INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Project-Team perception

Interpretation and Modelling of Images and Videos

Grenoble - Rhône-Alpes

THEME COG

Table of contents

1. Team

Research Scientist

Radu Horaud [Research Director (DR), HdR]
Emmanuel Prados [Research Associate (CR)]
Peter Sturm [Research Director (DR), HdR]

Faculty Member

Elise Arnaud [Université Joseph Fourier Grenoble]
Edmond Boyer [Université Joseph Fourier Grenoble, HdR]

Technical Staff

Hervé Mathieu [Research Engineer (IR), Until September 2007]
Florent Lagaye [From November 2008]
Bertrand Holveck [Development Engineer]
David Knossow [Development Engineer]

PhD Student

Amaël Delaunoy [INRIA grant]
Mauricio Diaz [Alban-EU grant]
Diana Mateus [Marie-Curie grant, until September 2008]
Julien Morat [CIFRE funding with Renault, until June 2008]
Ramya Narasimha [INRIA grant]
Régis Perrier [INRIA grant]
Benjamin Petit [INRIA grant]
Kiran Varanasi [INRIA grant]
Daniel Weinland [Marie-Curie Grant, until September 2008]
Andrei Zaharescu [Marie-Curie grant]
Visesh Chari [INRIA grant]
Avinash Sharma [INRIA grant]
Jamil Draréni [Co-supervision agreement with Université de Montréal]
Yalin Bastanlar [Visiting PhD student, Middle East Technical University, Ankara, Turkey]
Luis Puig [Visiting PhD student, University of Zaragoza, Spain]
Ketut Fundana [Visiting Marie Curie PhD student, Malmo University, Sweden]
Yan-Chen Lu [Visiting PhD student, University of Sheffield, UK]

Post-Doctoral Fellow

Fabio Cuzzolin [Marie Curie grant, until August 2008]
Simone Gasparini
Miles Hansard
Clément Ménier [INRIA industrial grant, until September 2008]
Kuk-Jin Yoon [until September 2008]

Administrative Assistant

Anne Pasteur [Secretary (SAR) Inria]

2. Overall Objectives

2.1. Introduction

The overall objective of the PERCEPTION research team is to develop theories, models, methods, and systems in order to allow computers to see and to understand what they see. A major difference between classical computer systems and computer vision systems is that while the former are guided by sets of mathematical and logical rules, the latter are governed by the laws of nature. It turns out that formalizing interactions between an artificial system and the physical world is a tremendously difficult task.

A first objective is to be able to gather images and videos with one or several cameras, to calibrate them, and to extract 2D and 3D geometric information from these images and videos. This is an extremely difficult task because the cameras receive light stimuli and these stimuli are affected by the complexity of the objects (shape, surface, color, texture, material) composing the real world. The interpretation of light in terms of geometry is also affected by the fact that the three dimensional world projects onto two dimensional images and this projection alters the Euclidean nature of the observed scene.

A second objective is to analyse articulated and moving objects. The real world is composed of rigid, deformable, and articulated objects. Solutions for finding the motion fields associated with deformable and articulated objects (such as humans) remain to be found. It is necessary to introduce prior models that encapsulate physical and mechanical features as well as shape, aspect, and behaviour. The ambition is to describe complex motion as "events" at both the physical level and at the semantic level.

A third objective is to describe and interpret images and videos in terms of objects, object categories, and events. In the past it has been shown that it is possible to recognize a single occurrence of an object from a single image. A more ambitious goal is to recognize object classes such as people, cars, trees, chairs, etc., as well as events or *objects evolving in time*. In addition to the usual difficulties that affect images of a single object there is also the additional issue of the variability within a class. The notion of statistical shape must be introduced and hence statistical learning should be used. More generally, learning should play a crucial role and the system must be designed such that it is able to learn from a small training set of samples. Another goal is to investigate how an object recognition system can take advantage from the introduction of non-visual input such as semantic and verbal descriptions. The relationship between images and meaning is a great challenge.

A fourth objective is to build vision systems that encapsulate one or several objectives stated above. Vision systems are built within a specific application. The domains at which vision may contribute are numerous:

- Multi-media technologies and in particular film and TV productions, database retrieval;
- Visual surveillance and monitoring;
- Augmented and mixed reality technologies and in particular entertainment, cultural heritage, telepresence and immersive systems, image-based rendering and image-based animation;
- Embedded systems for television, portable devices, defense, space, etc.

2.2. Highlights of the year

2.2.1. 3D and free-viewpoint television

This year we started an European collaboration within the MEDEA+/Eureka program: iGLANCE. The iGLANCE project addresses problems associated with 3DTV (three-dimensional television) and FVT (free viewpoint television). The project aims at the design of the next generation processing chip for HDTV compatible with both 3DTV and FVT. We will be particularly concerned with the production of rich video content that is boht 3D and iteractive, using multiple cameras. The iGLANCE project started in October 2008 for a duration of 36 months. Project partners: ST Microelectronics (F), INPG (F), UNILOG (F), 4D View Solutions (F), Philips (NL), Eindhoven University of Technology (NL).

2.2.2. Organization of the European Conference on Computer Vision 2008

Organization of the European Conference on Computer Vision 2008, held in Marseille. Edmond Boyer and Peter Sturm have been the Organization Chairs of this major bi-annual conference, which this year had a record attendance of 900 participants. Elise Arnaud and Anne Pasteur have been involved in the administration of this conference and Emmanuel Prados has been Tutorial Chair. Most of PERCEPTION's PhD students and post-docs have been active as student helpers.

2.2.3. BMVC'08 CRS Industrial award

The CRS Industrial Prize, sponsored by Computer Recognition Systems, was awarded to Amaël Delaunoy, Emmanuel Prados, Pau Gargallo, Jean-Philippe Pons and Peter Sturm for their paper entitled "Minimizing the Multi-view Stereo Reprojection Error for Triangular Surface Meshes", [21]

2.2.4. AMDO'08 best paper award

The paper "Inverse Kinematics using Sequential Monte Carlo Methods" by Elise Arnaud and co-author Nicolas Courty received the best paper award at AMDOâ08 (Conference on Articulated Motion and Deformable Objects), [19].

2.2.5. PhD award

Srikumar Ramalingam has received the INPG PhD Award, was runner-up for the AFRIF PhD Award (French Association for Pattern Recognition and Interpretation) and finalist of the European Cor Baayen PhD award (ERCIM, European Research Consortium for Informatics and Mathematics).

3. Scientific Foundations

3.1. The geometry of multiple images

Computer vision requires models that describe the image creation process. An important part (besides e.g. radiometric effects), concerns the geometrical relations between the scene, cameras and the captured images, commonly subsumed under the term "multi-view geometry". This describes how a scene is projected onto an image, and how different images of the same scene are related to one another. Many concepts are developed and expressed using the tool of projective geometry. As for numerical estimation, e.g. structure and motion calculations, geometric concepts are expressed algebraically. Geometric relations between different views can for example be represented by so-called matching tensors (fundamental matrix, trifocal tensors, ...). These tools and others allow to devise the theory and algorithms for the general task of computing scene structure and camera motion, and especially how to perform this task using various kinds of geometrical information: matches of geometrical primitives in different images, constraints on the structure of the scene or on the intrinsic characteristics or the motion of cameras, etc.

3.2. The photometry component

In addition to the geometry (of scene and cameras), the way an image looks like depends on many factors, including illumination, and reflectance properties of objects. The reflectance, or "appearance", is the set of laws and properties which govern the radiance of the surfaces . This last component makes the connections between the others. Often, the ÃappearanceÃ of objects is modeled in image space, e.g. by fitting statistical models, texture models, deformable appearance models (...) to a set of images, or by simply adopting images as texture maps.

Image-based modelling of 3D shape, appearance, and illumination is based on prior information and measures for the coherence between acquired images (data), and acquired images and those predicted by the estimated model. This may also include the aspect of temporal coherence, which becomes important if scenes with deformable or articulated objects are considered.

Taking into account changes in image appearance of objects is important for many computer vision tasks since they significantly affect the performances of the algorithms. In particular, this is crucial for feature extraction, feature matching/tracking, object tracking, 3D modelling, object recognition etc.

3.3. Shape Acquisition

Recovering shapes from images is a fundamental task in computer vision. Applications are numerous and include, in particular, 3D modeling applications and mixed reality applications where real shapes are mixed with virtual environments. The problem faced here is to recover shape information such as surfaces, point positions, or differential properties from image information. A tremendous research effort has been made in the past to solve this problem and a number of partial solutions had been proposed. However, a fundamental issue still to be addressed is the recovery of full shape information over time sequences. The main difficulties

are precision, robustness of computed shapes as well as consistency of these shapes over time. An additional difficulty raised by real-time applications is complexity. Such applications are today feasible but often require powerful computation units such as PC clusters. Thus, significant efforts must also be devoted to switch from traditional single-PC units to modern computation architectures.

3.4. Motion Analysis

The perception of motion is one of the major goals in computer vision with a wide range of promising applications. A prerequisite for motion analysis is motion modelling. Motion models span from rigid motion to complex articulated and/or deformable motion. Deformable objects form an interesting case because the models are closely related to the underlying physical phenomena. In the recent past, robust methods were developed for analysing rigid motion. This can be done either in image space or in 3D space. Image-space analysis is appealing and it requires sophisticated non-linear minimization methods and a probabilistic framework. An intrinsic difficulty with methods based on 2D data is the ambiguity of associating a multiple degree of freedom 3D model with image contours, texture and optical flow. Methods using 3D data are more relevant with respect to our recent research investigations. 3D data are produced using stereo or a multiple-camera setup. These data (surface patches, meshes, voxels, etc.) are matched against an articulated object model (based on cylindrical parts, implicit surfaces, conical parts, and so forth). The matching is carried out within a probabilistic framework (pair-wise registration, unsupervised learning, maximum likelihood with missing data).

