IN PARTNERSHIP WITH:
CNRS

Institut polytechnique de Grenoble

Université Joseph Fourier (Grenoble)

Activity Report 2013

Team PERCEPTION

Interpretation and Modeling of Images and Sounds

IN COLLABORATION WITH: Laboratoire Jean Kuntzmann (LJK)

RESEARCH CENTER
Grenoble - Rhône-Alpes

THEME
Vision, perception and multimedia interpretation

Table of contents

Team PERCEPTION

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

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

1. Members

Research Scientist
Radu Horaud [Team leader, Inria, Senior Researcher, HdR]

External Collaborator
Laurent Girin [Grenoble INP, Professor, HdR]

Engineers
Pierre Arquier [Inria, industrial contract]
Quentin Pelorson [Inria, industrial contract]

PhD Students
Xavier Alameda-Pineda [Grenoble University]
Antoine Deleforge [Grenoble University]
Israel Dejene-Gebru [Inria, since November, Cordi-S]
Maxime Janvier [Inria, DGA-Inria]
Jordi Sanchez-Riera [Inria, until June, EU project]
Kaustubh Kulkarni [Inria, EU project]

Post-Doctoral Fellow
Georgios Evangelidis [Inria, industrial contract]

Administrative Assistant
Nathalie Gillot [Inria]

Others
Vincent Drouard [Inria]
Dionyssos Kounades Bastian [Inria]
Gurkirt Singh [Inria]

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

2.2.1. European project HUMAVIPS.

The European project HUMAVIPS – Humanoids with Auditory and Visual Abilities in Populated Spaces – is a 36-month FP7 STREP project coordinated by Radu Horaud and which started in 2010. The project addressed multimodal perception and cognitive issues associated with the computational development of a social robot. The objective was 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 emphasized the role played by multimodal perception within principled models of human-robot interaction and of humanoid behavior. The HUMAVIPS project was successfully terminated in January 2013.

An article about *Integrating Smart Robots into Society* refers to HUMAVIPS. The article stresses the role of cognition in human-robot interaction and refers to HUMAVIPS as one of the FP7 projects that has paved the way towards the concept of audio-visual robotics. The article was published in HORIZON, which is Europe's Research & Innovation Magazine.

2.2.2. ERC Advanced Grant VHIA.

The PERCEPTION team is pleased to announce that Radu Horaud was awarded an ERC Advanced Grant for his project "Vision and Hearing in Action" (VHIA). This five year project (2014-2019) will develop the concept of social robots.

2.2.3. Best Paper Award at IEEE MMSP'13.

The article received the "Best Paper Award" at the IEEE International Workshop on Multimedia Signal Processing (MMSP'13), Pula, Italy, September-October 2013. The paper addresses the problem of aligning visual and auditory data using a sensor that is composed of a camera-pair and a microphone-pair. The original contribution of the paper is a method for audio-visual data aligning through estimation of the 3D positions of the microphones in the visual centred coordinate frame defined by the stereo camera-pair. Please consult http://www.mmsp2013.org/mmsp2013_awards.php and [24].

BEST PAPER AWARD :

[24] **Alignment of Binocular-Binaural Data Using a Moving Audio-Visual Target in MMSP 2013 - IEEE International Workshop on Multimedia Signal Processing**. V. KHALIDOV, F. FORBES, R. HORAUD.

3. Research Program

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 and Platforms

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 were 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.

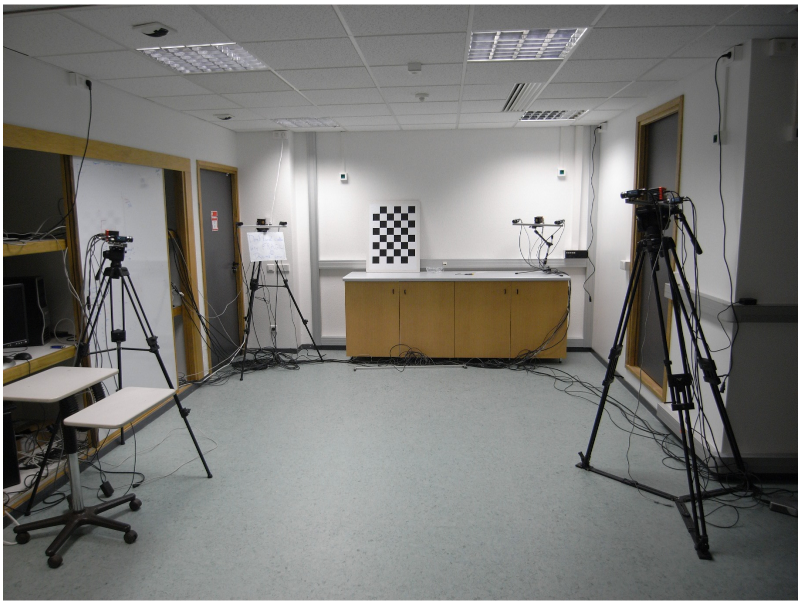

Figure 1. The mixed multi-camera system composed of four TOF-stereo sensor units.

