IN PARTNERSHIP WITH:
CNRS

Institut polytechnique de Grenoble

Université Joseph Fourier (Grenoble)

Activity Report 2012

Team PERCEPTION

Interpretation and Modeling of Images and Videos

IN COLLABORATION WITH: Laboratoire Jean Kuntzmann (LJK)

RESEARCH CENTER
Grenoble - Rhône-Alpes

THEME
Vision, Perception and Multimedia Understanding

Table of contents

Team PERCEPTION

Keywords: Computer Vision, Auditory Signal Analysis, Machine Learning, Audio-Visual Fusion, Human-Robot Interaction

Creation of the Team: September 01, 2006 *, Updated into Project-Team:* January 01, 2008 .

1. Members

Research Scientists
Radu Horaud [Team leader, Research Director (DR), HdR]
Laurent Girin [Professor at Grenoble INP, HdR]

Engineer
Michel Amat [Development Engineer]

PhD Students
Xavier Alameda-Pineda [MESR grant]
Antoine Deleforge [MESR grant]
Maxime Janvier [DGA-Inria grant]
Kaustubh Kulkarni [Inria grant]
Jordi Sanchez-Riera [Inria grant]
Avinash Sharma [Inria grant]

Post-Doctoral Fellows
Jan Cech [Research Engineer]
Georgios Evangelidis [Post-doctoral Researcher]

2. Overall Objectives

2.1. Introduction

The overall objective of the PERCEPTION group is to develop theories, models, methods, and systems allowing computers to see, to hear and to understand what they see and what they hear. A major difference between classical computer systems and computer perception 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. This is difficult 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 gather sounds using several microphones, to localize and separate sounds composed of several auditory sources, and to analyse and interpret them. Sound localization, separation and recognition is difficult, especially in the presence of noise, reverberant rooms, competing sources, overlap of speech and prosody, etc.

A third objective is to analyse articulated and moving 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 fourth objective is to combine vision and hearing in order to disambiguate situations when a single modality is not sufficient. In particular we are interested in defining the notion of *audio-visual object* (AVO) and to deeply understand the mechanisms allowing to associate visual data with auditory data.

A fifth objective is to build vision systems, hearing systems, and audio-visual systems able to interact with their environment, possibly in real-time. In particular we are interested in building the concept of an audio-visual robot that communicates with people in the most natural way.

2.2. Highlights of the Year

Figure 1. Audio-visual interaction between a person and the humanoid robot NAO developed under the HUMAVIPS project.

2.2.1. The European project Humavips – Humanoids with Auditory and Visual Abilities in Populated Spaces

HUMAVIPS (http://humavips.inrialpes.fr) is a 36 months FP7 STREP project coordinated by Radu Horaud and which started in 2010. The project addresses multimodal perception and cognitive issues associated with the computational development of a social robot. The ambition is to endow humanoid robots with audiovisual (AV) abilities: exploration, recognition, and interaction, such that they exhibit adequate behavior when dealing with a group of people. Research and technological developments emphasize the role played by multimodal perception within principled models of human-robot interaction and of humanoid behavior.

2.2.2. Collaboration with SAMSUNG – 3D Capturing and Modeling from Scalable Camera Configurations

In 2010 started a multi-year collaboration with the Samsung Advanced Institute of Technology (SAIT), Seoul, Korea. Whithin this project we develop a methodology able to combine data from several types of visual sensors (2D high-definition color cameras and 3D range cameras) in order to reconstruct, in real-time, an indoor scene without any constraints in terms of background, illumination conditions, etc. In 2012 we developed a novel TOF-stereo algorithm.

2.2.3. Book on Time-of-Flight Cameras

A book on Time-of-Flight Cameras was published in 2012 in the collection *Springer Briefs in Computer Science*. The book stems from the scientific collaboration between the PERCEPTION team and SAIT. The book describes a variety of recent research into time-of-flight imaging. Time-of-flight cameras are used to estimate 3D scene-structure directly, in a way that complements traditional multiple-view reconstruction methods. The first two chapters of the book explain the underlying measurement principle, and examine the associated sources of error and ambiguity. Chapters three and four are concerned with the geometric calibration of time-of-flight cameras, particularly when used in combination with ordinary colour cameras. The final chapter shows how to use time-of-flight data in conjunction with traditional stereo matching techniques. The five chapters, together, describe a complete depth and colour 3D reconstruction pipeline. This book will be useful to new researchers in the field of depth imaging, as well as to those who are working on systems that combine colour and time-of-flight cameras. The publisher's url of the book is http://www.springer.com/computer/image+processing/book/978-1-4471-4657-5#.