Challenging problems are the detection and segmentation of multiple moving objects and of complex articulated objects, such as human-body motion, body-part motion, etc. It is crucial to be able to detect motion cues and to interpret them in terms of moving parts, independently of a prior model. Another difficult problem is to track articulated motion over time and to estimate the motions associated with each individual degree of freedom.

3.5. Multiple-camera acquisition of visual data

Modern computer vision techniques and applications require the deployment of a large number of cameras linked to a powerful multi-PC computing platform. Therefore, such a system must fulfill the following requirements: The cameras must be synchronized up to the millisecond, the bandwidth associated with image transfer (from the sensor to the computer memory) must be large enough to allow the transmission of uncompressed images at video rates, and the computing units must be able to dynamically store the data and/to process them in real-time.

Until recently, the vast majority of systems were based on hybrid analog-digital camera systems. Current systems are all-digital ones. They are based on network communication protocols such as the IEEE 1394. Current systems deliver 640×480 grey-level/color images but in the near future 1600×1200 images will be available at 30 frames/second.

Camera synchronization may be performed in several ways. The most common one is to use special-purpose hardware. Since both cameras and computers are linked through a network, it is possible to synchronize them using network protocols, such as NTP (network time protocol).

4. Application Domains

4.1. 3D modeling and rendering

3D modeling from images can be seen as a basic technology, with many uses and applications in various domains. Some applications only require geometric information (measuring, visual servoing, navigation) while more and more rely on more complete models (3D models with texture maps or other models of appearance) that can be rendered in order to produce realistic images. Some of our projects directly address potential applications in virtual studios or “edutainment” (e.g. virtual tours), and many others may benefit from our scientific results and software.

4.2. Mixed and Augmented Reality

Mixed realities consist in merging real and virtual environments. The fundamental issue in this field is the level of interaction that can be reached between real and virtual worlds, typically a person catching and moving a virtual object. This level depends directly on the precision of the real world models that can be obtained and on the rapidity of the modeling process to ensure consistency between both worlds. A challenging task is then to use images taken in real-time from cameras to model the real world without help from intrusive material such as infrared sensors or markers.

Augmented reality systems allow an user to see the real world with computer graphics and computer animation superimposed and composited with it. Applications of the concept of AR basically use virtual objects to help the user to get a better understanding of her/his surroundings. Fundamentally, AR is about augmentation of human visual perception: entertainment, maintenance and repair of complex/dangerous equipment, training, telepresence in remote, space, and hazardous environments, emergency handling, and so forth. In recent years, computer vision techniques have proved their potential for solving key-problems encountered in AR: real-time pose estimation, detection and tracking of rigid objects, etc. However, the vast majority of existing systems use a single camera and the technological challenge consisted in aligning a prestored geometrical model of an object with a monocular image sequence.

4.3. Human Motion Capture and Analysis

We are particularly interested in the capture and analysis of human motion, which consists in recovering the motion parameters of the human body and/or human body parts, such as the hand. In the past researchers have concentrated on recovering constrained motions such as human walking and running. We are interested in recovering unconstrained motion. The problem is difficult because of the large number of degrees of freedom, the small size of some body parts, the ambiguity of some motions, the self-occlusions, etc. Human motion capture methods have a wide range of applications: human monitoring, surveillance, gesture analysis, motion recognition, computer animation, etc.

4.4. Multi-media and interactive applications

The employment of advanced computer vision techniques for media applications is a dynamic area that will benefit from scientific findings and developments. There is a huge potential in the spheres of TV and film productions, interactive TV, multimedia database retrieval, and so forth.

Vision research provides solutions for real-time recovery of studio models (3D scene, people and their movements, etc.) in realistic conditions compatible with artistic production (several moving people in changing lighting conditions, partial occlusions). In particular, the recognition of people and their motions will offer a whole new range of possibilities for creating dynamic situations and for immersive/interactive interfaces and platforms in TV productions. These new and not yet available technologies involve integration of action and gesture recognition techniques for new forms of interaction between, for example, a TV moderator and virtual characters and objects, two remote groups of people, real and virtual actors, etc.

4.5. Car driving technologies

In the long term (five to ten years from now) all car manufacturers foresee that cameras with their associated hardware and software will become parts of standard car equipment. Cameras' fields of view will span both outside and inside the car. Computer vision software should be able to have both low-level (alert systems) and high-level (cognitive systems) capabilities. Forthcoming camera-based systems should be able to detect and recognize obstacles in real-time, to assist the driver for manoeuvering the car (through a verbal dialogue), and to monitor the driver's behaviour. For example, the analysis and recognition of the driver's body gestures and head motions will be used as cues for modelling the driver's behaviour and for alerting her or him if necessary.

4.6. Defense technologies

The PERCEPTION project has a long tradition of scientific and technological collaborations with the French defense industry. In the past we collaborated with Aérospatiale SA for 10 years (from 1992 to 2002). During these years we developed several computer vision based techniques for air-to-ground and ground-to-ground missile guidance. In particular we developed methods for enabling 3D reconstruction and pose recovery from cameras on-board of the missile, as well as a method for tracking a target in the presence of large scale changes.

5. Software

5.1. Platforms

5.1.1. The Grimage platform

The Grimage platform is an experimental laboratory dedicated to multi-media applications of computer vision. It hosts a multiple-camera system connected to a PC cluster, as well as to a multi-video projection system. This laboratory is shared by several research groups, most proeminently PERCEPTION and MOAIS. In particular, Grimage allows challenging real-time immersive applications based on computer vision and interactions between real and virtual objects, Figure 1.

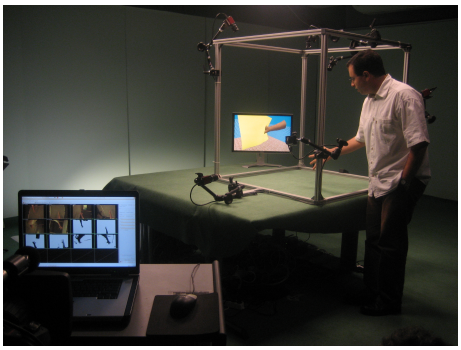


Figure 1. Left: The Grimage platform allows immersive/interactive applications such as this one. The real character is reconstruced in real-time and immersed in a virtual world, such that he/she can interact with virtual objects. Right: The mini-Grimage platform holds on a table top. It uses six cameras connected to six mini-PCs and to a laptop.

5.1.2. The mini-Grimage platform

We also deveoped a miniaturized version of Grimage. Based on the same algorithms and software, this mini-Grimage platform can hold on a desk top and/or can be used for various experiments involving fast and realistic 3-D reconstruction of objects, Figure 1.

5.1.3. POPEYE: an audiovisual robotic head

We have developed an audiovisual (AV) robot head that supports software for AV fusion based on binocular vision and binaural audition (see below). The vision module is composed of two digital cameras that form a stereoscopic pair with control of vergence (one rotational degree of freedom per camera). The auditory module is composed of two microphones. The head can perform pan and tilt rotations as well. All the sensors are linked to a PC. POPEYE computes ITD (interaural time difference) signals at 100 Hz and stereo disparities

at 15 Hz. These audio and visual observations are then fused by a AV clustering technique. POPEYE has been developed within the European project POP (http://perception.inrialpes.fr/POP in collaboration with the project-team MISTIS and with two other POP partners: the Speech and Hearing group of the University of Sheffield and the Institute for Systems and Robotics of the University of Coimbra.

5.2. Software packages

5.2.1. TransforMesh: Mesh evolution with applications to dense surface reconstruction

We completed the development of TransforMesh, started in 2007. It is a mesh-evolution software developed within the thesis of Andrei Zaharescu [13]. It is a provably correct mesh-based surface evolution method, see figure 2. It is able to handle topological changes and self-intersections without imposing any mesh sampling constraints. The exact mesh geometry is preserved throughout, except for the self-intersection areas. Typical applications, including mesh morphing and 3-D reconstruction using variational methods, are currently handled. TransforMesh will soon be publicly available as open source with LGPL on http://gforge.inria.fr

5.2.2. Dense volume- and surface-registration

We started to develop a software package that registers shapes based on either their volumetric (voxels) or surface (meshes) representations. The software implements a spectral graph matching method combined with non-linear dimensionality reduction and with rigid point registration, as described in [27] as well as in the forthcoming PhD thesis of Diana Mateus. The SpecMatch software package will soon be publicly available as open source with GPL on http://gforge.inria.fr.

5.2.3. Real-time shape acquisition and visualization

This software can be paraphrased as *from pixels to meshes*. It is a complete package that takes as input uncompressed image sequences grabbed with synchronized cameras. The software typically handles between 8 and 20 HDTV cameras, i.e., 2 million pixels per image. The software calibrates the cameras, segments the images into foreground (silhouettes) and background, and converts the silhouettes into a 3D meshed surface. The latter is smoothed and visualized using an image-based rendering technique. Currently this software package is commercialized by our start-up company, 4D View Solutions (http://www.4dviews.com). We continue to collaborate with this company. The latest version of the software is available for INRIA researchers and it runs on the GrImage platform.

5.2.4. Audio-visual localization of speakers

We developed a software package that uses binocular vision and binaural audition to spatially localize speakers. The software runs on the POPEYE platform (see above) and it has been developed in collaboration with the MISTIS project-team and with the Speech and Hearing group of the University of Sheffield. The software combines stereo, interaural time difference, and expectation-maximization algorithms. It is developed within the European project POP.