5.2. Audiovisual robot heads

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 with 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 (https://team.inria.fr/perception/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 was 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.

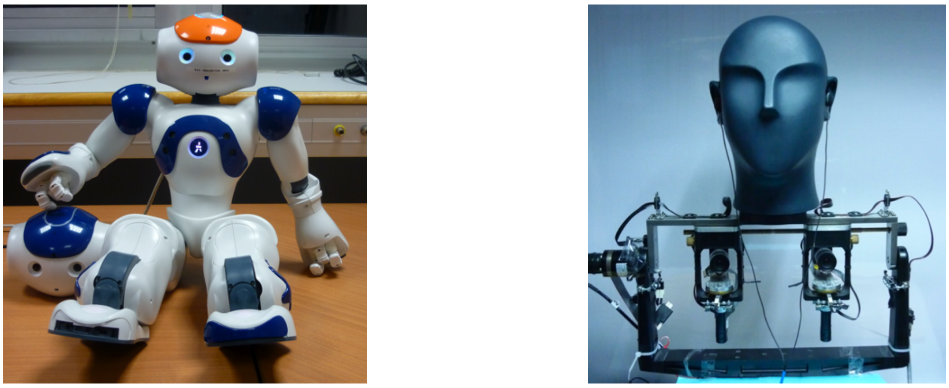

Figure 2. Left: The consumer humanoid robot NAO is equipped with a binocular-binaural head specially designed for human-humanoid interaction; Right: The binocular-binaural robot head POPEYE equipped with a four degrees of freedom stereo camera pair and with a dummy head.

For more information on POPEYE and on NAO please visit https://team.inria.fr/perception/popeye/ and https://team.inria.fr/perception/nao/.

6. New Results

6.1. 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. Two journal papers were submitted in 2013 and currently they are under review.

6.2. Continuous action recognition

Continuous action recognition is more challenging than isolated recognition because classification and segmentation must be simultaneously carried out. We build on the well known dynamic time warping (DTW) framework and devise a novel video alignment technique, dynamic *frame* warping (DFW), which performs isolated recognition based on a per-frame representation of videos and on aligning a test sequence with a model sequence. Next we devise two extensions which are able to perform action recognition and video segmentation in a concomitant manner, namely one-pass DFW and two-pass DFW. Both these algorithms have their roots in the continuous speech recognition domain but, to the best of our knowledge, their extension to visual recognition of actions and activities has been overlooked. We test and illustrate the proposed methods with several public-domain datasets and we compare both the isolated and continuous recognition algorithms with several recently published methods. One journal paper was submitted in 2013 and currently is under review.

6.3. High-dimensional regression

We addressed the problem of approximating high-dimensional data with a low-dimensional representation. We make the following contributions. We propose an inverse regression method which exchanges the roles of input and response, such that the low-dimensional variable becomes the regressor, and which is tractable. We introduce a mixture of locally-linear probabilistic mapping model that starts with estimating the parameters of inverse regression, and follows with inferring closed-form solutions for the forward parameters of the high-dimensional regression problem of interest. Moreover, we introduce a partially-latent paradigm, such that the vector-valued response variable is composed of both observed and latent entries, thus being able to deal with data contaminated by experimental artifacts that cannot be explained with noise models. The proposed probabilistic formulation could be viewed as a latent-variable augmentation of regression. We devise expectation-maximization (EM) procedures based on a data augmentation strategy which facilitates the maximum-likelihood search over the model parameters. We propose two augmentation schemes and we describe in detail the associated EM inference procedures that may well be viewed as generalizations of a number of EM regression, dimension reduction, and factor analysis algorithms. The proposed framework is validated with both synthetic and real data. We provide experimental evidence that our method outperforms several existing regression techniques. See [26], [12].

6.4. 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 analyze 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 [21], [12]. A journal paper was submitted in 2013 and accepted with minor revisions.