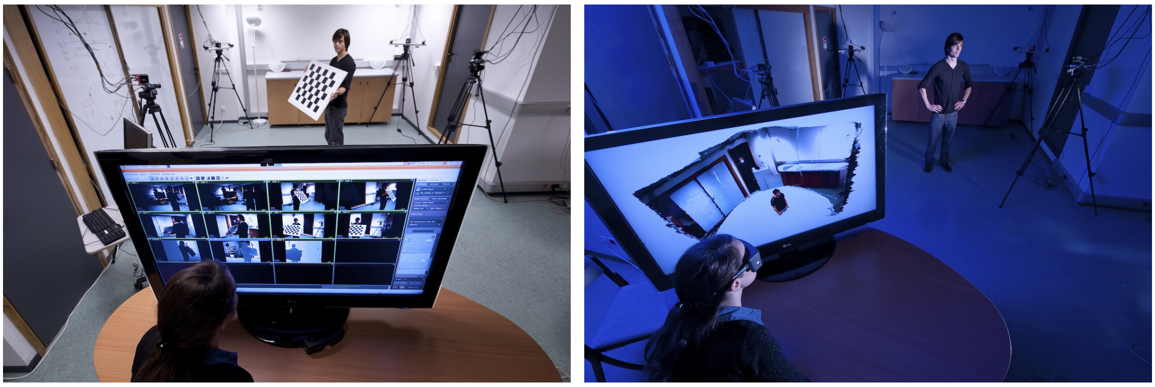

Figure 2. The mixed TOF-stereo multiple-camera system developed in collaboration with Samsung Electronics. Left: Geometric calibration of the camera system. Right: Live 3D display.

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 “appearance” 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.

Current camera acquisition systems are all-digital ones. They are based on standard network communication protocols such as the IEEE 1394. Recent systems involve as well depth cameras that produce depth images, i.e. a depth information at each pixel. Popular technologies for this purpose include the Time of Flight Cameras (TOF cam) and structured light cameras, as in the very recent Microsoft's Kinect device.

3.6. Auditory and audio-visual scene analysis

For the last two years, PERCEPTION has started to investigate a new research topic, namely the analysis of auditory information and the fusion between auditory and visual data. In particular we are interested in analyzing the acoustic layout of a scene (how many sound sources are out there and where are they located? what is the semantic content of each auditory signal?) For that purpose we use microphones that are mounted onto a human-like head. This allows the extraction of several kinds of auditory cues, either based on the time difference of arrival or based on the fact that the head and the ears modify the spectral properties of the sounds perceived with the left and right microphones. Both the temporal and spectral binaural cues can be used to locate the most proeminent sound sources, and to separate the perceived signal into several sources. This is however an extremely difficult task because of the inherent ambiguity due resemblance of signals, and of the presence of acoustic noise and reverberations. The combination of visual and auditory data allows to solve the localization and separation tasks in a more robust way, provided that the two stimuli are available. One interesting yet unexplored topic is the development of hearing for robots, such as the role of head and body motions in the perception of sounds.

4. Application Domains

4.1. Human action recognition

We are particularly interested in the analysis and recognition of human actions and gestures. The vast majority of research groups concentrate on isolated action recognition. We address continuous recognition. The problem is difficult because one has to simultaneously address the problems of recognition and segmentation. For this reason, we adopt a per-frame representation and we develop methods that rely on dynamic programming and on hidden Markov models. We investigate two type of methods: one-pass methods and two-pass methods. One-pass methods enforce both within-action and between-action constraints within sequence-to-sequence alignment algorithms such as dynamic time warping or the Viterbi algorithm. Two-pass methods combine a per-action representation with a discriminative classifier and with a dynamic programming post-processing stage that find the best sequence of actions. These algorithms were well studied in the context of large-vocabulary continuous speech recognition systems. We investigate the modeling of various per-frame representations for action and gesture analysis and we devise one-pass and two-pass algorithms for recognition.

4.2. 3D reconstruction using TOF and color cameras

TOF cameras are active-light range sensors. An infrared beam of light is generated by the device and depth values can be measured by each pixel, provided that the beam travels back to the sensor. The associated depth measurement is accurate if the sensed surface sends back towards the sensor a fair percentage of the incident light. There is a large number of practical situations where the depth readings are erroneous: specular and bright surfaces (metal, plastic, etc.), scattering surfaces (hair), absorbing surfaces (cloth), slanted surfaces, e.g., at the bounding contours of convex objects which are very important for reconstruction, mutual reflections, limited range, etc. The resolution of currently available TOF cameras is of 0.3 to 0.5MP. Modern 2D color cameras deliver 2MP images at 30FPS or 5MP images at 15FPS. It is therefore judicious to attempt to combine the active-range and the passive-stereo approaches within a mixed methodology and system. Standard stereo matching methods provide an accurate depth map but are often quite slow because of the inherent complexity of the matching algorithms. Moreover, stereo matching is ambiguous and inaccurate in the presence of weakly textured areas. We develop TOF-stereo matching and reconstruction algorithms that are able to combine the advantages of the two types of depth estimation technologies.

