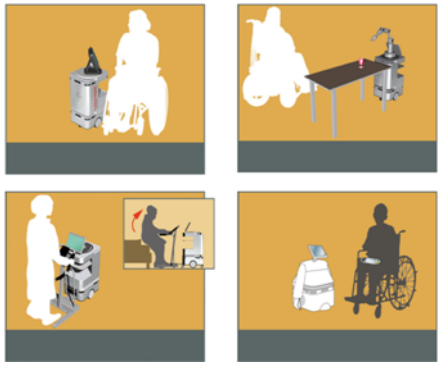

DOMEO Project
AAL-2008-1-159

D4.2 Advanced Functions Lab Tests Report

Document Information	
Title	Advanced Functions Lab Tests Report
Workpackage/Deliverable	ROBUMATE
Responsible	ISIR
Due Date	T0+36
Actual Date	25/12/2012
Type	Delivrable
Status	Version 1.0
Dissemination Level	Public
Authors	Philippe Bidaud
Project URL	www.aal-domeo.eu

Abstract:
This document describes the advanced functions developed by ISIR, and tested by TUW. According to the DoW revision, these functions have been tested only in laboratory, without patients.

Keyword List:
Advanced functions, trials

Table of Content

Summary

The developments realized in the context of the DOMEO WP4 aim to increase interaction capabilities of the system by integrating advanced features whose main objective are: the perception of the subject, his location and the motion control of the robot for interactive tracking and focusing interfaces on the subject.

In this document, we describe the methods and the algorithms developed and integrated into the robot Kompaï. The latter relate more specifically to:

- the detection of the person
- the location of the person in the environment and in the coordinate robot
- the robot motion control tracking
- the detection of the person's face for recognition and extraction of non-verbal signals
- the detection of the speaker
- the focus on the speaker to strengthen the robustness of the voice interaction

These so-called advanced functions have been evaluating their performance in laboratory tests at the ISIR and some of them under conditions reproducing a living at TUW.

1. Introduction

Personal robotic systems can be used in our households, offices, hospitals, shopping malls, and others in-door cluttered and human-populated environments. One of the main functions they have to fulfill is localizing, coming close to and tracking a person as for instance for focusing the person by the sensors onboard sound and vision, this taking into account the movements of the person, the environment in which it operates and the tasks that the robot is supposed to achieve for any individual. The synthesis of interactive and relevant behaviors is one of the central issues for these systems and is particularly challenging. Decide the action to do in full autonomy requires first that the robot can perceive its fixed and moving environment, localize, recognize and take into account both the context and the person activity. In addition, it may react to possible interactions with the user (inputs from different interfaces: audio, tactile, video, etc.). It must also consider the state of the person, predict its movements and anticipate its path if necessary.

We have developed a decision making engine that uses the fuzzy logic mechanisms to select a particular robot behavior for a given context and with which it is possible to handle a wide range of scenarios and of specific robot functions. In the context of DOMEO project, we focused the basic robot functions essential for a natural and efficient interaction with humans: searching a person in the environment, follow him/her, focalize him/her for face-to-face interaction. The realization of these functions requires the exploitation of a large amount of information collected by robot's sensors: the webcam on the Kompaï's head for human body detection, the webcam on the touch screen for face detection and the laser scan for autonomous navigation, obstacle avoidance and for human legs detection.

In what follows, we develop the technical principles of the main functions developed and integrated in the experimental platform DOMEO. All these functions are illustrated by a video document which is accessible at the URL: http://youtube/yllJtS_S7tk

2. Human Detection

Laser based human detection

Purpose of function:Robust laser based leg detector to identify the presence in the near field of the robot and to locate him in the robot local frame.

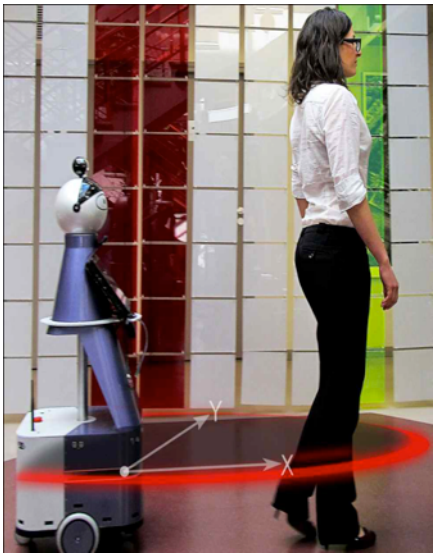

Figure 1 Robot detection plane (laser plane).

Method: The laser based detector operates from a model of the legs developed in the form of two Gaussian distributions of the legs size and posture (distance between 2 legs) obtained experimentally from data collected from a sample of 100 scans capturing several persons static or moving close and far.
Using these models, the proposed approach is composed by the 3 following steps (Figure 2).

1. We first look interest points by detecting "blobs" in the laser scan. A blob is a set of consecutive laser reflection points supposed appertain to the same object. Two points belong to the same blob if their distances are in the range of 0.1 m.
2. For the recognition of a blob as leg, we apply the defined Gaussian model of the leg.
3. Two legs are considered as a pair of legs if their distance is in the Gaussian model leg posture.

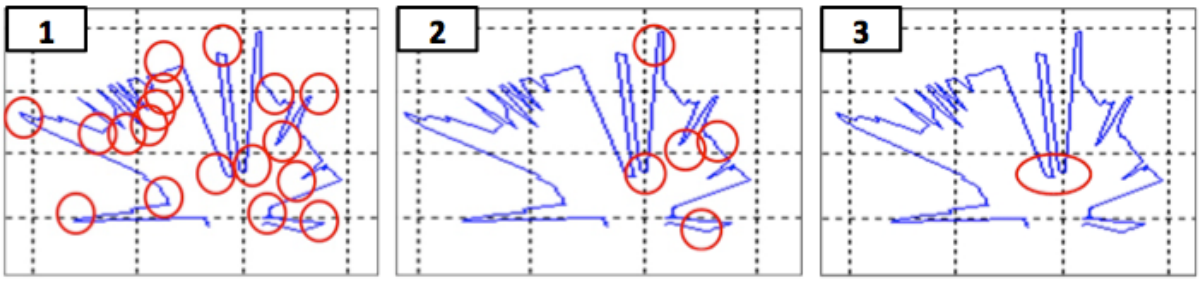

Figure 2: The three steps proposed for the laser based person detector.

Laboratory test at ISIR: The reliability and the robustness of the detector are been tested in real and natural conditions by moving the robot in front of 15 persons evolving naturally along various trajectories. The detector shows experimentally a correct detection rate >30% within 3 m (this means an average of 4 detections per second) (Figure 3).

