Project Final Report

Grant agreement number 247525

HUMAVIPS
HUManoids with Auditory and VIsual abilities in Populated Spaces
`http://humavips.inrialpes.fr`

STREP

Project period: 1 February 2010 – 31 January 2013 (36 months)

Project coordinator:

Radu Horaud
INRIA Grenoble Rhône-Alpes
655, avenue de l'Europe
38330 Montbonnot Saint-Martin, FRANCE
Radu.Horaud@inria.fr
tel: +33476615226
fax: +33476615454

Contents

1 Executive Summary

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.

2 Project Context and Objectives

HUMAVIPS addressed a partially unexplored scientific terrain on audio-visual processing for humanoid robotics. Working towards an interesting and relevant social robot demonstrator, the consortium produced a considerable amount of contributions to the state of the art. The major objectives and contributions of HUMAVIPS relate to human-robot interaction, robot vision, robot audition, and multimodal perception (fusion of audition and vision). A particular strength of the demonstrated prototypes was that the environment did not have to be strongly modified for the robot demonstrations (the input contains relatively unconstrained physical stimuli or *dirty data*) and that a commercially available and inexpensive humanoid robot with corresponding limitations was used. These practical limitations motivated the development of a special-purpose middleware that allowed easy integration of the various software modules implementing sophisticated situation understanding and interactive capabilities. It also consumed a lot of effort and were a reason why the demonstrations were not as convincing as they could have been given the impressive research results. This refers to, e.g., a more elaborate walking behavior or navigation without artificial markers and under more natural light conditions. The problems due to the bad sound quality (because of an unfortunate placement of the processor fans close to the microphones) would not be an issue with a auditory-dedicated platform; Nevertheless, it allowed the development of robust sound-source separation and localization algorithms that are ready to be used with an updated version of the humanoid robot. Apart from the impact through the HUMAVIPS demonstrator, which integrated sound, speech, vision and action, the main impact of HUMAVIPS is expected through the impressive amount of

publications produced during the project. The project will also impact through its contribution to the development of the new NAO head.

The HUMAVIPS project achieved the following objectives:

- **A humanoid that explores an unstructured environment.** HUMAVIPS investigated methods to dynamically build a 3D description of its surrounding objects through unsupervised extraction and fusion of relevant sensor information gathered with a few microphones and cameras, i.e., spatial hearing and stereoscopic vision. The emphasis was on the detection and localization of humans and on the characterization of their motion patterns and status (silent, emitting sounds, speaking, etc.). In particular, HUMAVIPS developed methodologies allowing a humanoid to robustly deal with very general situations such as a varying number of people that wander around, gesticulate, emit speech and non-speech sounds, all in the presence of reverberations or other auditory sources, other objects, etc.

- **A humanoid that recognizes, understands and interacts with people.** HUMAVIPS explored the roles of *multimodality* and *active sensing* to design a robust speech-, prosody- and gesture-based **humanoid-human interface**. Emphasis was put on informal settings where several people are present. The humanoid is able to: (i) select a person who is available (easy to reach, not committed in a private interaction with another person, etc.), (ii) optimally place itself in front of the selected person in order to robustly perform the humanoid-to-human AV interactions, and (iii) communicate with that person using verbal modalities (e.g., speech recognition and simple dialog) and non-verbal modalities (e.g., sounds and simple gestures).

- **A humanoid with a memory-centered cognitive architecture.** HUMAVIPS developed an architecture needed for the challenge of a humanoid being engaged in interaction with several people in parallel. Indeed, a continuous balancing between active multisensor exploration, on one side, and adequate synthesis of behavior, on the other side, demands for a systemic architectural approach. HUMAVIPS hence investigated the potential of memory-centered architecture comprising short- and long-term memories with a special focus on fusing *social* and *perceptual* abilities in cognitive models. This resulted in a novel humanoid-focused robot architecture informed by cognitive foundations such as associative retrieval of knowledge, active perception loops, as well as arbitration on the basis of action primitives such as particular body movements and gestures.

- **Demonstration of achievements using a realistic and challenging scenario.** To demonstrate the achievements of HUMAVIPS we have decided to use a specific scenario, termed the "Vernissage", e.g., figure 1. In this scenario, the humanoid (NAO) acts as a guide robot in a small art gallery taking care of a corner of the room where several paintings are exhibited. The robot is waiting in front of the paintings, being aware of the visitors in the room. Based on the appropriateness given the current situation either the robots takes the initiative to give explanations for some paintings or NAO waits until being addressed by visitors. This decision as well as keeping up the ongoing conversation

Figure 1: The "Vernissage" scenario has been chosen to demonstrate the project's achievements because it is representative of a wide spectrum of situations were a small "toy" robot interacts with a group of people.

require good knowledge about the people and their state inside the scene. Besides being able to focus on single persons, we have specifically chosen this scenario as it allows to demonstrate the robot's ability to understand group constellations and react appropriately when multiple people are interacting with the robot.

- **Design, implementation and demonstration of an audio-visual head compatible with the project roadmap.** In the context of the HUMAVIPS, Aldebaran Robotics designed and provided to all the partners a new NAO head featuring a stereoscopic camera pair. While stereoscopic vision is a well investigated topic, it appeared that, from an industrialization point of view, it has not been easy to find a solution to this problem. Nevertheless, a new robot head was delivered to all the partners at month 18 of the project. The second and third (final) project demonstrators successfully used this new head through the RSB (robotics service bus) that integrates the Naoqi middleware and that allows remote and real-time processing of both the visual and auditory data provided by the head's two cameras and four microphones.

3 Scientific and Technological Results

3.1 The audio-visual fusion model

The problem of data fusion and multi sensory integration has been recognized for a long time as being a key ingredient of an intelligent system. Among all possible applications using audio-

visual data, we are interested in detecting multiple speakers in informal scenarios. The robot's primary task (prior to speech recognition, language understanding, and dialog) consists in retrieving the auditory status of several speakers along time. This allows the robot to concentrate its attention onto one of the speakers, *i.e.,* turn its head towards in the speaker's direction to optimize the emitter-to-receiver pathway, and attempt to extract the relevant auditory and visual data from the incoming signals. We note that this problem cannot be solved within the traditional human-computer interaction paradigm which is based on *tethered* interaction (the user must wear a close-range microphone) and which primarily works in the single-person-to-robot communication case. This considerably limits the range of potential interactions between robots and people engaged in a cooperative task or simply in a multi-party dialog. In this paper we investigate *untethered* interaction thus allowing a robot with its *onboard sensors* to perceive the status of several people at once and to communicate with them in the most natural way.

The original contribution of HUMAVIPS is a complete real-time audio-visual speaker detection and localization system that is based binocular and binaural robot perception as well as on a generative probabilistic model able to fuse data gathered with camera and microphone pairs. Because of on-line and on-board processing constraints associated with humanoid platforms, the computational load and complexity are constraints that need to be taken into account. Furthermore, the robot does not have a distributed sensor network, but merely a few sensors, which are all located in its head – *an agent-centered sensor architecture*. Hence, one should achieve a trade-off between performance and complexity.

We present both a novel method and an original system approach to tackle the problem of on-line audio-visual detection of multiple speakers using the companion humanoid robot NAO[1]. The proposed method uses data coming from a stereo pair of cameras and two microphones. Implemented on a hardware- and sensor-independent middleware, the software runs on-line with good performance. The 3D positions of the speakers' heads are obtained from the stereo image pair, and interaural time difference (ITD) values are extracted from the binaural signals. These features are then fused in a probabilistic manner in order to compute, over time, the probability of each person's speaking activity.

The approach exhibits a number of novelties with respect to previous work addressing audio-visual fusion for speaker detection: (i) visual features are obtained from a stereoscopic setup and thus represented in 3D, (ii) auditory features are obtained from only two microphones, while most of previous work uses an array of microphones, (iii) the software is reusable with other robot sensor architectures, due to the flexibility of the underlying middleware layer, and (iv) good on-line performance in a complex environment, *e.g.,* echoic rooms, simultaneous auditory sources, background noise, uncontrolled lighting, cluttered scenes, etc.

The overall goal is to retrieve the audio-visual (AV) state of the speakers in front of the robot. That is, the number of speakers as well as their positions and their speaking state. In order to reach this goal, we adopted the framework proposed in [**?**]. Based on a multimodal Gaussian mixture model (mGMM), this method is able to detect and localize audio-visual events from auditory and visual observations. We chose this framework because it is able to account for several issues: (i) the observation-to-speaker assignment problem, (ii) observation

[1] `http://www.aldebaran-robotics.com`

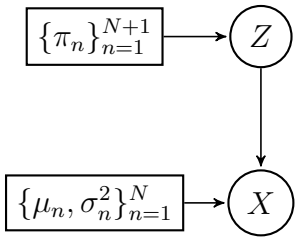


Figure 2: Graphical model generating the audio-visual observations. The hidden variable Z follows a multinomial distribution with parameters $\pi_1, \ldots, \pi_{N+1}$. The audio-visual observations X follow the law described by the probability density function in Equation 2.

noise and outliers, (iii) the possibility to weight the relevance of the two modalities, (iv) a generative formulation linking the audio and visual observation spaces, and (v) the possibility to deal with a varying number of speakers through a principled model selection method.

In a first stage, the low-level auditory and visual features are extracted. While the former correspond to the interaural time differences (ITDs), the latter correspond to interest points in image regions related to motion which are further reconstructed in the 3D space using a stereo algorithm. These 3D points will be referred to as the visual features.

The following direct sound propagation model:

$$\mathrm{ITD}(\boldsymbol{S}) = \frac{\|\boldsymbol{S} - \boldsymbol{M}_1\| - \|\boldsymbol{S} - \boldsymbol{M}_2\|}{\nu}, \tag{1}$$

is assumed. In this equation $\boldsymbol{S}$ corresponds to the sound source positions in the 3D space, *e.g.,* a speaker, $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are the 3D coordinates of the microphones in some robot-centered frame, and ν denotes the sound speed. Equation (1) maps 3D points onto the 1D space of ITD observations. The key aspect of our generative audio-visual model [**?**, **?**] is that (1) can be used to map 3D points (visual features) onto the ITD space associated with two microphones, on the premise that the cameras are aligned with the microphones [**?**]. Hence the fusion between binaural observations and binocular observations is achieved in 1-D.

The underlying multimodal GMM (mGMM) is a one-dimensional mixture of Gaussians. Each mixture component is associated with an *audio-visual object* centered at μ_n and with variance σ_n^2. This mixture has the following probability density function:

$$\mathrm{prob}(x\,;\,\Theta) = \sum_{n=1}^{N} \pi_n \mathcal{N}(x\,;\,\mu_n, \sigma_n^2) + \pi_{N+1}\mathcal{U}(x), \tag{2}$$

where N is the number of components, *i.e.,* audio-visual objects, π_n is the weight of the n^{th} component, $\mathcal{N}(x\,;\,\mu_n, \sigma_n^2)$ is the value of the Gaussian distribution at x, $\mathcal{U}$ is the value of the uniform distribution accounting for outliers, and $\Theta = \{\pi_n, \mu_n, \sigma_n^2\}$. In this equation, x stands for a realization of the random variable X, shown in the corresponding graphical model on Fig. 2, that could be either an auditory observation, *i.e.,* an ITD value or an observed 3D point, *i.e.,* a visual feature, mapped with (1). Notice that both Θ and the hidden variable

Z (modeling the observation-to-object assignments) need to be estimated. This is done using an Expectation-Maximization (EM) algorithm, derived from the probabilistic graphical model. Notice that with this formulation the number of AV objects N can be estimated from the observed data by maximizing a Bayesian information criterion (BIC) score [**?**]. However, this implies to run the EM algorithm several times with different values of N, which is prohibitive in the case of an on-line implementation. From a practical point of view the problem of estimating N can be overcome by replacing the 3D visual points with *3D faces* as described below.

The initial implementations of the nGMM EM algorithm was using 3D points, as just described. Alternatively, one can replace 3D points with 3D faces, more precisely with 3D face centers which are fair approximations of 3D mouth positions, *i.e.,* the 3D acoustic emitters. In practice we start by detecting faces in images. Face centers are then detected in the left image of the stereo camera pair. Each left-image face center is then correlated with the right image along an epipolar line in order to obtain a stereo correspondence between the face center and a associated point along the epipolar line in the right image. This allows to reconstruct a 3D point, $\boldsymbol{S}_n$, that can be viewed as 3D face center. See [**?**] for more details. The use of faces drastically simplifies the complexity of the approach because a single semantically-meaningful face center replaces a cloud of points associated with a, possibly moving, 3D object. Initial means can be easily obtained from (1), *i.e.,* $\mu_n = \text{ITD}(\boldsymbol{S}_n)$ while N, the number of AV objects can be easily estimated using any face detector.

As already mentioned we use ITDs, *i.e.,* the time delay between the signals received at the left and right microphones. Notice that, due the symmetric nature of the ITD function, there is a front/back ambiguity, which is however slightly attenuated by the transfer function of the robot head. There are several methods to estimate ITDs; We chose the cross-correlation method, since it optimizes a trade-off between performance and complexity. ITD values are obtained in real-time by computing the cross-correlation function between the left and right perceived signals during an integration time window of length W, expressed in number of time samples, or frames. The time delay τ corresponding to the maximum of the cross-correlation function in the current integration window is computed as follows:

$$\tau = \frac{1}{F_s} \underset{d \in [-d_M, d_M]}{\text{argmax}} \sum_{t=1}^{W} l(t) r(t+d) \tag{3}$$

where l and r are the left and right audio signals, F_s is the sampling frequency and d_M denotes the maximum possible delay between microphones, *i.e.,* $d_M = \|\boldsymbol{M}_1 - \boldsymbol{M}_2\|/\nu$. The time window W is a trade-off between reliability and significance. On one hand, a high W value implies more reliable ITD values, since the effect on the local maxima of the cross-correlation function is reduced. On the other hand, a small W value speeds up the computation. The parameter f denotes the shift of the sliding window used to compute the ITD. In order to extract one ITD value, two conditions need to be satisfied. First, there should be enough samples available within the integration window W. Second, the mean energy of the signals in the integration window should be higher than a given threshold E_A. In this way, we avoid to compute ITD values when the audio stream contains nothing but noise. Notice the method does not assume that the perceived sound signals are associated with some semantic *i.e.,* speech, pulse-resonance sounds, etc.