4.3. Sound-source separation and localization

We explore the potential of binaural audition in conjunction with modern machine learning methods in order to address the problems of sound source separation and localization. We exploit the spectral properties of interaural cues, namely the interaural level difference (ILD) and the interaural phase difference (IPD). We have started to develop a novel supervised framework based on a training stage. During this stage, a sound source emits a broadband random signal which is perceived by a microphone pair embedded into a dummy head with a human-like head related transfer function (HRTF). The source emits from a location parameterized by azimuth and elevation. Hence, a mapping between a high-dimensional interaural spectral representation and a low-dimensional manifold can be estimated from these training data. This allows the development of various single-source localization methods as well as multiple-source separation and localization methods.

4.4. Audio-visual fusion for human-robot interaction

Modern human-robot interaction systems must be able to combine information from several modalities, e.g., vision and hearing, in order to allow high-level communication via gesture and vocal commands, multimodal dialogue, and recognition-action loops. Auditory and visual data are intrinsically different types of sensory data. We have started the development of a audio-visual mixture model that takes into account the heterogenous nature of visual and auditory observations. The proposed multimodal model uses modality specific mixtures (one mixture model for each modality). These mixtures are tied through latent variables that parameterize the joint audiovisual space. We thoroughly investigate this novel kind of mixtures with their associated efficient parameter estimation procedures.

5. Software

5.1. Mixed camera platform

We started to develop a multiple camera platform composed of both high-definition color cameras and low-resolution depth cameras. This platform combines the advantages of the two camera types. On one side, depth (time-of-flight) cameras provide relatively accurate 3D scene information. On the other side, color cameras provide information allowing for high-quality rendering. The software package developed during the year 2011 contains the calibration of TOF cameras, alignment between TOF and color cameras, and image-based rendering. These software developments are performed in collaboration with the Samsung Advanced Institute of Technology. The multi-camera platform and the basic software modules are products of 4D Views Solutions SAS, a start-up company issued from the PERCEPTION group.

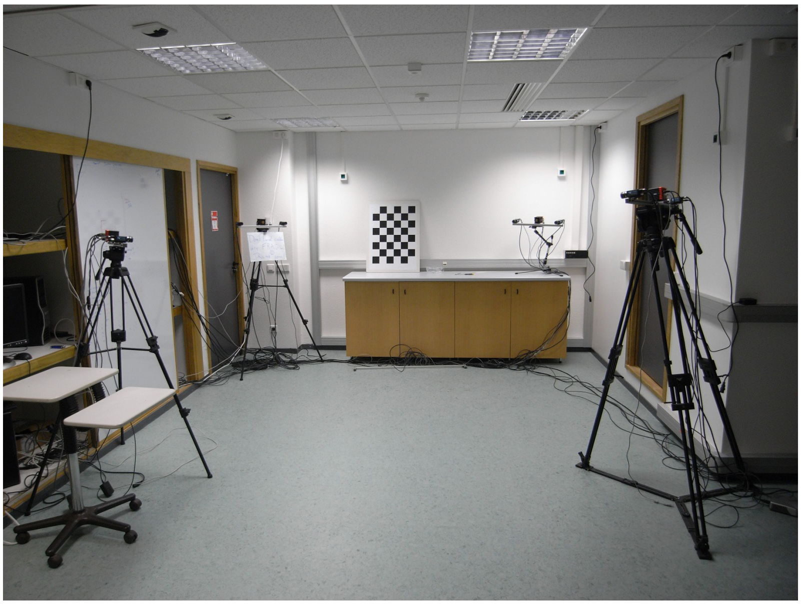

Figure 3. The mixed multi-camera system composed of four TOF-stereo sensor units.

5.2. Audiovisual robot head

We have developed two audiovisual (AV) robot heads: the POPEYE head and the NAO stereo head. Both are equipped with a binocular vision system and four microphones. The software modules comprise stereo matching and reconstruction, sound-source localization and audio-visual fusion. 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. The NAO stereo head is being developed under the European project HUMAVIPS (http://humavips.inrialpes.fr) in collaboration with Aldebaran Robotics (which manufactures the humanoid robot NAO) and with the University of Bielefeld, the Czech Technical Institute, and IDIAP. The software modules that we develop are compatible with both these robot heads.