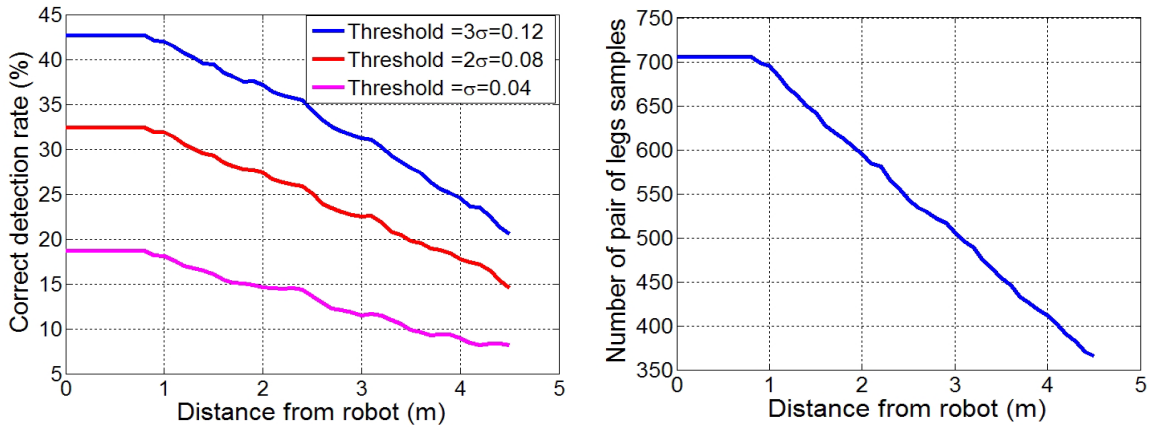

Figure 3 Left: percentage of correct detection of legs (pairs) with 3 detection thresholds; Right: ground truth (mean number of pairs of legs obtained by using sliding window method).

We can remark in Figure 4 that the rate of false positives detections drops considerably at a distance of about 1.2 m. It can be explained by the fact that the false detections are naturally filtered during the coupling phase.

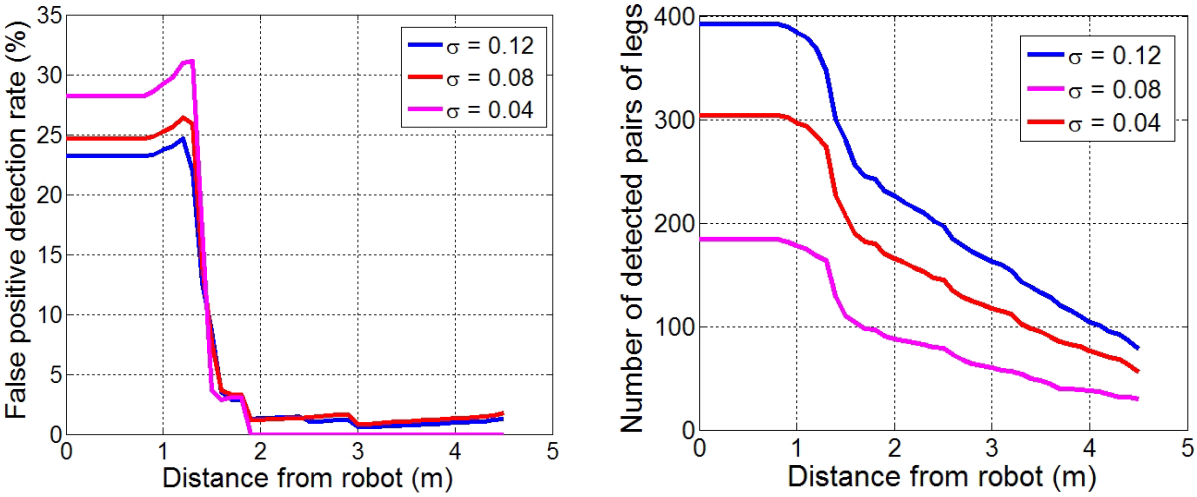

Figure 4 Left: percentage of false positives in detecting legs (pairs). Right: total number of pairs of legs detected (obtained by using sliding windows method).

Laser+camera based human detection

Purpose of function:Detecting a person with a depth of field greater than that of the laser and increase the robustness of detection in the field of laser based on the geometric characteristics 2D vertical.

Method: Embedded on wheeled mobile robots, laser sensors and cameras are often used respectively for legs detection and for face or body detection. We propose to merge the data from the two sensors using probability grid (sampling in cells the robot's environment). The probability of human presence is computed for each cell. Our grid is based on polar split of the space: each cell is identified by its polar coordinates (ρ, θ). The grid is denser close to the robot (where we need more accurate position estimation) (Figure 5). The geometric relation between the camera and the laser scanner is estimated in order to project the cells of the grid in laser and vision space (Figure 6). Each cell in grid is passed through the laser and vision based detectors. The two detectors' outputs are normalized using distances of Mahalanobis, and then summed to get the final score. A threshold rule is applied on the final score.

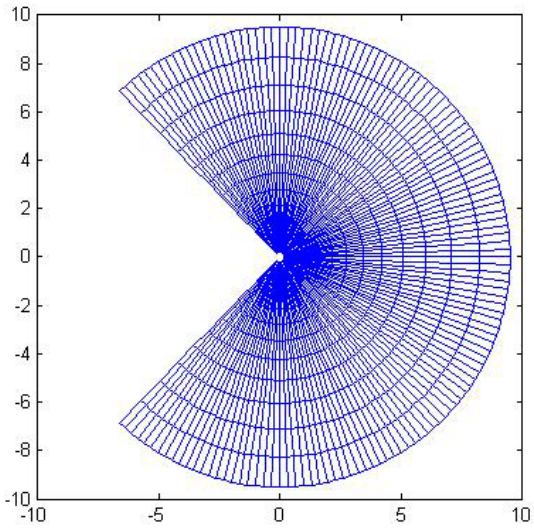

Figure 5 Probability grid representation for laser sensor.

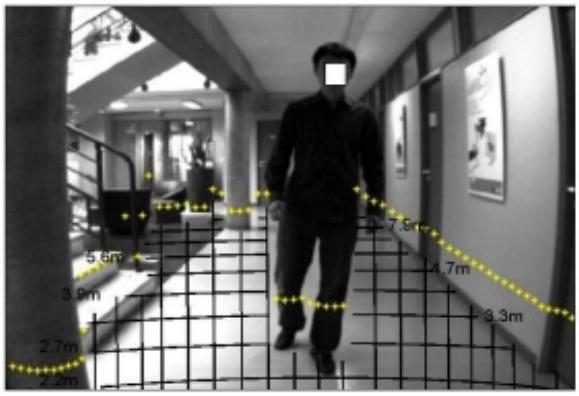

Figure 6 Example of grid projected in vision space.

For the legs detection we have used the same Gaussian models defined for the laser based detector. For the body detection we have used a cascade of boosted classifiers: Real

AdaBoost algorithm[1] (an improved version of the Adaboost algorithm). Furthermore, two types of features are employed: Haar-Like (Haar-like features or Haar features) and HoG (Histogram of Oriented Gradients). These two features are frequently used in the domain of object detection and recognition. The output of the cascade is finally fitted to a sigmoid probabilistic function using the learning database. This last process allows us to allocate a probability of human presence to each cell.
The output probabilities of the two detectors are combined by using a grid based approach and Gaussian Mixture Models (GMM). For the grids fusion, the distance between the robot and the detection is explicitly taken into account. In fact, close to the robot (< 1.4 m), only the laser based detector works because of the camera field of view. At distances greater than 3m, the total body and the upper body detectors are more reliable than the legs detector, this because the probability of intersection of laser rays and a leg decreases drastically farther 3m. At intermediate distances, a merging process can exploit the information redundancy to improve the performances of each detector.

Laboratory test at ISIR: We have selected 15 persons of different heights, origins and dressed with different clothes (Figure 7 and Table 1). We have asked to the participants to walk in front of the robot. We have collected 271 data (image and laser frame) in the 3 different indoor environments showed in Figure 8 (that represents 471 potential human observations).

Figure 7 Pictures of the 15 subjects selected for the testing data base.

