Data fusion and eigenface based tracking dedicated to a Tour-Guide Robot

Thierry Germa[†], Ludovic Brèthes[†], Frédéric Lerasle[†‡], Thierry Simon[†¶]

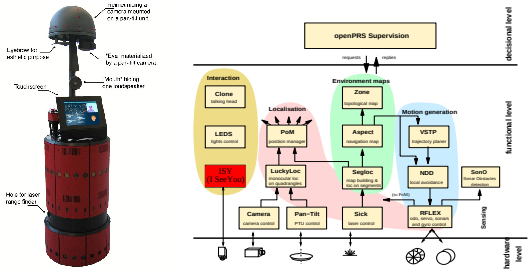
[†] LAAS-CNRS, 7 avenue du Colonel Roche, 31077 Toulouse Cedex 4, France
[‡] Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse Cedex, France
[¶] IUT Figeac, avenue de Nayrac, 46100 Figeac, France

Abstract This article presents a key-scenario of H/R interaction for our tour-guide robot. Given this scenario, three visual modalities, the robot deals with, have been outlined, namely the "search of visitors" attending the exhibition, the "proximal interaction" through the robot interface and the "guidance mission". The paper focuses on the two last ones which involves face recognition and visual data fusion in a particle filtering framework. Evaluations on key-sequences in a human centred environment show the tracker robustness to background clutter, sporadic occlusions and group of persons. The tracker is able to cope with target loss by detecting and re-initializing automatically thanks to the face recognition outcome. Moreover, the multi-cues association proved to be more robust to clutter than any of the cues individually.

1 Introduction and framework

The development of autonomous robots acting as human companions is a motivating challenge and a considerable number of mature robotic systems have been implemented which claim to be companions, servants or assistants (see a survey in [5]). The autonomy of such robots are fully oriented towards navigation in human environments and human-robot interaction but few of them exhibit advanced visual capabilities. In this context, we recently developed a tour-guide mobile robot whose role is to help people attending an exhibition, and guide them by proposing either group or personalized tours. Our robot is equipped with one Sony camera EVI-D70, one digital camera mounted on a Directed Perception pan-tilt unit, one ELO touch-screen, a pair of loudspeakers, an optical fiber gyroscope and wireless Ethernet (Figure 1(a)).

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de



(a) (b)

Figure 1. Our mobile robot Rackham and its software architecture.

Its software architecture (Figure 1(b)) includes a supervisor which controls a distributed set of functional modules relative to perception, decision, action and interface issues. The paper focuses on the module called ISY [1] which integrates visual modalities involving detection, recognition and tracking of persons.

Embedding visual modalities on a mobile robot dedicated to human centred environments impose several constraints. First, on board processing power is limited and care must be taken to design efficient algorithms dedicated to recognition/tracking of persons. As the robot's evolution takes place into cluttered environments subjected to illumination changes, several hypotheses must be handled at each instant concerning the parameters to be estimated in trackers. Robust integration of multiple informations is required to cope with the environment and to keep locking on the targeted person throughout the interaction session.

The literature proposes a plethora of approaches dedicated to face detection and recognition [7]. Techniques can be effectively organized into two broad categories, feature-based approaches and image-based approaches which uses training algorithms without feature derivation and analysis[2]. Regarding 2D tracking, many paradigms involving a single camera have been proposed in the literature which we shall not attempt to review here [12]. Particle filters [4] constitute one of the most powerful framework for tracking purpose. Their popularity stems from their simplicity, modeling flexibility (over a wide variety of applications), and ease of combination/fusion of diverse kinds of measurements. Nevertheless, it can be argued that data fusion using particle filters has been fairly seldom exploited in the robotics context, for it has often been onfined to a restricted number of visual cues [9].

The paper is organized as follows. Section 2 briefly sums up the well-known particle filtering formalism and principles to fuse several cues in the filters. Section 3 presents the key-scenario and the on-board visual modalities. Sections 4 and 5 focus on two modalities involving image-based face recognition and particle filtering based person tracking. Experiments and results are presented and discussed for each modality. Last, section 6 summarizes our contribution and discuss future extensions.

[1] Details regarding the other modules can be found in [2].