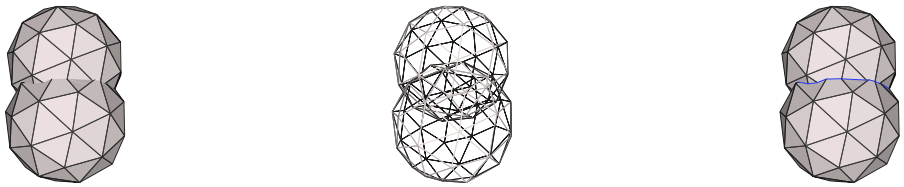
6. New Results

6.1. Satellite Imaging

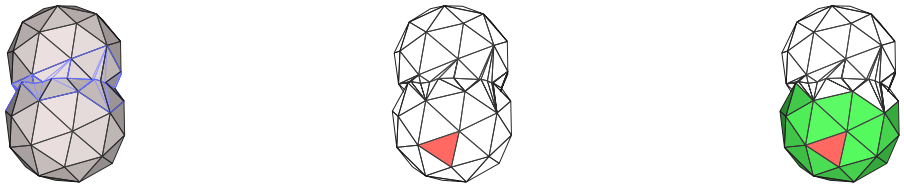
6.1.1. Modeling, estimating and compensating the vibrations

Within a collaboration with ASTRIUM, work on satellite imaging has started this year. High-resolution satellite imagery is typically based on so-called push-broom cameras (one or a few rows of pixels, covering different spectral bands). High-resolution images are generated by stitching individual push-broom images, taken at successive time instants, together. This is based on measuring the satellite's displacement using various sensors. Due to unavoidable vibrations of the satellite, estimated displacements are not accurate and have to be compensated. The main goal of our work is to develop theoretical models and practical approaches for modeling, estimating and compensating these vibrations, using information contained in the images. This is work in progress.

(a) Input (shaded) (b) Input (wireframe) (c) Facet Intersections



(d) 2-D Delaunay Segments (e) Seed Triangle Finding (f) Valid Triangle Expansion



(g) Partial Triangle Expansion (h) End of Region Growing (i) Final Output (shaded)

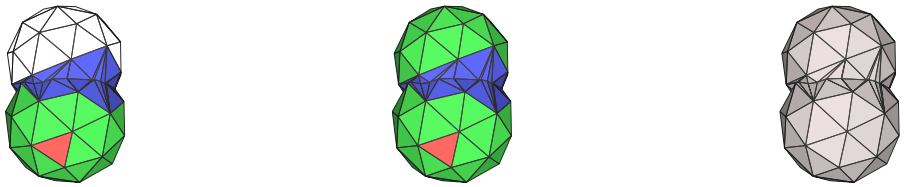


Figure 2. An example of TransforMesh applied to two meshes that intersect each other (a), (b). The algorithm starts by computing all the self-intersections (c), followed by the valid region growing, which consists of first identifying a seed triangle (e) (marked in red), expanding on the neighbouring valid triangles (f) (marked in green) and selecting the correct subparts of the partially valid triangles (g) (marked in blue). The geometry of partial triangles is locally defined using a 2-D Delaunay triangulation (d). The procedure ends when all the valid and partially valid triangles have been visited (h).

6.1.2. Multiview stereovision for push-broom cameras

Within a collaboration with the IPGP, "Institut de Physique du Globe de Paris", we have started to work on an adaptation of some of our multiview stereovision algorithms to push-broom cameras. This is work in progress.

6.2. Audio-visual perception

This work takes place in the context of the POP European project and includes further collaborations with researchers from University of Sheffield, UK. The context is that of multi-modal sensory signal integration. We focus on audio-visual integration. Fusing information from audio and video sources has resulted in improved performance in applications such as tracking. However, crossmodal integration is not trivial and requires some cognitive modelling because at a lower level, there is no obvious way to associate depth and sound sources. Combining our expertise with expertise both from project-team MISTIS and from the University of Sheffield's Speech and Hearing Group, we address the difficult problems of integrating spatial and temporal audio-visual stimuli using a combined geometrical and probabilistic framework and attack the problem of associating sensorial descriptions with representation of prior knowledge.

6.2.1. Multi-speaker localization

We address the problem of speaker localization within the framework of maximum likelihood with missing data. Both auditory and visual observations are available. We gather observations over a short time interval. We assume that within this short interval the speakers can be described by their 3-D locations. A mixture-model component (a cluster) is associated with each speaker, Figure 3. In practice we consider $N + 1$ possible components corresponding to the addition of an extra outlier category to the N speakers.

Within each time interval we observe both visual and auditory observations. Each visual observation corresponds to a binocular disparity. Note that such a binocular disparity corresponds to the location of a physical point lying onto an object that is visible in both the left and right images of the stereo pair. Similarly, each auditory observation corresponds to an audio disparity, namely the interaural time difference, or ITD.

We developed a method that recovers speaker localizations and that can be seen as a parameter estimation issue in a missing data framework. The parameters to be estimated are the speaker locations, and the missing variables are the assignement variables associating each individual (visual and auditory) observation to one of the N speakers or to the outlier class. We implemented the method within the framework of the *generalized* expectation-maximization algorithm [25], [26].

6.2.2. An audio-visual database

Two POP partners (University of Sheffield and INRIA) have gathered synchronized auditory and visual datasets for the study of audio-visual fusion. The idea was to record a mix of scenarios where the audio-visual tasks of tracking the speaking face, where either the visual or auditory cues add disambiguating information; or more varied scenarios (eg. sitting in at a coffee break meeting) with a large amount of challenging audio and visual stimuli such as multiple speakers, varied amount of background noise, occulting objects, faces turned away and getting obscured, etc. Central to all scenarios is the state of the audio-visual perceiver and we have been very interesed in getting hold of some data recored with an active perceiver, so we propose that the perceiver is either static, panning or moving (probably limited to rotating its head) so as to mimic attending to the most interesting source at the moment [17]

To achieve the acquisition of such a data collection, the following setup has been developed, (let us note that this setup is designed to be easily plugged with the audio-visual robot head). The audio-visual perceiver is either a person or the dummy head/torso wearing earbud microphones. The perceiver is also fitted with a helmet on which is mounted a pair of stereo cameras. On top of the head, a 4 point tracking device is attached. This has to be viewable from the tracking camera, which is to be placed above; either suspended from the ceiling or similar. The three cameras (stero pair and tracking) are controlled with a software package and the raw image sequences are recorded on to a PC. The audio is recorded on to a laptop or PC. The three cameras are synchronized with the audio signal using NTP network. The calibrated data collection will is freely accessible for research purposes at http://perception.inrialpes.fr/CAVA_Dataset/Site/

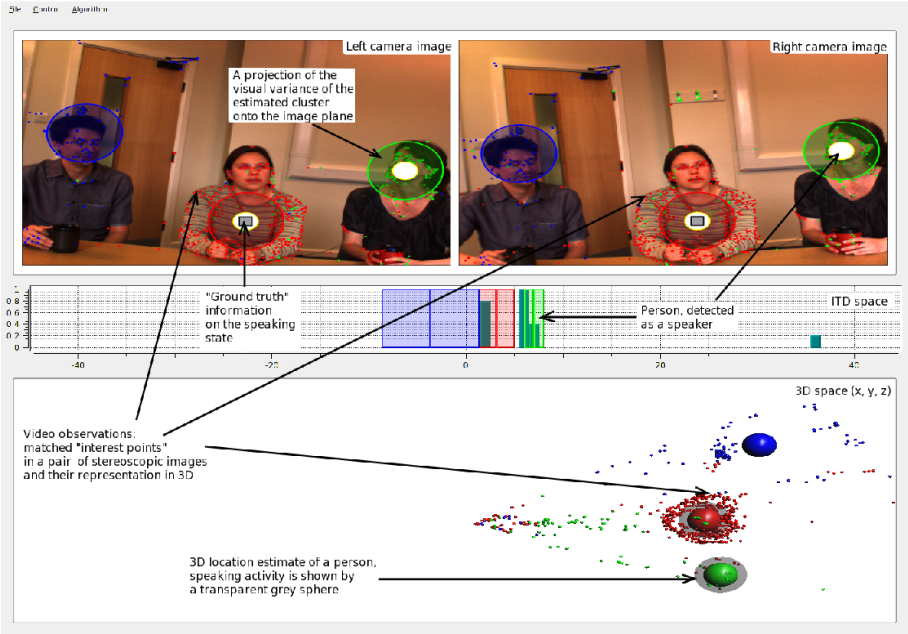


Figure 3. A typical output of the audio-visual clustering algorithm: a stereoscopic image pair (top), the ITD observation space (middle), and 3D clustering (bottom).

6.3. Stereoscopic vision

6.3.1. Dense stereo and Markov random fields

Current approaches to dense stereo matching estimate the disparity by maximizing its a posteriori probability, given the images and the prior probability distribution of the disparity function. This is done within a Markov random field model that makes tractable the computation of the joint probability of the disparity field. In practice the problem is analogous to minimizing the energy of an interacting spin system plunged into an external magnetic field. Statitistical thermodynamics provide the proper theoretical framework to model such a problem and to solve it using stochastic optimization techniques. However the latter are very slow. Alternative deterministic methods were recently used, such as deterministic annealing, mean-field approximation, graph cuts, and belief propagation. Basic assumptions of all these approaches are that the two images are properly rectified (such that the epipolar lines coincide with the image rows, that the illumination is homogeneous and the surfaces are lambertian (such that corresponding pixels have identical intensity values), and that there are not too many occluded or half-occluded surfaces. We investigated the link between intensity-based stereo and contour-based stereo. We carry out cooperatively both disparity and object boundary estimations by setting the two tasks in a unified Markovian framework. We define an original joint probabilistic model that allows to estimate disparities through a Markov random field model. Boundary estimation is then not reduced to a second independent step but cooperates with disparity estimation to gradually and jointly improve accuracy. The feedback from boundary estimation to disparity estimation is made through the use of an additional auxiliary field referred to as a displacement field. This field suggests the corrections that need to be applied at disparity discontinuities in order that they align with object boundaries. The joint model reduces to a Markov random field model when considering disparities while it reduces to a Markov chain when focusing on the displacement field. This work has been published in [28]. One drawback of such an approach, and of traditional stereo algorithms, is the use of the frontal parallel assumption that bias the resultÂ towards frontal parallel

plane solution. To overcome this issue, Â we are currently investigating the use of a joint random Markov field, so that to each pixel is associated a disparity value and a surface normal. The estimation of the two field is done alternatively using minimization methods described above.

