Proceedings of the 2007 IEEE/RSJ International
Conference on Intelligent Robots and Systems
San Diego, CA, USA, Oct 29 - Nov 2, 2007

Human / Robot visual interaction for a Tour-Guide Robot

Thierry Germa*, Frédéric Lerasle*, Patrick Danès*, Ludovic Brèthes*

Abstract— **This paper deals with visual recognition and tracking of people and gestures from a camera mounted on a tour-guide robot in a human, cluttered, environment. The particle filtering framework enables the fusion of visual cues, both into an importance function from which the particles are sampled, and into a measurement model used for weights definition. The multi-cues associations prove to be more robust than any of the cues individually. For the purpose of gestures recognition, a tracker is proposed which handles multiple hand configurations templates. Finally, implementation and experiments on a tour-guide robot are presented in order to highlight the relevance and complementarity of the developed visual functions. Extensions are finally discussed.**

I. Introduction and framework

The development of robots acting as museum tour-guides is a motivating challenge, so that a considerable number of mature robotic systems have been developed during the last decade (see a survey in [2]). Their dedicated hardware and software classically consist of three main components: mobility, safety, and interactivity.

To our knowledge, Rhino [3] was the first robot to be deployed in a densely populated museum. Rhino and the second generation robot Minerva [14] infer people's location during an interaction session from laser scan data and distance filtering. Minerva as well as Mobot [10], are able to generate a deep inner understanding of their environments but they do not emphasize the interaction part so much.

Though these and others tour-guide robots have led to remarkable results in terms of interaction, their vision-based capabilities remain surprisingly limited. We recently developed a mobile robot named Rackham whose role is also to guide visitors by proposing either group or personalized tours. In this context, the autonomy capacities of Rackham are fully oriented towards navigation in human environments but also vision-based human-robot interaction. This paper focuses on several monocular visual modalities, namely (1) recognition and tracking of persons so as to interpret their motion in the exhibition, and (2) interpretation of commanding gestures in order to improve the communication capabilities between the robot and its tutors.

People or gestures tracking from a platform operating in a museum is a very challenging task. As the robot's evolution takes place into cluttered and densely crowded environments, several hypotheses concerning the tracking parameters to be estimated must be handled at each instant, and a robust integration of multiple visual cues together with automatic re-initialization capabilities are required. The aim is to define computationally efficient strategies, yet discriminatory enough to detect and coarsely track either the whole human body or body parts in complex scenes. Monte Carlo simulation methods, also known as particle filters [4] constitute one of the most powerful frameworks for tracking. Their popularity stems from their simplicity, ease of implementation, and modeling flexibility over a wide variety of applications. The principle is to represent the posterior distribution by a set of samples—or particles—with associated importance weights. At initial time, this weighted particle set is defined from the state vector initial probability distribution. Its propagation between two consecutive sampling consists in two steps: the particles are first drawn from an importance function which aims at exploring "relevant areas" of the state space, *e.g.* by mixing measured data with prior knowledge and dynamics; then, they are properly weighted, often entailing their likelihoods defined from the measurement function, so that the point-mass distribution they define is a consistent approximation of the posterior.

This framework is well-suited to the aforementioned requirements. Indeed, it makes no restrictive assumption on the probability distributions entailed in the characterization of the problem, and enables an easy fusion of diverse kinds of measurements. Last, some of the numerous particle filtering strategies proposed in the literature are expected to fit the specifications of the different modalities which compose the Rackham interaction mechanisms. Another observation concerns data fusion. It can be argued that data fusion using particle filtering schemes has been fairly seldom exploited within this tracking context [13]. Using multiple cues simultaneously, both into the importance and measurement functions of the underlying estimation scheme, allows to use complementary and redundant information but also enables a more robust tracking and automatic target recovery.

The paper is organized as follows. Section II describes Rackham and outlines its embedded visual modalities. Section III focuses on a proximal interaction modality involving image-based face recognition. Section IV describes the setups which best fulfill the requirements for the people tracking modalities in terms of filtering strategies and visual cues. Section V details the commanding gestures interpretation modalities. Section VI reports on the implementation of all these modalities on our robot. Last, section VII summarizes our contribution and puts forward some future extensions.

II. Rackham and its on-board visual modalities

A. Characteristics and typical tasks

Rackham is an iRobot B21r mobile platform. Its standard equipment has been extended with one pan-tilt Sony camera EVI-D70, one digital camera mounted on a Directed Perception PTU, one ELO touch-screen, a pair of loudspeakers, an

*LAAS-CNRS, Université de Toulouse, Toulouse, France FirstName.Name@laas.fr

optical fiber gyroscope and wireless Ethernet (Figure 1(a)).

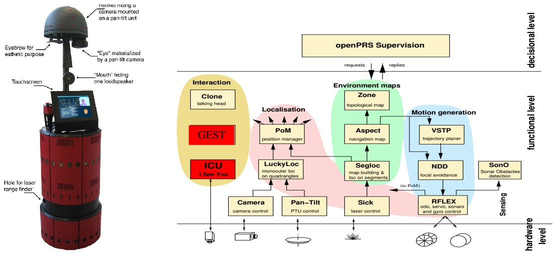
All the functions are embedded into the "LAAS" layered software architecture [1], see Figure 1(b).