[2] and so requires *a priori* less on-line computational load.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

$[\{x_k^{(i)}, w_k^{(i)}\}]_{i=1}^N = \mathrm{SIR}([\{x_{k-1}^{(i)}, w_{k-1}^{(i)},\}]_{i=1}^N, z_k)$

1: **IF** $k = 0$**, THEN** Draw $x_0^{(1)}, \ldots, x_0^{(i)}, \ldots, x_0^{(N)}$ i.i.d. according to $p(x_0)$, and set $w_0^{(i)} = \frac{1}{N}$ **END IF**

2: **IF** $k \geq 1$ **THEN** $\{$—$[\{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}]_{i=1}^N$ being a particle description of $p(x_{k-1}|z_{1:k-1})$—$\}$

3: **FOR** $i = 1, \ldots, N$, **DO**

4: "Propagate" the particle $x_{k-1}^{(i)}$ by independently sampling $x_k^{(i)} \sim q(x_k|x_{k-1}^{(i)}, z_k)$

5: Update the weight $w_k^{(i)}$ associated to $x_k^{(i)}$ according to the formula $w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(z_k|x_k^{(i)})p(x_k^{(i)}|x_{k-1}^{(i)})}{q(x_k^{(i)}|x_{k-1}^{(i)}, z_k)}$, prior to a normalization step so that $\sum_i w_k^{(i)} = 1$

6: **END FOR**

7: Compute the conditional mean of any function of x_k, *e.g.* the MMSE estimate $\mathrm{E}_{p(x_k|z_{1:k})}[x_k]$, from the approximation $\sum_{i=1}^N w_k^{(i)} \delta(x_k - x_k^{(i)})$ of the posterior $p(x_k|z_{1:k})$

8: At any time or depending on an "efficiency" criterion, resample the description $[\{x_k^{(i)}, w_k^{(i)}\}]_{i=1}^N$ of $p(x_k|z_{1:k})$ into the equivalent evenly weighted particles set $[\{x_k^{(s^{(i)})}, \frac{1}{N}\}]_{i=1}^N$, by sampling in $\{1, \ldots, N\}$ the indexes $s^{(1)}, \ldots, s^{(N)}$ according to $P(s^{(i)} = j) = w_k^{(j)}$; set $x_k^{(i)}$ and $w_k^{(i)}$ with $x_k^{(s^{(i)})}$ and $\frac{1}{N}$

9: **END IF**

Table 1. Generic particle filtering algorithm (SIR)

2 Particle filtering algorithms for data fusion

2.1 Generic algorithm

The aim is to recursively estimate the posterior density function of the state vector $\mathbf{x}_k$ at time k conditioned on the knowledge of past measurements [1]. The key idea is to represent this probability density function (pdf) by a set of Gaussian random samples with associated weights and to compute estimates based on these samples and weights. Let $z1 : k = z_1, \ldots, z_k$ term the available measurements from time 1 to k. At each time k, the density $p(\mathbf{x}_k|z1 : k)$ is depicted by a set of particles $\mathbf{x}_k^{(i)}$ - which are samples of the state vector- affected by weights $w_k^{(i)}$. The idea is to get $p(\mathbf{x}_k|z_{1:k}) \approx \sum_i w_k^{(i)} \delta(\mathbf{x}_k - \mathbf{x}_k^{(i)})$ *i.e.* to approximate random sampling from the pdf $p(\mathbf{x}_k|z_{1:k})$ by the selection of a particle with a probability equal to its associated weight.

A generic particle filter or SIR is shown on Table (1). The particles $\mathbf{x}_k^{(i)}$ evolve stochastically over the time, being sampled from an importance density $q(.)$ which aims at adaptively exploring "relevant" areas of the state space. Their weights $w_k^{(i)}$ are updated thanks to $p(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)})$ and $p(z_k|\mathbf{x}_k^{(i)})$, resp. the state dynamics and measurement functions, so as to guarantee the consistency of the above approximation. In order to limit the degeneracy phenomenon, which says that after few instants all but one particle weights tend to zero, step 8 inserts a resampling process. Another solution to limit this effect in addition to re-sampling, is the choice of a good importance density.