6.3.2. Computational modeling of binocular vision

We continued our work that investigates the links between computational and biological stereopsis. This year we further investigated a model that takes into account eye/camera movements. This leads us to *cyclopean parameterizations of visual direction and binocular disparity* [14]. We provide a useful description of binocular geometry; not to construct a detailed model of biological stereopsis. For this reason, the estimation of scene and gaze variables is not considered in detail. Indeed, the present geometric account is compatible with a range of algorithmic models. It is not assumed that the orientation of the eyes is known. Rather, the binocular disparity field will be parameterized by a set of gaze variables, as well as by the scene structure. If the visual system is to recover the unknown gaze parameters from the observed disparity field, then this is the required representation. Although the orientation of the eyes is unknown, some qualitative constraints on oculomotor behaviour is observed. For example, it is assumed here that the left and right visual axes intersect at a point in space. This is approximately true, and moreover, in the absence of an intersection, it would be possible to define an appropriate chord between the left and right visual axes, and to choose a notional fixation point on this segment. In particular, it would be straightforward to extend the analysis of the disparity field to allow for mis-alignment of the eyes. In addition to the fixation constraint, it will be assumed that each eye rotates in accordance with Dondersâ law, meaning that the cyclo-rotation of the eyes can be estimated from the gaze direction. The âsmall baselineâ assumption (that the inter-ocular separation is small with respect to the viewing distance) will not be required here. Nor will it be assumed that the disparity function is continuous from point to point in the visual field.

6.4. Human-body tracking and human-motion capture

6.4.1. Human-body tracking using an implicit surface, 3D points, and surface normals.

We developed a method for tracking human motion based on fitting an articulated implicit surface to 3-D points and normals. There are two important contributions of this work to the state of the art. First, we introduce a new distance between an observation (a point and a normal) and an ellipsoid. We show that this can be used to define an implicit surface as a blending over a set of ellipsoids which are linked together to form a kinematic chain. Second, we exploit the analogy between the distance from a set of observations to the implicit surface and the observed-data log-likelihood of a mixture of Gaussian distributions. This allows us to cast the problem of implicit surface fitting into the problem of maximum likelihood estimation with missing variables. We argue that outliers are best described by a uniform component that is added to the mixture, and we formally derive the associated EM algorithm.

Casting the data-to-model association problem into unsupervised clustering has already been addressed in the past within the framework of point registration. We appear to be the first to apply EM clustering to the problem of fitting a blending of ellipsoids to a set of 3-D observations and to explicitly model outliers within this context.

6.4.2. Human-body tracking using the kinematics of extremal contours.

We also address the problem of human motion tracking from 2-D features available with image sequences [16]. The human body is described by an articulated mechanical chain and human body-parts are described by volumetric primitives with curved surfaces.

An extremal contour appears in an image whenever a curved surface turns smoothly away from the viewer. We have developed a method that relies on a kinematic parameterization of such extremal contours. The apparent motion of these contours in the image plane is a function of both the rigid motion of the surface and the relative position and orientation of the viewer with respect to the curved surface. The method relies onto the following key features: A parameterization of an extremal-contour point, and its associated image velocity, as a function of the motion parameters of the kinematic chain associated with the human body; The

Figure 4. From left to right: A set of ellipsoids and their associated articulated implicit surface; Observations composed of 3D points and normals; the pose of the articulated model that fits the observations.

zero-reference kinematic model and its usefulness for human-motion modelling; The chamfer-distance used to measure the discrepancy between predicted extremal contours and observed image contours; Moreover the chamfer distance is used as a differentiable multi-valued function and the tracker based on this distance is cast in an optimization framework. We have implemented a practical human-body tracker that may use an arbitrary number of cameras. One great methodological and practical advantage of our method is that it relies neither on model-to-image, nor on image-to-image point matches. In practice we model people with 5 kinematic chains, 19 volumetric primitives, and 54 degrees of freedom; We observe silhouettes in images gathered with several synchronized and calibrated cameras.

6.4.3. Inverse Kinematics using Sequential Monte Carlo Methods.

We proposed a new and original approach to solve the inverse kinematics problem. Our approach has the advantages to avoid the classical pitfalls of numerical inversion methods such as singularities and to accept arbitrary types of constraints. As shown fig 5 – where we compared the average time per iteration of two numerical IK solutions (the Jacobian transpose method and the damped pseudo-inverse methods) and our method – our approach exhibits a linear complexity with respect to degrees of freedom which makes it far more efficient for articulated figures with a high number of degrees of freedom. Our framework is based on Sequential Monte Carlo Methods that were initially designed to filter highly non-linear, non-Gaussian dynamic systems. They are used here in an online motion control algorithm that allows to integrate motion priors. The effectiveness of our method is shown fig 6 for a human figure animation application and fig 7 for an exemple of hand animation. Future work will consist in integrating measurements from image sequences to constrain the algorithm. This work received the best paper award at AMDO'08 [19].

6.4.4. Human-body tracking from a single camera

Tracking objects in 3D using as input a video sequence captured using a single camera has been known to be a very under-contstrained problem. This is especially valid if the target to be tracked is a human body. Other difficulties may be caused by cluttered backgrounds and poor image resolution. In this context, we address the problem of retrieving the 3D pose of a walking person using single viewpoint sequences, shot with a moving camera in everyday life environments. To this end we first initialize the body pose with the help of a walking motion model. This motion model is learnt after embedding the set of possible poses in a low dimensional space using Principal Component Analysis. This initialised pose is then refine using a novel Generalized Expectation Maximization algorithm [23]. This algorithms has the task of assigning the contour pixels, obtained from the input images after a few pre-processing steps, to the corresponding body part. It also

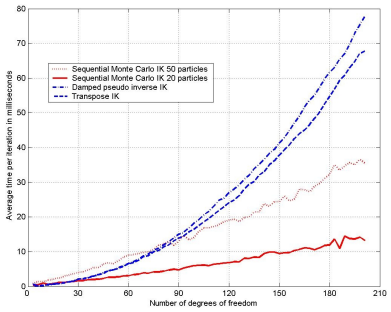


Figure 5. ***Performances of our method*** *Comparison with state-of-the-art IK methods. One can can effectively see the linear nature of the complexity of our method (instead of exponential with numeric IK).*

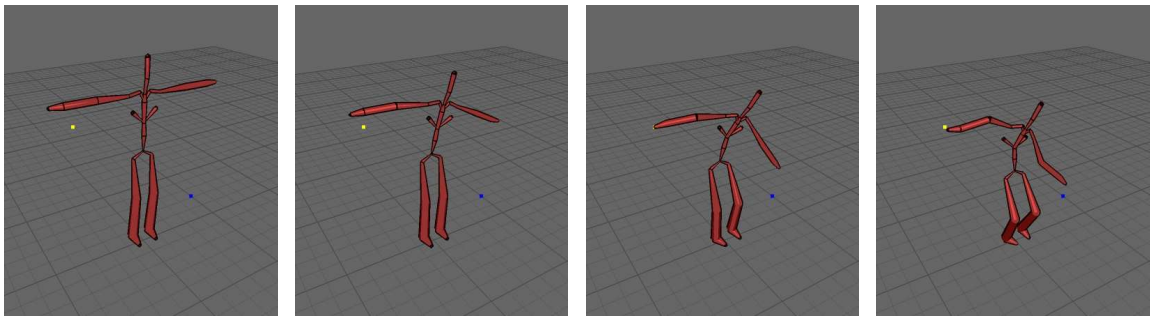


Figure 6. ***human figure animation*** *In this animation, feet are constrained to lie on the floor, the right hand is linked with the yellow dot while the left arm has the blue dot as target. Notice how the knees bend for the task to be achieved*

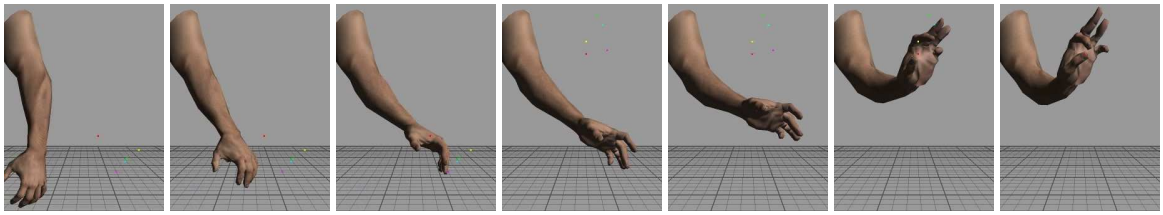


Figure 7. ***Hand animation*** *In this animation the fingers were given a target position represented as colored dots in the images. The two strips correspond to two different tasks that were chained along the animation.*

correctly finds the outlier pixels. The pose that gives the best match between the image measurements and the body parts is finally kept as output. This framework is promising and gives good results in the walking case. We now plan to extend it to track persons performing different activities.

6.5. Multiple camera reconstruction

6.5.1. Point-based reconstruction using robust factorization

The problem of 3-D reconstruction from multiple images is central in computer vision. Bundle adjustment provides a general method and practical algorithms for solving this reconstruction problem using maximum likelihood. Nevertheless, bundle adjustment is non-linear in nature and sophisticated optimization techniques are necessary, which in turn require proper initialization. Moreover, the combination of bundle adjustment with robust statistical methods to reject outliers is not clear both from the points of view of convergence properties and of efficiency.