Subject	Height(m)	Gender	Origin	Clothes*
1	1.56	F	Asia	LS
2	1.66	F	Europe	SS + TP
3	1.69	M	Europe	LS + TP
4	1.70	F	Asia	SS
5	1.70	M	Europe	LS
6	1.70	F	Europe	SS
7	1.73	M	Europe	LS
8	1.73	M	Asia	LS + TP
9	1.76	M	Asia	LS + TP
10	1.78	M	Colombia	LS
11	1.78	M	Europe	LS
12	1.78	M	Half-cast	LS
13	1.80	M	Europe	LS
14	1.86	M	Europe	LS
15	1.98	M	Europe	LS

*(SS: snug slacks, LS: large slacks, TP: textured pullover)

Table 1 data base subject's characteristics

1 Schapire, R. E., and Singer, Y. Improved boosting algorithms using confidence rated predictions. Machine Learning 37, 3 (1999).

To analyze the detector, we have grouped the observations following the human proxemics categorization proposed by Hall[2] (Figure 9). We can observe that the weakness of the leg detector observed in the global ROC curve, is partially due to the non-homogeneous distribution of the data. The most of them (>50% of the potential observations) have been collected in the Public Space, where the body detector is more reliable and the legs one works badly (because the probability that a laser ray intersects a leg decreases drastically in the beginning of the Public Space). In the Personal Space, only the laser based detector works: the camera cannot detect the whole body of a close person. In the Social Space, the information provided by the two sensors has to be merged.

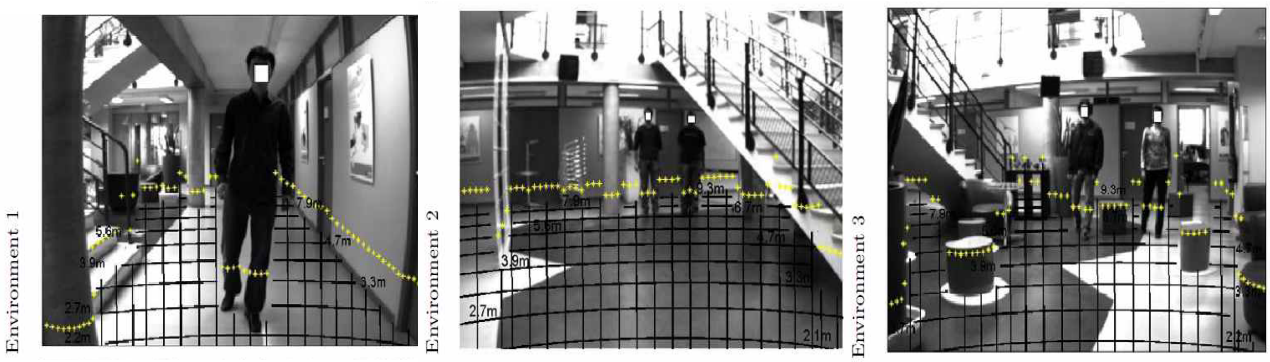

Figure 8 The testing environment at ISIR.

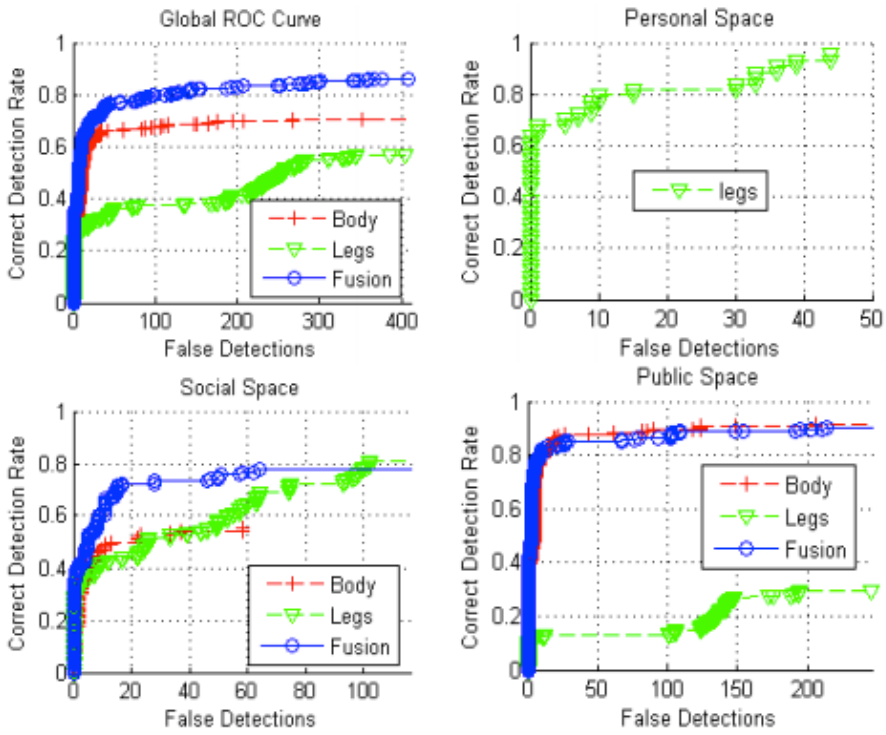

Figure 9 ROC curves t different spaces.

2 Hall, E. Proxemics. Current Anthropology 9, 2-3 (1968).

3. Human Localization

Purpose of function: The position of the person cannot be determined just by direct and punctual observation, this for obvious reasons related to false or lack of detection and noise in the measurement. To determine the person evolution and to predict its trajectory, the information provided by the human detector are processed by an EKF (Extended Kalman Filter).

Method: Assuming that humans move with constant velocity in a small interval of time (the time elapsed between 2 laser scans) and that the system and measurement noises have zero-mean Gaussian distribution, we can describe the human motion in polar coordinates by the discrete state variable of the following equation.

$$\begin{cases} x_{k+1|k} = F_k x_{k|k} + w_k \\ y_k = C_k x_k + v_k \end{cases}$$

$$x_k = \begin{bmatrix} \rho_k \\ \theta_k \\ \dot{\rho}_k \\ \dot{\theta}_k \end{bmatrix} K_k = \begin{bmatrix} 1 & 0 & \Delta t_k & 0 \\ 0 & 1 & 0 & \Delta t_k \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} C_k = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{bmatrix}$$

where $x_{k+1|k}$ is the state equation, y_k is the measurements equation, ρ_k and θ_k are the polar coordinates at the sampling time k ; $\dot{\rho}_k$and $\dot{\theta}_k$ are the velocities of ρ_k and θ_k at k ; F_k is the state transition matrix; C_k is the observation matrix and w_k and v_k are respectively the system noise due to mis-modeling and the observation noise due to the sensor and to the detector errors. At each sampling time k, the velocities$\dot{\rho}_k$and $\dot{\theta}_k$ are updated in the state equation computing the person velocities from its motion during the 10 last sampling times.

Therefore, we investigated the use of a human locomotion oriented prediction by relying on Hitcheur's studies[3] that have shown that humans tend to keep a constant speed when walking straight and to reduce their speed when turning.

Laboratory test at ISIR: Since the first test session has appeared as a model of locomotion at constant speed is not suitable for predicting the human trajectory, especially if occultations and losses of detection occur (Figure 10).