The CONDENSATION algorithm is instanced from the SIR algorithm as $q(\mathbf{x}_k|\mathbf{x}_{k-1}^{(i)}, z_k) = (\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)})$. Another difference is that the re-sampling step 8 is applied on every cycle. Resampling by itself cannot efficiently limit the degeneracy phenomenon as the state-space is blindly explored without any knowledge of the observations. On the other side, the ICONDENSATION algorithm [8], consider importance density (.) which classically relates to importance function $\pi(\mathbf{x}_k^{(i)}|z_k)$ defined from the current image.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

However, if a particle drawn exclusively from the image is inconsistent with its predecessor from the point of view of the state dynamics, the update formula leads to a small weight. An alternative consists in sampling the particles according the measurements, dynamics and the prior so that, with $\alpha, \beta \in [0; 1]$

$$q(\mathbf{x}_k^{(i)}|\mathbf{x}_{k-1}^{(i)}, z_k) = \alpha\pi(\mathbf{x}_k^{(i)}|z_k) + \beta p(\mathbf{x}_k|\mathbf{x}_{k-1}^{(i)}) + (1 - \alpha - \beta)p_0(\mathbf{x}_k). \quad (1)$$

3 Key-scenario and on-board visual modalities

Three visual modalities have been outlined which the robot must deal with (Figure 2):

1. **the "search for interaction",** (Figure 2(a)) where the robot, static and left alone, visually tracks visitors thanks to the camera mounted on its helmet, in order to heckle them when they enter the exhibition.
2. **the "proximal interaction",** (Figure 2(b)) where the interlocutors select the area to visit through the ELO touch-screen. Here, the robot remains static and possibly learns their faces thanks to the camera materializing its eye.
3. **the "guidance mission",** (Figure 2(c)) where the robot drives the lonely visitor or the group to the selected area, keeping the visual contact with any memberof the guided group even if som of them can move away. The modality must enable automatic target recovery when the tracker fails or/and the targeted person reappears after an occlusion or out-of-sight.

(a) (b) (c)

Figure 2. The three visual modalities of the Rackham robot : (a) search for interaction, (b) proximal interaction, (c) guidance mission.

We focus in this paper on the two last EVI-D70 camera based modalities which involve persons detection/recognition and tracking at mid-range H/R distances.

4 User recognition dedicated to the "proximal interaction"

This modality aims to classify facial regions $\mathcal{F}$, segmented from the input image, into either one class C_t out of the set $\{C_t\}_{1 \leq t \leq M}$ of M tutors faces using training algorithms. For detecting faces, we apply the well known window scanning technique introduced by Viola *et al.* [11] which covers a range of $\pm 45°$ out-of plane rotation. The

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

bounding boxes of faces segmented by the Viola's detector are then fed to the recognition process. For each class C_t, we perform PCA and keep as eigenface bases the first $N_{B(C_t)}$ eigenvectors accounting on average for η of the total class variance. The basis is noted for the next $B(C_t) = \{B_k(C_t), k \in \{1, ..., N_{B(C_t)}\}\}$. Let $\mathcal{F}(j)$ be an input image, written as $nm \times 1$ vector, to be compared to the eigenfaces of a given class C_t. Recall that eigenspace method constructs an approximation $\mathcal{F}_{r,t}$ from the input face $\mathcal{F}$ by projecting it onto basis $B(C_t)$. $\mathcal{F}$ is linked to the class C_t by its error norm

$$\mathscr{D}(C_t|\mathcal{F}) = \frac{1}{n \times m} \sum_{j=1}^{n \times m} ((\mathcal{F}(j) - \mathcal{F}_{r,t}(j)) - \mu)^2,$$

where $\mathcal{F} - \mathcal{F}_{r,t}$ is the difference image, given that $|\mathcal{F} - \mathcal{F}_{r,t}|$ terms the DFFS[3], and μ the mean of $\mathcal{F} - \mathcal{F}_{r,t}$, and its associated likelihood

$$\mathscr{L}(C_t|\mathcal{F}) = \mathcal{N}(\mathscr{D}(C_t|\mathcal{F}); 0, \sigma_t)$$

where σ_t terms the standard deviation of distances of $B(C_t)$ training set and $\mathcal{N}(.; \mu, \sigma)$ is a Gaussian distribution with mean μ and covariance σ.