We addressed the problem of building a class of robust factorization algorithms that solve for the shape and motion parameters (i.e., 3-D reconstruction) with both affine (weak perspective) and perspective camera models. We introduce a Gaussian/uniform mixture model and its associated EM algorithm. This allows us to address robust parameter estimation with an unsupervised clustering approach. We devise both an affine factorization algorithm and an iterative perspective factorization algorithm which are robust in the presence of a large number of outliers. We carry out numerous experiments to validate our algorithms and to compare them with existing ones. We also compare our approach with factorization methods that use M-estimators, [13], chapter 3.

6.5.2. Surface reconstruction based on mesh evolution

The point-based reconstruction algorithm just described provides sparse 3-D points that are impractical for rendering. Nevertheless, they can be used to build a rough mesh. We developed a method that starts with such a rough description and which consists in an evolution towards a very accurate description.

Most of the algorithms dealing with image based 3-D reconstruction involve the evolution of a surface based on a minimization criterion. The mesh parametrization, while allowing for an accurate surface representation, suffers from the inherent problems of not being able to reliably deal with self-intersections and topology changes. As a consequence, an important number of methods choose implicit representations of surfaces, e.g. level set methods, that naturally handle topology changes and intersections. Nevertheless, these methods rely on space discretizations, which introduce an unwanted precision-complexity trade-off. In this paper we explore a new mesh-based solution that robustly handles topology changes and removes self intersections, therefore overcoming the traditional limitations of this type of approaches. To demonstrate its efficiency, we present results on 3-D surface reconstruction from multiple images and compare them with state-of-the art results, [13].

6.5.3. Multi-view stereo with meshes

We propose a variational multi-view stereo vision method based on meshes for recovering 3D scenes (shape and radiance) from images. Our method is based on generative models and minimizes the reprojection error (difference between the observed images and the images synthesized from the reconstruction). Our contributions are twofold. 1) For the first time, we rigorously compute the gradient of the reprojection error for non smooth surfaces defined by discrete triangular meshes. The gradient correctly takes into account the visibility changes that occur when a surface moves; this forces the contours generated by the reconstructed surface to perfectly match with the apparent contours in the input images. 2) We propose an original modification of the Lambertian model to take into account deviations from the constant brightness assumption without explicitly modelling the reflectance properties of the scene or other photometric phenomena involved by the camera model. Our method is thus able to recover the shape and the diffuse radiance of non Lambertian scenes.

See [21] for more details. This work awards the CRS Industrial Prize, sponsored by Computer Recognition Systems.

6.6. Analysis and Exploitation of the reflectance properties and lighting

6.6.1. Image-based modeling of reflectance properties

We develop a variational method to recover both the shape and the reflectance of a scene surface(s) using multiple images, assuming that illumination conditions are fixed and known in advance. Scene and image formation are modeled with known information about cameras and illuminants, and scene recovery is achieved by minimizing a global cost functional with respect to both shape and reflectance. Contrary to previous works which consider specific individual scenarios, our method applies to a number of classical scenarios – classical stereovision, multiview photometric stereo, and multiview shape from shading. In addition, our approach naturally combines stereo, silhouette and shading cues in a single framework and, unlike most previous methods dealing with only Lambertian surfaces, the proposed method considers general dichromatic surfaces. For more detail see [43].

6.6.2. Shape from ambient shading

We study the mathematical and numerical aspects of the estimation of the 3-D shape of a Lambertian scene seen under diffuse illumination. This problem is known as "shape from ambient shading" (SFAS), and its solution consists of integrating a strongly non-local and non-linear Integro-Partial Differential Equation (I-PDE). We provide a first analysis of this global I-PDE, whereas previous work had focused on a local version that ignored effects such as occlusion of the light field. We also design an original approximation scheme which, following Barles and Souganidis' theory, ensures the correctness of the numerical approximations, and discuss about some numerical issues. This work has been submitted to SSVM 2009 [39].

6.6.3. Reflectance segmentation on meshes

We have started to considered the problem of segmenting data on meshes in order to introduce segmentation constraints on the reflectance properties in 3D shape reconstruction problems. We have formulated the problem in the form of a convex optimization based on the Chan-Vese segmentation models. Since our model is convex we avoid local minima and we obtain global minima via a simple gradient descent. Also this make our method robust to initialization. The optimization is made over a discrete domain consisting of the nodes in the mesh giving efficient implementations. To demonstrate the robustness of the proposed model, we present some numerical results on various synthetic data and real data from computer vision applications. This work has been submitted to SSVM 2009 [38]. The actual method being limited to two regions segmentation and having been used only for segmenting the radiance (instead of reflectance), this work is still in progress.

6.7. Motion Segmentation

We developed a novel tool for body-part segmentation and tracking in the context of multiple camera systems. Our goal is to produce robust motion cues over time sequences, as required by human motion analysis applications. Given time sequences of 3D body shapes, body-parts are consistently identified over time without any supervision or a priori knowledge. The approach first maps shape representations of a moving body to an embedding space using locally linear embedding. While this map is updated at each time step, the shape of the embedded body remains stable. Robust clustering of body parts can then be performed in the embedding space by k-wise clustering, and temporal consistency is achieved by propagation of cluster centroids. The contribution with respect to methods proposed in the literature is a totally unsupervised spectral approach that takes advantage of temporal correlation to consistently segment body-parts over time. Comparisons on real data are run with direct segmentation in 3D by EM clustering and ISOMAP-based clustering: the way different approaches cope with topology transitions is discussed in [20].

6.8. Articulated shape matching

Matching articulated shapes described as meshes or, more generally, as sets of 3-D points reduces to maximal sub-graph isomorphism when representing each set of points as a weighted graph. Spectral graph theory can be used to map these graphs onto lower dimensional isometric spaces and match shapes by aligning their

embeddings in virtue of their invariance to change of pose. Classical graph isomorphism schemes relying on the ordering of the eigenvalues to align Laplacian eigenvectors fail when handling large data-sets or noisy data. We derive a new formulation equivalent to finding the best alignment between two congruent K-dimensional sets of points, where the dimension K of the embedded space results from the selection of the best subset of eigenfunctions of the Laplacian operator. This set is detected by matching the signatures of those eigenfunctions expressed as histograms, and provides a smart initialization for the alignment problem with a considerable impact on the overall performance. Dense matching then reduces to embedded point registration under orthogonal transformations, a task we cast into the framework of unsupervised clustering and solve using the EM algorithm. Maximal subset matching of non identical shapes is handled by defining an appropriate outlier class. Experimental results on challenging examples show how the algorithm naturally treats changes of topology, shape variations and different sampling densities, Figure 8 and [27].

Figure 8. From left to right: A mannequin, a person, and their matched voxel representations.

6.9. Temporal surface tracking

We address the problem of surface tracking in multiple camera environments and over time sequences. In order to fully track a surface undergoing significant deformations, we cast the problem as a mesh evolution over time [13]. Such an evolution is driven by 3D displacement fields estimated between meshes recovered independently at different time frames. Geometric and photometric information is used to identify a robust set of matching vertices. This provides a sparse displacement field that is densified over the mesh by Laplacian diffusion. In contrast to existing approaches that evolve meshes, we do not assume a known model or a fixed topology. The contribution is a novel mesh evolution based framework that allows to fully track, over long sequences, an unknown surface encountering deformations, including topological changes. Results on very challenging and publicly available image based 3D mesh sequences demonstrate the ability of our framework to efficiently recover surface motions [33].

6.10. Action representation and recognition

Traditional approaches to action recognition model actions as space-time representations which explicitly or implicitly encode the dynamics of an action through temporal dependencies. In contrast, we have proposed a new compact and efficient representation which does not account for such dependencies. Instead, motion sequences are represented with respect to a set of discriminative static key pose-exemplars and without modeling any temporal ordering. The interest is a time-invariant representation that drastically simplifies learning and recognition by removing time related information such as speed or length of an action. The proposed representation is equivalent to embedding actions into a space defined by distances to key poses [34].

6.11. Omnidirectional vision

6.11.1. Unified Imaging Geometry for Catadioptric Cameras.

Catadioptric devices consist of usually curved mirrors and cameras; their main goal is to provide a large field of view. Together with fisheyes, they are currently the most popular omnidirectional cameras. The geometry of such cameras has been rather well understood in recent years, e.g. it has been known how points and lines are imaged by them and what the epipolar geometry looks like. However, for most essential building blocks of imaging geometry, general closed-form algebraic expressions have not been known. In our recent work [31], we provide such expressions, for the main building blocks: projection of 3D points, backprojection of image points, projection of conics and quadrics, epipolar geometry (fundamental matrix), and homographies between images of planar objects. These results are based on Veronese mappings of coordinate vectors and matrices. It is also shown that these expressions can be similarly factorized as the corresponding ones for perspective cameras, e.g. the decomposition of projection matrices into intrinsic and extrinsic parameters. On the one hand, this work provides the first complete algebraic formulation of the imaging geometry of catadioptric cameras, on the other hand it also clarifies and deepens the geometric understanding of it. This work has been done in collaboration with Joao Barreto from Coimbra University (Portugal).

One of the results provided in [31] consists in the formulation of a projection matrix for catadioptric cameras and its decomposition into intrinsic and extrinsic parameters. In [18], we propose a calibration approach based on this, that works analogously to the standard DLT (Direct Linear Transform) approach known for perspective cameras. This work has been done in collaboration with Yalin Bastanlar (Ankara, Turkey), Luis Puig and Josechu Guerrero (Zaragoza, Spain), and Joao Barreto (Coimbra, Portugal), mainly during visits of the PhD students Bastanlar and Puig in PERCEPTION.

6.11.2. Matching of Omnidirectional Images.

In collaboration with Luis Puig and Josechu Guerrero from Zaragoza University (Spain), we have investigated approaches for matching images taken by an omnidirectional camera with images taken by an omnidirectional or perspective camera [29]. Two main aspects have been considered. First, a generalization of the epipolar geometry for such image pairs, that allows to define geometric matching constraints (e.g., the epipolar curves in the considered omnidirectional images are conics). Second, descriptors (here, SIFT) used to characterize and match points in omnidirectional images, are computed in images warped in order to remove non-perspective distortions. This was found to improve matching performance significantly. Figure 9 shows a results for pair consisting of one omnidirectional and one perspective image.