6.5. The geometry of non-coplanar microphone arrays

We addressed the problem of sound-source localization from time-delay estimates using arbitrarily-shaped non-coplanar microphone arrays. A novel geometric formulation is proposed, together with a thorough algebraic analysis and a global optimization solver. The proposed model is thoroughly described and evaluated. The geometric analysis, stemming from the direct acoustic propagation model, leads to necessary and sufficient conditions for a set of time delays to correspond to a unique position in the source space. Such sets of time delays are referred to as *feasible sets*. We formally prove that every feasible set corresponds to exactly one position in the source space, whose value can be recovered using a closed-form localization mapping. Therefore we seek for the optimal feasible set of time delays given, as input, the received microphone signals. This time delay estimation problem is naturally cast into a programming task, constrained by the feasibility conditions derived from the geometric analysis. A global branch-and-bound optimization technique is proposed to solve the problem at hand, hence estimating the best set of feasible time delays and, subsequently, localizing the sound source. Extensive experiments with both simulated and real data are reported; we compare our methodology to four state-of-the-art techniques. This comparison clearly shows that the proposed method combined with the branch-and-bound algorithm outperforms existing methods. These in-depth geometric understanding, practical algorithms, and encouraging results, open several opportunities for future work. See [18], [25], [11].

6.6. Audiovisual calibration and alignment

We addressed the problem of aligning visual (V) and auditory (A) data using a sensor that is composed of a camera-pair and a microphone-pair. The original contribution of the paper is a method for AV data aligning through estimation of the 3D positions of the microphones in the visual-centred coordinate frame defined by the stereo camera-pair. We exploit the fact that these two distinct data sets are conditioned by a common set of parameters, namely the (unknown) 3D trajectory of an AV object, and derive an EM-like algorithm that alternates between the estimation of the microphone-pair position and the estimation of the AV object trajectory. The proposed algorithm has a number of built-in features: it can deal with A and V observations that are misaligned in time, it estimates the reliability of the data, it is robust to outliers in both modalities, and it has proven theoretical convergence. We report experiments with both simulated and real data. See [24] (this work received the best paper award).

6.7. Audiovisual fusion for human-robot interaction

Natural human-robot interaction in complex and unpredictable environments is one of the main research lines in robotics. In typical real-world scenarios, humans are at some distance from the robot and the acquired signals are strongly impaired by noise, reverberations and other interfering sources. In this context, the detection and localisation of speakers plays a key role since it is the pillar on which several tasks (e.g.: speech recognition and speaker tracking) rely. We address the problem of how to detect and localize people that are both seen and heard by a humanoid robot. We introduce a hybrid deterministic/probabilistic model. Indeed, the deterministic component allows us to map the visual information into the auditory space. By means of the probabilistic component, the visual features guide the grouping of the auditory features in order to form AV objects. The proposed model and the associated algorithm are implemented in real-time (17 FPS) using a stereoscopic camera pair and two microphones embedded into the head of the humanoid robot NAO. We performed experiments on (i) synthetic data, (ii) a publicly available data set and (iii) data acquired using the robot. The results we obtained validate the approach and encourage us to further investigate how vision can help robot hearing. See [19], [20], [27], [11], [13]

7. Partnerships and Cooperations

7.1. European Initiatives

7.1.1. FP7 Projects

7.1.1.1. 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
Abstracrt: Humanoids expected to collaborate with people should be able to interact with them in the most natural way. This involves significant perceptual and interactive skills, 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 analyze a populated space, to localize people, and to determine whether they are looking at the robot and are speaking to it. Humans appear to solve these tasks routinely by integrating the often complementary information provided by multi-sensory data processing, from 3D object positioning and sound-source localization to gesture recognition. Understanding the world from unrestricted sensorial data, recognizing people's intentions and behaving like them are extremely challenging problems. The objective of HUMAVIPS has been 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. Developed research and technological developments have emphasized the role played by multimodal perception within principled models of human-robot interaction and of humanoid behavior. An adequate architecture has implemented auditory and visual skills onto a fully programmable humanoid robot (the consumer robot NAO). A free and open-source software platform has been developed to foster dissemination and to ensure exploitation of the outcomes of HUMAVIPS beyond its lifetime.

7.2. International Initiatives

7.2.1. Inria International Partners

7.2.1.1. Declared Inria International Partners

- Bielefeld University (Germany),
- The Czech Technical University of Prague (Czech Republic),
- IDIAP Institute (Switzerland),
- Aldebaran Robotics (France).
- University of Patras (Greece).

7.2.1.2. Informal International Partners

- The Technion (Israel Institute of Technology),
- Bar Ilan University.

7.3. International Research Visitors

7.3.1. Visits of International Scientists

- Professor Sharon Gannot (Bar Ilan University),
- Professor Yoav Schechner (The Technion),
- Professor Michael Bronstein (University of Lugano),
- Professor Vasek Hlavac (Czech Technical University),
- Professor Geoff McLachlan (University of Queensland, Australia),
- Professor Josep Ramon Casas, (Technical University of Catalonia).