6. New Results

6.1. 3D shape analysis and registration

We address the problem of 3D shape registration and we propose a novel technique based on spectral graph theory and probabilistic matching. Recent advancement in shape acquisition technology has led to the capture of large amounts of 3D data. Existing real-time multi-camera 3D acquisition methods provide a frame-wise reliable visual-hull or mesh representations for real 3D animation sequences The task of 3D shape analysis involves tracking, recognition, registration, etc. Analyzing 3D data in a single framework is still a challenging task considering the large variability of the data gathered with different acquisition devices. 3D shape registration is one such challenging shape analysis task. The main contribution of this chapter is to extend the spectral graph matching methods to very large graphs by combining spectral graph matching with Laplacian embedding. Since the embedded representation of a graph is obtained by dimensionality reduction we claim that the existing spectral-based methods are not easily applicable. We discuss solutions for the

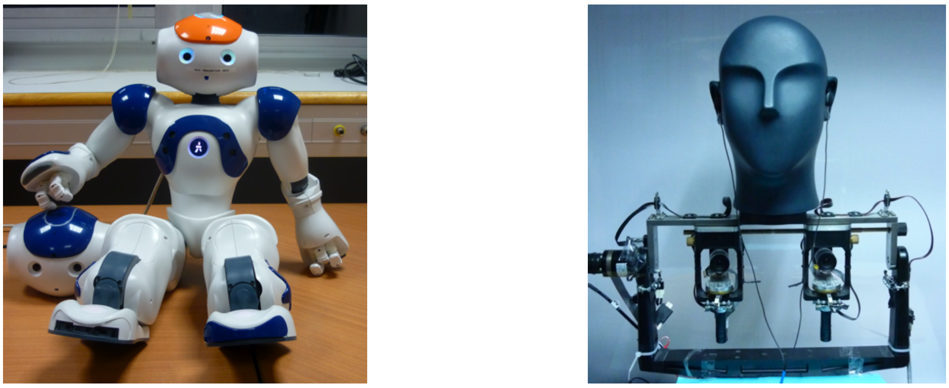

Figure 4. Left: The consumer humanoid robot NAO is equipped with a binocular-binaural head specially designed for human-humanoid interaction; Right: The binocular-binaural robot head POPEYE equipped with a four degrees of freedom stereo camera pair and with a dummy head.

exact and inexact graph isomorphism problems and recall the main spectral properties of the combinatorial graph Laplacian; We provide a novel analysis of the commute-time embedding that allows us to interpret the latter in terms of the PCA of a graph, and to select the appropriate dimension of the associated embedded metric space; We derive a unit hyper-sphere normalization for the commute-time embedding that allows us to register two shapes with different samplings; We propose a novel method to find the eigenvalue-eigenvector ordering and the eigenvector sign using the eigensignature (histogram) which is invariant to the isometric shape deformations and fits well in the spectral graph matching framework, and we present a probabilistic shape matching formulation using an expectation maximization point registration algorithm which alternates between aligning the eigenbases and finding a vertex-to-vertex assignment. See [22], [34], [19] for more details.

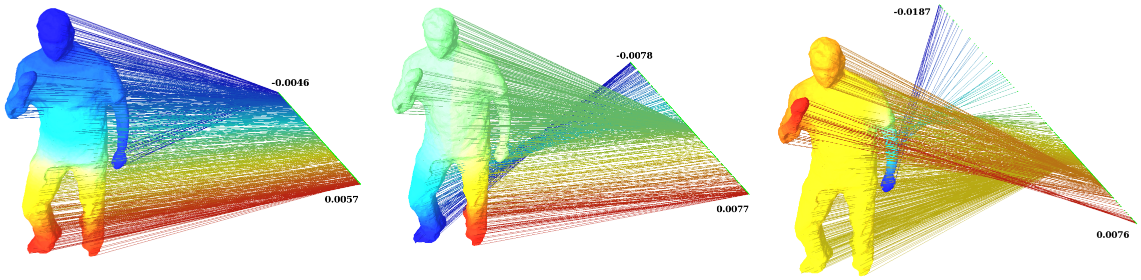

Figure 5. This is an illustration of the concept of the PCA of a shape embedding. The shapes's vertices are projected onto the second, third and fourth eigenvectors of the Laplacian matrix. These eigenvectors can be viewed as the principal directions of the shape (see [34] for more details).