The aforementioned likelihood $\mathscr{L}$ have to be thresholded in order to match the input face $\mathcal{F}$ with an already learnt individual C_t. This threshold τ is deduced by computing likelihoods $\mathscr{L}$ between test image database with their own class C_t but also with the other classes noted $\neg C_t$. Let $p(\mathscr{L}|C_t)$ and $p(\mathscr{L}|\neg C_t)$ be the probability densities which are approximated by their corresponding likelihood distributions (Figure 3). To specify an optimal threshold τ, we minimize:

$$S(\tau) = \lambda \underbrace{\int_0^{\tau} p(\mathscr{L}|C_t)\mathrm{d}\mathscr{L}}_{S_t(\tau)} + \gamma \underbrace{\int_{\tau}^{\infty} p(\mathscr{L}|\neg C_t)\mathrm{d}\mathscr{L}}_{\neg S_t(\tau)}$$

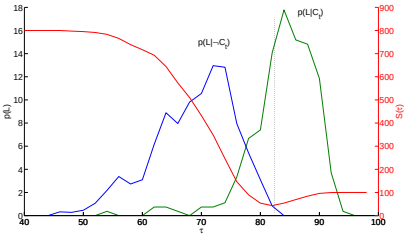


Figure 3.
Graph $S(\tau)$ for threshold determination.

where the weights λ and γ respectively balance the false rejection $S_t(\tau)$ and false acceptance $\neg S_t(\tau)$ results from test image database acquired by the robot in a wide range of typical conditions: illumination changes, variations in facial orientation and expression. The choice $\gamma = \frac{1}{4}\lambda$ allows to give more importance to false acceptances than to false rejections as false acceptances cannot be accepted and so are more important to be avoided when determining the threshold τ.

Besides, Heseltine *et al.* in [6] have outlined a range of image preprocessing techniques which improve the recognition accuracy. We now continue this line of investigation by applying some pertinent pre-processing techniques prior to training and testing each face image of the database. Our face database is composed of 6000 examples of $M = 10$ individuals acquired by the robot in a wide range of typical conditions: illumination changes, variations both in out-of plane rotation ($\pm 45°$) and expression, etc.

[3] Distance From Face Space

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

Distance	Preproc.	FAR	Sensitivity	η
Euclidean	None	4.38%	4.46%	0.40
	Equal.	5.22%	6.40%	0.80
	S+C	4.58%	7.52%	0.90
DFFS	None	3.17%	18.44%	0.35
	Equal.	1.50%	41.28%	0.90
	S+C	2.45%	10.40%	0.35
Error Norm	None	1.92%	19.44%	0.35
	Equal.	0.95%	48.08%	0.70
	S+C	2.03%	10.06%	0.30

Table 2. Analysis of some image preprocessing methods and distance measurement. The first column terms the distance measurement (Euclidean, DFFS, Error Norm) and the second terms the image preprocessing (None, Histogram equalization, Smooth and Contour filter).

The database is separated into two disjoint sets : i) the training set (dedicated to PCA) containing 100 images per class, ii) the test set containing 500 images per class. Each image is cropped to a size of 30×30 pixels.

A crossed evaluation enables the selection of both most meaningful image preprocessing and distance measurement. As shown in Table 4, histogram equalization coupled to our Error Norm are shown to outperform the other techniques for our database. In fact, the sensitivity is increased of 6.8% compared to the DFFS, while the False Acceptance Rate is very low (0.95%).

For a set of M learnt tutors (classes) noted $\{C_t\}_{1 \leq t \leq M}$ and a detected face $\mathcal{F}$, we can define for each class C_t, the likelihood $\mathscr{L}_t = \mathscr{L}(\mathcal{F}, C_t)$ and an *a priori* probability $P(C_t|\mathcal{F})$ of labeling to C_t

$$\begin{cases} P(C_\emptyset|\mathcal{F}) = 1 \text{ and } \forall l \ P(C_t|\mathcal{F}) = 0 \text{ when } \forall l \ \mathscr{L}_t < \tau \\ P(C_\emptyset|\mathcal{F}) = 0 \text{ and } \forall l \ P(C_t|\mathcal{F}) = \frac{\mathscr{L}(C_t|\mathcal{F})}{\sum\limits_p \mathscr{L}(C_p|\mathcal{F})} \text{ otherwise} \end{cases}$$

where $C_\emptyset$ refers the void class.