The audio-visual fusion model outlined above, and 1 in particular, implies that the visual and auditory observations are computed in a common reference frame. This allows visual data to be *aligned* with auditory data. In practice it means that the cameras' extrinsic calibration parameters (position and orientation) and the microphones' positions are expressed in a common reference frame. Extrinsic camera calibration is performed using the state-of-the-art algorithm of [**?**].

Audio-visual calibration can be achieved using (1). A sound-source is placed in a known position $\boldsymbol{S}$ while $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ are unknown and hence must be estimated. The method (i) uses an audio-visual target (a loud-speaker emitting white noise coupled with a small red-light bulb) to precisely position the sound source in the camera-pair reference frame, and (ii) estimates the unknown parameters $\boldsymbol{M}_1$ and $\boldsymbol{M}_2$ by considering several target positions and by solving a non-linear system of equations of the form of (1).

This calibration procedure does not take into account the fact that the microphones are plugged into the robot head, as already mentioned above. To account for head effects we introduce two corrective parameters, α and β, to form of an affine transformation.

$$\mathrm{ITD}_{\mathrm{AD}}(\boldsymbol{S}) = \alpha \frac{\|\boldsymbol{S} - \boldsymbol{M}_1\| - \|\boldsymbol{S} - \boldsymbol{M}_2\|}{\nu} + \beta, \tag{4}$$

These parameters are estimated using the same audio-visual target mentioned above. The audio-visual target is freely moved in front of the robot thus following a zigzag-like trajectory. The use of white noise greatly facilitate the task of cross-correlation, *i.e.,* there is single shapr peak, and hence, makes the ITD computation extremely reliable. The reverberant components are suppressed by the direct component of the long lasting white noise signal. However, it is possible to set up the experimental conditions such as to reduce the effects of reverberation, *e.g.,* the room size is much larger than the target-to-robot distance. If the microphone positions are estimated in advance, the estimation of α and β can be carried out via a linear least-square estimator derived from 4.

3.2 Face and gender recognition

The task is to discriminate still images depicting human faces into two classes - males and females. The gender classification task is a challenging problem for several reasons. Namely, there is a huge variability in images belonging to the male and female class which is caused by changes in illumination (direction and intensity), background, pose of the face, expression and so on. A robust gender classifier must take these variations into account. In addition, our application requires to select a reasonable trade-off between the classification accuracy and the computational complexity of the classifier. The computational complexity is an issue as the final gender classifier is required to run in real-time on a robot equipped with a low-end CPU comparable to that of current mobile devices.

We use a standard Ada-Boost based face-detector to find a rough positions of faces in the input image. Images cropped around the face detections whose bounding boxes are enlarge by some margin serve as the input to the gender classifier. We use a pyramid of Local

Binary Patterns (LBP) as a low-level feature description of the images. The LBP features provide invariance against monotonic changes in the intensity values at a low computational cost. In addition, the LBP feature description is rich enough to discriminate image manifolds corresponding to the male class the female class. We learn the manifolds from a training set containing thousands of examples of male and female images. The major problems is that face-detector does not provide a reliable pose estimate. If applied to real-life images the size and position of a bounding box estimated by the face detector fluctuates in order of tens of percents of the actual face size. The same holds for the estimate of the in-plane rotation of the face which, if implemented, considerably prolongs the detection time. This implies that the gender classifier must be sufficiently robust against changes in scale, position and rotation. We describe two different classification models to cope with the problem which offer different classification accuracy and computational complexity.

1. *Gender classifier trained from virtual examples.* The input image is described by a high dimensional feature vector fed into two-class linear SVM classifier. The classifier is rich enough to separate image manifolds containing male and female images in all admissible poses. To make the classifier robust against the three image transformations – rotation, translation and scale – we train the classifier parameters from a huge set of virtual examples which are generated by applying the transformations to a set of registered images. The main problem of this approach are very high computational and memory demands on the training algorithm which are caused by the huge set (tens of millions) of the virtual examples. We propose a new framework for training linear Support Vector Machines (SVM) classifier which makes the virtual example method feasible even on standard notebooks. The considerable advantage of this approach is a very low computational complexity of the resulting gender classifier which makes it a perfect choice for the low-end hardware of the robot.

2. *Structured output gender classifier.* In this case, the uncertainty in the position, scale and rotation is modeled explicitly by the designed classification model. To this end, an additional hidden parameter is introduced which accounts for the actual pose of the face within the input image. The resulting classifier simultaneously estimates pose of the face as well as the gender of the displayed person. This approach differs from the current state-of-the-methods when the pose estimation and the classification itself is carried out in two independent stages which inevitably increases the classification error. Training and classification times of the proposed structured output classifier are significantly higher compared to the first approach. On the other hand, the classifier provides higher classification accuracy.

Let X be a set of images containing a human face and $Y = \{1, 2\}$ be a set of image labels where $y = 1$ denotes male and $y = 2$ female face. Let (x, y) be a realization of random variables with a p.d.f. $p(x, y)$ defined over $X \times Y$. We are interested in a classifier $h\colon X \to Y$ which estimate a class label y from a given input image x. Given a set of training examples $\{(x_1, y_1), \ldots, (x_m, y_m)\} \in (X \times Y)^m$ i.i.d. from the unknown p.d.f. $p(x, y)$, the goal is to learn

a classifier whose expected classification error

$$R[h] = \sum_x \sum_y p(x,y)[\![y \neq h(x)]\!]$$

Let $\Phi\colon X \to \mathbb{R}^n$ be a feature map which assigns n-dimensional feature vector $\Phi(x)$ to the input image x. We assume that a reasonably good classifier should be in the set of linear classifiers

$$h(x) = \begin{cases} 1 & \text{if} \quad \langle \Phi(x), \boldsymbol{w} \rangle \geq 0\,, \\ 2 & \text{if} \quad \langle \Phi(x), \boldsymbol{w} \rangle < 0\,, \end{cases}$$

where $\boldsymbol{w} \in \mathbb{R}^n$ is an unknown parameter vector to be learned from the training examples. We use the Support Vector Machine (SVM) algorithm to train the parameter vector $\boldsymbol{w}$ from the examples. The SVM algorithm finds $\boldsymbol{w}$ by solving a convex optimization task

$$\hat{w} = \underset{\boldsymbol{w} \in \mathbb{R}^n}{\operatorname{argmin}} \left[\frac{\lambda}{2} \|\boldsymbol{w}\|^2 + \frac{1}{2} \sum_{i=1}^{m} \max\{0, 1 - y_i \langle \boldsymbol{w}, \Phi(x_i) \rangle \right] \tag{5}$$

where $\lambda > 0$ is a prescribed regularization constant used to control over-fitting of the classifier. In practice, the value of λ is tuned on an independent set of examples. The problem (5) is well understood and there exists a plethora of optimization algorithms for its solution. We use the Optimized Cutting Plane Algorithm (OCA) which is guaranteed to find ε-optimal solution in $\mathcal{O}(\frac{1}{\varepsilon})$ iterations. The computational complexity of each iteration of the OCA algorithm scales linearly with both the number of examples m and the dimensionality of feature vector n.

In our application we have a prior knowledge that makes it possible to generalize from the training examples to novel test examples. Specifically, we know that there are transformations (translation, rotation and scale) of the image $x \in X$ which leave its class membership y invariant. A commonly used way to incorporate prior knowledge into SVM classifiers is to augment the set of training examples with virtual examples (VE) that are created by applying a set of transformations (against which we want invariance) to the training examples [**?**].

To put it formally, our prior knowledge is described by a set $\mathcal{T}$ which contains a finite number of transformations $T\colon X \mapsto X$. We require that

$$h(\Phi(Tx_i)) = y_i\,, \qquad \forall T \in \mathcal{T}, i = 1, \ldots, m$$

where $\{(x_i, y_i)\}_{i=1}^m$ are given training examples. Training of h can be expressed as training of a standard SVM classifier from $|\mathcal{T}|m$ virtual training examples

$$\{(x,y) \mid x = Tx_i, y = y_i, T \in \mathcal{T}, i = 1, \ldots, m, \}$$

The VE method has two important advantages. First, it does not impose any constraints on the transformations $\mathcal{T}$. Second, existing SVM solvers can be used to train the invariant classifier. However, the cardinality of $\mathcal{T}$ may increase exponentially when the transformation T is composed of s simpler ones, $T = T_1 \odot \cdots \odot T_s$ and thus $\mathcal{T} = \mathcal{T}_1 \times \cdots \times \mathcal{T}_s$. Thus, VE are computationally demanding because they (a) significantly increases the number of training examples and (b) pose huge memory requirements to store all $m|\mathcal{T}|$ virtual examples.

To alleviate the computational demands of the VE method we have proposed the computational framework for linear SVMs (COFFIN). Here we give only a brief describe the framework. The core idea of the COFFIN is simple – instead of pre-computing the VE in advance, we generate them *on demand*. Since only the original examples need to be stored in memory, this approach drastically reduces memory requirements. The on demand generation of the VE can be implemented efficiently thanks to the following key observation about the SVM solvers: All existing SVM solvers do not require direct access to elements of the parameter vector $\boldsymbol{w}$ or the feature vectors $\Phi(x_i)$, but merely require the following two operations:

1. Dot product between feature vector and the vector $\boldsymbol{w}$:
   $r \leftarrow \langle \Phi(x), \boldsymbol{w} \rangle$ **DOT**
2. Multiplication with a scalar $\lambda \in \mathbb{R}$ and addition to the vector $\boldsymbol{v} \in \mathbb{R}^n$:
   $\boldsymbol{v} \leftarrow \alpha\Phi(x) + \boldsymbol{v}$. **ADD**

By well organizing the computation of the operations **DOT** and **ADD** for the on demand generated examples one can significantly save memory and even the computational demands.

We consider exactly the same formulation of the classification task as defined in the previous section. That is, the goal is to find a classifier $h\colon X \to Y$ which minimizes the expected classification error (**??**) where X is a set of images and $Y = \{1, 2\}$ is a set of image labels. In this section, however, we consider very different classification model. Namely, the approach described in the previous section achieves robustness of the classifier by incorporating the variation in the face pose into the training set. By contrast, the approach of this section accounts for the pose explicitly in the construction of classification model.

We assume that pose of the face contained within the input image x is know only approximately. To model this uncertainty, we introduce a set $Z = \{(u, v, \varphi, s) \mid u \in U, v \in V, \varphi \in \Phi, s \in S\}$ containing all possible locations of a rectangular region within the input image x. We assume that the face region has a fixed aspect ratio and its location is determined by an upper left corner coordinates (u, v), a rotation φ and a scale s.

We assume that a reasonably good classifier should have the following form

$$h(x) = \operatorname*{argmax}_{y \in Y} \max_{z \in Z} f(x, y, z; \boldsymbol{w}) \quad \text{and} \quad f(x, y, z) = \langle \boldsymbol{w}, \Psi(x, y, z) \rangle$$

where $f(x, y, z; \boldsymbol{w})$ is a score function measuring a match between the triplet (x, y, z). The score function is defined as a dot product between a parameter vector $\boldsymbol{w} \in \mathbb{R}^n$ and a mapping $\Psi\colon X \times Y \times Z \to \mathbb{R}^n$. We construct the mapping Ψ as follows. Let $\Phi\colon X \times Z \to \mathbb{R}^{\frac{n}{2}}$ be a feature description of a rectangular region with location z cut off from the input image x. In particular, we normalize the cropped region to a fixed size and compute a pyramid of Local Binary Pattern (LBP) features which stacked to a column vector form the feature description. The advantage of the LPB features is their invariance to monotonic transformation of the image intensity and relatively low computational cost. However, any other feature description can be readily used. Having the feature map Φ defined, we define the mapping Ψ as follows

$$\Psi(x, y \overset{\Phi(x,z)}{\mathbf{0}} \text{and} \quad \Psi(x, y = 2, z) = \begin{bmatrix} \mathbf{0} \\ \Phi(x, z) \end{bmatrix},$$

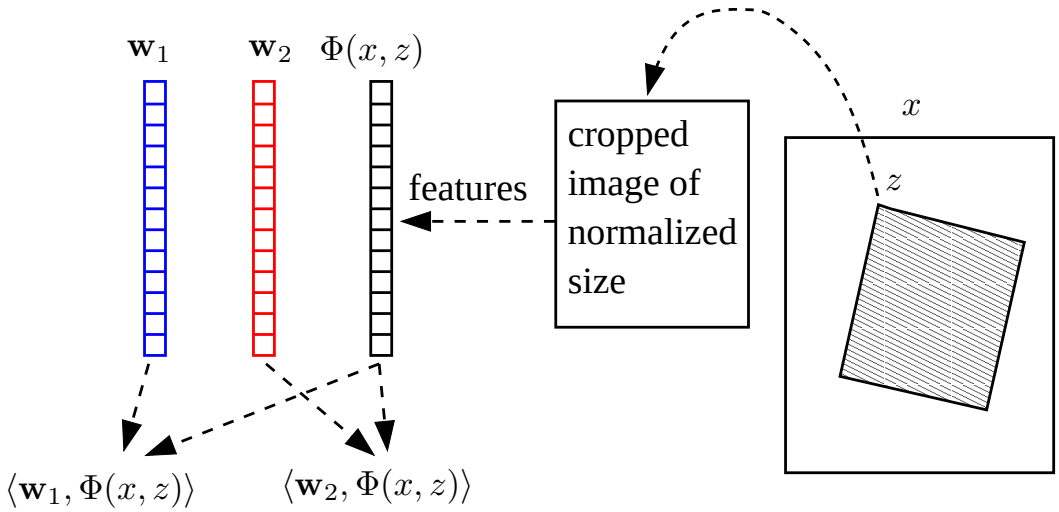


Figure 3: The chart illustrates the process of computing the score function of the gender classifier.

where $\mathbf{0} \in \mathbb{R}^{\frac{n}{2}}$ is a vector of all zeros. Let us split the parameter vector $\boldsymbol{w} \in \mathbb{R}^n$ into $\boldsymbol{w}_1 \in \mathbb{R}^{\frac{n}{2}}$ and $\boldsymbol{w}_2 \in \mathbb{R}^{\frac{n}{2}}$, i.e. $\boldsymbol{w} = [\boldsymbol{w}_1; \boldsymbol{w}_2]$. With this definition the score function equals to $\langle \boldsymbol{w}_1, \Phi(x,z) \rangle$ for the class $y = 1$ and it is $\langle \boldsymbol{w}_2, \Phi(x,z) \rangle$ for the class $y = 2$. In other words, the score function $f(x,y,z)$ evaluates the match between the prototype feature vector $\boldsymbol{w}_y$ of the class y and the feature description of the rectangular region with location z cropped out of the input image x. With these definitions, the classifier (5) boils down to

$$h(x) = \begin{cases} 1 & \text{if } \max_{z \in Z} \langle \boldsymbol{w}_1, \Phi(x,z) \rangle \geq \max_{z \in Z} \langle \boldsymbol{w}_2, \Phi(x,z) \rangle \\ 2 & \text{otherwise} \end{cases}$$

Figure 5 illustrates the procedure to compute the score function.