[3] H. Hicheur, S. Vieilledent, M. J. E. Richardson, T. Flash and A. Berthoz. Velocity and Curvature in Human Locomotion along Complex Curved Paths : a Comparison with Hand Movements. Experimental Brain Research, vol. 162, issue 2, pages145-154 (2005).

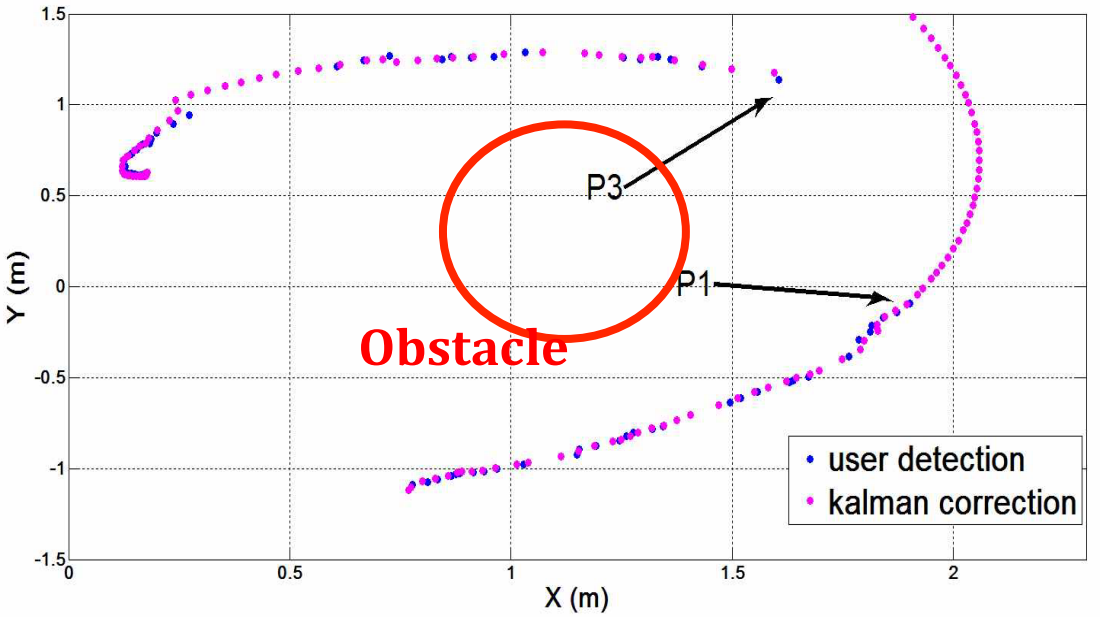

Figure 10 Example of person state prediction in a case of important loss of detection.

The introduction of the human locomotion model has greatly improved the EKF performances, even in case of important occlusions (Figure 11).

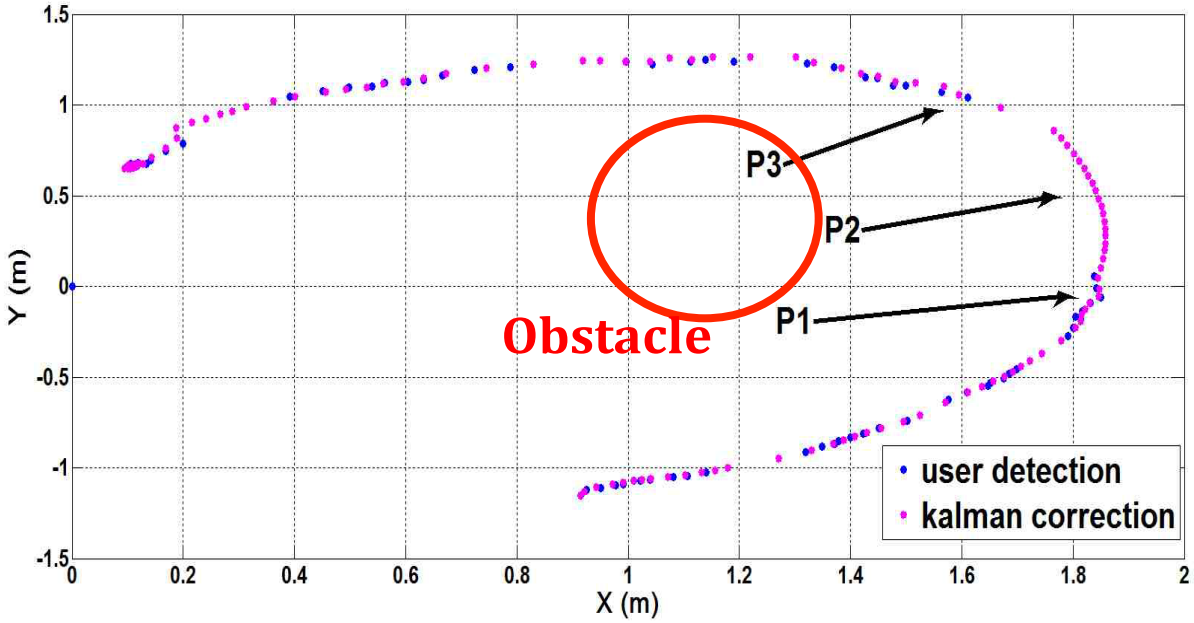

Figure 11 Example of person state prediction in case of important loss of detection and considering the human locomotion model.

4. Human Searching

Purpose of function: When the robot has to start an interaction with the person from an initial configuration which is far from him/her, it has at first to search and localize him/her in the global environment.

Method: We have developed software architecture for the synthesis of interactive robot behavior. This architecture combines three layers (Figure 12). A perception layer which is connected to a number of sensors and interfaces to identify and analyze the behavior of the person (human detection and localization) and requests (by using both voice based and graphical based interfaces). A decision layer, which from the entries, can be inferred on the actions and finally a control layer, which generates and controls the actions. For the decision layer we propose a decisional engine that uses the fuzzy logic mechanisms to select a particular strategy for a given context (exploiting SpirOps software as development tool). In the case of the person searching function; it exploits an exploration strategy, which makes use of the SLAM integrated on the Kompaï robot. The criteria used to select the point to explore among a predefined set of points is based on the time information of the last visit: the chosen point is located to the place that has been visited less recently.

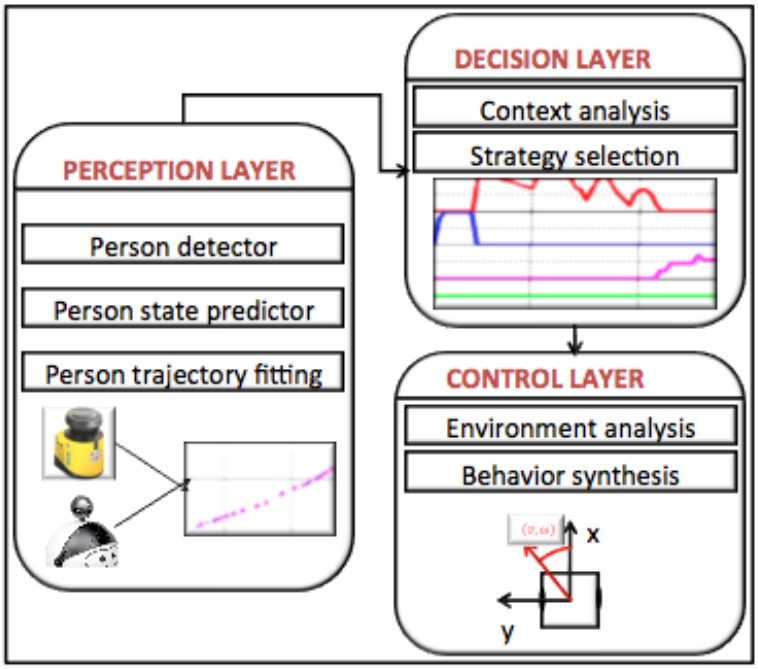

Figure 12 Framework architecture combining perception, decision and control layers.

