This has been made possible by using a representation in the algebra of quaternions. The resulting method is linear and of low complexity. The filter-response includes local phase, local amplitude and local orientation of intrinsically one-dimensional neighborhoods in the signal. As for the structure tensor, the structure multivector field can be used to apply special filters to it for detecting two-dimensional features like corners. Experiments and comparisons with other approaches have been made.

Learning Group Actions from Visual Data

SVEN BUCHHOLZ
University of Kiel

Geometric transformations are one of the mathematical principal items in computer vision. With the study of the corresponding transformation groups group theoretical methods entered many fields of computer vision successfully. One particular area is that of object recognition, in which besides classical approaches problems are also often treated as learning problems. In this talk we introduce neural networks in Clifford algebras as a considerable learning scheme for such tasks. Special emphasis is given on geometrical transformations that can be computed by single Clifford neurons. The required material of the representation theory of classical groups is presented. Finally, the applicability of Clifford neural networks to object recognition problems is illustrated by some experimental results.

Reasoning with Uncertainty

MOHAMMED R. AHMADZADEH, MARIA PETROU and K. R. SASIKALA
University of Surrey

It has been established recently that combining classifiers improves the classification accuracy for many problems. This has been established both theoretically, mainly within

the framework of probability theory, and experimentally by many researchers. In this work we address two major problems of classifier combination:

1) The case when the classifiers combined are not of the same type. For example, one approach used to demonstrate classifier combination is to use different sources of data and apply a Bayesian classifier to each source and then combine the classifications of the two classifiers to obtained an improved final answer. In this work we combine the results obtained by classifiers of different nature not only from the point of view of yielded confidences in the classification results, but also from the more fundamental philosophical point of the approach they use, namely a probabilistic classifier and a Fuzzy logic-based classifier.

2) The classifiers combined are expected to use the same classes to classify the objects in question. In this work we address the problem of different classes, which however span the same classification space.

We address both these problems using the Dempster-Shafer orthogonal combination rule, where the results of the two classifiers are considered as items of evidence in support of a certain proposition. The problem of different classes is solved by using a superset of finer classes which can be combined to produce classes according to either of the two classifiers.

Our method is demonstrated in conjunction with the problem of predicting the risk of soil erosion of burned forests in the Mediterranean region using data concerning relevant factors like soil depth, ground slope and aspect, rock permeability, etc. This problem has been solved in the past using Pearl-Bayes networks and Fuzzy Logic. The results of these classifiers are combined to produce a more reliable classification.

Appendix: Reports of Working Groups I - V

WG I: Learning, Action and Development

Chairs: CLAUS-E. LIEDTKE
University of Hannover
JOHN WENG
Michigan State University

The working group has discussed the topic in two official and one unofficial meeting with a different number of participants. Some additional questions came up and were discussed during the final presentation.

The topic of learning requires a definition of the term. According to Michalski et.al. one possible definition is the following: "Learning denotes changes in the system that are adaptive in the sense that they enable the system to do the same task drawn from the same population more efficiently and more effectively the next time".

The members of the working group work presently themselves or are interested in

- learning of robot actions using reinforcement strategies,
- learning receptive fields,
- learning object recognition from multiple views of objects,
- learning explicit formulated knowledge contents from images like the the rules of a rule-based system or the scene structure represented by semantic nets,
- real-time learning from sensors by active robots, etc.

Aspects of learning which have been discussed included the model representations of the models to be learned and learning paradigms. Technical challenges which were recognized included

- slow convergence of reinforcement learning,
- automated learning, direct from images,
- autonomous learning,
- automated learning of scene structure,
- learning using a small learning set, especially learning from one sample.

Machine learning is often seen in competition with human learning abilities. In this connection the working group discussed a new approach called developmental approach. It is motivated by human mental development from infancy to adulthood. Development includes learning but requires more. Some major differences between conventional learning and the automatic development for mental skills are:

1. In conventional machine learning, the task is given, but in automatic development, the tasks that the machine will learn are unknown to the system designer.
2. In the former, the human designer defines features, but in automatic development, the system automatically derives features from sensory inputs.
3. In the former, the internal representation is pre-designed by the human according to the task given but in the latter, the representation is automatically generated according to the developmental scheme and the sensory data.
4. In the former, invariance depends on the predefined features and the rules that the human designer has specified. In the latter, the invariance is automatically achieved through experience as a generalization capability.

More information about development is available at http://www.cse.msu.edu/dl/ on an upcoming Workshop on Development and Learning (WDL) funded by National Science Foundation (NSF) and Defense Advanced Research Project Agency (DARPA) to be held at Michigan State University.