6.2. High-resolution depth maps based on TOF-stereo fusion

The combination of range sensors with color cameras can be very useful for a wide range of applications, e.g., robot navigation, semantic perception, manipulation, and telepresence. Several methods of combining range- and color-data have been investigated and successfully used in various robotic applications. Most of these systems suffer from the problems of noise in the range-data and resolution mismatch between the range sensor and the color cameras, since the resolution of current range sensors is much less than the resolution of color cameras. High-resolution depth maps can be obtained using stereo matching, but this often fails to construct accurate depth maps of weakly/repetitively textured scenes, or if the scene exhibits complex self-occlusions. Range sensors provide coarse depth information regardless of presence/absence of texture. The use of a calibrated system, composed of a time-of-flight (TOF) camera and of a stereoscopic camera pair, allows data fusion thus overcoming the weaknesses of both individual sensors. We propose a novel TOF-stereo fusion method based on an efficient seed-growing algorithm which uses the TOF data projected onto the stereo image pair as an initial set of correspondences. These initial "seeds" are then propagated based on a Bayesian model which combines an image similarity score with rough depth priors computed from the low-resolution range data. The overall result is a dense and accurate depth map at the resolution of the color cameras at hand. We show that the proposed algorithm outperforms 2D image-based stereo algorithms and that the results are of higher resolution than off-the-shelf color-range sensors, e.g., Kinect. Moreover, the algorithm potentially exhibits real-time performance on a single CPU. See [27], [33] for more details.

6.3. Simultaneous sound-source separation and localization

Human-robot communication is often faced with the difficult problem of interpreting ambiguous auditory data. For example, the acoustic signals perceived by a humanoid with its on-board microphones contain a mix of sounds such as speech, music, electronic devices, all in the presence of attenuation and reverberations. We proposed a novel method, based on a generative probabilistic model and on active binaural hearing, allowing a robot to robustly perform sound-source separation and localization. We show how interaural spectral cues can be used within a constrained mixture model specifically designed to capture the richness of the data gathered with two microphones mounted onto a human-like artificial head. We describe in detail a novel expectation-maximization (EM) algorithm that alternates between separation and localization, we analyse its initialization, speed of convergence and complexity, and we assess its performance with both simulated and real data. Subsequently, we studied the *binaural manifold*, i.e., the low-dimensional space of sound-source locations embedded in the high-dimensional space of perceived interaural spectral features, and we provided a method for mapping interaural cues onto source locations. See [25], [24], [26]

6.4. Sound localization and recognition with a humanoid robot

We addressed the problem of localizing recognizing everyday sound events in indoor environments with a consumer robot. For localization, we use the four microphones that are embedded into the robot's head. We developed a novel method that uses four non-coplanar microphones and that guarantees that for each set of pairwise TDOA (time difference of arrival) there is a unique 3D source location. For recognition, sounds are represented in the spectrotemporal domain using the stabilized auditory image (SAI) representation. The SAI is well suited for representing pulse-resonance sounds and has the interesting property of mapping a time-varying signal into a fixed-dimension feature vector space. This allows us to map the sound recognition problem into a supervised classification problem and to adopt a variety of classifications schemes. We developed a complete system that takes as input a continuous signal, splits it into significant isolated sounds and noise, and classifies the isolated sounds using a catalogue of learned sound-event classes. The method is validated with a large set of audio data recorded with a humanoid robot in a typical home environment. Extended experiments showed that the proposed method achieves state-of-the-art recognition scores with a twelve-class problem, while requiring extremely limited memory space and moderate computing power. A first real-time embedded implementation in a consumer robot show its ability to work in real conditions. See [23], [28] for more details.

6.5. Audiovisual fusion based on a mixture model

The problem of multimodal clustering arises whenever the data are gathered with several physically different sensors. Observations from different modalities are not necessarily aligned in the sense there there is no obvious way to associate or to compare them in some common space. A solution may consist in considering multiple clustering tasks independently for each modality. The main difficulty with such an approach is to guarantee that the unimodal clusterings are mutually consistent. In this paper we show that multimodal clustering can be addressed within a novel framework, namely conjugate mixture models. These models exploit the explicit transformations that are often available between an unobserved parameter space (objects) and each one of the observation spaces (sensors). We formulate the problem as a likelihood maximization task and we derive the associated expectation-maximization algorithm. The algorithm and its variants are tested and evaluated within the task of 3D localization of several speakers using both auditory and visual data. See [36], [30], [29] for more details.

7. Bilateral Contracts and Grants with Industry

7.1. Bilateral Contract with Samsung Electronics

We continued a collaboration with the Samsung Advanced Institute of Technology (SAIT), Seoul, South Korea. Within this project we develop a methodology able to combine data from several types of visual sensors (2D high-definition color cameras and 3D range cameras) in order to reconstruct, in real-time, an indoor scene without any constraints in terms of background, illumination conditions, etc. A software package was successfully installed in December 2012 at Samsung.