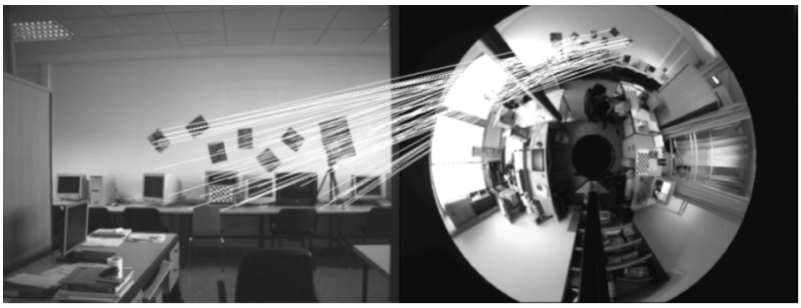


Figure 9. Result of matching between an omnidirectional and a perspective image.

6.11.3. Plane-Based Calibration of Linear Cameras.

In this work, we consider cameras consisting of a one-dimensional image sensor that moves along a line while acquiring images, which are stacked together in order to form a complete 2D image. Most prominent examples are pushbroom cameras, which have been a dominant technology for satellite imaging, and flatbed scanners. We show how to calibrate such sensors from images of planar calibration grids [22]. Calibration consists in the estimation of the 1D sensor's parameters, as well as the direction and speed of its displacement. Our approach is based on a generalization of plane homographies, from perspective to linear cameras.

6.11.4. Minimal Solutions for Generic Camera Calibration.

In previous work [41], [40] we have introduced a generic concept for camera calibration, allowing to calibrate practically any camera type, based on a generic camera model. The provided algorithms start with a linear least-squares solution which is then optimized using non-linear optimization. In practice, it is usually beneficial to replace the suboptimal linear least squares part which uses more than the actual minimal required data, by a method that uses that amount of minimal data. This is for two main reasons: first, if the method is to be embedded in a sampling scheme in order to make it robust to outliers (e.g. RANSAC), then the fewer data are required, the less samples are needed. Second, a solution obtained from minimal data and that fits these data exactly, is often more accurate than a solution obtained from a linear equation system and more data.

In [30], we propose minimal solutions for the generic calibration concept. The main result is a method requiring 4 matches across the images of 3 calibration grids (3 grids are always required for generic calibration). The solution is obtained *via* the computation of the Gröbner basis of an underlying polynomial equation system. A sample result for the distortion correction applied to a fisheye image, based on the computed calibration, is shown in figure 10.

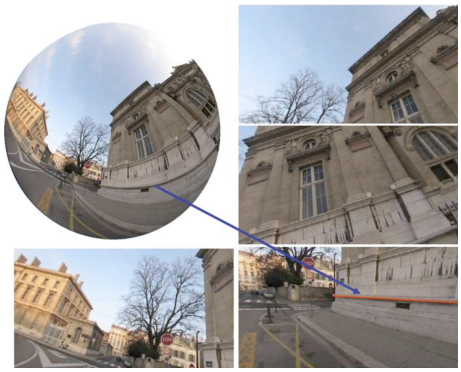


Figure 10. Distortion correction of a fisheye image using the cali- bration obtained with our 4-point algorithm [30].

6.11.5. General Matching Tensors for Line Images.

Geometrical constraints for the matching of points and lines in perspective images have been well understood since the last decade. They can be represented *via* bifocal, trifocal or quadrifocal so-called matching tensors, that relate matching primitives in two to four views. In previous work [42], we have shown how to obtain such matching tensors for any camera model, i.e. not necessarily perspective cameras, provided the cameras are calibrated. This has been done for the matching of points. As for line images, the situation is rather more complicated since without a given parametric camera model, there are no closed-form representations for line images. In [24], we show how to still obtain matching tensors for line images. In order to do so, we consider image points that have been determined as lying on line images, and formulate matching tensors based on them. The difference with the previous matching tensors for points is that here, it is not required to determine matching points in different images, but that it is enough to consider points in different images that lie on the

projections of a same 3D line. We show how to estimate the tensors and how to extract the camera motion from them.

6.12. Other results

6.12.1. Multiperspective Images.

In collaboration with Jingyi Yu (U Delaware) and Leonard McMillan (U North Carolina), we have published a State-of-the-Art-Report (STAR) at Eurographics 2008 [35]. It reviews several aspects concerning so-called multiperspective images, i.e. images generated by a camera that does not have a single optical center. Main motivations are artistic rendering of objects (see an example in figure 11) or the generation of depictions of objects in a single image that are more complete than what could be achieved with e.g. a perspective image. Multiperspective images and cameras have also been used for motion estimation and 3D modeling.

Figure 11. Left: Nusch Eluard by Pablo Picasso. Right: A multiperspective image synthesized using the GLC framework proposed by Yu and McMillan.

7. Contracts and Grants with Industry

7.1. Contract with ASTRIUM

This year we started a three year contract with ASTRIUM. High-resolution satellite imagery is typically based on so-called push-broom cameras (one or a few rows of pixels, covering different spectral bands). High-resolution images are generated by stitching individual push-broom images, taken at successive time instants, together. The main goal of our work is to develop both theoretical models and algorithms for modeling, estimating and compensating these vibrations, using information contained in the images.

8. Other Grants and Activities

8.1. National initiatives

8.1.1. ANR project CAVIAR

The global topic of the CAVIAR project (http://www.anr-caviar.org/) is to use omnidirectional cameras for aerial robotics. Our team implements calibration software for various kinds of omnidirectional cameras. We will develop approaches for matching images obtained with such cameras, as well as for performing camera motion estimation and 3D reconstruction of environments from them. This information is to be used for aiding an aerial robot's navigation and for 3D map generation.

This 3-year project started in December 2005. The partners are CREA (Amiens, coordinator), LAAS (Toulouse), ICARE (INRIA Sophia-Antipolis), and LE2I (Le Creusot). The current team members who are involved in this project are Peter Sturm and Simone Gasparini.

8.1.2. ANR project FLAMENCO

FLAMENCO is a 3-year project that has started on January 1, 2007. This project deals with the challenges of spatio-temporal scene reconstruction from several video sequences, i.e. from images captured from different viewpoints and at different time instants. This project tackles the following three important factors which limit the major problems in computer vision so far:

- the computational time / the poor resolution of the models: the acquisition of video sequences from multiple cameras generates a very large amount of data, which makes the design of efficient algorithms very important. The high computational cost of existing methods has limited the spatial resolution of the reconstruction and has allowed to handle video sequences of a few seconds only, which is prohibitive in real applications.
- the lack of spatio-temporal coherence: to our knowledge, none of the existing methods has been able to reconstruct coherent spatio-temporal models: Most methods build threedimensional models at each time step without taking advantage of the continuity of the motion and of the temporal coherence of the model. This issue requires elaborating new mathematical and algorithmic tools dedicated to four-dimensional representations (three space dimensions plus the time dimension).
- the simplicity of the models: the information available in multiple video sequences of a scene are not restricted to geometry and motion. Most reconstruction methods disregard such information as the illumination of the scene, and the reflectance, the materials and the textures of the objects. Our goal is to build more exhaustive models, by automatically estimating these parameters concurrently to geometry and motion. For example, in augmented reality, reflectance properties allow to synthesize novel views with higher photo-realism.

In this project, we are collaborating with the CERTIS laboratory (Ecole Nationale des Ponts et Chaussees) and the PRIMA group (INRIA Rhone-Alpes) via Frederic Devernay.

The team members directly involved in this project are Peter Sturm, Emmanuel Prados (INRIA researchers) and Amael Delaunoy (PhD thesis). During 2007, they have focused on the illumination and the reflectance models.

8.1.3. ARC–FANTASTIK

In 3D control animation, one of the main difficulty is to take into account both the kinematic and dynamic constraints to obtain a physically plausible motion. Classical approaches are based on a global spacetime optimisation. The fact that they are both time consuming and non sequential make them difficult to use in practice. As an alternative, within this project, we propose to investigate the use of statistical tools, such as sequential Monte Carlo approaches combined with dimension reduction techniques, to the problem of motion control, where the evolution law will be defined using dynamic constraints, and the data collected from a motion cature system will constraint the solution sequentially.

The partners of this project are INRIA Rhône-Alpes (PERCEPTION and EVASION teams), the university of Bretagne Sud (équipe SAMSARA), and ENS Cachan (Centre de Mathématiques et de Leurs Applications). The team member involved in this project is Elise Arnaud.

8.1.4. ADT GrimDev

GrimDev is an ADT (Action de Developpement Technologique) proposed in the context of the Grimage interactive and immersive platform. The objective of GrimDev is to organize and manage software developements around the Grimage platform in order to ensure their reusabilities and durations. GrimDev was proposed by the Perception team and involves the following teams from INRIA Grenoble Rhône-Alpes: Perception, Evasion, Moais and SED.

8.2. Projects funded by the European Commission

8.2.1. FP6/Marie-Curie EST Visitor

Visitor is a 4 year European project (2004-2008) under the Marie-Curie actions for young researcher mobility – Early Stage Training or EST. Within these actions, VISITOR has been selected to host PhD students granted by the European commission. The PERCEPTION team actively participated in the project elaboration. Edmond Boyer is the coordinator of this project and we host two PhD students from this program.

8.2.2. FP6/Marie-Curie RTN VISIONTRAIN

VISIONTRAIN is a 4 year Marie Curie Research Training Network, or RTN (2005-2009) coordinated by Radu Horaud. This network gathers 11 partners from 11 European countries and has the ambition to address foundational issues in computational and cognitive vision systems through an European doctoral and post-doctoral program.