To obtain the parameter vector $\boldsymbol{w}$, we propose a discriminative learning algorithm which is inspired by the Structured Output Support Vector Machine (SO-SVM) classifier [cite Tsochantaridies]. Unlike the original SO-SVM, we need to cope with two hidden variable y and z. The hidden variable y is the class label whose value should be estimated by the classifier. The misclassifications of the class label are penalized by the 0/1-loss function. On the other hand, the hidden parameter z modeling the unknown pose of the face is actually not required. This implies that misclassifications of z should not be penalized. Note, that the original SO-SVM framework does not allow to cope with the setting when the hidden state splits to penalized and non-penalized part. We require that each training image x is annotated with its actual class label y and also the face pose z. That is, the training set is a collection of examples of triplets $\{(x_1, y_1, z_1), \ldots, (x_m, y_m, z_m)\} \in (X \times Y \times Z)^m$. Having such training set, we formulate learning of the parameter vector $\boldsymbol{w}$ as a convex minimization problem

$$\hat{\boldsymbol{w}} = \operatorname*{argmin}_{\boldsymbol{w} \in \mathbb{R}^n} F(\boldsymbol{w}) := \frac{\lambda}{2} \|\boldsymbol{w}\|^2 + R(\boldsymbol{w}) \tag{5}$$

Training set	Validation set	Test set	Total
12,822	5,682	5,093	23,597

Table 1: The split of the image database into training, validation and test set.

where

$$R(\boldsymbol{w}) = \frac{1}{m}\sum_{i=1}^{m}\left[\max\left\{0, 1 - \left\langle \boldsymbol{w}, \Psi(x_i, y_i, z_i)\right\rangle + \max_{z\in Z}\left\langle \boldsymbol{w}, \Psi(x_i, \hat{y}_i, z)\right\rangle\right\}\right]$$

and $\hat{y}_i$ denotes the complement of label y_i, that is, $\hat{y}_i = 2$ if $y_i = 1$ and $\hat{y}_i = 1$ otherwise. The risk function $R(\boldsymbol{w})$ evaluates performance of the classifier (5) on the training examples. It is easy to see that $R(\boldsymbol{w}) \geq \frac{1}{m}\sum_{i=1}^{m}[\![h(x_i) \neq y_i]\!]$ holds which implies that $R(\boldsymbol{w})$ is a convex (piece-wise linear) upper bound on the empirical risk defined by the 0/1-loss function. The term $\frac{\lambda}{2}\|\boldsymbol{w}\|^2$ is a quadratic regularizer introduced to control over-fitting which is achieved by constraining the parameter space (and thus class of the classifiers) whose volume is inversely proportional to the constant λ. The optimal value of λ is not known and, in practice, it is tuned on validation data.

The objective function F of the problem (5) is strictly convex and non-differentiable. Though the problem can be equivalently transformed into a standard convex quadratic program its optimization would not be feasible by off-the-shelf algorithms. Instead, we minimize directly the non-smooth objective F by the Bundle Method for Risk Minimization Algorithm (BMRM) [cite teo] which is a variant of the bundle methods – a workhorse of the non-smooth optimization. The BMRM algorithm transforms the original hard problem (5) into a series of small-size quadratic problems with simple constraints. The BMRM algorithms is guaranteed to finds ε-optimal solution in $\mathcal{O}(\frac{1}{\varepsilon})$ iterations. The BMRM algorithm only require procedures to compute the function value of the risk $R(\boldsymbol{w})$ and its sub-gradient $R'(\boldsymbol{w})$ at given points $\boldsymbol{w}$. By Danskin's theorem, the formula for the sub-gradient of R reads

$$R'(\boldsymbol{w}) = \frac{1}{m}\Big[\Psi(x_i, \hat{y}_i, \hat{z}_i)\big\rangle - \Psi(x_i, y_i, z_i)\Big]$$

where

$$\hat{z}_i \in \operatorname*{argmax}_{z\in Z}\left\langle \boldsymbol{w}, \Psi(x_i, \hat{y}_i, z)\right\rangle \tag{5}$$

The main computational burden in evaluation of the risk R and its sub-gradient R' is in solving the task (5). It is seen that solving (5) is also required in the classification stage (c.f. formula (5)). Hence, it is beneficial to spend time in efficient implementation of (5) as it speeds up both training stage and the classification.

We use a large database of images containing faces of 23,597 people with varying ethnicity, age, resolution and background clutter. We split the images into training, validation and test set using the proportions summarized in Table 1.

Each face is manually annotated with gender of the depicted person. In addition, the training and the validation images are endowed with the face pose derived from a manually annotated positions of eyes, nose and mouth.

Method	Classification error	Classification time [sec/image]
Virtual Examples SVM	13.7%	0.2×10^{-3}
Structured Output SVM	12.3%	1.2

Table 2: Average classification error and the classification time of the gender classifiers based on the structured output SVM and two-class linear SVM trained from virtual examples.

We compare the gender classifiers based on the two-class linear SVM trained from the virtual examples (VE-SVM) and the structured output SVM (SO-SVM). Both methods are trained on the same training set and the SVM regularization constants λ are tuned on the same validation set. Both methods use exactly the same code for computing the low-level features, i.e., the Local Binary Patterns. The comparison was done on a notebook with Intel(R) Core(TM)2 Duo CPU at 2.66 GHz with 4GB RAM. In the case of the VE-SVM, we applied 735 transformations (3 scales, 5 rotations, 49 translations) to each training image which generated around 9.4×10^6 virtual examples (dimension 723,712, sparsity 0.4%) which would require around 99GB memory. Thanks to the proposed COFFIN framework, training SVM classifier from such huge amount of examples is tractable on the standard notebook with only 4GB RAM.

Here we compare VE-SVM and SO-SVM classifiers in terms of the classification accuracy and the classification time which are the characteristics of interest in our application. The average classification accuracy and the time are measured on the test set. Table 2 summarizes the obtained results. The receiver operating characteristics of both classifiers is displayed in Figure 4. It is seen that the SO-SVM has by 1.4% lower classification error compared to the VE-SVM as expected. However, the gain in performance is paid by significantly longer classification time. In particular, the SO-SVM requires 1.2 seconds while the VE-SVM needs only 200 microseconds. There is still a room for improving the efficiency of the structured output classifier, however, the classification time will not be probably lower than 100 milliseconds on the same hardware.

3.3 Visual focus of attention based on head localization and tracking

The Humavips vernissage scenario represents an archetype of robot-to-human and robot-to-group situation. There, as a person or a group of persons addresses the robot, or the robot detects that somebody is interested in a particular painting, it should be able to propose the guide service and provide information on paintings. This involves both, detection of situations where visual attention is focused on the robot or on a painting, and keeping track whether people follow the explanations. Several tasks are crucial in achieving this goal: *detect* people that enter the robot field of view, *track* them in a dynamic environment where people move, rotate their head and the robot himself walks and performs gestures, and *stop tracking* people as they move out of the scene. Secondly, the *head rotation angles* of tracked persons should be estimated to approximate their visual gaze direction.

The most straightforward approach to solve this problem is to employ a face detector [13].

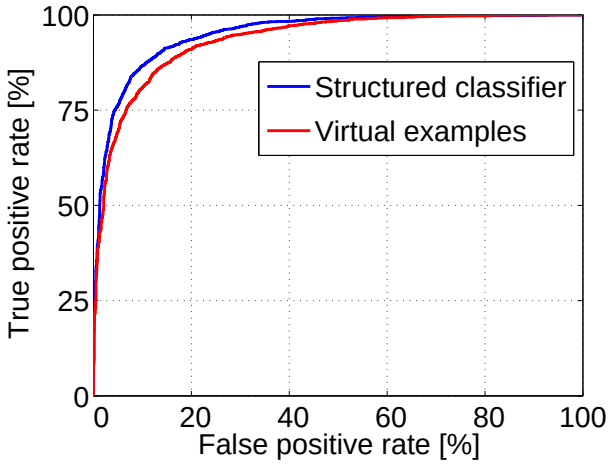


Figure 4: Receiver operating characteristics of the gender classifiers based on the structured output SVM (blue) and two-class linear SVM trained from virtual examples (red).

However, despite much progress performed on multi-view face detection, this is not sufficient, and 30 to 40% of faces are missed even in simple scenario, for instance due to variability in face appearance or lighting conditions. Above all, detection failures are the consequence of less common head poses that people naturally take e.g. when looking at other people in the same room, or looking down, which often involves large head tilts. Unfortunately, the missed detections do not happen at random time, since for the above reasons, the difficult head postures can last for long periods. In practice, this means that face detection algorithms have to be complemented by robust tracking approaches; not only to interpolate detection results or filter out spurious detection as is often assumed, but also to allow head localisation over extended periods of time.

Tracking people's head all (most of) the time is very important to avoid the robot becoming 'blind' under under non-frontal poses, making it difficult to understand for it what is going on, and for instance, check that people look at the painting it is curretnly explaining.

- we have developed a method that exploits multiple dynamics for the head: both state-based dynamics that accounts for regular motions, and image-based that accounts for abrupt motion changes and variable frame sampling. It is described in Subsection **??**.

- we have developed a principle framework to address track creation and track deletion. For instance, how do we know at each point in time that a tracker is doing fine or that there is a failure? This an important issue in practice since a false positive (detecting a track failure while there was none) may mean losing a person track for a long period until the detector finds the face again. In most algorithms, assessing tracking failure is often left to the (sudden) drop of objective or likelihood measures which are not that easy to control in practice.

- code has been optimized to achieve real-time performance, as well as to address the

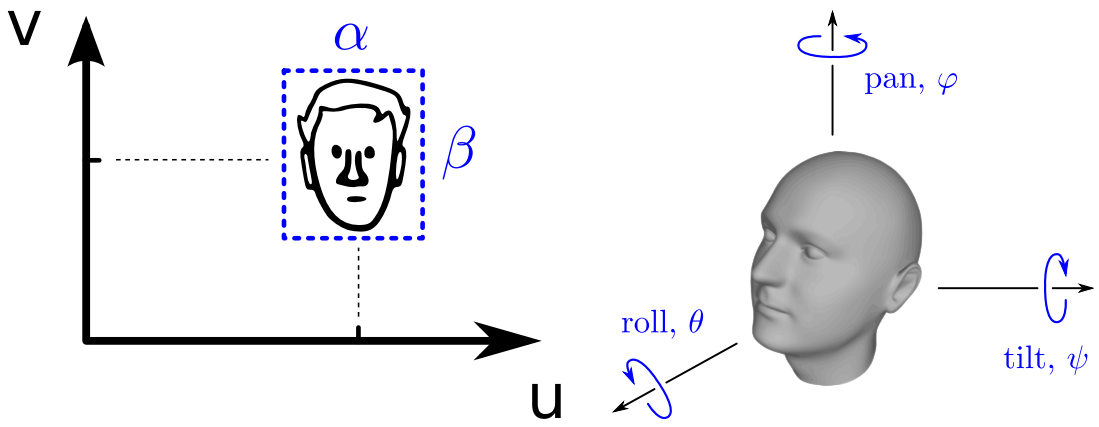


Figure 5: Head parameters $\boldsymbol{s}$ used in the tracking model: location (u, v), scale α and excentricity β of a head in a 2D image, and head pose defined by pan φ, tilt ψ and roll θ.

problem of irregular frame sampling due to the variability in the robot computational load that affects frame acquisition. The current algorithm is integrated and running on the HUMAVIPS framework.

Experiments on Nao data, including those from the recently collected data, demonstrate the algorithm performance.

We adopted an approach where head tracking was coupled with head pose estimation, as it allows to avoid state drifts that are due to head rotations, better localize head in the image and perform more precise head pose estimation, as would be shown further.