8. Partnerships and Cooperations

8.1. European Initiatives

8.1.1. FP7 Projects

8.1.1.1. HUMAVIPS

Title: Humanoids with audiovisual skills in populated spaces

Type: COOPERATION (ICT)

Defi: Cognitive Systems and Robotics

Instrument: Specific Targeted Research Project (STREP)

Duration: February 2010 - January 2013

Coordinator: Inria (France)

Others partners: CTU Prague (Czech Republic), University of Bielefeld (Germany), IDIAP (Switzerland), Aldebaran Robotics (France)

See also: http://humavips.inrialpes.fr

Abstract: Humanoids expected to collaborate with people should be able to interact with them in the most natural way. This involves significant perceptual, communication, and motor processes, operating in a coordinated fashion. Consider a social gathering scenario where a humanoid is expected to possess certain social skills. It should be able to explore a populated space, to localize people and to determine their status, to decide to join one or two persons, to synthetize appropriate behavior, and to engage in dialog with them. Humans appear to solve these tasks routinely by integrating the often complementary information provided by multi sensory data processing, from low-level 3D object positioning to high-level gesture recognition and dialog handling. Understanding the world from unrestricted s

8.2. International Research Visitors

8.2.1. Visits of International Scientists

8.2.1.1. Internships

Charlotte CLARK (from Apr 2012 until Jul 2012)
Subject: Piecewise Planar Reconstruction of a Scene from Depth Data
Institution: Massachusetts Institute of Technology (United States)

Siva KUMAR (from May 2012 until Jul 2012)
Subject: Visual Matching Using Kernel Canonical Correlation Analysis
Institution: IIT Delhi (India)

Ravi Kant MITTAL (from May 2012 until Jul 2012)
Subject: Finding Audio Visual Objects (AVO) with the Kinect
Institution: IIT Delhi (India)

Christopher STOCK (from May 2012 until Aug 2012)
Subject: Detection of keypoints on 2D manifolds
Institution: Harvard University (United States)

9. Dissemination

9.1. Scientific Animation

- Radu Horaud is a member of the following editorial boards:
  - advisory board member of the *International Journal of Robotics Research*,
  - associate editor of the *International Journal of Computer Vision*, and
  - area editor of *Computer Vision and Image Understanding*.
- Radu Horaud was a PC member of the IEEE International Conference on Humanoid Robotics (HUMANOIDS 2012), Osaka, Japan.
- Georgios Evangelidis was a PC member of the ACCV'12 Workshop on Color Depth Fusion for Computer Vision, Dajeon, Korea.
- Xavier Alameda-Pineda organized the D-META challenge in conjunction with the 2012 ACM/IEEE International Conference on Multimodal Interaction, Santa Monica, CA, USA.

9.2. Teaching - Supervision - Juries

9.2.1. Teaching

Doctorat : Radu Horaud, Data Analysis and Manifold Learning, 30 hours, Université de Grenoble, France

9.2.2. Supervision

PhD: Avinash Sharma, Représentation, Segmentation et Appariement de Formes Visuelles 3D Utilisant le Laplacient et le Noyau de la Chaleur, Grenoble INP, 29 October 2012, Radu Horaud

PhD in progress: Kaustubh Kulkarni, Continuous Action Recognition, November 2009, Radu Horaud

PhD in progress: Jordi Sanchez-Rieira, 3D Human-Robot Interaction with NAO, November 2009, Radu Horaud

PhD in progress: Antoine Deleforge, Sound-Source Separation and Localization, October 2010, Radu Horaud

PhD in progress: Xavi Alameda-Pineda, Audio-Visual Fusion for HRI, October 2010, Radu Horaud

PhD in progress: Maxime Janvier, Sound Recognition for Humanoids, November 2012, Radu Horaud

10. Bibliography

Major publications by the team in recent years

[1] N. ANDREFF, B. ESPIAU, R. HORAUD. *Visual Servoing from Lines*, in "International Journal of Robotics Research", August 2002, vol. 21, n^{o} 8, p. 679-700, http://perception.inrialpes.fr/Publications/2002/AEH02.

[2] A. Bartoli, N. Dalal, R. Horaud. *Motion Panoramas*, in "Computer Animation and Virtual Worlds", 2004, vol. 15, p. 501-517, http://perception.inrialpes.fr/Publications/2004/BDH04.

[3] Y. Dufournaud, C. Schmid, R. Horaud. *Image Matching with Scale Adjustment*, in "Computer Vision and Image Understanding", February 2004, vol. 93, n^{o} 2, p. 175-194, http://perception.inrialpes.fr/Publications/2004/DSH04.

[4] J.-S. Franco, E. Boyer. *Efficient Polyhedral Modeling from Silhouettes*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", March 2009, vol. 31, n^{o} 3, p. 414–427,, http://perception.inrialpes.fr/Publications/2009/FB09.

[5] 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.

[6] M. Hansard, R. Horaud. *Cyclopean Geometry of Binocular Vision*, in "Journal of the Optical Society of America A", September 2008, vol. 25, n^{o} 9, 2357Ð2369, http://perception.inrialpes.fr/Publications/2008/HH08.