Laboratory test at ISIR: In this experiment we want to demonstrate the robot ability to find a person in a cluttered environment and without any a priori information about the person localization. The robot and the person are in a corridor and an armchair hides the legs of the person. We have repeated the experience twice. The robot uses the laser based detector (only the laser data are considered for the detection) at first time. At the second time it uses the laser+camera based detector. In the first case it is impossible for the robot to detect the person from its initial pose. The robot goes searching for the person into two rooms, R1 and R2 (Figure 13). In these rooms there are tables and chairs that are considered likely persons because of the similar shape (Figure 14). In order to validate the detection the robot approaches, but it rejects the false detections and continues the global search.

The real person is detected when the robot is moving along the corridor after about 106 s. Note the use of colors: purple when the robot detects the person, blue when it doesn't. In the second case the robot uses the laser+camera based detector and the person is detected immediately (Figure 15). The total duration of the research has been only 24 s. This experience proves the clear advantage of using a more reliable and powerful detector in cluttered environments.

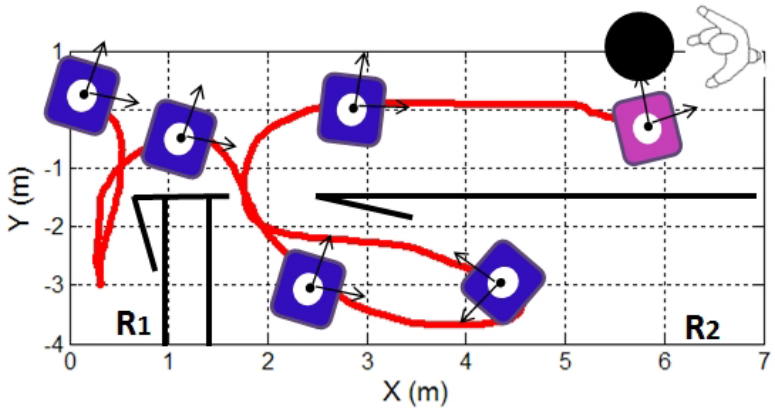

Figure 13 Robot trajectory during the research of the person and using just the laser based detector.

Figure 14 Picture of a laboratory test environment (ISIR).

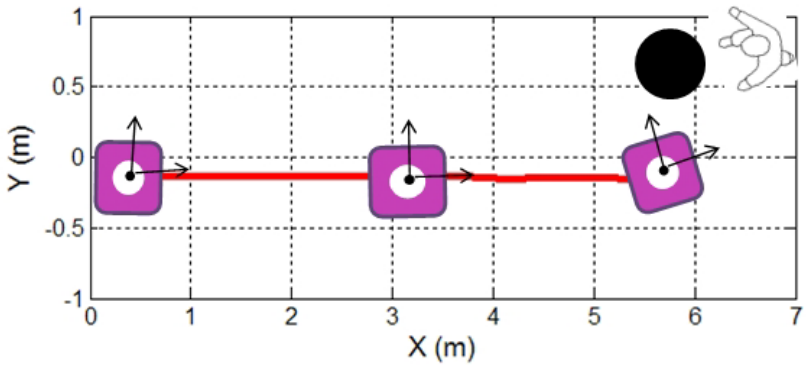

Figure 15 Robot trajectories during the research of the person and using the laser+camera detector.

5. Human Following

Purpose of function: The person following for HRI is a central function, even essential, because it carries out the robot ability to move autonomously in order to bring all its functionalities at the service of the person.

Method: This function is treated with the same framework architecture explained for the human searching function, but it involves the alternation of 3 strategies: go close to the person, stay connected to him/her while he/she is moving (with "predation" trajectories) and search him/her near the location where the person has been detected the most recently.

Living space test at TUW: The first experiment was carried out in a closed area, to avoid disturbances and to check the general behavior in a quite simple environment. The test room was prepared to be robot safe (e.g. avoid problems with door sill, open wooden panels, a cable channels,) and is shown in Figure 16:

Figure 16 : 360° picture of the test room.

The test showed some general problems to be avoided during further tests. E.g. the clearance between door and wall was detected as a person (see Figure 17 left) and was patched (see Figure 17 right):

Figure 17 Clearance between door frame and door leaf

The second experiment was carried out in the environment shown in Figure 18 (left) and by using the laser based detector. The robot stands at starting position and is turned on. At the start of the experiment, no person is in the room, thus the robot starts doing a patrol of the environment. Then, a person appears (from several entry points in the room) and starts to walk around the room (purple circle). The robot should detect the person (*target person*) and follow him/her. From time to time another person (*perturbation person*) walks through the room passing between the robot and the followed person (the green and blue lines show the path of other people traversing the scene) in order to introduce perturbations. Robot must not be distracted from following the *target person*. Two *target persons* were followed by the robot for a total time of 2 hours each. During each hour several *perturbation persons* walked through the room 12 times (some of them unintentional). The speed of the *target person* was about 0.6 m/s. If the *target person* walked on a circular path and no disturbance occurs, the robot followed him/her in 10 out of 10 cases without a problem. If a *perturbation person* crossed and distracted the robot walking more quickly than the *target person*, the robot continued to follow the target 8 out of 10 times. If the *perturbation person* crossed the robot walking at about the same speed of the *target person*, the robot was distracted and switched the target 9 out of 10 times. That happened because when the *perturbation person* walked enough slowly between the robot and the target, the robot was forced to slow down and it lost the target detection for too long. Since in the tested configuration the trajectories of the *target* and the *perturbation person* were enough close, the EKF accepted the new detection (the *perturbation person*) as target. This is a compromise choice between the flexibility and the precision of the EKF.

Figure 18 Pictures of test environments in TUW. Left: second test environment. Right: third test environment.

The third test session was done in a hallway (Figure 16 right) by first walking up and

down the hallway (pink path). Then a disturbing walking by people was added (green path). The procedure was repeated by 2 different target persons, 10 times each. At each time, 1 or 2 *perturbation person* disturbed along the green path walking at different speeds. The robot was never distracted by the people on the green path. This proves the prediction system robustness in the situations where the disturbances occur close to the target person, but not between the target persons and the robot.

Laboratory test at ISIR: This experiment was conducted in the environment of Figure 19 and shows the anticipating skill of the robot. For this test the robot detected the person to follow by using the laser based detector. The robot started moving toward the person until the loss of detection, when an obstacle occulted the person. Thus, the strategy of local search was activated and the robot moved to reach the predicted person position. In this case the collision-free generated path was a shortcut (Figure 20). This robot behavior only occurs under particular conditions: the robot as to be at a certain distance from the person (not too close) and the path of the shortcut has to be shorter than the path which should be covered skirting the obstacle.

Figure 19 Picture of a laboratory test environment (ISIR).