Thus we consider the task of simultaneous head tracking and head pose estimation based on a set of consecutive images. A head is described by a set of parameters $\boldsymbol{s}$: its location (u, v) on the 2D image plane, scale α and excentricity β paratemers that define its shape, and head pose parameters (φ, ψ) that stand for *pan* and *tilt* head rotation angles in 3D:

$$\boldsymbol{s} = \{u, v, \alpha, \beta, \varphi, \psi\}. \tag{5}$$

Schematic representation of these parameters is given in Figure 5. We note that the head roll angle θ is not included into the state, as soon as it is irrelevant to the gaze direction which we use to estimate the VFOA.

We assume that head state evolution $\{\boldsymbol{s}_0, t_0; \boldsymbol{s}_1, t_1; \ldots; \boldsymbol{s}_n, t_n; \ldots\}$ is described by a first order Markov process with transition probability $P_{1|1}(\boldsymbol{S}_n, t_n | \boldsymbol{s}_{n-1}, t_{n-1})$ and the initial distribution $P_1(\boldsymbol{S}_0, t_0)$. Here and in what follows we write capital letters for random variables and small letters for their realisations. Above, t_n denotes the time instant at which the n^{th} image has been captured. Note that unlike fixed-infrastructure systems that address real-time face and VFOA tracking, the sequences of these time instants is not regular in our real-time system since the computational load on the robot can highly vary in function of its activities (eg is it moving or not, is it synthetizing speech, etc), and thus impact the visual data acquisition and processing. The observations $\{\boldsymbol{Y}_0, \boldsymbol{Y}_1, \ldots, \boldsymbol{Y}_n, \ldots\}$ corresponding to the state sequences are assumes to

be independent given the object states. Alltogether, our dynamical bayesian network is thus fully defined by the transition probability $P_{1|1}$ (that we refer to as *dynamic model*) and to the observation probability $P(\boldsymbol{Y}_n|\boldsymbol{s}_n)$ that forms our *observation model*.

Given the model above, we use *particle filtering*, a well-known technique to approximate posterior density $P(\boldsymbol{S}_0, t_0; \boldsymbol{S}_1, t_1; \ldots; \boldsymbol{S}_n, t_n \mid \boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ or the filtering density $P(\boldsymbol{S}_n, t_n|\boldsymbol{y}_0, \boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ by a set of weighted particles. More precisely, we adopt the sequential importance sampling strategy and take the dynamic model $P_{1|1}$ as the importance density. Hence, the weight w_n^i of a particle i at time t_n after sampling is simply given by its likelihood, ie given by $w_n^i = P(\boldsymbol{Y}_n|\boldsymbol{s}_n^i)$. In our experiments we take 200 particles per tracker with the total number of trackers N_{tr} limited to 10, which guarantees real-time performance of our model.

As prior on the state process, the choice of dynamics is important to constrain the estimation and avoid tracking failure. In our tracking framework, it is even more important since it used as well as proposal to explore the state space during optimization. However, people's motion are difficult to predict: they may remain relatively static when interacting with Nao. When they move around, they can have a constant speed. Finally, we can also observe abrupt motion changes at motion transitions, or due to sudden and fast motion of Nao's head. Accordingly, to handle all these situations, we have defined the dynamical model as a mixture of different elements:

$$P(\boldsymbol{S}_n, t_n|\boldsymbol{s}_{n-1}, t_{n-1}) = \gamma_{RS} P_{\text{RS}}(\boldsymbol{S}_n, t_n|\boldsymbol{s}_{n-1}, t_{n-1}) + \gamma_{SB} P_{\text{SB}}(\boldsymbol{S}_n, t_n|\boldsymbol{s}_{n-1}, t_{n-1}) + \gamma_I P_{\text{I}}(\boldsymbol{S}_n, t_n|\boldsymbol{s}_{n-1}, t_{n-1}) \quad (5)$$

which are defined below. The first one defines a *random search* which accounts for a no-motion situation:

$$P_{\text{RS}}(\boldsymbol{S}_n, t_n|\boldsymbol{s}_{n-1}, t_{n-1}) = \mathcal{N}(\boldsymbol{S}_n;\ \boldsymbol{s}_{n-1}, \Delta t_n \boldsymbol{\Sigma}(\boldsymbol{s}_{n-1})),$$

and the second one a *random search with state-based velocity estimates*:

$$P_{\text{SB}}(\boldsymbol{S}_n, t_n|\boldsymbol{s}_{n-1}, t_{n-1}) = \mathcal{N}(\boldsymbol{S}_n;\ \boldsymbol{s}_{n-1} + \Delta t_n \boldsymbol{\mu}_n, \Delta t_n \boldsymbol{\Sigma}(\boldsymbol{s}_{n-1})),$$

that accounts for constant speed motion, where $\Delta t_n = t_n - t_{n-1}$ is a time interval between the two states, $\mathcal{N}$ is a probability density of a multivariate Gaussian distribution and $\boldsymbol{\mu}_n$ is an estimated velocity vector $\boldsymbol{\mu}_n = (\hat{\boldsymbol{s}}_{n-1} - \hat{\boldsymbol{s}}_{n-2})/(t_{n-1} - t_{n-2})$, for the state estimates $\hat{\boldsymbol{s}}_{n-1}$ and $\hat{\boldsymbol{s}}_{n-2}$ obtained at t_{n-1} and t_{n-2} respectively. In both the above cases, $\boldsymbol{\Sigma}(\boldsymbol{s})$ is a covariance matrix that depends on head scale α, and thus implicitely on the distance to a person and its potential visual speed[2].

Image-based dynamics. These proposal distributions are entirely based on random search and state statistics. However, as mentionned above, there are also abrupt speed changes that are difficult to predict based only on past information, and which are the situation that often lead to failure. Indeed, in these cases, it is more appropriate to directly exploit the information contained in the images, and which are of two different natures: instantaneous observations

[2]The assumption is that person closer to a robot will exhibit faster visual motion than people in the background.

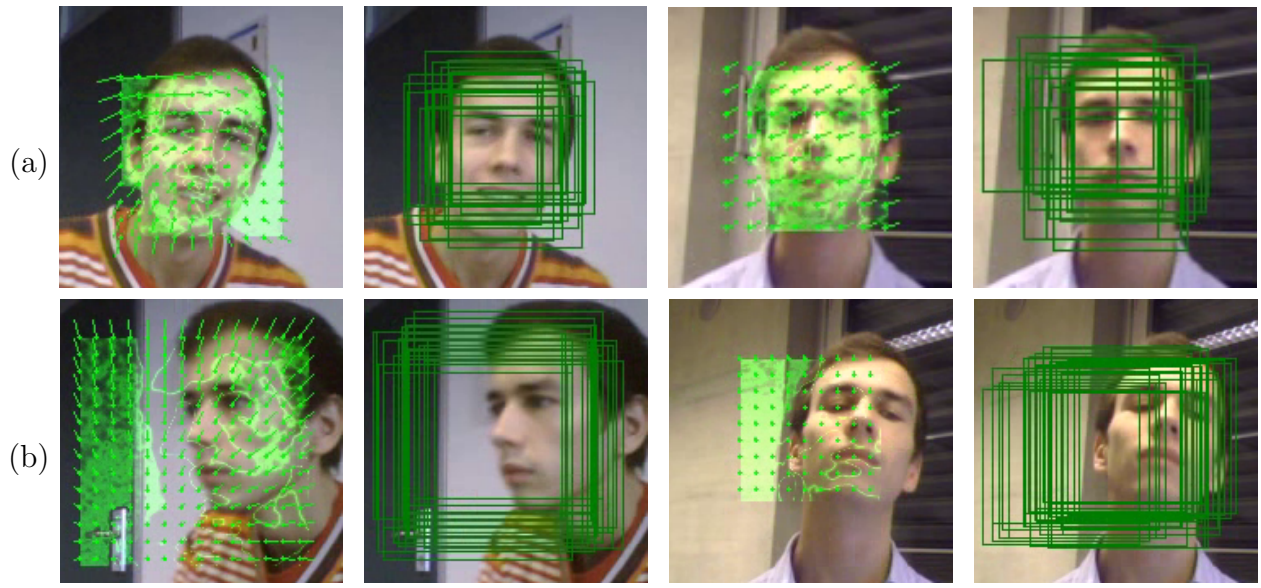


Figure 6: Estimated displacement field and support weights obtained from the parametric motion model [**?**] with the associated sampling results for the cases of (a) proper support choice; and (b) improper support choice. Green arrows depict motion vectors evenly distributed over the support, green colour intensity encodes motion support weights: the more intensive the colour is, the more compliant is the pixel with the estimated model.

reflecting the presence of the object, as produced by a *face detector*; and sequential observations reflecting observed *image-based motion* between frames. Thus, the image based dynamics is defined as $P_{\mathrm{I}} = \gamma_{ID} P_{\mathrm{ID}} + \gamma_{IM} P_{\mathrm{IM}}$ with

$$P_{\mathrm{ID}}(\boldsymbol{S}_n, t_n) = \mathcal{N}(\boldsymbol{S}_n;\ \boldsymbol{s}_{\mathrm{ID}}, \boldsymbol{\Gamma}), \tag{5}$$

where $\boldsymbol{s}_{\mathrm{ID}}$ denotes the closest face detection associated with the track (when it exists) and

$$P_{\mathrm{IM}}(\boldsymbol{S}_n, t_n | \boldsymbol{s}_{n-1}, t_{n-1}) = \mathcal{N}(\boldsymbol{S}_n;\ \boldsymbol{s}_{\mathrm{IM}}, \boldsymbol{\Gamma}_{\mathrm{IM}}), \tag{5}$$

where $\boldsymbol{s}_{\mathrm{IM}} = \boldsymbol{s}_{n-1} + \boldsymbol{\mu}_n^{\mathrm{IM}} \Delta t_n$ is a state predicted from the image motion. This motion is measured using a robust parametric motion model [**?**] estimator applied to an image patch around the estimated head location in the previous frame to calculate the displacement field at every pixel and derive $\boldsymbol{\mu}_n^{\mathrm{IM}}$. However, a proper choice of the estimation support is important for the reliability of the algorithm.

Examples of the estimated motion with support pixel weights, as well as the associated samples are shown in Figure 6. The images are taken from sequences that were acquired on Nao robot at approximately 7 fps. Green arrows depict motion vectors evenly distributed over the support (10 pixels apart), green colour intensity encodes motion support weights: the more intensive the colour is, the more compliant is the pixel with the estimated model. Top row shows the case when the support is properly chosen - head represents the major part of the support. Bottom row demonstrates bad performance of the predictor distribution which is caused by support misplacement - significant part of its area belongs to background.

3.4 Robot-to-group interaction and dialog

From NAO's point of view, knowing who the current addressee is i.e. 'to whom a spoken utterance is addressed at' is important in multi-party interactions. This information is useful for him to decide automatically if he 'has to' or 'should not' or 'can respond'. Though gaze information about 'who the current speaker is looking at' carries valuable information, previous research has shown that this cue is not always sufficient for addressee prediction. Therefore inbvestigation of other cues that could benefit the task is of interest to improve the performance. In this view, we have performed the following works:

- **Addressee Estimation:** Third, we defined several interesting features that capture different contextual information, such as the gaze cues not only from the speaker but also from the partner, the subjective difficulty of the quiz question, and if the speaker spoke the previous utterance as well. The VFOA used was from the ground-truth as well as automatically estimated from the vicon head pose. Our results shows that results up to 85-90 % accuracy (using the ground-truth VFOA) and 75-80% (using automatic VFOA) could be obtained. Interestingly, to the contrary of what people used dominantly, only using looking at Nao as a single cue is not the best gaze performing feature. Furthermore, results showed that context can help in obtaining better performance.
- **Learning from interactions.** Finally, we explored the possibility to adapt over time (from one session to another) an addressee recognizer that would be different for each quizz question. We proposed a method that adapted a feature online (when are people addressing the robot after the question was set) to build models that can be adapted to learn with time, instead of just using static models. Experiments did not show convincing results, demonstrating the difficulty of the task and of online learning for the addressee estimation scenario.

Dataset scenario: We use nine interactions from the Vernissage corpus, where a humanoid robot NAO explains a set of paintings and gives a quiz to two human participants. All participants are involved in only one interaction. For the analysis below, we use only the quiz part, which consists of nine questions (or **quiz episodes**) in art and culture, which are the same across the participant set. Some of the questions are about a set of paintings that NAO introduces to the participants before the quiz. In general, participants discuss among themselves before answering a question, but this is not always the case (eg when questions are 'easy').

Annotations: All utterances and corresponding addressee (either Nao or the Partner) were annotated following the method described above. The visual focus of attention of each participants was also annoated with labels: *NAO*, *Ptr*, the three paintings *Pai1*, *Pai2*, and *Pai3*, *Unfocussed*, and a catch-all-class *Don't know*.