[7] R. Horaud, G. Csurka, D. Demirdjian. *Stereo Calibration from Rigid Motions*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", December 2000, vol. 22, n^{o} 12, p. 1446-1452, http://perception.inrialpes.fr/Publications/2000/HCD00.

[8] R. Horaud, M. Niskanen, G. Dewaele, E. Boyer. *Human Motion Tracking by Registering an Articulated Surface to 3-D Points and Normals*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", January 2009, vol. 31, n^{o} 1, p. 158-164, http://perception.inrialpes.fr/Publications/2009/HNDB09.

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

[10] D. Mateus, R. 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.

[11] S. Ramalingam, S. Lodha, P. Sturm. *A Generic Structure-from-Motion Framework*, in "Computer Vision and Image Understanding", sep 2006, vol. 103, n^{o} 3, p. 218–228, http://perception.inrialpes.fr/Publications/2006/RLS06.

[12] C. Sminchisescu, B. Triggs. *Estimating Articulated Human Motion with Covariance Scaled Sampling*, in "International Journal of Robotics Research", 2003, http://perception.inrialpes.fr/Publications/2003/ST03.

[13] J.-P. Tardif, P. Sturm, M. Trudeau, S. Roy. *Calibration of Cameras with Radially Symmetric Distortion*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", 2009, vol. 31, n^{o} 9, p. 1552-1566, http://perception.inrialpes.fr/Publications/2009/TSTR09.

[14] D. WEINLAND, R. RONFARD, E. BOYER. *Free Viewpoint Action Recognition using Motion History Volumes*, in "Computer Vision and Image Understanding", November/December 2006, vol. 104, n^{o} 2-3, p. 249–257, http://perception.inrialpes.fr/Publications/2006/WRB06a.

[15] M. WILCZKOWIAK, P. STURM, E. BOYER. *Using Geometric Constraints Through Parallelepipeds for Calibration and 3D Modelling*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", feb 2005, vol. 27, n^{o} 2, p. 194-207, http://perception.inrialpes.fr/Publications/2005/WSB05.

[16] K.-J. YOON, E. PRADOS, P. STURM. *Joint Estimation of Shape and Reflectance using Multiple Images with Known Illumination Conditions*, in "International Journal of Computer Vision", 2010, vol. 86, n^{o} 2-3, p. 192–210, http://perception.inrialpes.fr/Publications/2010/YPS10.

[17] A. ZAHARESCU, E. BOYER, K. VARANASI, R. HORAUD. *Surface Feature Detection and Description with Applications to Mesh Matching*, in "Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition", Miami Beach, Florida, June 2009, http://perception.inrialpes.fr/Publications/2009/ZBVH09.

[18] A. ZAHARESCU, R. HORAUD. *Robust Factorization Methods Using A Gaussian/Uniform Mixture Model*, in "International Journal of Computer Vision", March 2009, vol. 81, n^{o} 3, p. 240-258, http://perception.inrialpes.fr/Publications/2009/ZH09.

Publications of the year

Doctoral Dissertations and Habilitation Theses

[19] A. SHARMA. *Représentation, Segmentation et Appariement de Formes Visuelles 3D Utilisant le Laplacient et le Noyau de la Chaleur*, Institut National Polytechnique de Grenoble - INPG, October 2012, http://hal.inria.fr/tel-00768768.

Articles in International Peer-Reviewed Journals

[20] X. ALAMEDA-PINEDA, J. SANCHEZ-RIERA, J. WIENKE, V. FRANC, J. CECH, K. KULKARNI, A. DELEFORGE, R. HORAUD. *RAVEL: An Annotated Corpus for Training Robots with Audiovisual Abilities*, in "Journal on Multimodal User Interfaces", 2012 [*DOI :* 10.1007/S12193-012-0111-Y], http://hal.inria.fr/hal-00720734.

[21] M. SAPIENZA, M. HANSARD, R. HORAUD. *Real-time Visuomotor Update of an Active Binocular Head*, in "Autonomous Robots", September 2012 [*DOI :* 10.1007/S10514-012-9311-2], http://hal.inria.fr/hal-00768615.

[22] A. ZAHARESCU, E. BOYER, R. HORAUD. *Keypoints and Local Descriptors of Scalar Functions on 2D Manifolds*, in "International Journal of Computer Vision", 2012, vol. 100, n^{o} 1, p. 78-98 [*DOI :* 10.1007/S11263-012-0528-5], http://hal.inria.fr/hal-00699620.

International Conferences with Proceedings

[23] X. ALAMEDA-PINEDA, R. HORAUD. *Geometrically-constrained Robust Time Delay Estimation Using Non-coplanar Microphone Arrays*, in "Proceedings of the 20th European Signal Processing Conference (EUSIPCO)", Bucharest, Romania, August 2012, p. 1309 - 1313, http://hal.inria.fr/hal-00768763.