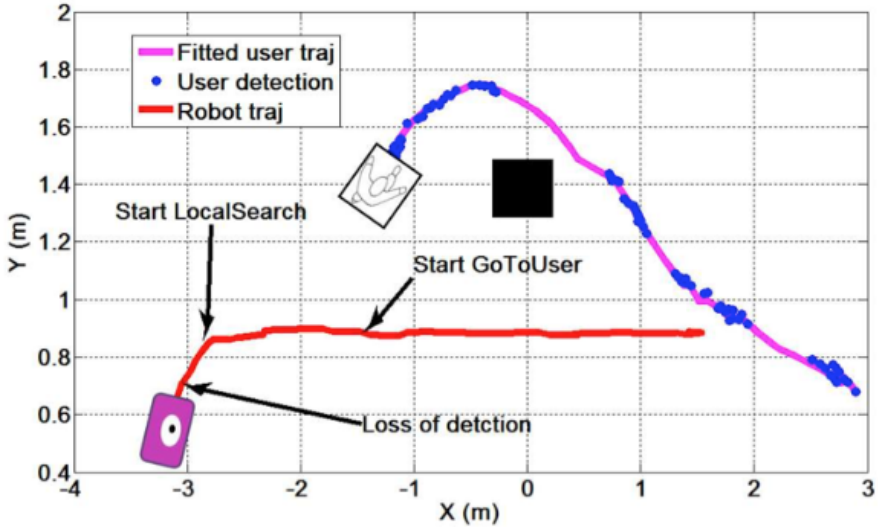

Figure 20 The robot takes a shortcut to reach the person.

Living space test at TUW (1): The same experiment described above was conducted using the laser+camera based detector. This time, the person following function has proved no stable enough to be evaluated in a quantitative way. Anyway, that has allowed us to classify different issues that must be fixed concerning the developed visual detector. First of all, the frequency of detection on the embedded tablet PC was too low (around one detection per second) due to the computing power of the tablet PC. This caused a lack of robustness due to false detections that were too large with respect to the positive detections. The frequency should be increased if the algorithms were implemented on another PC than the tablet or by using a more powerful tablet. Moreover, the field of view of the used camera was low. When the person was in motion as well as the robot, the duration of the presence of the person in the field of the sensor was very low. Considering of the low frequency of treatment, the number of detections remained low.

We could account for the low frequency of detection in SpirOps (by changing the parameters of fuzzy rules). However, strength gains are not obvious.
In order to have better condition for visual detection in far field it was necessary that the movement of the person relative to the robot was adapted to provide some continuing presence in the field of the camera.

To increase the performance of detection in the far field, several solutions are possible:

1) The first one is to consider the low frequency in SpirOps.
2) The second would be to replace the existing treatment by a more powerful PC.
3) A third source of improvement would be to use a camera with an optical high field.

Then, the solution we are considering is the development of a multimodal sensor that combines information from various sensors installed at various focal lengths of the robot:

- Kinect for body detection (using the camera for the detection of RGV in body image and infrared cameras for depth information)
- Kinect combined with laser sensor to detect sound source (the Kinect is used to locate the direction of the source and the laser to know the distance to the robot).
- Webcam for face detection (see following paragraph).

The merger detections from different sensors are made on the basis of a spatial field that considers the detection and performance of each sensor. The implementation of these detectors requires a modification of the platform to install Kinect sensor but also an additional computer with the Ubuntu OS. This OS is required because the new detection modules have been developed on ROS (Robot Operating System), which has as a prerequisite to run on Linux (the version for Microsoft is still unstable). The evaluation of the person following function by using this new multimodal detector will continue in the next few months.

6. Face Detection

Purpose of function: Locking for a face in front of the robot for detection of human engagement during an interaction.

Method: For this function, we exploit the images captured by the robot webcam by using an algorithm based on 2 steps. First, we detect the faces on each frame through the Haar features (by using the OpenCV library). This classifier is frequently used in the domain of object detection and recognition.

Then, we use a filter with the aim to track robustly the detected face. This tracking is very important, because the person can move and the robot needs to be able to follow him/her while being sure it is tracking the same person. The chosen tracking method is a particle filter. Our particle filter is based on the particle filter developed by Kyle Brocklehurst[4].

This filter is a variant of sequential importance sampling (SIS). In an SIS, at each interaction, a set of points is generated. These points are moved randomly and their new weights are computed. In a sampling method based on importance resampling (SIR), the points with a weight below a threshold are resampled.

4 Kyle Brocklehurst, Project for Penn State CSE 598 - Object Tracking.

The method developed by Brocklehurst creates recovery points at each interaction (Figure 21). This method is a hybrid between the SIS and the SIR, there is no resampling, but the result is close to a method with resampling, since points are generated at each iteration.

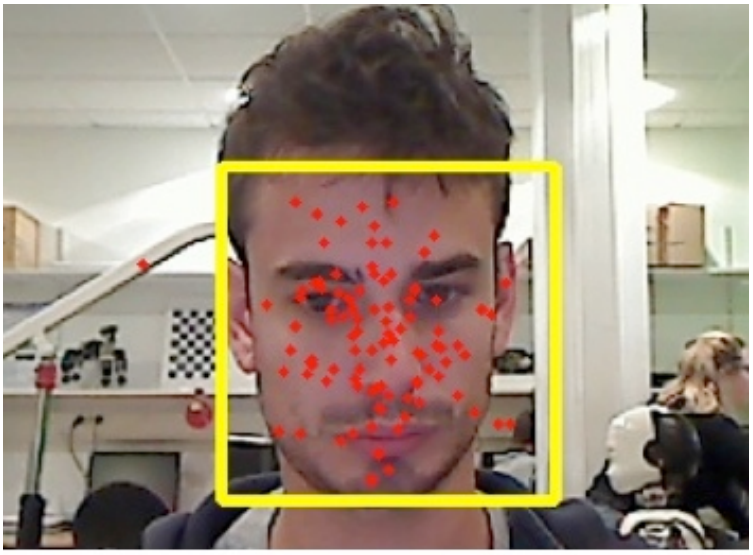

Figure 21 Example of draw of points by particle filter.

The following 4 steps compose the face tracking method:

1. A set of points (we have considered 100 points) with a Gaussian distribution centred on the detected face is drawn.
2. For each point, we compute the weight (from a colour histogram).
3. We consider, as coordinates of the new face, the point with biggest weight.
4. The distance between the old and the new coordinate is used in order to compute the movement and the speeds of the face on the axis x and y.

Laboratory test at ISIR: The disadvantage of using the OpenCV library is that the classifier was trained on images of frontal faces. The classifier does not detect faces whose rotation is more than 60 degrees with respect to the camera. Moreover, the classifier detects false positives when an object has the same mathematical properties that a face (Figure 22).

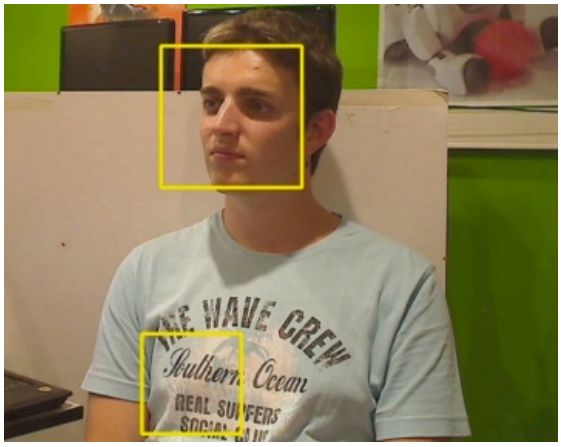

Figure 22Example of false positive detection.