A discussion about some basics of learning from vision sensors came to the following conclusion:

Learning is always related to actions to be taken or tasks to be performed. Therefore learning from images is related to a model, which maps visual images onto actions. The model can be structured into different layers, which are again substructured into different domains. This seems to be true for any recognition process. The bottom layer is the sensor layer, the highest layer is the intelligent action layer. For machine learning from images the question is how to identify the layers and domains and how to interrelate them. One approach which has been used is to model the layers and domains as a semantic net. There we have different abstraction levels from the sensor layer at the bottom, which is directly connected to the images, up to the most abstract layer at the top, which represents the understanding of the visualized scene. Within the layers there are the nodes of the semantic net which we call concepts. The concepts are linked by relations, which do not have merely some descriptive property but influence actively the automated interpretation process. Right now the levels of the semantic net, the concepts and the relations have to be determined by a human from his expert knowledge. Autonomous learning would mean to find the layers and concepts automatically, and this is still an unsolved problem.

As was mentioned, learning is based on a model which links sensor signals with tasks. The sensor information has to be processed such, that the task can be performed. If many vision based tasks have successfully been completed through many experiments, then one can observe, that certain processing steps and certain intermediate processing results appear again and again. It can further be observed, that some steps and some intermediate results differ from application to application and from situation to situation. We call the latter the variants and the first ones the invariants. One objective of learning is to to identify the variants and invariants on the different levels of abstraction. In this connection the question was raised to which extent should unsupervised learning be used and to which extent supervised learning. Improvement of learning, i.e. learning on a higher level, is the search for variants and invariants not only on the sensor level

but on the level of previously identified "higher" invariants. So learning may become more efficient.

In order to implement vision based tasks by learning one need not to learn everything from scratch. We have already lots of experience and know how to solve some sensor based action problems. Referring to the visual sensor the variants and invariants are related to spatial-temporal effects. So we know that some useful invariants are gradients, surfaces, vector-fields, flow, etc.. We know that a good approach to describe the invariants of the world is to assume a 3D geometrical representation of the world. So the use of 3D-transforms like translations and rotations might be useful elements in the interrelations between the concepts. If we want to develop a powerful vision based action tool, with a high degree of flexibility, we have probably to start out developing a toolbox of known invariants first.

The working group consisted of the following participants: Yiannis Aloimonos, Sven Buchholz, Fay Huang, Yukiko Kenmochi, Volker Krüger, Claus-E. Liedtke, Josef Pauli, Christian Perwass, Nikolay Petkov, Moshe Porat, Gerald Sommer, Shou-Kang Wei, and John Weng.

WG II: Image/Video Databases: Browsing, Searching, Learning and Summarization

Chair: KAP LUK CHAN
Nanyang Technological University, Singapore

The group discussed problems and challenges related to the organization of image/video databases, their indexing and retrieval. The group recognized the importance of identifying potential application areas of image/video databases. This gives the motivation and justification of the effort to be devoted to the research and development of image/video database organization, indexing and retrieval methodologies. The functions, that an image/video databases should have, were discussed. Browsing, searching, learning and summarization of databases and their related issues were also discussed. The state-of-the-art development in this area was also mentioned. The MPEG-7 standardization effort was considered to provide a common set of protocols for the development of image/video databases for various applications. The group also looked at the promising research directions in the development and application of image/video databases.

The potential application areas of image/video databases include medicine, sports, mass media, education, traffic monitoring and so on in which image/video retrieval of relevant scenes or clip sequences is desired. Such R&D effort will have a significant impact on the industry and the society just as the Internet technology itself. The availability of the image/video databases over the Internet can bring a lot benefits to many people. Image databases have been created. Many studies on its organization, indexing and retrieval have been on-going and such databases for commercial use have also emerged. However, there is relatively few video databases available. With the identification of the likely application areas, more video databases will emerge.

The problems in developing methodologies for the organization, indexing and retrieval image/video databases are many. These problems are very challenging. The representation for content-based description is a key issue. Querying database can be by keywords, key phrases, image content, video content, and even audio content. Algorithms for processing each mode of query have to be developed. In view of the inherent ambiguities of using a particular mode of querying method, the use of multi- modal query should help to reduce if not resolving such ambiguities. On the particular issue of content representations, there are a number of low level content descriptors, which have proven to be useful. However, some form of standardization will help the development of the image/video retrieval systems. The MPEG-7 proposals considered several such representations for describing image/video content by color, texture and shape. There are also proposed description schemes. The acceptance of MPEG-7 standards will facilitate the development of various algorithms for image/video database organization, indexing and retrieval so that browsing and search an database will use a commonly agreed set of protocols and descriptors.