Annotation analysis: Considering the quiz part in the nine interactions, there are 331 utterances of human participants in total, of which 172 were directed towards NAO, whereas 159 were directed towards a **human partner (denoted Ptr henceforth)**. This dataset is denoted DatasetFULL henceforth. We also experimented with a subset of utterances, that occur before the first interruption from NAO (denoted DatasetSUB). This subset contains 200 utterances of

	FULL dataset			SUB dataset		
	NAO	Ptr	Total	NAO	Ptr	Total
occurences	172	159	331	107	93	200
Turn duration (mean) in sec.	1.43	1.25	1.34	1.48	1.38	1.43
Turn duration (std) in sec.	1.3	1.1	1.2	1.3	1.2	1.3

Table 3: Statistics: FULL and SUB datasets

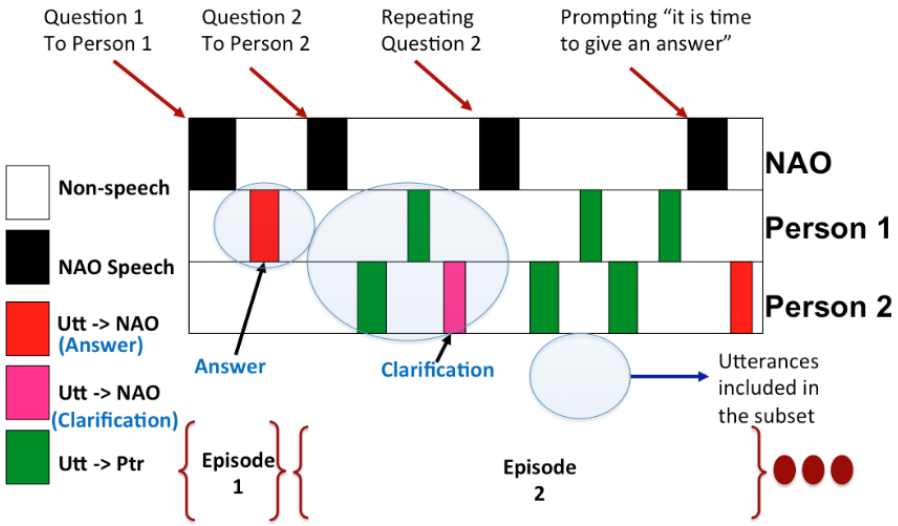


Figure 7: Illustration: Dataset-FULL and Dataset-SUB. Speech segmentation of a typical interaction. Utterances directed at NAO (Utt-> NAO) could be an answer or a clarification (like 'could you please repeat the question'). Dataset-FULL includes all utterances of Person 1 and 2. Dataset-SUB includes only those utterances before the first interruption from NAO after asking the question.

human participants in total, of which 107 were directed towards NAO, whereas 93 were directed towards a human partners. Table 3 gives some basic statistics about the utterances directed to NAO and the partner for the two datasets. We have also marked the start and end of quiz questions as episode units (see Fig. 7 for the illustration about the **episodes** and the subset of utterances). Later, we use some attributes of the episodes as contextual cues.

Episode statistics analysis. Fig. 8 gives some basic statistics about the episodes. The first three episodes are relatively trivial questions about the paintings that people usually answer quite fast. Apart from these three, we also observe from the first row of Fig. 8 that the seventh episode is also a short episode. We have labeled these as 'easy' episodes. An interesting aspect of these episodes is that due to the simplicity of the question, partners are hardly addressed in these easy episodes, as seen in the second row of Fig. 8. Such information could be useful for the addressee estimartion task by biasing the odds in favor of the Nao addressee label.

As compared to existing addressee literature in HRI in our work, we estimate addressees in a realistic scenario, where a humanoid robot with significant nonverbal displays induces unconstrained nonverbal behaviors in human partners. We predict the addressees on manually annotated utterances form VFOA (both manually labeled and automatically estimated using

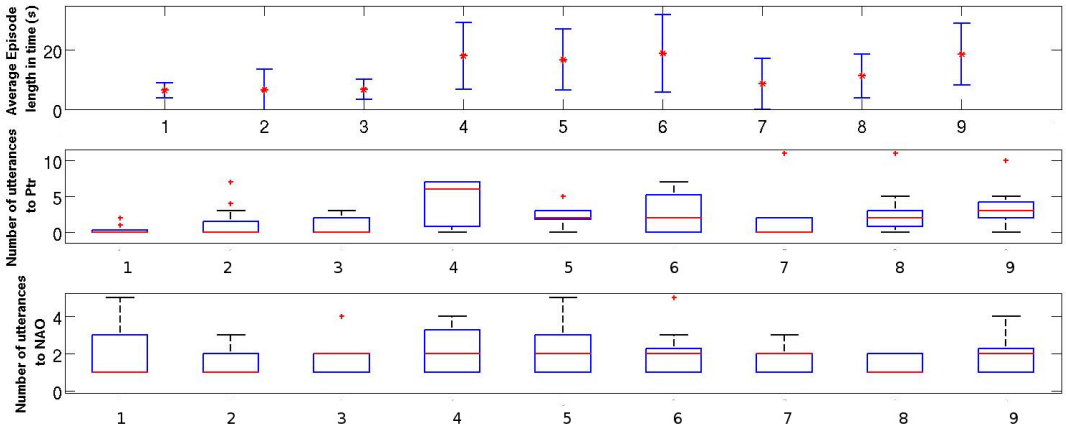


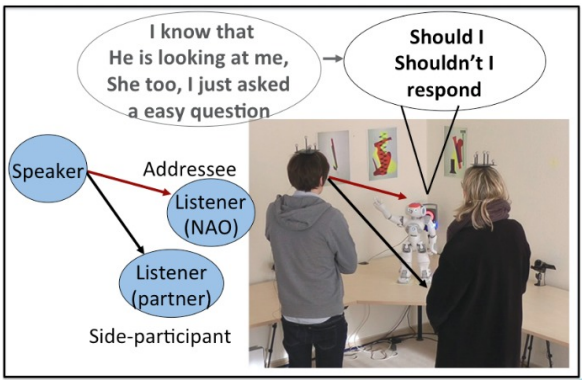
Figure 8: Statistics related to Episodes: First row shows the duration of episodes (mean in red cross and symmetric error bars that are two standard deviation units in length). Second row gives the number of utterances in each episode addressed to Ptr (Median in red central mark, 25th and 75th percentile as edges of the blue box). Third row gives the number of utterances in each episode addressed to NAO.

vicon head pose) to investigate the best case performance. As contextual cues, we investigate not only the cues from the speaker, but also gaze cues from the side-participant, and contextual prior information about the current activity (here the quiz), and the current dialog context (previous speaker). In the following, we first detail the feature we used, then present the recognizer used, and finally present our results.

For every utterance, we defined the following features, summarized in Fig. 9:

- SpkrL@NAO (the proportion [%] of time when the speaker looked at NAO during the last one second of an utterance) and SpkrL@Ptr (% of time when the speaker looked at the partner);
- PtrL@NAO (% of time when the partner looked at NAO) and PtrL@Spkr (% of time when the partner looked at the speaker);
- TimeSinceQ (the time difference between the end of a quiz question and the start of the utterance);
- PrevSpkrSame (whether the previous speaker is the current speaker coded as 2 and 1 if not);
- EpType, the episode type that roughly indicate the difficulty of the quiz question: 1 being easy and 2 being difficult.

In this work, we assigned the difficulty of the question manually, but it it could also be learned over multiple sessions i.e. with experience, as we implicitly propose towards the end of the addressee experiments. A question could be difficult because the listeners do not follow what the robot is saying or they follow the question but do not know the answer. We do not distinguish between these two cases in this work. While PtrL@NAO and PtrL@Spkr are contextual cues



	Utterance	
(f_1) PrevSpkrSame	{1,2}	discrete
(f_2) EpType	{1,2}	discrete
(f_3) TimeSinceQ	Time(St. Utt – End.Q)	continuous
(f_4) SpkrL@NAO	P(VFOASpkr = NAO)	continuous
(f_5) SpkrL@Ptr	P(VFOASpkr = Ptr)	continuous
(f_6) PtrL@NAO	P(VFOAPtr = NAO)	continuous
(f_7) PtrL@Spkr	P(VFOAPtr = Spkr)	continuous
Feature	Value	Cont / Disc

Figure 9: Left: Overview of the addressee estimation task. Right: Addressee estimation task: Tested features and their encoding.

from the side-participant, EpType is a task-related long term context, and PrevSpkrSame is a short-term context about the dialog.

We used two supervised models to predict the addressee. The first is a Gaussian Naive-Bayes classifier, which assumes that the features are independent given the class. Univariate Gaussians were used as conditional densities for the continous variables, and probability tables for discrete ones. Also, if $f_{1:N} = (f_1, f_2, ...f_N)$ denote the set of individual features f_i, the log-likelihood ratio is given, by using Bayes' theorem and cancelling the common terms as follows:

$$LR_{NB} = \log(\frac{P(Add = NAO|f_{1:N})}{P(Add = Ptr|f_{1:N})}) = \log(\prod_{k=1}^{N} \frac{P(f_k|Add = NAO)}{P(f_k|Add = Ptr)}) + \log(\frac{P(Add = NAO)}{P(Add = Ptr)}) \quad (5)$$

where the probabilities $P(f_k|A)$ or $P(f_k|B)$ are estimated by fitting a Gaussian to the data from the respective class and the ratio of the priors are inferred from the data. When this ratio LR_{NB} is greater than zero, then the estimated addressee is NAO.

The second classifier was a discriminative one, the Logistic Regression. In this case, the log-ratio of the probability of addressing the partner vs NAO is a linear function of the features:

$$LR_{LogR} = \log(\frac{P(Add = NAO|f_{1:N})}{P(Add = Ptr|f_{1:N})}) = \beta_0 + \beta_1 f_1 + \beta_2 f_2 + ...\beta_N f_N \quad (5)$$

where the β parameters are estimated during training and indicate the relative importance of the features. Similarly, when LR_{LogR} is greater than 0, then the estimated addressee is NAO.

In this report, results will be shown using the Gaussian Naive-Bayes classifier, as we also use this classifier for experiments on online classifier adaptation.

The results of the addressee estimation task are given in Fig. 10 and 11. We did a leave-an interaction-and-quiz question-out evaluation. First, we report the difference in performance using the Dataset-FULL and Dataset-SUB in Fig. 10. The following comments can be made:

- **Single Features:** The best single feature SpkrL@Ptr performed at an accuracy of 84.0% and 87.5% respectively. Predicting the majority class would be 52.0% and 53.5% respec-

tively. Interestingly, we can notice that only using how much the speaker look at Nao (SpkrL@NAO) as most people do, provides slightly worse result than the looking at the partner indicator (SpkrL@Ptr).

- **Feature fusion:** Fusing the single features helped improve the performance. The results with feature fusion are shown in the figure. The feature combination of SpkrL@NAO, SpkrL@Ptr, PtrL@Spkr, and EpType for example an accuracy of 86.4% and 90.0% on the two tasks.
- **FULL vs SUB:** We hypothesize that the improvement in performance in the Dataset-SUB is due to the fact when NAO interrupts (rather the Wizard of Oz) saying "It is time to give an answer", then the participant provide a forced response and hence affecting the VFOA to addressee mapping. Before the interruption, the response is rather natural.
- **Automatic VFOA:** Next, we report the results with automatic VFOA estimation using the head pose from the VICON motion capture system in Fig. 11. The procedure to infer the VFOA from the head pose is described in Deliverable D.4.3. Using the automatically estimated VFOA (with the head-pose from VICON) resulted in a drop of 6-7% accuracy on the addressee estimation task. We tried using both the hard-decisions of VFOA as well as posteriors to compute the features. This shows that the contextual information is complementary to the gaze cues from the speaker. With the feature combination of SpkrL@NAO, SpkrL@Ptr, PtrL@Spkr, and EpType and with automatic VFOA (and manual VFOA) there were 32 (and 20) misclassifications and the confusion was assymmetric i.e. 22 (13) times addressing the partner was confused with NAO and 10 (7) times NAO confused with addressing the partner.

Finally, we also observe that for a real-time implementation, thresholding SpkrL@Ptr itself is a good classifer. It works better than thresholding SpkrL@NAO. (see Fig. 12). When there are more than one partner the effectiveness of thresholding SpkrL@Ptr remains to be seen.

Online adaptation: Motivated by online adaptation techniques in related literatures (speaker adaptation for speech recognition online apperance adaptation for visual tracking), we explored the possibility of adapting addressee models for different questions. The idea being, instead of using the same parameters for all the questions, why not adapt the parameters for different questions. In the case of the Gaussian Naive Bayes Classifier, this can be achieved by adapting the mean of the Gaussian for instance.
To begin with, we hypothesized, that our TimeSinceQ feature could be the candidate for adaptation. If we imagine a binary hidden variable such as the difficulty of the question, depending on this variable, the relationship between TimeSinceQ and the addressee label could be different.

Towards the goal of automatically inferring a hidden variable that could help to improve the addressee estimation, we first performed experiments where we learnt question-dependent parameters, a different mean for each question. More precisely, for a given question, let μ_{NAO} and μ_{Ptr} denote the mean of the Gaussian for the addressee being NAO and Ptr respectively learnt from other questions. Then, we adapted these parameters for the question q, estimating

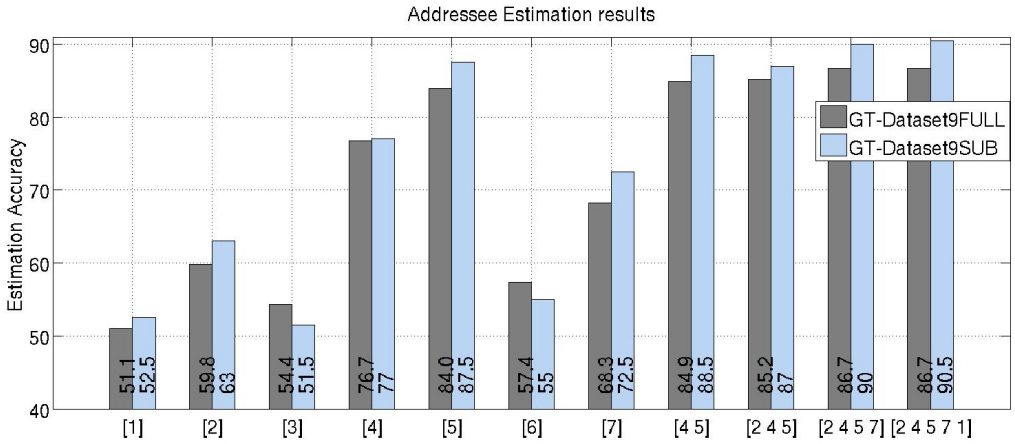


Figure 10: Result: Addressee estimation task - with two different datasets. We show the estimation accuracies for the single and multiple cues used for classification (the numbers in brackets indicate the cues used, e.g. [5] means that cue f_5 was used; see Fig. 9 for feature description).