To know who is present (or not) at time k, the probability of the presence of each class C_t is updated by applying the following recursive Bayesian scheme from the classifier ouputs in the p previous frames, *i.e.*

$$P(C_t|z_{k-p}^k) = \left[1 + \frac{1 - P(C_t|z_k)}{P(C_t|z_k)} . \frac{1 - P(C_t|z_{k-p}^{k-1})}{P(C_t|z_{k-p}^{k-1})} . \frac{p(C_t)}{1 - p(C_t)}\right]^{-1},$$

where

$$p(C_t) = \frac{1}{M}, P(C_t|z_k) = \frac{1}{B}\sum_{j=1}^{B} P(C_t|\mathcal{F}_j),$$

and B is the number of detected faces at time k.

Figure 4 shows some snapshots of recognized faces where the detector marked – in red color – the detected faces but only those in green color are recognized from the previously learnt faces.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

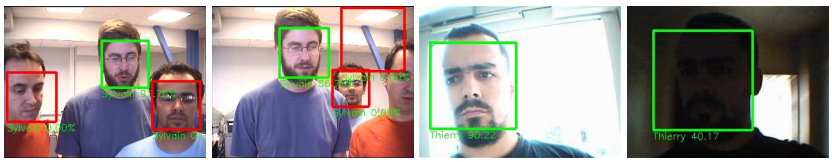


Figure 4. Snapshots of detected/recognized faces with associated probabilities. The target is Sylvain (resp. Thierry) for the two first (resp. last) frames.

5 User tracking dedicated to the "guidance mission"

5.1 Overview

In our human centred environment, more than the beforehand identified person can be in the robot vicinity while the visual modality must lock onto the guided person as soon as he/she enters the scene, and track him/her throughout the guidance mission. This modality involves logically the previous face recognition process as well as the tracking of the targeted person.

This tracker is inspired from previously developed ones detailled in [3]. We aim to estimate at time k the state vector $\mathbf{x}_k = [u_k, v_k, s_k]^{'}$ which is composed of location $\mathbf{x}_k^{'} = [u_k, v_k]^{'}$, and scale of the targeted person. With regard to the dynamics model $p(\mathbf{x}_k|\mathbf{x}_{k-1})$, the image motions of observed people are difficult to characterize over time. The state vector entries are assumed to evolve according to mutually independent random walk models, viz. $p(\mathbf{x}_k|\mathbf{x}_{k-1}) = \mathcal{N}(\mathbf{x}_k; \mathbf{x}_{k-1}, \Sigma)$, where $\mathcal{N}(.; \mu, \Sigma)$ is a Gaussian distribution with mean μ and covariance $\Sigma = \mathrm{diag}(\sigma_u^2 \sigma_v^2, \sigma_s^2)$. Regarding the filtering strategy, we opt for the ICONDENSATION scheme depicted in section 2. Let us characterize both importance and measurement functions involved in the tracker.

5.2 Importance and measurement functions

The importance function $\pi(.)$ in (1) offers a mathematically principled way of directing search according to the face verification process inspired from section 4. Taking into account the face recognition, and the probability $P(C_l|\mathcal{F}_j)$ of each face $\{\mathcal{F}_j\}_{1 \leq j \leq B}$, the importance function becomes, with B the number of detected faces and $p_j = (u_j, v_j)$ the centroid coordinate of each face

$$\pi(\mathbf{x}_k|z_k) = \sum_{j=1}^{B} P(C_l|\mathcal{F}_j).\mathcal{N}(\mathbf{x}; p_i, diag(\sigma_{u_i}^2, \sigma_{v_i}^2))$$

Let us characterize the measurement function. We consider multi-patches of distinct color distribution related to the head and the torso[4], each with its own N_{bi}-bin normalized reference histograms models in channel c (annoted resp. $h_{ref,1}^c$, $h_{ref,2}^c$ for the next). Let the union $B_x = \bigcup_{p=1}^{2} B_{x,p}$ for any state $\mathbf{x}_k$ be associated with the set of