VISIONTRAIN addresses the problem of understanding vision from both computational and cognitive points of view. The research approach is based on formal mathematical models and on the thorough experimental validation of these models. We intend to reduce the gap that exists today between biological vision (which performs outstandingly well and fast but not yet understood) and computer vision (whose robustness, flexibility, and autonomy remain to be demonstrated). In order to achieve these ambitious goals, 11 internationally recognized academic partners work cooperatively on a number of targeted research topics: computational theories and methods for low-level vision, motion understanding from image sequences, learning and recognition of shapes, categories, and actions, cognitive modelling of the action of seeing, and functional imaging for observing and modelling brain activity. There are three categories of researchers involved in this network: doctoral students, post-doctoral researchers, as well as highly experienced researchers. The work includes participation to proof-of-concept achievements, annual thematic schools, industrial meetings, attendance of conferences, etc.

8.2.3. FP6 IST STREP project POP

We are coordinators of the POP project (Perception on Purpose) involving the MISTIS and the PERCEPTION INRIA groups, as well as 4 other partners: University of Osnabruck (cognitive neuroscience), University Hospital Hamburg-Eppendorf (neurophysiology), University of Coimbra (robotics), and University of Sheffield (hearing and speech). POP proposes the development of a fundamentally new approach, perception on purpose, which is based on 5 principles. First, visual and auditory information should be integrated in both space and time. Second, active exploration of the environment is required to improve the audiovisual signal-to-noise ratio. Third, the enormous potential sensory requirements of the entire input array should be rendered manageable by multimodal models of attentional processes. Fourth, bottom-up perception should be stabilized by top-down cognitive function and lead to purposeful action. Finally, all parts of the system should be underpinned by rigorous mathematical theory, from physical models of low-level binocular and binaural sensory processing to trainable probabilistic models of audiovisual scenes.

8.2.4. FP6-IST STREP project INTERACT

The INTERACT project considers Human Machine Interfaces based on both speech and hand motion. The objective is the capability to manipulate virtual 3D objects using hands and speech. The resulting system will be based on computer vision techniques for the capturing hand motion and on speech recognition. 4 partners were involved in this two-year project: PERCEPTION (INRIA Rhone-Alpes), Holographica (Hungary),Total-Immersion (France) and Vecsys (France).

9. Dissemination

9.1. Editorial boards and program committees

- Radu Horaud is a member of the editorial boards of the *International Journal of Robotics Research* and of the *International Journal of Computer Vision*, he is an *area editor* for *Computer Vision and Image Understanding*, *Image and Vision Computing* and an *associated editor* for *Machine Vision Applications* and *IET Computer Vision*.
- Edmond Boyer has been a member of the program committees of: cvpr2008, eccv2008, bmvc2008, cvmp2008, omnivis2008, ISUVR2008
- Peter Sturm is a member of the editorial boards of the *Image and Vision Computing* journal, the *Journal of Computer Science and Technology* and the Transactions on Computer Vision and Applications – IPSJ (Information Processing Society of Japan)
- Peter Sturm has been a member of the Program Committees of: CVPR, ICPR, Congress of the International Society for Photogrammetry and Remote Sensing, OMNIVIS, NORDIA, VISAPP, ICVGIP, ISVC.
- Emmanuel Prados has been a member of the Program Committees of: 3DFP'08 program committee (Workshop in 3D Face Processing) : in conjunction with CVPR 2008 - Anchorage, Alaska on June 27th 2008.

9.2. Services to the Scientific Community

- Edmond Boyer: Eccv Area Chair Workshop Organization in June 2008, Eccv'08 Organisation in October 2008, Eccv'08 Video Chair.
- Edmond Boyer is coordinator of the Marie-Curie Visitor Project and member of the Visitor Scientific Committe.
- Radu Horaud is the coordinator of the Visiontrain Marie Curie Research Training Network.
- Emmanuel Prados is the coordinator of the Flamenco Project (ANR-MDCA-2007-2010).
- Peter Sturm is the Co-chairman of the Working Group III/1 "Automatic Calibration and Orientation of Optical Cameras" of the ISPRS (International Society for Photogrammetry and Remote Sensing), 2004-2008.
- Peter Sturm is Chairing of the Working Group "Imaging and Geometry" of the GdR ISIS, since 2006.

9.3. Teaching

- Representation de connaissance et inference, m2r, ujf, 15h, E. Arnaud
- Tutorat d'apprentis, m2p, ujf, 45 h, E. Arnaud
- image retrieval, m2p, ujf, 15h, E. Arnaud
- Outils mathematiques, m1, ujf, 10h, E. Arnaud
- Methodes statistiques pout la biologie, l2, ujf, 36h, E. Arnaud
- Informatique instrumentale et multimedia, l1, ujf, 66h, E. Arnaud
- Decouverte des mathematiques appliquees, l1, ujf, 18h, E. Arnaud
- 3D Modelling from Images or Videos, m2r, ujf, 18h, P. Sturm and E. Boyer
- Synthese d'images, m1, ujf, 60h, E. Boyer
- Projet, m1, ujf, 15h, E. Boyer
- Vision par Ordinateur, m2p, ujf, 40h, E. Boyer
- Synthese d'images, m1, Polytech, 36h, E. Boyer
- Modelisation 3D, m2r, inpg, 18h, E. Boyer
- Introduction aux techniques de l'images, l3, ujf, 15h, E. Boyer

9.4. Tutorials and invited talks

- Emmanuel Prados has given an invited talk to the Workshop âmathematical methods for image analysisâ, OrlÃ©ans, France, April 2008.
- Peter Sturm has given an invited talk on *Calibrage de caméras à des fins métrologiques.* Journée Thématique Métrologie 3D par vision, Lyon, France, March 2008.
- Peter Sturm has given a Keynote lecture on *General Imaging – Design, Modelling and Applications* at VISAPP – International Conference on Computer Vision Theory and Applications, Funchal, Madeira, Portugal, January 2008.
- Peter Sturm has given a Seminar entitled *General Imaging – Design, Modelling and Applications* within the VIBOT Erasmus Mundus programme, Le Creusot, France, November 2008.
- Edmond Boyer: AMDO 2008, Keynote speaker July 2008, Deutch Telecom Berlin, invited talk, January 2008.

9.5. Thesis

- Julien Morat [11]
- Andrei Zaharescu [13]
- Daniel Weinland [12]

10. Bibliography

Major publications by the team in recent years

[1] P. GARGALLO, E. PRADOS, P. STURM. *Minimizing the Reprojection Error in Surface Reconstruction from Images*, in "Proceedings of the International Conference on Computer Vision, Rio de Janeiro, Brazil", IEEE Computer Society Press, 2007, http://perception.inrialpes.fr/Publications/2007/GPS07.

[2] R. P. HORAUD, G. CSURKA, D. DEMIRDJIAN. *Stereo Calibration from Rigid Motions*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", vol. 22, n^{o} 12, December 2000, p. 1446-1452, ftp://ftp.inrialpes.fr/pub/movi/publications/HoraudCsurkaDemirdjian-pami2000.ps.gz.

[3] S. LAZEBNIK, E. BOYER, J. PONCE. *On How to Compute Exact Visual Hulls of Object Bounded by Smooth Surfaces*, in "Proceedings of the Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii, USA", IEEE Computer Society Press, Dec 2001, http://perception.inrialpes.fr/publication.php3?bibtex=LBP01.

[4] S. PETITJEAN, E. BOYER. *Regular and Non-Regular Point Sets: Properties and Reconstruction*, in "Computational Geometry - Theory and Application", vol. 19, n^{o} 2-3, 2001, p. 101-126, http://perception.inrialpes.fr/publication.php3?bibtex=PB01.

[5] E. PRADOS, O. FAUGERAS. *Shape from Shading: a well-posed problem ?*, in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, California", vol. II, IEEE, jun 2005, p. 870–877, http://perception.inrialpes.fr/Publications/2005/PF05a.

[6] S. RAMALINGAM, P. STURM, S. LODHA. *Towards Complete Generic Camera Calibration*, in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, California", vol. 1, jun 2005, p. 1093-1098, http://perception.inrialpes.fr/Publications/2005/RSL05a.

[7] A. RUF, R. P. HORAUD. *Visual Servoing of Robot Manipulators, Part I: Projective Kinematics*, in "International Journal of Robotics Research", vol. 18, n^{o} 11, November 1999, p. 1101-1118, http://hal.inria.fr/inria-00073002.

[8] P. STURM, S. MAYBANK. *On Plane-Based Camera Calibration: A General Algorithm, Singularities, Applications*, in "Proceedings of the Conference on Computer Vision and Pattern Recognition, Fort Collins, Colorado, USA", June 1999, p. 432-437.

[9] P. STURM, S. RAMALINGAM. *A Generic Concept for Camera Calibration*, in "Proceedings of the European Conference on Computer Vision, Prague, Czech Republic", vol. 2, Springer, May 2004, p. 1-13, http://perception.inrialpes.fr/Publications/2004/SR04.

[10] P. STURM. *Critical Motion Sequences for Monocular Self-Calibration and Uncalibrated Euclidean Reconstruction*, in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Puerto Rico", Juin 1997, p. 1100-1105.

Year Publications

Doctoral Dissertations and Habilitation Theses

[11] J. MORAT. *Vision stéréoscopique par ordinateur pour la détection et le suivi de cibles pour une application automobile*, Ph. D. Thesis, Institut National Polytechnique de Grenoble, Grenoble, France, July 2008, http://perception.inrialpes.fr/Publications/2008/Mor08.

[12] D. WEINLAND. *Action Representation and Recognition*, Ph. D. Thesis, INPG, October 2008, http://perception.inrialpes.fr/Publications/2008/Wei08.