μ^q_{NAO} and μ^q_{Ptr} as a weighted combination between the prior μ_{NAO} and μ_{Ptr} and the sample mean of the measured observations for that question, leaving out the interaction to be processed.

$$\mu^q_{NAO} = \frac{N_1}{N_1 + P}.(\frac{\sum_i f_3}{N_1}) + \frac{P}{N_1 + P}.\mu_{NAO} \tag{5}$$

$$\mu^q_{Ptr} = \frac{N_2}{N_2 + P}.(\frac{\sum_j f_3}{N_2}) + \frac{P}{N_2 + P}.\mu_{Ptr} \tag{5}$$

We started with a feature combination of TimeSinceQ, SpkrL@NAO, SpkrL@Ptr, PtrL@Spkr (denoted [3 4 5 7]), one of the best performing combination without the episode type. Only TimeSinceQ was adapted according to Equation 3, 4. Let N_1 and N_2 be the number of samples of data for a particular question q when NAO and Ptr is being addressed respectively. P is a parameter that was varied to change the weightage given to the sample mean of the measured observations and the prior. Table 4 shows the results, for varying P and no adaptation. Small P means more adaptation. The last column shows the results with the feature combination EpType, SpkrL@NAO, SpkrL@Ptr, and PtrL@Spkr for comparison (denoted [2 4 5 7]), which includes EpType as one of the features. Our results do not show significant improvement (and no improvement) with adaptation using ground-truth VFOA (and automatic VFOA) respectively. These are preliminary results and we need to rethink what approach could yield better results, say introducing a hidden variable in the modeling or choosing a different feature for adaptation.

We have reconfirmed in our setting that gaze cues from the speaker is the most important feature for addressee estimation. We also show that additional contextual features from the fellow participant, short-term and long-term dialog-context features helps improve the estimation accuracy. We observed improvements for both using ground-truth VFOA and automatic

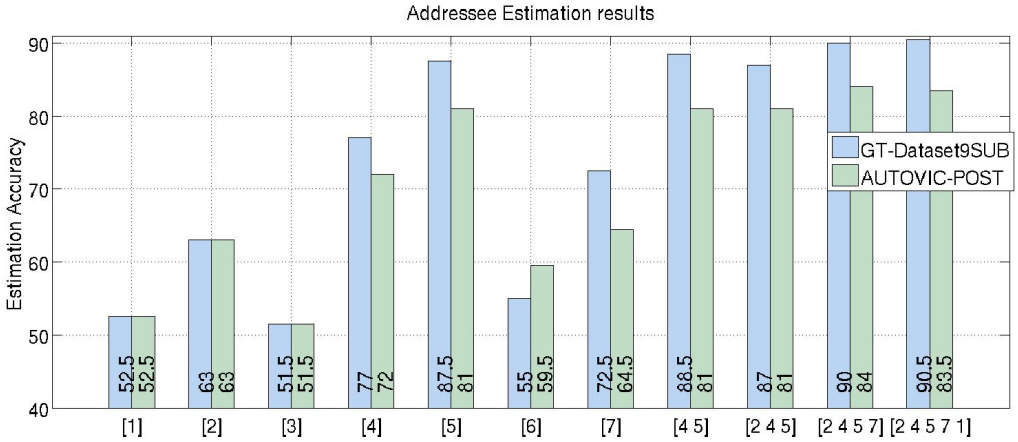


Figure 11: Result: Addressee estimation task, using either VFOA ground-truth or automatic VFOA.

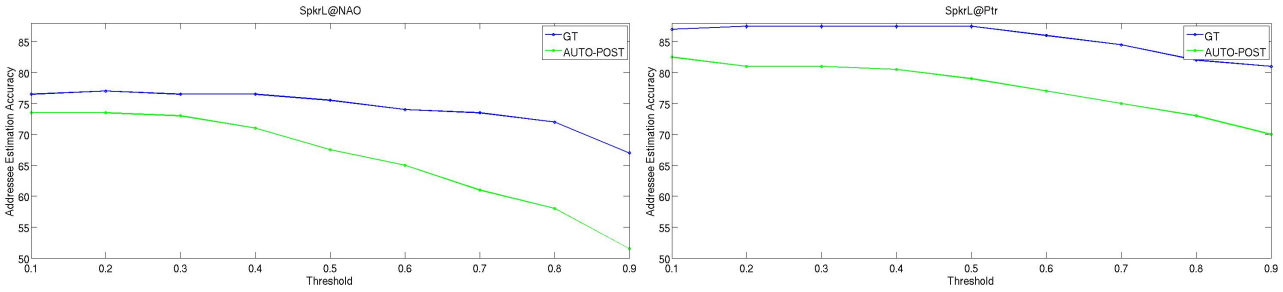


Figure 12: Accuracy vs threshold values for a simple rule-based classifer

VFOA with VICON head-pose. The loss in accuracy between ground-truth VFOA and automatic VFOA with VICON head-pose was around 6-7% accuracy. In the future, we plan to use automatically estimated VFOA using head-tracking and head-pose estimation to check the loss in accuracy as compared the results reported here. We hope context can play a more important role in the case of degraded VFOA estimation due to tracking difficulties. We also want to implement our addressee estimator on a real-time NAO platform and perform subjective user studies.

	Without EpType [3 4 5 7]				With EpType [2 4 5 7]
Using VFOA	P = 1	P = 5	P = 10	No adaptation	
Ground-truth	88%	88%	87%	87%	90%
Automatic(VICON)	82%	81%	80.5%	82%	81%

Table 4: Adaptation results: Shows the results of experiments where TimeSinceQ feature was adapted, with varying P and no adaptation. Ref. text for more details.

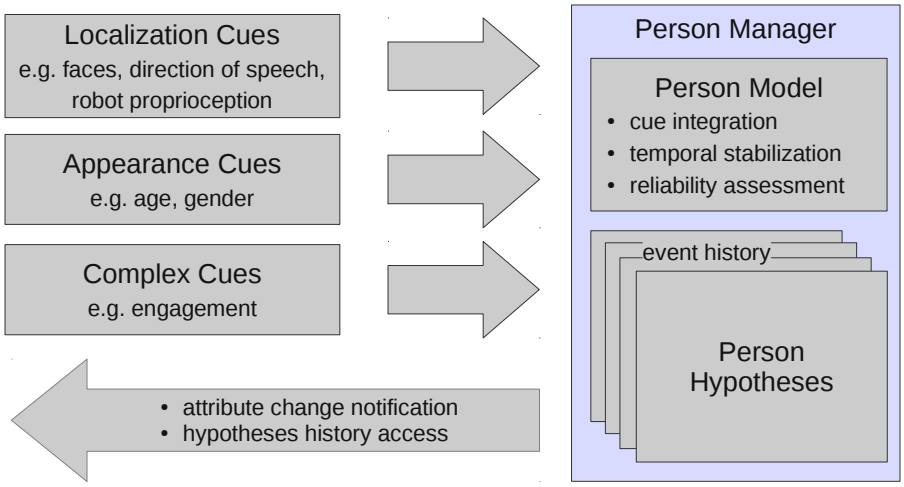


Figure 13: Overview of the person manager component

3.5 Memory architecture for a situation-aware robot

For robust interaction with a group of people, the stable perception of individuals is a strong necessity. Even though individual cues like face detection or sound events have been available before, their integration into consistent perception results is a major challenge and on the other hand an important benefit for the generation of advanced robotic behavior in interaction situations. This task has been addressed with the development of the person manager, a key component for the HUMAVIPS system. With its results, the person manager provides an important part of the situation-awareness of the system with respect to the external world and for the functioning of this advanced component, the architectural support components like temporal buffers ad timesync are essential tools, greatly simplifying the development process (either directly or in the underlying low-level perception layer).

The person manager provides stabilized and aggregated person hypotheses to the system as well as visualizing these hypotheses. The person manager component realizes these capabilities by maintaining an internal representation of persons in the environment which is continuously updated by integrating sensory cues available within the system. This method allows the person manager to derive results which exceed the capabilities of modules processing individual cues. A simple example is that the person manager is able to transform head tracking results to a robot centric representation of persons which is independent from the robot's head rotation. This result is achieved by integrating proprioception information from the robot with results of the head tracking component. Despite the simplicity of this example, it points to many requirements of the person manager component which will be analyzed in the following.

The person manager provides stabilized and aggregated person hypotheses to the system as well as visualizing these hypotheses. The person manager component realizes these capabilities by maintaining an internal representation of persons in the environment which is continuously updated by integrating sensory cues available within the system. This method allows the person manager to derive results which exceed the capabilities of modules processing individual cues. An overview of different cues which are integrated by the person manager is depicted

in Figure 13. For the analysis of requirements we refer to a simple example: The person manager is able to transform head tracking results to a robot centric representation of persons which is independent from the robot's head rotation. This result is achieved by integrating proprioception information from the robot with results of the head tracking component. Despite the simplicity of this example, it points to many requirements of the person manager component which will be analyzed after the following review of related work.

Robust Handling of Sparse Sensory Cues It cannot be expected that all parts of the system run at all time. There are several reasons for this assumption. First, during development it must be possible to test the person manager with subsystems. This allows for faster testing and debugging of the component. Furthermore, missing sensory input due to a failing component should cause as few successive errors as possible. Additionally, different configurations of the system might not require all cues. Thus, the person manager component can be used flexibly to avoid resource intensive system configurations if not necessary.

Temporal Synchronization The person manager needs to temporally synchronize sensory cues provided by different modules. For the input provided via the RSB framework this is achieved by using the temporal buffer component provided by the framework. With respect to the the requirement of handling sparse cues it is necessary to utilize the component in a way that does not lead to blocking in case sensory cues are missing.

Multi-Perspective Visualization To stabilize person hypotheses over time the person manager needs to maintain an internal model. To debug and analyze this model it must be possible to visualize its current state. A complete visualization of the person manager's current state not only requires a view from the robot body's perspective but also needs to visualize information about the robots position in the scene. These visualizations are required to efficiently debug and evaluate the system. It allows to quickly assess problems in the accuracy of sensory input or during integration multiple cues. Since temporal synchronization of events is crucial for the person manager component, a visualization of the temporal flow of events provides necessary means to detect such problems while running the system.

Event History The architecture of the person manager approximately follows the model view controller paradigm. RSB listeners receive events containing the relevant input cues (see Figure 14). The current implementation distinguishes between events which may lead to the instantiation of new person hypotheses and events which never lead to an instantiation of a person hypothesis but still are relevant to a person hypothesis. An example for events that can lead to the instantiation of a new person hypothesis are events from the head tracker. If such an event cannot be associated to an existing hypothesis based on attributes like ID, location or detection results, a new hypothesis is instantiated. All other events are buffered using a RSB temporal buffer component to make them available for query operations during further hypothesis processing.

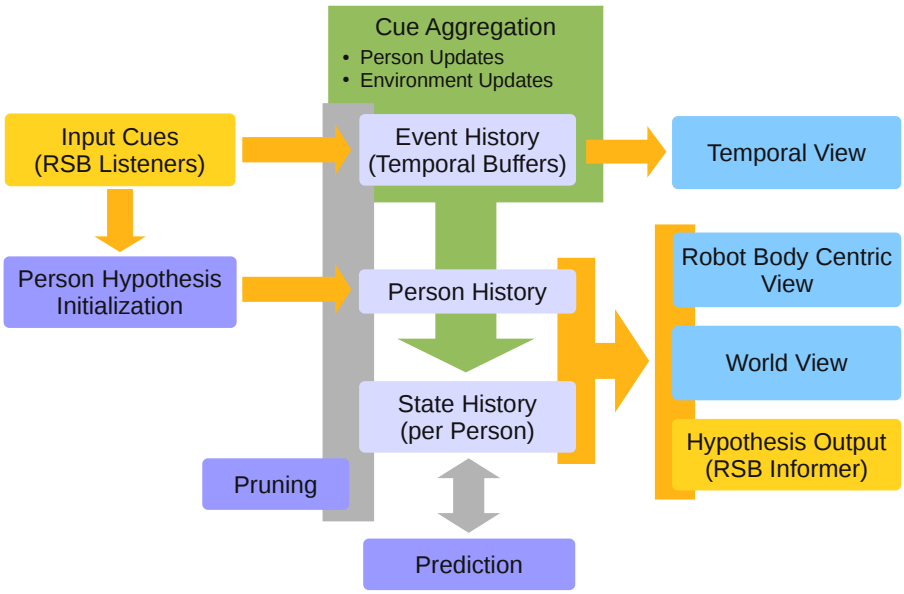


Figure 14: Architecture overview

Aggregation of Sensory Cues The model of persons is continuously updated. At each update step a cue aggregation algorithm tries to further complete states in the state history of a person hypothesis. A state can be completed if an event from each sensory cue is available and no events with a smaller temporal distance are possible. Since the RSB framework guarantees that the events are received in order this can be decided as soon as one event in the cue to be associated has arrived which is newer than the timestamp of the current state. At each time step any given state is completed as far as possible. This ensures that all information which can be aggregated in a person hypothesis is available with the lowest latency possible. Furthermore, this approach ensures that missing sensory input does not lead to person states being unavailable. Since the temporal position of an event in the history can be determined with a time complexity of $O(log(n))$ where n is the number of elements in the history. Therefore the aggregation algorithm is very efficient and can thus be executed at a high frequency.