[4] of the guided person.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

reference histograms $\{h^c_{ref,p} : c \in \{R,G,B\}, p = 1,2\}$. By assuming conditional independence of the color measurements, the likelihood $p(z^c_k|\mathbf{x}_k)$ becomes

$$p(z^c_k|\mathbf{x}_k) \propto \exp\left(-\sum_c \sum_{p=1}^{2} \frac{D^2(h^c_{x,p}, h^c_{ref,p})}{2\sigma^2_c}\right), \tag{2}$$

provided that D terms the Bhattacharyya distance [10] and σ_c the standard deviation being determined *a priori*. This multi-part extension is more accurate thus avoiding the drift, and possible subsequent loss, experienced sometimes by the single-part version. To overcome the ROIs' appearance changes in the video stream, the target reference models are updated at time k from the computed estimates through a first-order filtering process *i.e.*

$$h^c_{ref,k} = (1-\kappa).h^c_{ref,k-1} + \kappa.h^c_{E[\mathbf{x}_k]}, \tag{3}$$

where κ weights the contribution of the mean state histogram $h^c_{E[\mathbf{x}_k]}$ to the target model $h^c_{ref,k-1}$ and index p has been omitted for compactness reasons. This models updating can lead to drifts with the consequent loss of the target. To avoid such tracker failures, we also consider a shape-based likelihood $p(z^s_k|\mathbf{x}_k)$ that depends on the sum of the squared distances between N_p points uniformly distributed along a head silhouette template corresponding to $\mathbf{x}_k$ and their nearest image edges *i.e.* the shape-based likelihood is given by

$$p(z^s_k|\mathbf{x}_k) \propto exp\left(-\frac{D^2}{2\sigma^2_s}\right), D = \sum_{l=0}^{N_p} |x(l) - z(l)|,$$

where l indexes the N_p template point $x(l)$ and associated closest edge $z(l)$ in the image. Finally, assuming the cues to be mutually independant, the unified measurement function can then be formulated as $p(z^s_k, z^c_k|\mathbf{x}_k) = p(z^s_k|\mathbf{x}_k).p(z^c_k|\mathbf{x}_k)$.

5.3 Implementation and evaluations

The initializations of the histograms $h^c_{ref,1}, h^c_{ref,2}$ are achieved[5] according to frames which lead to $P(C_l|\mathcal{F})$ probabilities egal to one (Figure 5). In the tracking loop, the histogram model $h^c_{ref,2}$ (torso) is re-initialized with the previous values when the user verification is highly confident, typically $P(C_l|\mathcal{F}_j) = 1$. Numerical values for the dynamical and measurement parameters used in our tracker are given in Table 3.

Figure 5. The template.

Figure 6 and Figure 7 shows snapshots of two typical sequences in our context. All regions – centred on the yellow dots – close to detected faces with high recognition probabilities corresponding to the person on the foreground are continually explored. Those – in blue color – that do not comply with the targeted person are discarded

[5] during "proximal interaction".

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

Symbol	Meaning	Value
(α, β)	coeff. in the importance function $q(x_k \vert x_{k-1}, z_k)$	$(0.3, 0.6)$
$(\sigma_u, \sigma_v, \sigma_s)$	standard deviation in random walk models	$(10, 4, \sqrt{0.1})$
σ_s	standard deviation in shape-based likelihood $p(z^s \vert \mathbf{x}_k)$	28
σ_c	standard deviation in color-based likelihood $p(z^c \vert \mathbf{x}_k)$	0.03
N_{bi}	number of color bins per channel involved in $p(z^c \vert \mathbf{x}_k)$	32
κ	coeff. for reference histograms $h^c_{ref,1}, h^c_{ref,2}$ update in (3)	0.1

Table 3. Parameter values used in our upper human body tracker.

during the importance sampling step. Recall that, for large range out-of-plane face rotations ($> |45°|$), the proposal continues to generate pertinent hypotheses from the dynamic and the skin blobs detector. The green (resp. red) rectangles represent the MMSE estimate in step 7 of Table 1 with high (resp. low) confidence in the face recognition process. The proposal generates hypotheses (yellow dots) in regions of significant face recognition probabilities. The first scenario (Figure 6), involving sporadic target disap-