7.3.1.1. Internships

- Dionyssos Kounades-Bastien, University of Patras (Master student),
- Israel Dejene-Gebru, University of Trento (Master student).

8. Dissemination

8.1. 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*.

8.2. Teaching - Supervision - Juries

8.2.1. Teaching

Doctorat (EEATS) : Radu Horaud, Manifold Learning for Signal and Image Analysis, 18 hours, Université de Grenoble, France.

Doctorat (MSTII) : Radu Horaud, Three-Dimensional Sensors, 12 hours, Université de Grenoble, France.

8.2.2. Supervision

PhD: Jordi Sanchez-Rieira, 3D Human-Robot Interaction with NAO, November 2009 – June 2013, Radu Horaud

PhD: Antoine Deleforge, Sound-Source Separation and Localization, October 2010 – December 2013, Radu Horaud

PhD: Xavi Alameda-Pineda, Audio-Visual Fusion for HRI, October 2010 – November 2013, Radu Horaud

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

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

PhD in progress: Israel Dejene-Gebru, Multimodal human-robot dialog based on audio and visual features, October 2013, Radu Horaud

9. 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", 2002, vol. 21, n^{o} 8, pp. 679–700, http://hal.inria.fr/hal-00520167

[2] Y. Dufournaud, C. Schmid, R. Horaud. *Image matching with scale adjustment*, in "Computer Vision and Image Understanding", February 2004, vol. 93, n^{o} 2, pp. 175–194 [*DOI :* 10.1016/j.cviu.2003.07.003], http://hal.inria.fr/inria-00548555

[3] 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, pp. 2357-2369 [*DOI :* 10.1364/JOSAA.25.002357], http://hal.inria.fr/inria-00435548

[4] 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, pp. 1446–1452 [*DOI :* 10.1109/34.895977], http://hal.inria.fr/inria-00590127

[5] R. Horaud, F. Forbes, M. Yguel, G. Dewaele, J. Zhang. *Rigid and Articulated Point Registration with Expectation Conditional Maximization*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", March 2011, vol. 33, n^{o} 3, pp. 587-602 [*DOI :* 10.1109/TPAMI.2010.94], http://hal.inria.fr/inria-00590265

[6] 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, pp. 158-163 [*DOI :* 10.1109/TPAMI.2008.108], http://hal.inria.fr/inria-00446898

[7] V. KHALIDOV, F. FORBES, R. HORAUD. *Conjugate Mixture Models for Clustering Multimodal Data*, in "Neural Computation", February 2011, vol. 23, n^{o} 2, pp. 517-557 [*DOI :* 10.1162/NECO_A_00074], http://hal.inria.fr/inria-00590267

[8] 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} 3, pp. 247-269 [*DOI :* 10.1007/s11263-007-0116-2], http://hal.inria.fr/inria-00590247

[9] A. ZAHARESCU, E. BOYER, R. HORAUD. *Topology-Adaptive Mesh Deformation for Surface Evolution, Morphing, and Multi-View Reconstruction*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", April 2011, vol. 33, n^{o} 4, pp. 823-837 [*DOI :* 10.1109/TPAMI.2010.116], http://hal.inria.fr/inria-00590271

[10] A. ZAHARESCU, E. BOYER, R. HORAUD. *Keypoints and Local Descriptors of Scalar Functions on 2D Manifolds*, in "International Journal of Computer Vision", October 2012, vol. 100, n^{o} 1, pp. 78-98 [*DOI :* 10.1007/s11263-012-0528-5], http://hal.inria.fr/hal-00699620

Publications of the year

Doctoral Dissertations and Habilitation Theses

[11] X. ALAMEDA-PINEDA. , *Analyse Égocentrique de Scènes Audio-Visuelles. Une approche par Apprentissage Automatique et Traitement du Signal*, Université Joseph-Fourier - Grenoble I, October 2013, http://hal.inria.fr/tel-00880117

[12] A. DELEFORGE. , *Projection d'espaces acoustiques: une approche par apprentissage automatisé de la séparation et de la localisation*, Université de Grenoble, November 2013, http://hal.inria.fr/tel-00913965

[13] J. SANCHEZ-RIERA. , *Développement d'aptitudes audio-visuelles pour le robot humanoïde NAO*, Université de Grenoble, June 2013, http://hal.inria.fr/tel-00863461

Articles in International Peer-Reviewed Journals

[14] 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", March 2013, vol. 7, n^{o} 1-2, pp. 79-91 [*DOI :* 10.1007/s12193-012-0111-Y], http://hal.inria.fr/hal-00720734