The particle filter for face tracking is very efficient. The only limitation is in the Haar classifier trained with frontal faces. The algorithm does not detect profile persons, but tracks very well frontal faces.

7. Speaker Detection

Purpose of function: Speech activity detection with lips movement image signals for speaker detection is achieved for interaction engagement analysis.

Method: The lips activity detection algorithm correlates lips movement visual information acquired via a camera with speech audio information acquired via a microphone from a human speaker in order to locate the speaker and if required move the robot toward him/her.

The process used for speaker detection is shown in Figure 23.

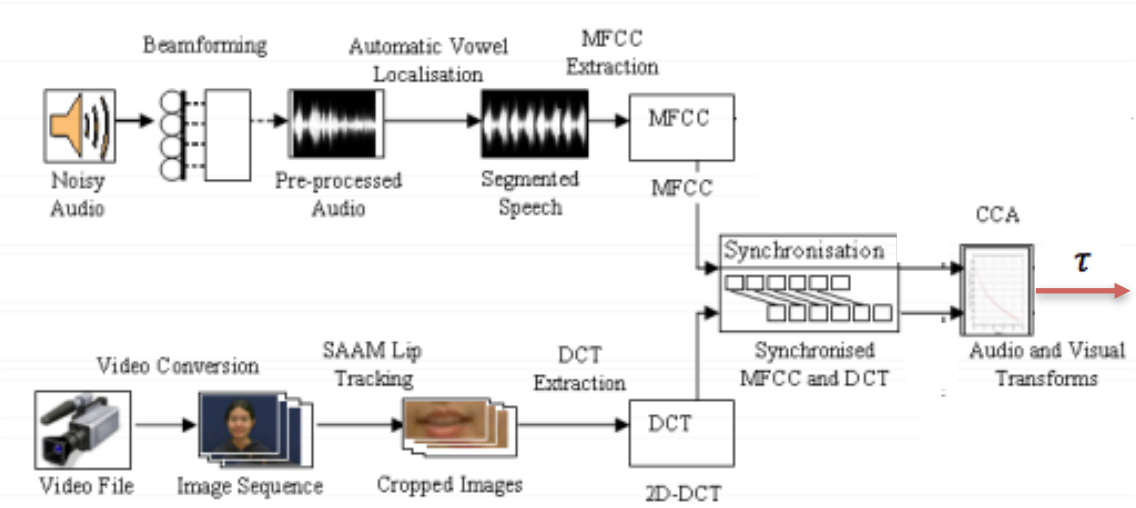

Figure 23 Proposed multimodal speech enhancement system.

The vision-based speaker detection involves a mouth detector. In order to facilitate the detection, the search zone is limited to the lower face, which allows to lighten the computation and to avoid false positives. Visual lips features are extracted by using a Semi Adaptive Appearance Models (SAAMs)[5]. Lips tracking essentially deals with non-stationary data, as the appearance of a target object may alter drastically over time due to factors like pose variation and illumination changes. First, we use the Adaptive Appearance Models (AAMs) to update the mean and Eigen vectors of d-dimensional observation vectors. Then, we extend the AAMs by inserting a supervisor model[6] that satisfies AAM performance at each frame, by using a Support Vector Machine (SVM) to filter the AAM result for an individual frame. Finally, shape models are constructed to allow SAAM to track feature points in video sequences. The tracking algorithm is based on the maximization of the following cost function:

$$p^* = \arg\max_p(d_e)$$

Where d_e is a negative exponential of projection error between x_t (trained vector) and the Principal Component Analysis (PCA) subspace created by earlier observations. By using this technique we find the 2D-DCT vector f_x. The first 30 2D-DCT components of each image are vectorized in a zigzag order to produce the vector for a single frame in an image sequence.

For audio features extraction, we have developed a system both speaker and language independent based on vowel detection method described in[7]. The detection of vowel like segments is done by analyzing the characterization of the spectral envelope. The "Reduce Energy Cumulating" (REC) is used as measure for vowel spectrum characterization by comparing the energy computed from Mel bank filters. For a given sentence, peak detection on the smoothed REC curve (by introducing a Simple Moving Average filter) allows vocalic nucleus detection. To reject low energy peaks, which can be due both to spectral noises and low energy vowels, only those higher than the half mean of the REC values are considered. If two vowels are detected closer than 150 ms, only the single highest peak is kept. 22 Mel Frequency Cepstral Coefficients (MFCC) are computed on each frame, producing MFCC matrices for full sentences. To extract vowel only data, vowel localization results are used to group vowel only segments together into a single MFCC file, providing the feature vector f_y.

5 Q.D. Nguyen, M. Milgram. Semi Adaptive Appearance Models For Lip Tracking. In Proc. International Conference on Image Processing (ICIP) 2009, pp 2437-2440.

6 G. H. Golub, C. F. Van Loan. Matrix Computations (3rd Edition). John Hopkins Uni. Press (1996)

7 F. Ringeval, M. Chetouani. A Vowel Based Approach For Acted Emotion Recognition. In Proc. Interspeech, 2008, pp. 2763–2766 (2008).

We use a CCA for the analysis of the relationships between multidimensional audio and visual speech variables. The CCA has been preferred to the others correlation analysis for the independence of the analysis from the coordinate system describing the variables. Taking f_x and f_y, respectively the multidimensional visual and audio signal variables, we use the QR Decomposition method to calculate the projection matrices u_x and u_y, that mutually maximizes the projections of f_y and f_x onto their respective basis vectors.

We consider the linear combinations.

$$\hat{f}_x = u_x^T f_x \text{ and } \hat{f}_y = u_y^T f_y$$

The aim is to maximize ρ defined as follows:

$$\rho = \frac{E\left[\hat{f}_x \hat{f}_y^T\right]}{\sqrt{E\left[\hat{f}_x \hat{f}_x^T\right] E\left[\hat{f}_y \hat{f}_y^T\right]}}$$

We have defined the correlation measure as follows:

$$\tau = \sum_{i=1}^{N} \lambda^2$$

Where N represents the number of canonical correlations found and

$$\lambda_x = \lambda_y^{-1} = \sqrt{\frac{u_y^T C_{yy} u_y}{u_x^T C_{xx} u_x}}$$

Note that C_{xx} and C_{yy} are the elements on the diagonal of the covariance block matrix computed as follow:

$$C = \begin{bmatrix} C_{xx} & C_{xy} \\ C_{yx} & C_{yy} \end{bmatrix} = E\left[\begin{bmatrix} f_x \\ f_y \end{bmatrix} \quad \begin{bmatrix} f_x \\ f_y \end{bmatrix}^T\right]$$

Laboratory test at ISIR: The performance of the proposed multimodal algorithm was assessed on speech from the VidTIMIT Corpus[8]. VidTIMIT contains a number of image sequences of sentences recorded at 25 fps. Sentences are segmented into 32 ms frames with an overlap ratio equal to 50%. Initially, synchrony between audio and visual signals was assessed. In the literature, we have found that the maximum audiovisual correlation is with an audio delay of 40 ms [9]. To confirm this, CCA was applied to a 24 sentences dataset from VidTIMIT. τ was taken when shifting the visual data in relation to the equivalent audio data. The mean synchronization results are shown in Figure 24 (Left), confirming that audiovisual correlation is maximized when there is a small degree of asynchrony, in line with results found by Sargin et al. [9].