The group identifies a few promising research directions. It is generally recognized that many low level descriptors for color, texture and shape exist. However, there are missing links between the low level descriptors and their association to some high level

concepts. The clustering or structuring of large databases is an issue to be addressed. The approach to the structuring can vary from static pre- clustering to adaptive or evolutionary clustering. In here, the issue is not only on the perceptual feature sets and their association to high level concept but also the approach to learning, such as learning from users. There is a whole spectrum of possibilities for dynamic clustering of databases. The important possibility of human intervention in the searching and browsing of the image/video database should be recognized. The system should have the intelligence of learning from query and continuously re-structuring the database. Filtering of information during a search should facilitate faster convergence the satisfactory results. Relevance feedback is an important mechanism not only for filtering but also plays a role in evolutionary structuring of databases. The summarization of video database is an important functionality of the system, and is also a very challenging problem. Summarization allows extraction and collection of key information embedded in video and facilitates fast search and retrieval.

The emphasis of future research should not be on a particular application of database, but hope to develop generic frameworks and tools for the creation of image/video databases. Such frameworks and tools provide the means for building these databases according the chosen representations and structures and the ways in which such databases would be used. Collaboration between the computer vision and patter recognition community and the database community is important in developing such information systems.

The working group consisted of the following participants: Lothar Hermes, Thomas Huang, Songde Ma, Tieniu Tan, Sofia Tsekeridou, and Thomas Zöller.

WG III: Visualization, Animation, Manipulation

Chair: KIYOHARU AIZAWA
University of Tokyo

The WG III discussed the goal, requirements, bottlenecks and limitations regarding CV and IP technologies for visualization, animation. The filed covers, for example, topics such as scene reconstruction, video editing, image manipulation, face & body animation, panoramic imaging, interactive systems, VR systems, human machine interfaces etc. The common goal for them is REPRODUCTION OF REALITY: modeling, analyzing and producing the real world in the virtual environment. CV and IP technologies are extending their role into this new field.

The reproduction quality requires different preciseness for the representation of modeling technique because the scene is finally visualized on a screen. Three different dimensions required for the representation are 2D/3D, high/low detail and static/dynamic. Depending on the relative position between the viewer and the object, the requirements for the representations differ as shown below.

	Very close ($< 1m$)	Close ($< 12m$)	Distant
Static	3D, high detail, static	3D, low detail, static	2D images
Dynamic	3D, low detail, dynamic	3D, low detail, dynamic	2D movies

The above categories of requirements should be taken into consideration for the CV & IP tasks aimed at visualization.

One of the visible examples is VIRTUAL OFFICE, in which 3D scenes of the office is visualized onto a white wall as if the person could feel he were in a big office and could talk to his colleagues in a physically distant place. However, today, there still exist difficulties for its realization. In general, modeling natural objects is still difficult. The technical bottlenecks (challenges) are

1) modeling natural complex motion (ex. human motion)
2) modeling natural detail (ex. human hair)
3) building natural sharing environment (ex. natural sense, eye contact)
4) 3D & high resolution display (ex. "young pixel" problem).

From the optimistic point of view, the technical bottlenecks will be solved by any means in future.

The physical limits (boundaries) which can be never overcome by technologies were also discussed. Such limits finally become the constraint for realization of system. They are listed below:

1) speed of light (ex. delay),
2) time (ex. time difference),
3) gravity,
4) physical existence.

Among the aboves, the most important to the virtual environment is physical existence. For example, suppose the person in the virtual office would like to grab a book

visualized in the screen, he can never grab it, although the system could do compensate by down loading its digital content.

The working group consisted of the following participants: Mohammad Reza Ahmadzadeh, Kiyoharu Aizawa, Chia-Yen Chen, Jan-Michael Frahm, Chie-Feng Huang, Jörn Ostermann, Dana Segal, Shou-Kang Wei, and Jeff Yang.

WG IV: Color and Photometry

Chair: MORITZ STÖRRING
Computer Vision & Media Technology, University of Aalborg

This working group discussed applications and problems of using color and photometry/radiometry in computer vision and in visualization.

The four main areas of interest to the group were (1) shape recovery using, e.g., colored, structured light or dynamic photometric stereo, (2) image analysis and recognition using color, e.g., color edge detection, color-based segmentation, and color indexing of image databases, (3) image processing and enhancement such as filtering and morphological operations, and (4) visualization, e.g., panoramic view or 3D scene modeling from multiple input images to generate photometrically-accurate new virtual views.