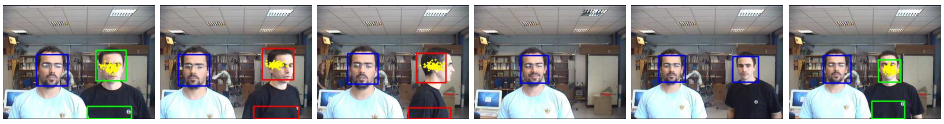


Figure 6. Tracking scenario including two persons with target's out-of-sight. Target loss detection and automatic re initialization.

pearance, shows that our probabilistic tracker is correctly positioned on the desired person (on the background) throughout the sequence. Allthought the later disappears, the tracker doesn't lock onto the undesired person thanks to a low recognition likelihood. The tracker re initializes automatically as soon as the target re-appears. The second sce-

Figure 7. Tracking scenario involving full occlusions between persons. Target recovery.

nario (Figure 7) involves occlusion of the target by another person traversing the field of view. The combination of multiple cues based likelihood and face recognition allows to keep track of the region of interest even after the complete occlusion.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de

6 Conclusion and future works

This article presents a key-scenario of H/R interaction for our tour-guide robot. Given this scenario, we have outlined visual modalities the robot must deal with.

Face detection and recognition based on Haar functions and eigenfaces enable the recognition of the robot user during proximal interaction and then in the tracking loop during the guidance mission session. Our tracker mixes face recognition and visual data fusion in a stochastic way. Evaluations on a database sequences acquired from the robot in a human centred environment show the tracker robustness to background clutter, sporadic occlusions and group of persons. Image pre processing enables to improve the face recognition process while the multi-cues associations proved to be more robust than any of the cues individually.

Future investigations concern tests on partially occluded faces and active vision to actively adapt the focal length with respect to the H/R distance and the current robot status.

References

1. S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp. A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking. *Trans. on Signal Processing*, 2(50):174–188, 2002.
2. G. Bailly, L. Brèthes, R. Chatila, A. Clodic, J. Crowley, P. Danès, F. Elisei, S. Fleury, M. Herrb, F. Lerasle, P. Menezes, and R. Alami. HR+ : Towards an interactive autonomous robot. In *Journées ROBEA*, pages 39–45, Montpellier, March 2005.
3. L. Brèthes, F. Lerasle, and P. Danès. Data fusion for visual tracking dedicated to human-robot interaction. In *Int. Conf. on Robotics and Automation (ICRA'05)*, 2005.
4. A. Doucet, S. J. Godsill, and C. Andrieu. On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing*, 10(3):197–208, 2000.
5. T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and Autonomous Systems (RAS'03)*, 42:143–166, 2003.
6. T Heseltine, N. Pears, and J. Austin. Evaluation of image pre-processing techniques for eigenface based recognition. In *Int. Conf. on Image and Graphics, SPIE*, pages 677–685, 2002.
7. E. Hjelmas. Face detection: a survey. *Computer Vision and Image Understanding (CVIU'01)*, 83(3):236–274, 2001.
8. M. Isard and A. Blake. I-CONDENSATION: Unifying low-level and high-level tracking in a stochastic framework. In *European Conf. on Computer Vision (ECCV'98)*, pages 893–908, 1998.
9. P. Pérez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proc. IEEE*, 92(3):495–513, 2004.
10. P. Pérez, J. Vermaak, and M. Gangnet. Color-based probabilistic tracking. In *European Conf. on Computer Vision (ECCV'02)*, pages 661–675, Berlin, 2002.
11. P. Viola and M. Jones. Rapid Object Detection using a Boosted Cascade of Simple Features. In *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'01)*, 2001.
12. S. Wachter and H. Nagel. Tracking persons in monocular image sequences. *Computer Vision and Image Understanding (CVIU'99)*, 74(3):174–192, 1999.

Proceedings of the 5th International Conference on Computer Vision Systems (ICVS 2007)
Published in 2007 by Applied Computer Science Group, Bielefeld University, Germany, ISBN 978-3-00-020933-8
This document and other contributions archived and available at: http://biecoll.ub.uni-bielefeld.de