[24] A. DELEFORGE, R. HORAUD. *A Latently Constrained Mixture Model for Audio Source Separation and Localization*, in "10th International Conference on Latent Variable Analysis and Signal Separation", Tel Aviv, Israel, LNCS, Springer, March 2012, vol. 7191, p. 372–379 [*DOI :* 10.1007/978-3-642-28551-6_46], http://hal.inria.fr/hal-00768660.

[25] A. DELEFORGE, R. HORAUD. *The Cocktail Party Robot: Sound Source Separation and Localisation with an Active Binaural Head*, in "7th ACM/IEEE International Conference on Human Robot Interaction (HRI)", Boston, United States, March 2012, p. 431 - 438, http://hal.inria.fr/hal-00768668.

[26] A. DELEFORGE, R. HORAUD. *2D Sound-Source Localization on the Binaural Manifold*, in "IEEE Workshop on Machine Learning for Signal Processing", Santander, Spain, IEEE, September 2012 [*DOI :* 10.1109/MLSP.2012.6349784], http://hal.inria.fr/hal-00768657.

[27] V. GANDHI, J. CECH, R. HORAUD. *High-Resolution Depth Maps Based on TOF-Stereo Fusion*, in "IEEE International Conference on Robotics and Automation", Saint-Paul Minnesota, United States, IEEE Robotics and Automation Society, May 2012, p. 4742 - 4749 [*DOI :* 10.1109/ICRA.2012.6224771], http://hal.inria.fr/hal-00725616.

[28] M. JANVIER, X. ALAMEDA-PINEDA, L. GIRIN, R. HORAUD. *Sound-Event Recognition with a Companion Humanoid*, in "IEEE International Conference on Humanoid Robotics (Humanoids)", Osaka, Japan, December 2012, http://hal.inria.fr/hal-00768767.

[29] J. SANCHEZ-RIERA, X. ALAMEDA-PINEDA, R. HORAUD. *Audio-Visual Robot Command Recognition*, in "Proceedings of the 14th ACM international conference on Multimodal interaction (ICMI'12)", Santa-Monica, CA, United States, October 2012, p. 371-378 [*DOI :* 10.1145/2388676.2388760], http://hal.inria.fr/hal-00768761.

[30] J. SANCHEZ-RIERA, X. ALAMEDA-PINEDA, J. WIENKE, A. DELEFORGE, S. ARIAS, J. CECH, S. WREDE, R. HORAUD. *Online Multimodal Speaker Detection for Humanoid Robots*, in "IEEE International Conference on Humanoid Robotics (Humanoids)", Osaka, Japan, December 2012, http://hal.inria.fr/hal-00768764.

[31] J. SANCHEZ-RIERA, J. CECH, R. HORAUD. *Action Recognition Robust to Background Clutter by Using Stereo Vision*, in "4th International Workshop on Video Event Categorization, Tagging and Retrieval", Firenze, Italy, October 2012, http://hal.inria.fr/hal-00768670.

[32] J. SANCHEZ-RIERA, J. CECH, R. HORAUD. *Robust Spatiotemporal Stereo for Dynamic Scenes*, in "21st International Conference on Pattern Recognition", Tsukuba Science City, Japan, December 2012, http://hal.inria.fr/hal-00768766.

Scientific Books (or Scientific Book chapters)

[33] M. HANSARD, S. LEE, O. CHOI, R. HORAUD. *Time of Flight Cameras: Principles, Methods, and Applications*, Springer Briefs in Computer Science, Springer, October 2012, 95, http://hal.inria.fr/hal-00725654.

[34] A. SHARMA, R. HORAUD, D. MATEUS. *3D Shape Registration Using Spectral Graph Embedding and Probabilistic Matching*, in "Image Processing and Analysing With Graphs: Theory and Practice", O. LEZORAY, L. GRADY (editors), CRC Press, 2012, p. 441-474, http://hal.inria.fr/inria-00590273.

Research Reports

[35] X. ALAMEDA-PINEDA, R. HORAUD. *Geometrically-constrained time delay estimation-based sound source localisation (gTDESSL)*, Inria, June 2012, n$^{\text{o}}$ RR-7988, 28, http://hal.inria.fr/hal-00704986.

[36] V. KHALIDOV, F. FORBES, R. HORAUD. *Calibration of A Binocular-Binaural Sensor Using a Moving Audio-Visual Target*, Inria, January 2012, n$^{\text{o}}$ RR-7865, 27, http://hal.inria.fr/hal-00662306.

[37] M. SAPIENZA, M. HANSARD, R. HORAUD. *Real-time Visuomotor Update of an Active Binocular Head*, Inria, August 2012, n$^{\text{o}}$ RR-7682, http://hal.inria.fr/inria-00608769.