[13] A. ZAHARESCU. *Contributions to Spatial and Temporal 3-D Reconstruction from Multiple Cameras*, Ph. D. Thesis, Institut National Polytechnique de Grenoble, Grenoble, France, November 2008, http://perception.inrialpes.fr/Publications/2008/Zah08.

Articles in International Peer-Reviewed Journal

[14] M. HANSARD, R. P. HORAUD. *Cyclopean Geometry of Binocular Vision*, in "Journal of the Optical Society of America A", vol. 25, n^{o} 9, September 2008, p. 2357-2369, http://perception.inrialpes.fr/Publications/2008/HH08.

[15] H. JIN, D. CREMERS, D. WANG, E. PRADOS, A. YEZZI, S. SOATTO. *3-D Reconstruction of Shaded Objects from Multiple Images Under Unknown Illumination*, in "International Journal of Computer Vision", vol. 76, n^{o} 3, March 2008, http://perception.inrialpes.fr/Publications/2008/JCWPYS08.

[16] D. KNOSSOW, R. RONFARD, R. P. HORAUD. *Human Motion Tracking with a Kinematic Parameterization of Extremal Contours*, in "International Journal of Computer Vision", vol. 79, n^{o} 2, September 2008, p. 247-269, http://perception.inrialpes.fr/Publications/2008/KRH08.

International Peer-Reviewed Conference/Proceedings

[17] E. ARNAUD, H. CHRISTENSEN, Y.-C. LU, J. BARKER, V. KHALIDOV, M. HANSARD, B. HOLVECK, H. MATHIEU, R. NARASIMHA, E. TAILLANT, F. FORBES, R. P. HORAUD. *The CAVA corpus: synchronised stereoscopic and binaural datasets with head movements*, in "ACM/IEEE International Confer-

ence on Multimodal Interfaces (ICMI'08)", October 2008, http://perception.inrialpes.fr/Publications/2008/ACLBKHHMNTFH08.

[18] Y. BASTANLAR, L. PUIG, P. STURM, J. GUERRERO, J. BARRETO. *DLT-Like Calibration of Central Catadioptric Cameras*, in "Proceedings of the Workshop on Omnidirectional Vision, Camera Networks and Non-Classical Cameras, Marseille, France", oct 2008, http://perception.inrialpes.fr/Publications/2008/BPSGB08.

[19] N. COURTY, E. ARNAUD. *Inverse Kinematics using Sequential Monte Carlo Methods*, in "Conference on articulated motion and deformable object", LNCS, july 2008, http://perception.inrialpes.fr/Publications/2008/CA08.

[20] F. CUZZOLIN, D. MATEUS, D. KNOSSOW, E. BOYER, R. P. HORAUD. *Coherent Laplacian 3-D Protrusion Segmentation*, in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition", 2008, http://perception.inrialpes.fr/Publications/2008/CMKBH08.

[21] A. DELAUNOY, E. PRADOS, P. GARGALLO, J.-P. PONS, P. STURM. *Minimizing the Multi-view Stereo Reprojection Error for Triangular Surface Meshes*, in "Proceedings of the 19th British Machine Vision Conference, Leeds, UK", Award: CRS Industrial Prize, BMVA, sept 2008, http://perception.inrialpes.fr/Publications/2008/DPGPS08.

[22] J. DRARÉNI, P. STURM, S. ROY. *Plane-Based Calibration for Linear Cameras*, in "Proceedings of the Workshop on Omnidirectional Vision, Camera Networks and Non-Classical Cameras, Marseille, France", oct 2008, http://perception.inrialpes.fr/Publications/2008/DSR08.

[23] A. FOSSATI, E. ARNAUD, R. P. HORAUD, P. FUA. *Tracking Articulated Bodies using Generalized Expectation Maximization*, in "Workshop on Non-Rigid Shape Analysis and Deformable Image Alignment (CVPR workshop, NORDIA'08)", june 2008, http://perception.inrialpes.fr/Publications/2008/FAHF08.

[24] S. GASPARINI, P. STURM. *Multi-View Matching Tensors from Lines for General Camera Models*, in "In Proceedings of the CVPR 2008 Workshop on Tensors in Image Processing and Computer Vision", IEEE Press, Jun 2008, http://perception.inrialpes.fr/Publications/2008/GS08.

[25] V. KHALIDOV, F. FORBES, M. HANSARD, E. ARNAUD, R. P. HORAUD. *Audio-Visual Clustering for Multiple Speaker Localization*, in "5th International Workshop on Machine Learning for Multimodal Interaction (MLMI'08)", LNCS, Springer, September 2008, p. 86–97, http://perception.inrialpes.fr/Publications/2008/KFHAH08.

[26] V. KHALIDOV, F. FORBES, M. HANSARD, E. ARNAUD, R. P. HORAUD. *Detection and Localization of 3D Audio-Visual Objects Using Unsupervised Clustering*, in "ACM/IEEE International Conference on Multimodal Interfaces (ICMI'08)", October 2008, http://perception.inrialpes.fr/Publications/2008/KFHAH08a.

[27] D. MATEUS, R. P. HORAUD, D. KNOSSOW, F. CUZZOLIN, E. BOYER. *Articulated Shape Matching Using Laplacian Eigenfunctions and Unsupervised Point Registration*, in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition", 2008, http://perception.inrialpes.fr/Publications/2008/MHKCB08.

[28] R. NARASIMHA, E. ARNAUD, F. FORBES, R. P. HORAUD. *Cooperative disparity and object boundary estimation*, in "IEEE International Conference on Image Processing", oct 2008, http://perception.inrialpes.fr/Publications/2008/NAFH08.

[29] L. PUIG, J. GUERRERO, P. STURM. *Matching of Omindirectional and Perspective Images using the Hybrid Fundamental Matrix*, in "Proceedings of the Workshop on Omnidirectional Vision, Camera Networks and Non-Classical Cameras, Marseille, France", oct 2008, http://perception.inrialpes.fr/Publications/2008/PGS08.

[30] S. RAMALINGAM, P. STURM. *Minimal Solutions for Generic Imaging Models*, in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage", jun 2008, http://perception.inrialpes.fr/Publications/2008/RS08.

[31] P. STURM, J. BARRETO. *General Imaging Geometry for Central Catadioptric Cameras*, in "Proceedings of the 10th European Conference on Computer Vision, Marseille, France", vol. 4, Springer, oct 2008, p. 609–622, http://perception.inrialpes.fr/Publications/2008/SB08.

[32] M. TOURNIER, L. REVERET, X. WU, N. COURTY, E. ARNAUD. *Motion Compression using Principal Geodesics Analysis*, in "ACM Siggraph/Eurographics Symposium on Computer Animation, SCA (Poster)", july 2008, http://perception.inrialpes.fr/Publications/2008/TRWCA08.

[33] K. VARANASI, A. ZAHARESCU, E. BOYER, R. P. HORAUD. *Temporal Surface Tracking Using Mesh Evolution*, in "Proceedings of the Tenth European Conference on Computer Vision, Marseille, France", LNCS, vol. Part II, Springer-Verlag, October 2008, p. 30–43, http://perception.inrialpes.fr/Publications/2008/VZBH08.

[34] D. WEINLAND, E. BOYER. *Action Recognition using Exemplar-based Embedding*, in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage", 2008, p. 1–7, http://perception.inrialpes.fr/Publications/2008/WB08.

[35] J. YU, L. MCMILLAN, P. STURM. *Multiperspective Modeling, Rendering, and Imaging*, in "Proceedings of Eurographics, Crete, Greece", STAR - State of the Art Report, apr 2008, http://perception.inrialpes.fr/Publications/2008/YMS08.

[36] A. ZAHARESCU, C. CAGNIART, S. ILIC, E. BOYER, R. P. HORAUD. *Camera Clustering for Multi-Resolution 3-D Surface Reconstruction*, in "ECCV 2008 Workshop on Multi Camera and Multi-modal Sensor Fusion Algorithms and Applications", 2008, http://perception.inrialpes.fr/Publications/2008/ZCIBH08.

National Peer-Reviewed Conference/Proceedings

[37] P. GARGALLO, E. PRADOS, P. STURM. *Minimiser l'erreur de reprojection en reconstruction de surfaces basée images*, in "RFIA'08, janvier 2008, Amiens, France", 2008, http://perception.inrialpes.fr/Publications/2008/GPS08.

References in notes

[38] A. DELAUNOY, K. FUNDANA, E. PRADOS, A. HEYDEN. *Total Variation Based Segmentation on Meshes*, in "Submitted to IEEE Computer Society Conference on Computer Vision and Pattern (CVPR)", 2009.

[39] E. PRADOS, N. JINDAL, S. SOATTO. *A global approaxch to the Shape From Ambient Shading problem*, in "Submitted to the Second International Conference on Scale Space Methods and Variational Methods in Computer Vision, Voss, Norway, June 1 - June 5", 2009.

[40] S. RAMALINGAM, P. STURM, S. LODHA. *Towards Complete Generic Camera Calibration*, in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, California", vol. 1, jun 2005, p. 1093-1098, http://perception.inrialpes.fr/Publications/2005/RSL05a.

[41] P. STURM, S. RAMALINGAM. *A Generic Concept for Camera Calibration*, in "Proceedings of the European Conference on Computer Vision, Prague, Czech Republic", vol. 2, Springer, May 2004, p. 1-13, http://perception.inrialpes.fr/Publications/2004/SR04.

[42] P. STURM. *Multi-View Geometry for General Camera Models*, in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, California", vol. 1, jun 2005, p. 206-212, http://perception.inrialpes.fr/Publications/2005/Stu05.

[43] K.-J. YOON, E. PRADOS, P. STURM. *Joint Estimation of Shape and Reflectance using Multiple Images with Known Illumination Conditions*, in "the International Journal of Computer Vision", to appear, 2009, http://perception.inrialpes.fr/Publications/2009/YPS09.