Several unsolved research problems in these areas were identified by the group. Three particular issues that were discussed were the following:

1. The same material appears as two different colors under two different illuminations. This problem is a part of color constancy research and occurs frequently in uncontrolled illumination conditions, e.g., outdoors and under mixed illumination. Color-based segmentation becomes unreliable if the illumination color is unknown.

2. Under certain conditions of direct illumination, material surfaces show specularities or highlights. Their appearance depends on the light source itself, material properties (optical roughness), illumination direction, the viewing angle, and the phase angle. This makes segmentation difficult in both grayscale and color images, and image-based modeling and rendering of realistic virtual views for visualization becomes complicated.

3. Many commercial cameras are designed to provide images appropriate for human viewing, not image analysis. It would be advantageous to have more and other spectral sensitivities in the visual and non-visual wavelengths. Furthermore, the dynamic range in existing cameras is, in many applications, not sufficient. If an image contains, e.g., pronounced highlights, these will be overexposed and at the same time non-highlight areas will be underexposed and noisy. Thus image information gets lost. More color depth is therefore needed.

Physics-based knowledge about the reflectance characteristics of materials in an image can be used to estimate illumination conditions. There are materials such as wood, human skin, and green foliage that appear in many images. The chromaticities of each of these classes of materials have few variations. For example, in the case of skin, no matter the amount of melanin (darkness), all skin tones cluster in chromatic color space. But they vary for different illuminations. Hence, the reflected light, e.g. of skin, can be used to estimate the illumination color.

Methods for estimating surface reflectance models and local surface normals are needed based on multiple views of a scene. While recovery of geometric scene models has become quite successful using a variety of techniques, estimating surface normals accurately and surface reflectance (i.e., BRDF) properties remains poor. This information is needed for more realistic visualization of real scenes from virtual viewpoints and when scenes are augmented with synthetic objects.

In all discussions the importance of physics-based modeling of the scene was emphasized. More awareness of color science research in computer graphics, engineering and other fields is needed so that computer vision researchers can use results in those fields. Color in computer vision research remains a relatively minor subject of interest in the field. More emphasis on color in both research and teaching is recommended.

The working group consisted of the following participants: Chia-Yen Chen, Charles R. Dyer, Nahum Kiryati, Reinhard Klette, Reinhard Koch, and Moritz Störring.

WG V: New Camera Technologies

Chairs: KONSTANTINOS DANIILIDIS
University of Pennsylvania
SHMUEL PELEG
The Hebrew University of Jerusalem

The working group has addressed the following topics:

Why do we need new cameras?

What are the fundamental questions we try to solve?

Which applications will benefit?

What are possible designs and practical questions (today) ?

If nano-technology made it possible what is the camera of my dreams?

Computer vision research has been dominated for decades by the traditional simple lens-CCD sensor paradigm. This dedication to a single type of sensor has restricted the range of tasks a vision system can accomplish in a robotics or communications framework. However, biological visual sensors come in a wide variety of eye designs differing in the optics, the photo-reception, and the wiring to the brain.

Fundamental questions: Motion, shape, segmentation, photometry

Possible Designs Today:

"The Distributed Eye": Camera cluster

Catadioptric systems

A camera that will create the illusion of having everywhere

high resolution there will be always bandwidth limits.

The camera of my dreams will have

1. Adaptive properties
2. Deformable CCD or mirror surface
3. varying and dynamic resolutions (both spatial resolution and dynamic resolution - made possible with CMOS technology)
4. Multiple spectra
5. will be able to perform microsaccades
6. Use additional sensing (Z, audio, smell(!), etc)

Some application (in addition to obvious entertainment, robotics, medicine):

- Enhance visual experience - see better through the camera than directly. Examples are night vision cameras, and cameras that "sees" in invisible spectrum.

- cloaking (making object invisible) by connecting an array of cameras in one side of an object to an array of displays in the other side of the object. This way we could "see through" the object.

The working group consisted of the following participants: Yiannis Aloimonos, John Barron, Steven Beauchemin, Alfred Bruckstein, Konstantinos Daniilidis, Michael Felsberg, Georgy Gimel'farb, Volker Krüger, Robyn Owens, Shmuel Peleg, Yael Pritch, Ralf Reulke, Bernd Rosenhahn, Yoav Schechner, Dana Segal, and Adi Shavit.