8 C. Sanderson. Biometric Person Recognition: Face, Speech and Fusion. VDM-Verlag (2008).

9 M. E. Sargin, Y. Yemez, E. Erzin, A. M. Tekalp. Audiovisual Synchronization and Fusion Using Canonical Correlation Analysis. Mult., IEEE Trans. on, vol. 9, no. 7, pp. 1396–1403 (2007).

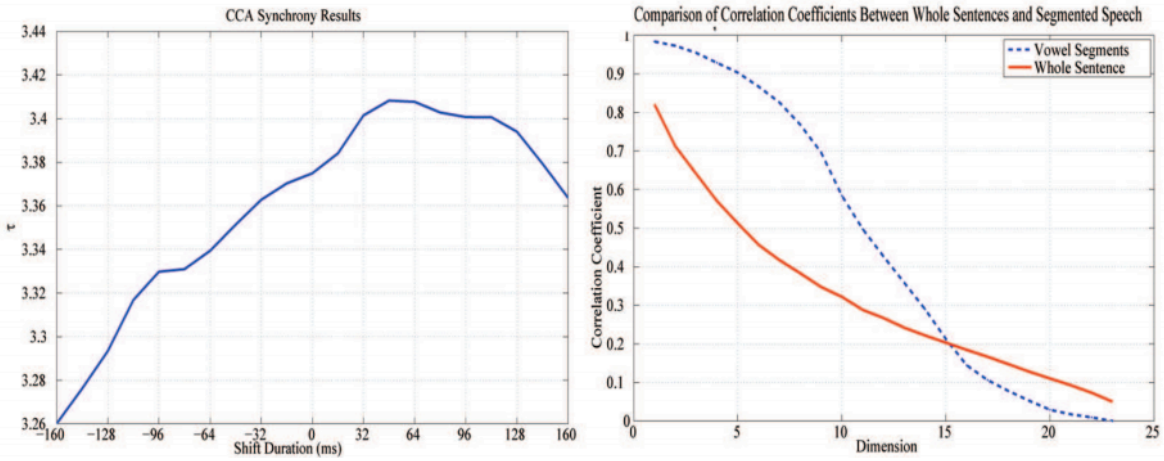

Figure 24 Left: CCA feature synchrony results. Right: Comparison of canonical correlation of complete sentences and vowel segments.

In Figure 24 (Right) the mean canonical correlations of sentences and segments are shown. The solid line indicates the mean canonical correlation coefficients for whole sentences, while the dashed line represents mean coefficients for vowel segments only. This proves a clear difference in correlation values and behavior, with segmented speech producing a significantly higher multimodal correlation, as it can be seen from Figure 24 (Left) that vowel segments produce much stronger initial canonical correlations. This is demonstrated by comparing the squared sum of canonical correlations (τ) for sentences and segments, with results of 3.41 and 8.48 respectively, confirming that speech segmentation significantly increases correlation.

In order to evaluate the performances of the proposed method for the speaker detection function, we have realized the following experience: three speakers (positioned in a triangle) talked about a topic. Three cameras filmed the scene, each one oriented to an interlocutor (Figure 25). The cameras used were webcams. Two of them were 2 million pixels cameras, the other one was just 1 million pixels. With this last camera it was possible to detected faces, but its resolution was not sufficient for lips features extraction.

Figure 25 The scene of the speaker detection experience.

The speech detection algorithm showed good performances. When a person was speaking, the probability of detection in a single frame was around of 50% (depending on the threshold chosen for the canonical correlation coefficient τ). When the person does not speak, the probability of detecting a false positive per frame was around 5%. This false positive detection often occurred when a person moved his head.

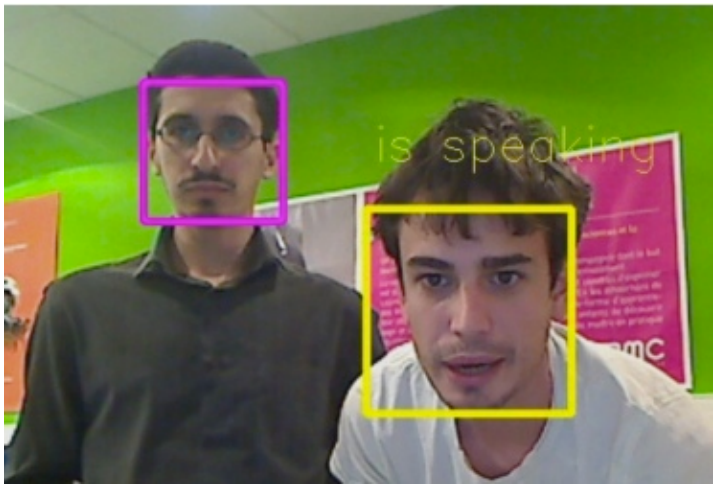

Figure 26 Speaker detection in presence of several persons.

8. Human Focalization

Purpose of function: Registration of the image coordinates given by the face detector in the laser plane and movement of the platform to focus on the detected person.

Method: The implementation of this function has been done by using SpirOps software. The final decision about the robot movement depends on the position of the detected face in the image (that gives the angular position of the person in the robot frame) and from the laser information relative to the face detection zone (that gives the distance of the person to the robot). By combining these two informations, the decisional engine is able to compute the robot interests to turn and to go forward in order to focus on the detected face. The robot linear and angular speeds are chosen proportionally to the computed interests.

Laboratory test at ISIR: The human focalization function has been tested in the ISIR premises. See the attached video for a demonstration.

9. Publications

Granata, C. and Bidaud, Ph. and Chetouani, M. and Melchior, N. (2012). Multimodal human detection and fuzzy decisional engine for interactive behaviors of a mobile robot, Proc. of 3rd IEEE Int. Conférence on Cognitive Info Communications.

Granata, C. and Melchior, N. and Bidaud, Ph. and Chetouani, M. (2012). Robust multimodal interactions with a mobile robot, 5th International Workshop on Human-Friendly Robotics (HFR 2012) October 18th-19th. Brussel.

Granata, C. and Bidaud, Ph. A framework for the design of person following behaviors for social mobile robots. In Proc of the International Conference on Intelligent Robots and Systems (IROS), 2012 IEEE/RSJ .

Granata, C., Bidaud, Ph., Beck, C. and Mayer, P. Experimental analysis of interactive behaviors for a personal mobile robot. In Proc of the 15th International Conference on Climbing and Walking Robots and the Support Technologies for Mobile Machines (CLAWAR2012). Pages 456-468. 2012.

Granata, C. Contribution to the design of interfaces and interactive behaviors for personal robots. 4 place jussieu, 75005 PARIS. These. Université Pierre et Marie Curie, Paris 6. 2012.

Wang, X. and Clady, X. and Granata, C. A Human Detection System for Proxemics Interaction, ACM/IEEE International Conference on Human-Robot Interaction (Late Breaking Results). Lausanne, Switzerland. 2011.

Granata, C. and Bidaud, Ph. Interactive person following for social robots. In Proc. of CLAWAR 2011, 11th International Conference on Climbing and Walking Robots and the Support Technologies for Mobile Machines, Paris, World Scientific, publisher. Pages 11-26. 2011.

Wang, X. and Granata, C. Caractérisation selon les proxémies d'un système de détection de personnes pour l'interaction homme-robot. Journées Jeunes Chercheurs En Robotique 2010.

Wang, X. Un système embarqué de détection de personnes considérant explicitement la distance Homme-Robot, Thèse de doctorat de l'Université Pierre et Marie Curie, Novembre 2012.