Prediction The history of states also allows for the integration of prediction algorithms which can compensate for missing sensory input concerning a person. Since the robot cannot always have all person in its field of view this step is necessary to estimate person positions over a small temporal duration even if no events corresponding to this person are received. If the robot looks again in the direction of this person, the estimated state of the person can be compared with the current sensory cues to maintain a hypothesis with the same identity over time.

Flexible Visualization The history of states within a person history might contain states with a different level of completeness. The necessary level of completeness can be specified when querying the model for visualization and hypothesis output. For example the robot centric view does not need the information about the robots current position. However, it still

needs proprioception information to relate sensory input to the robot's body coordinate system. By this approach several requirements can be realized. First, the person manager can be used in systems where for example localization is not necessary. The robot centric view will continue to work. Second, if a module such as the localization temporally fails the person manager is still able to maintain person hypotheses, although they cannot be related to world coordinates anymore.

The implementation of the person manager was realized in C++ using QT for developing the graphical user interface. For plotting the qcustomplot plotting component was chosen. The implementation language was chosen especially with regard to visualization with frequent updates which is typically much more resources intensive compared to interpreted languages. In the following a short overview of the person manager's different views will be given.

Figure 15 shows the robot-centric perspective which visualizes persons in relation to the robots body. The light blue triangle visualizes the robot's field of view. The orange triangle displays the relative position of a person to the robot as well as his head rotation.

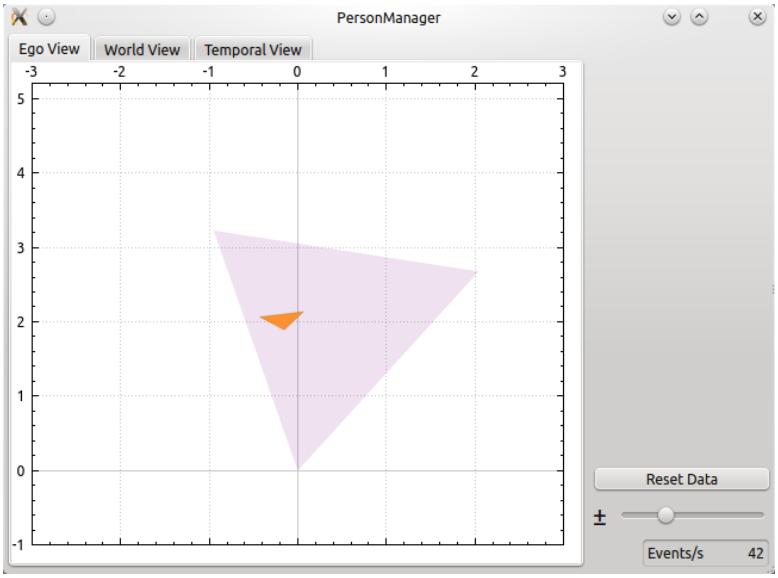


Figure 15: Robot-centric view of the person manager

Figure 16 shows the world-centric view which currently depicts the robot in relation to the table (light blue) within the scenario.

Figure 17 shows the temporal visualization of events. Events are plotted as points over time. The y-axis is used to display a relevant features as for example the number of heads encoded in an event originating from the headtracker. The more frequently events are received for a given sensory cue they are visualized darker due to their higher temporal density.

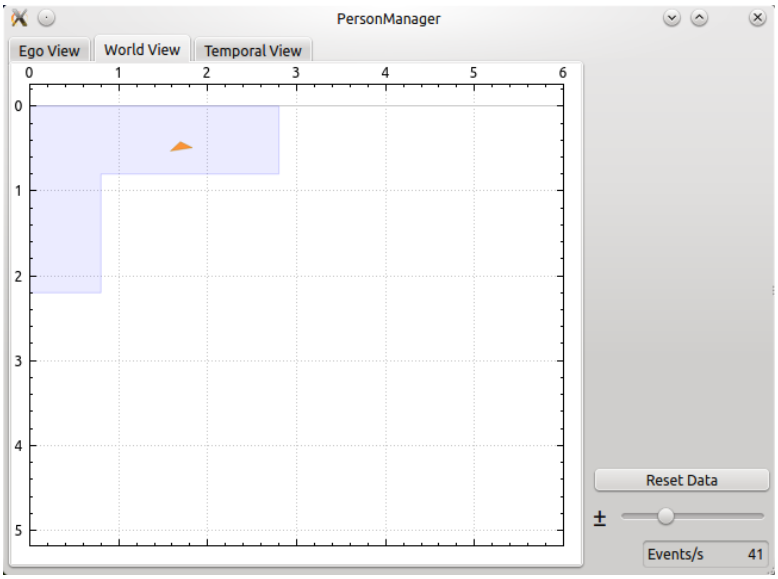


Figure 16: world-centric view of the person manager

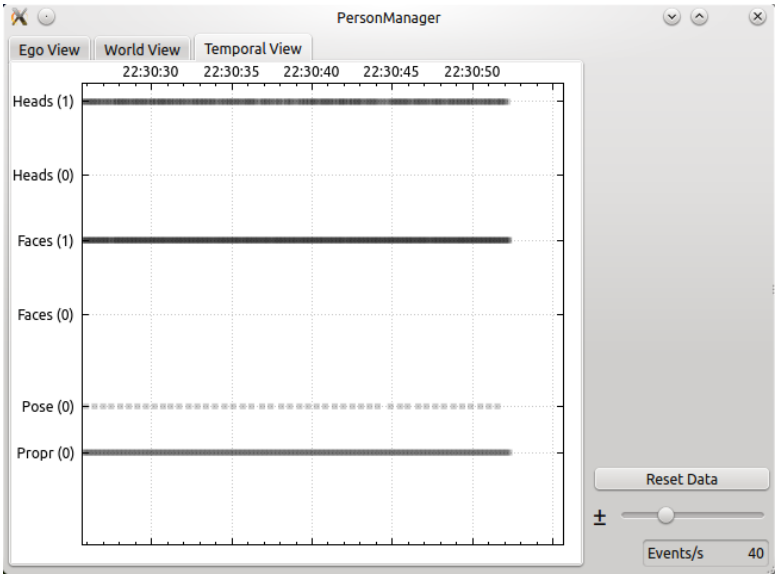


Figure 17: Temporal event view of the person manager

4 Potential Impact

The HUMAVIPS project has strong scientific, technological, and societal potential impacts, as detailed below:

4.1 Scientific Impact

HUMAVIPS was an interdisciplinary project that gathered the following disciplines: computer vision (face detection and recognition, vision-based localization and mapping, head tracking, gesture recognition), auditory signal processing and analysis (sound-source localization and separation, sound recognition), machine learning (mixture models, latent variable models, structured output classifiers, hidden Markov models), robotics (locomotion, navigation, human-robot interaction, hardware design, software architecture).

Such an interdisciplinary project is of high risk because it is difficult to combine scientific expertise from different topics, but it has higher scientific impact because it allows cross-fertilization. In particular, HUMAVIPS pioneered work in robot audition and in multimodal human-robot interaction. Robot audition is an emerging field. While robot vision is extremely well investigated, only a few teams in the world adventured into the difficult problem of endowing humanoid robots with advanced hearing capabilities.

The scientific impact of the project is demonstrated by the large number of scientific publications in all the topics addressed by the project (see below).

4.2 Technological Impact

From the technological point of view, HUMAVIPS advanced the state of the art in humanoid robotics by designing and building an *audiovisual* robot head. This new head is equipped with a stereoscopic camera pair and with four microphones. The head can deliver syncrhonized audio-visual data thus allowing audio-visual processing. Moreover, the head is fully compatible with the humanoid robot NAO commercialized worldwide by the project's partner Aldebaran Robotics. This simply means that NAO users can simply purchase this new head, plug it into their robot, and use it for any application involving the need of 3D vision and audio-visual integration. We note that currently there is no such integrated solution available on the humanoid market. Researchers in this field use an external 3D sensor, e.g. Kinect, that cannot be easily integrated into a stand-alone robotic head.

The software package developed under HUMAVIPS are fully compatible with the audiovisual NAO head and are available as open-source software packages. Hence, the project's demonstrators can be duplicated by other teams working with the same robot (NAO) and equipped with the new head just mentioned. This is likely to have an important impact because there are thousands of NAO users and developers in the world and hence they are potential users of the HUMAVIPS technology.

4.3 Societal Impact

The HUMAVIPS methodology and associated technology and software packages were implemented onto a consumer humanoid robot manufactured in Europe by one of the project partners. The choice of such a *cheap* robotic platform (of the order of 12,000 euros) is in strong contrast with the prevailing belief that expensive humanoids (of the order of 400,000 euros) are needed for research in this field and for developing novel robotic applications. The strong opinion of the HUMAVIPS consortium is that one needs to take into account hard constraints coming from the robotic platform being used (CPU power, memory size, limited electric power, limited weight, cheap components, etc.) for future commercialization of the technology.

During the review meeting, the project demonstrators were run in an unexpected environment for which the methods were neither trained nor tested. The trained/tested scenarios, while in an echoic environment, considered only a few people facing the robot. During the final demonstration there were 10-12 people wandering and chatting around, thus producing a level of background noise that was much higher than expected. Nevertheless, the outcomes were quite good and the interaction performances (visual and auditory) between the robot and a person were not too much affected but these severe conditions.

Of course, sometimes the robot was not able to correctly understand what was going on, but humans have also difficulties to understand each other in such situations. On the positive side, it was demonstrated that the fusion of visual information with audio (which was the main scientific goal of the project) drastically improves the standard audio-only and speech-based human-robot interaction paradigm.

Considering the ambitious objectives of the project, one may conclude that the presented demonstration is likely to foster the idea of *social robots*, more precisely, the following features will strongly impact the acceptance of robots among people:

- The humanoid robot NAO is able to interact with a group of people, based on fused audio and visual information, in a real physical world and in an unstructured environment;
- The robot is able to behave appropriately and to engage in a dialog with someone that is likely to be willing to interact with "him";
- The robot can react to both verbal and non-verbal signals generated by humans who do not wear any special device and who are not in a specially equipped room;
- The robot can recognize humans (based on face, age, and gender) and to name them once they have been seen for a while;
- The robot technology that was developed relies on a memory-centred and cognitive architecture, which is being provided to the research community as an open-source software platform.

All these features point out the advantages of an audio-visual robot. The methodology and technology developed by the project partners have been demonstrated in the *vernissage*

scenario. This goes well beyond a standard proof-of-concept demonstrator: it is already a real application for service robots. An interactive humanoid able to dialog with people and to have expressive gestures appears to be much more attractive, compared to the few guide robots available today.

The HUMAVIPS researchers are quite proud of their achievements. The integration of the demonstrations on NAO showed that the academic teams and the industrial partner can collaborate together to achieve an innovative application for a robot that is not just a laboratory prototype, but a commercial product.

5 Public Websites

The HUMAVIPS project has the following websites which remain active after the lifetime of the project:

- `http://humavips.inrialpes.fr`: Main project website (public deliverables, reports, scientific publications, events, etc.)
- `http://opensource.humavips.eu/`: The open-source software dissemination portal.
- `http://www.youtube.com/humavips`: Publicly available videos of the project's achievements
- `http://perception.inrialpes.fr/humavips/dataSet/`: Publicly available annotated datasets
- `https://code.humavips.eu/`: the HUMAVIPS Open Portal (HOP) hosting the projects Open Source Platform (OSS) and associated collaborative development tools.
- `https://ci.humavips.eu`: the continuous integration build service for the HUMAVIPS project.

6 Scientific Publications

- Peer-reviewed journal publications: [1] (book chapter), [2], [3], [4], [5], [6], [7]
- Paper that received a conference award: [8] (outstanding paper award), [9] (best student paper award), [10] (outstanding paper award), [11] (best student paper award).
- Peer-reviewed international conference and workshop publications: [12], [8], [13, 14], [15], [16, 17, 18, 19, 20, 21, 9] [22], [23], [24, 25, 26], [27], [28], [29], [30], [31], [32], [33], [34], [35, 36, 11, 37], [38, 39], [40, 10, 39, 41, 41, 42, 43, 44], [45], [18].
- Other publications: [46], [47, 48], [49].

7 Dissemination Activities

- The HUMAVIPS partners advertised the project towards several media. As a result, the following media articles emphasized the project:
  - An article published on wired.co.uk, February 2010: `http://www.wired.co.uk/news/archive/2010-02/12/humanoid-robots-to-gain-advanced-social-skills`
  - An article published in Hörakustik (Median-Verlag), January 2012, in German: `http://humavips.inrialpes.fr/files/2012/03/Cocktailparty-Effekt.pdf`
  - An article published on inriality.fr, August 2012, in French: `http://www.inriality.fr/informatique/robotique/interaction/les-robots-sinvitent/`
- The project's results were presented at the following international events: Automatica'12 (Istanbul, Turkey), ACM/IEEE ICMI'11 (Alicante, Spain), IEEE CVPR'11 (Colorado Springs, USA), IEEE ICASSP'11 (Prague, Czech Republic), IEEE Humanoids'12 (Osaka, Japan), ACM/IEEE HRI'12 (Boston, USA), ACM/IEEE ICMI'12 (Santa Monica, USA), IEEE ICASSP'13 (Vancouver, Canada), IEEE CVPR'13 (Providence, USA).

References

[1] V. Franc, S. Sonnenburg, and T. Werner, *Optimization for Machine Learning*. Cambridge, Massachusetts, USA: The MIT Press, 2012, ch. 7, pp. 185–218.

[2] V. Franc and P. Laskov, "Learning maximal margin markov networks via tractable convex optimization," *Control Systems and Computers*, no. 2, pp. 25–34, April 2011.