[15] G. EVANGELIDIS, C. BAUCKHAGE. *Efficient Subframe Video Alignment Using Short Descriptors*, in "IEEE Transactions on Pattern Analysis and Machine Intelligence", October 2013, vol. 35, n^{o} 10, pp. 2371-2386 [*DOI :* 10.1109/TPAMI.2013.56], http://hal.inria.fr/hal-00862002

[16] M. HANSARD, R. HORAUD, M. AMAT, G. EVANGELIDIS. *Automatic Detection of Calibration Grids in Time-of-Flight Images*, in "Computer Vision and Image Understanding", January 2014, http://hal.inria.fr/hal-00936333

[17] M. SAPIENZA, M. HANSARD, R. HORAUD. *Real-time Visuomotor Update of an Active Binocular Head*, in "Autonomous Robots", January 2013, vol. 34, n^{o} 1, pp. 33-45 [*DOI :* 10.1007/s10514-012-9311-2], http://hal.inria.fr/hal-00768615

International Conferences with Proceedings

[18] X. ALAMEDA-PINEDA, R. HORAUD, B. MOURRAIN. *The Geometry of Sound-Source Localization using Non-Coplanar microphone Arrays*, in "WASPAA 2013 - IEEE Workshop on Applications of Signal Processing to Audio and Acoustics", New Paltz, United States, October 2013, http://hal.inria.fr/hal-00848876

[19] X. ALAMEDA-PINEDA, J. SANCHEZ-RIERA, R. HORAUD. *Benchmarking Methods for Audio-Visual Recognition Using Tiny Training Sets*, in "ICASSP 2013 - IEEE International Conference on Acoustics, Speech, and Signal Processing", Vancouver, Canada, IEEE, September 2013, pp. 3662-3666 [*DOI :* 10.1109/ICASSP.2013.6638341], http://hal.inria.fr/hal-00861645

[20] J. CECH, R. MITTAL, A. DELEFORGE, J. SANCHEZ-RIERA, X. ALAMEDA-PINEDA, R. HORAUD. *Active-Speaker Detection and Localization with Microphones and Cameras Embedded into a Robotic Head*, in "Humanoids 2013 - IEEE-RAS International Conference on Humanoid Robots", Atlanta, United States, IEEE Robotics Society, September 2013, http://hal.inria.fr/hal-00861465

[21] A. DELEFORGE, F. FORBES, R. HORAUD. *Variational EM for Binaural Sound-Source Separation and Localization*, in "ICASSP 2013 - 38th International Conference on Acoustics, Speech, and Signal Processing", Vancouver, Canada, IEEE, 2013, pp. 76-80 [*DOI :* 10.1109/ICASSP.2013.6637612], http://hal.inria.fr/hal-00823453

[22] G. EVANGELIDIS, F. DIEGO, R. HORAUD. *From Video Matching to Video Grounding*, in "International Conference on Computer Vision, Workshop on Computer Vision in Vehicle Technology", Sidney, Australia, December 2013, http://hal.inria.fr/hal-00872517

[23] M. JANVIER, R. HORAUD, L. GIRIN, F. BERTHOMMIER, L.-J. BOË, C. KEMP, A. REY, T. LEGOU. *Supervised Classification of Baboon Vocalizations*, in "NIPS4B - Neural Information Processing Scaled for Bioacoustics", Lake Tahoe, Nevada, United States, December 2013, 10 p. , http://hal.inria.fr/hal-00910104

[24] *Best Paper*
V. KHALIDOV, F. FORBES, R. HORAUD. *Alignment of Binocular-Binaural Data Using a Moving Audio-Visual Target*, in "MMSP 2013 - IEEE International Workshop on Multimedia Signal Processing", Pula (Sardinia), Italy, IEEE, September 2013, pp. 242-247, Best Paper Award [*DOI :* 10.1109/MMSP.2013.6659295], http://hal.inria.fr/hal-00861482.

Research Reports

[25] X. ALAMEDA-PINEDA, R. HORAUD. , *A Geometric Approach to Sound Source Localization from Time-Delay Estimates*, November 2013, http://hal.inria.fr/hal-00910081

[26] A. DELEFORGE, F. FORBES, R. HORAUD. , *High-Dimensional Regression with Gaussian Mixtures and Partially-Latent Response Variables*, December 2013, http://hal.inria.fr/hal-00863468

Other Publications

[27] X. ALAMEDA-PINEDA, R. HORAUD. , *Vision-Guided Robot Hearing*, 2013, 26 p. , http://hal.inria.fr/hal-00911116