[3] I. Lütkebohle, R. Philippsen, V. Pradeep, E. Marder-Eppstein, and S. Wachsmuth, "Generic middleware support for coordinating robot software components: The task-state-pattern," *Journal of Software Engineering for Robotics*, vol. 2, no. 1, 2011.

[4] A. Shekhovtsov and V. Hlaváč, "A distributed mincut/maxflow algorithm combining path augmentation and push-relabel," *International Journal of Computer Vision*, 2012. [Online]. Available: http://link.springer.com/article/10.1007%2Fs11263-012-0571-2

[5] X. Alameda-Pineda, J. Sanchez-Riera, J. Wienke, V. Franc, J. Cech, K. Kulkarni, A. Deleforge, and R. P. Horaud, "Ravel: An annotated corpus for training robots with audiovisual abilities," *Journal on Multimodal User Interfaces*, vol. 7, no. 1-2, pp. 79–91, March 2013. [Online]. Available: http://perception.inrialpes.fr/Publications/2013/ASWFCKDH13/

[6] S. Duffner and J.-M. Odobez, "A track creation and deletion framework for long-term online multi-face tracking," *IEEE Transaction on Image Processing*, March 2013.

[7] D. Sanchez-Cortes, O. Aran, D. Jayagopi, M. Schmid Mast, and D. Gatica-Perez, "Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition," *Journal on Multimodal User Interfaces*, vol. 7, pp. 39–53, 2013. [Online]. Available: http://dx.doi.org/10.1007/s12193-012-0101-0

[8] X. Alameda-Pineda, V. Khalidov, R. P. Horaud, and F. Forbes, "Finding audio-visual events in informal social gatherings," in *Proceedings of the 13th International Conference on Multimodal Interfaces*. Alicante, Spain: ACM, November 2011, pp. 247–254, oustanding paper award. [Online]. Available: http://perception.inrialpes.fr/Publications/2011/AKHF11

[9] M. Uřičář, V. Franc, and V. Hlaváč, "Detector of facial landmarks learned by the structured output SVM," in *VISAPP '12: Proceedings of the 7th International Conference on Computer Vision Theory and Applications*, G. Csurka and J. Braz, Eds., vol. 1. Portugal: SciTePress — Science and Technology Publications, February 2012, pp. 547–556.

[10] D. Jayagopi, D. Sanchez-Cortes, K. Otsuka, J. Yamato, and D. Gatica-Perez, "Linking speaking and looking behavior patterns with group composition, perception, and performance," in *Proceedings of the 14th ACM International Conference on Multimodal interaction, Santa Monica, USA*, 2012, pp. 433–440, Outstanding paper award.

[11] K. Funes and J.-M. Odobez, "Gaze estimation from multimodal kinect data," in *CVPR Workshop on Face and Gesture and Kinect demonstration competition, Providence*, june 2012.

[12] J. Wienke and S. Wrede, "A middleware for collaborative research in experimental robotics," in *2011 IEEE/SICE International Symposium on System Integration, SII2011*, IEEE. Kyoto, Japan: IEEE, 2011.

[13] J. Cech, J. Sanchez-Riera, and R. P. Horaud, "Scene flow estimation by growing correspondence seeds," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, 2011. [Online]. Available: http://perception.inrialpes.fr/Publications/2011/CSH11

[14] J. Cech and R. P. Horaud, "Joint disparity and optical flow by correspondence growing," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Prague, Czech Republic: IEEE, May 2011, pp. 893 – 896. [Online]. Available: http://perception.inrialpes.fr/Publications/2011/CH11

[15] K. Pitsch, S. Wrede, J.-C. Seele, and L. Süssenbach, "Attitude of german museum visitors towards an interactive art guide robot," in *Proceedings of the 6th international conference on Human-robot interaction*, ser. HRI '11. New York, NY, USA: ACM, 2011, pp. 227–228. [Online]. Available: http://doi.acm.org/10.1145/1957656.1957744

[16] C. Scheffler and J.-M. Odobez, "Joint adaptive colour modelling and skin, hair and clothing segmentation using coherent probabilistic index maps," in *British Machine Vision Conference*, Sep. 2011.

[17] S. Duffner and J. Odobez, "Exploiting long-term observations for track creation and deletion in online multi-face tracking," in *IEEE Conference on Automatic Face and Gesture Recognition (F&G)*, mar 2011.

[18] M. Janvier, X. Alameda-Pineda, L. Girin, and R. P. Horaud, "Sound-event recognition with a companion humanoid," in *IEEE International Conference on Humanoid Robotics*, Osaka, Japan, November 2012. [Online]. Available: http://perception.inrialpes.fr/Publications/2012/JAGH12

[19] A. Shekhovtsov and V. Hlaváč, "A distributed mincut/maxflow algorithm combining path augmentation and push-relabel," in *Proceedings of the 8th International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR)*, ser. Lecture Notes in Computer Science, Y. Boykov, F. Kahl, V. Lempitsky, and F. Schmidt, Eds., vol. 6819. Berlin, Germany: Springer, July 2011, pp. 1–16.

[20] V. Franc, A. Zien, and B. Schölkopf, "Support vector machines as probabilistic models," in *Proceedings of the 28th Annual International Conference on Machine Learning (ICML 2011)*, L. Getoor and T. Scheffer, Eds. New York, USA: ACM, June/July 2011, pp. 665–672.

[21] M. Uřičář and V. Franc, "Efficient algorithm for regularized risk minimization," in *CVWW '12: Proceedings of the 17th Computer Vision Winter Workshop*, M. Kristan, R. Mandeljc, and L. Čechovin, Eds. Ljubljana, Slovenia: Slovenian Pattern Recognition Society, February 2012, pp. 57–64, cD-ROM.

[22] J. Wienke, D. Klotz, and S. Wrede, "A framework for the acquisition of multimodal human-robot interaction data sets with a whole-system perspective," in *LREC Workshop on Multimodal Corpora for Machine Learning: How should multimodal corpora deal with the situation?*, Istanbul, Turkey, 05/2012 2012.

[23] D. Jayagopi, S. Sheikhi, D. Klotz, J. Wienke, J.-M. Odobez, S. Wrede, V. Khalidov, L. Nguyen, B. Wrede, and D. Gatica-Perez, "The vernissage corpus: A conversational human-robot interaction dataset," in *8th ACM/IEEE International Conference on Human-Robot Interaction*, march 2013.

[24] J. Wienke, A. Nordmann, and S. Wrede, "A meta-model and toolchain for improved interoperability of robotic frameworks," in *SIMULATION, MODELING, and PROGRAMMING for AUTONOMOUS ROBOTS*, vol. 7628, Springer Berlin Heidelberg. Tsukuba, Japan: Springer Berlin Heidelberg, 11/2012 2012.

[25] F. Lier, F. Siepmann, T. Paul-Stüve, S. Wrede, S. Wachsmuth, and I. Lütkebohle, "Facilitating research cooperation through linking and sharing of heterogenous research artifacts," in *Proceedings of the 8th International Conference on Semantic Systems (I-SEMANTICS '12)*, H. Sack and T. Pellegrini, Eds. ACM, 2012, pp. 157–164. [Online]. Available: http://dx.doi.org/10.1145/2362499.2362521

[26] J. Moringen, A. Nordmann, and S. Wrede, "A cross-platform data acquisition and transformation approach for whole-systems experimentation – status and challenges," in *ERF Workshop Infrastructure for Robot Analysis and Benchmarking*, in press.

[27] A. Deleforge and R. P. Horaud, "The cocktail party robot: Sound source separation and localisation with an active binaural head," in *IEEE/ACM International Conference on Human Robot Interaction*, Boston, Mass, March 2012. [Online]. Available: http://perception.inrialpes.fr/Publications/2012/DH12

[28] ——, "A latently constrained mixture model for audio source separation and localization," in *Proceedings of the 10th International Conference on Latent Variable Analysis and Signal Separation*, vol. LNCS 7191. Tel Aviv, Israel: Springer-Verlag, March 2012, pp. 372–379. [Online]. Available: http://perception.inrialpes.fr/Publications/2012/DH12a

[29] ——, "2d sound-source localization on the binaural manifold," in *IEEE International Workshop on Machine Learning for Signal Processing*, Santander, Spain, September 2012. [Online]. Available: http://perception.inrialpes.fr/Publications/2012/DH12b

[30] J. Sanchez-Riera, J. Cech, and R. P. Horaud, "Action recognition robust to background clutter by using stereo vision," in *The Fourth International Workshop on Video Event Categorization, Tagging and Retrieval*, ser. LNCS. Springer, October 2012. [Online]. Available: http://perception.inrialpes.fr/Publications/2012/SCH12a

[31] J. Sanchez-Riera, X. Alameda-Pineda, and R. P. Horaud, "Audio-visual robot command recognition," in *ACM/IEEE International Conference on Multimodal Interaction*. ACM, November 2012. [Online]. Available: http://perception.inrialpes.fr/Publications/2012/SAH12

[32] X. Alameda-Pineda and R. P. Horaud, "Geometrically-constrained robust time delay estimation using non-coplanar microphone arrays," in *Proceeding of the 20th European Signal Processing Conference (EUSIPCO)*, Bucharest, Romania, August 2012, pp. 1309–1313. [Online]. Available: http://perception.inrialpes.fr/Publications/2012/AH12

[33] X. Alameda-Pineda, J. Sanchez-Riera, and R. P. Horaud, "Benchmarking methods for audio-visual recognition using tiny training sets," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vancouver, Canada: IEEE Signal Processing Society, May 2013. [Online]. Available: http://perception.inrialpes.fr/Publications/2013/ASH13

[34] A. Deleforge, F. Forbes, and R. P. Horaud, "Variational em for binaural sound-source separation and localization," in *IEEE International Conference on Acoustics, Speech, and Signal Processing*. Vancouver, Canada: IEEE Signal Processing Society, May 2013. [Online]. Available: http://perception.inrialpes.fr/Publications/2013/DFH13

[35] S. Sheikhi and J.-M. Odobez, "Investigating the midline effect for visual focus of attention recognition," in *ACM Int Conf. on Multimodal Interaction (ICMI)*, oct 2012.

[36] L. Nguyen, J.-M. Odobez, and D. Gatica-Perez, "Using self-context for multimodal detection of head nods in face-to-face interactions," in *ACM Int Conf. on Multimodal Interaction (ICMI), Santa Monica*, oct 2012.

[37] S. Sheikhi, V. Khalidov, and J.-M. Odobez, "Recognizing the visual focus of attention for human robot interaction," in *IROS workshop on Human Behavior Understanding, Vilamoura*, oct 2012.

[38] Š. Fojtů, M. Havlena, and T. Pajdla, "Nao robot localization and navigation using fusion of odometry and visual sensor data," in *Intelligent Robotics and Applications*, ser. Lecture Notes in Computer Science, C.-Y. Su, S. Rakheja, and H. Liu, Eds., vol. 7507, no. 2, Concordia University, Canada. Berlin, Germany: Springer Berlin Heidelberg, October 2012, pp. 427–438.

[39] D. Prusa and T. Werner, "Universality of the local marginal polytope," in *Proc. Computer Vision and Pattern Recognition (CVPR)*, 2013.

[40] D. Jayagopi and J.-M. Odobez, "Given that, should i respond? contextual addressee estimation in multi-party human-robot interactions," in *8th ACM/IEEE International Conference on Human-Robot Interaction*, march 2013.

[41] M. Uřičář, V. Franc, and V. Hlaváč, "Bundle methods for structured output learning – back to the roots," in *18th Scandinavian Conference on Image Analysis*, 2013.

[42] K. Antoniuk, V. Franc, and V. Hlaváč, "Learning markov networks by analytic center cutting plane method," in *ICPR '12: Proceedings of 21st International Conference on Pattern Recognition*, IAPR. New York, USA: IEEE, November 2012, pp. 2250–2253.

[43] L. Cerman and V. Hlaváč, "Tracking with context as a semi-supervised learning and labeling problem," in *ICPR '12: Proceedings of 21st International Conference on Pattern Recognition*, IAPR. New York, USA: IEEE, November 2012, pp. 2124–2127, cD-ROM.

[44] H. Cevikalp, B. Triggs, and V. Franc, "Face and landmark detection by using cascade of classifiers," in *Title: 10th IEEE International Conference on Automatic Face and Gesture Recognition*, 2013.

[45] J. Sanchez-Riera, X. Alameda-Pineda, J. Wienke, A. Deleforge, S. Arias, J. Cech, S. Wrede, and R. P. Horaud, "Online multimodal speaker detection for humanoid robots," in *IEEE International Conference on Humanoid Robotics*, Osaka, Japan, November 2012. [Online]. Available: http://perception.inrialpes.fr/Publications/2012/SAWDACWH12

[46] M. Uřičář and V. Franc, "Bundle method for structured output learning," Department of Cybernetics, Faculty of Electrical Engineering, Czech Technical University, Prague, Czech Republic, Research Report K333–46/12, CTU–CMP–2012–20, September 2012.

[47] M. Uřičář, "Detector of facial landmarks," MSc Thesis CTU–CMP–2011–05, Center for Machine Perception, K13133 FEE Czech Technical University, Prague, Czech Republic, June 2011.

[48] O. Fišar, "Structural classifier for gender recognition," MSc Thesis CTU–CMP–2011–09, Center for Machine Perception, K13133 FEE Czech Technical University, Prague, Czech Republic, September 2011.

[49] D. Klotz, J. Wienke, J. Peltason, B. Wrede, S. Wrede, V. Khalidov, and J.-M. Odobez, "Engagement-based multi-party dialog with a humanoid robot," in *Proceedings of the SIGDIAL 2011 Conference*. Portland, Oregon: Association for Computational Linguistics, June 2011, pp. 341–343. [Online]. Available: http://www.aclweb.org/anthology/W/W11/W11-2042