The envisaged typical tasks are as follows. When Rackham is left alone with no mission, it tries to find out people whom he could interact with, a behavior hereafter called "search for interaction". As soon as a lonely visitor or a group of individuals comes into its neighborhood, it introduces itself and tries to identify its interlocutors out of the detected faces. When no interlocutor is known, a learning session of all the detected faces inside the camera field of view is launched while a "guidance mission" is defined through the touch-screen. This way, the robot will further be able to switch between multiple persons appropriately during the mission execution. Whenever all known persons leave, the robot detects this and stops. If, after a few seconds, no interlocutor is re-identified, the robot restarts a "search for interaction" session. Otherwise, when at least one known user is re-identified, the robot proposes to continue the ongoing mission. Any mission can be stopped or selected by using simple communicative gestures, without any contact. Gestures are natural means that are particularly valuable in crowded environments where speech recognition may be garbled or drowned out.

B. Dedicated visual modalities

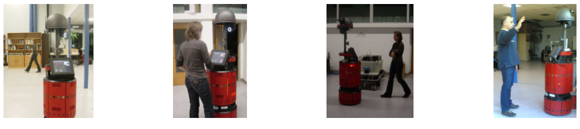
The design of visual modalities has been undertaken within the demonstration scenario depicted above. Four visual modalities, encapsulated in the modules ICU or GEST, have been outlined which the robot must basically embed:

1) **The "proximal interaction",** where the interlocutors select the area to visit through the touch-screen. Here, the robot remains static and possibly learns their faces thanks to the camera EVI-D70. This modality involves face detection and recognition at short H/R distances ($(< 1m)$) but no tracking mechanism.
2) **The "guidance mission",** where the robot drives the visitors to the selected area, keeping the visual contact with any member of the guided group even if some of them may move away. This modality involves both face recognition and upper human body tracking at medium H/R distances ($([1; 3]m)$).
3) **The "interaction through static commanding gestures",** where the aim is to recognize a number of well-defined purposeful hand postures performed by the users at medium H/R distances in order to communicate a limited set of commands to the robot. This way, the user can modify the goal of the ongoing mission, stop the robot, drive it towards another area to visit, etc.



(a) (b)

Fig. 1. (a) Rackham, (b) Rackham's layered software architecture.



(a) (b) (c) (d)

Fig. 2. The four visual modalities of the Rackham robot repeated in our lab: (a) search for interaction, (b) proximal interaction, (c) guidance mission, (d) interaction by gestures.

4) **The "search for interaction",** where the robot, static and left alone, tracks visitors in order to heckle them when they enter the exhibition. This modality involves either the whole human body tracking at long H/R distances ($(> 3m)$) or the upper human body tracking/face recognition at medium H/R distances.

III. Face recognition

This function aims to classify bounding boxes $\mathcal{F}$ of detected faces from Viola's detector [16] into either one class C_t out of the set $\{C_l\}_{1 \leq l \leq M}$ – corresponding to M users faces presumably learnt offline – or into the void class $C_{\emptyset}$. Our approach consists in performing PCA and keeping as an eigenfaces basis $B(C_t)$ the first eigenvectors accounting for a predefined ratio η of the total class variance. The approach was evaluated on a face database composed of 6000 examples of $M = 10$ individuals acquired by the robot in a wide range of typical conditions (illumination changes, variations in facial orientation and expression, etc). A crossed evaluation enables the selection of the most meaningful image preprocessing and error norms association in terms of False Acceptance Rate (FAR), and sensitivity. One evaluated error norm is inspired from the Distance From Face Space (DFFS). A given face $\mathcal{F} = \{\mathcal{F}(i), i \in \{1, \ldots, nm\}\}$ is linked to the class C_t by its error norm

$$\mathscr{D}(C_t, \mathcal{F}) = \textstyle\sum_{i=1}^{nm} (\mathcal{F}(i) - \mathcal{F}_{r,t}(i) - \mu)^2,$$

and its associated likelihood

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

where $\mathcal{F} - \mathcal{F}_{r,t}$ is the difference image of mean μ, σ_t terms the standard deviation of the error norms within the C_t's training set, and $\mathcal{N}(.; m, \sigma)$ is the Gaussian distribution with moments m and covariance σ.

As shown in Table I, 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%).

From a set of M learnt tutors (classes) noted $\{C_l\}_{1 \leq l \leq M}$ and a detected face $\mathcal{F}$, we can define for each class C_t the likelihood $\mathscr{L}_k^l = \mathscr{L}(C_t | \mathcal{F})$ for the detected face $\mathcal{F}$ at time k

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
Our 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 I

ANALYSIS OF SOME IMAGE PREPROCESSING METHODS (NONE, HISTOGRAM EQUALIZATION, SMOOTH AND CONTOUR FILTER) AND DISTANCE MEASUREMENTS.

and the posterior probability $P(C_t|\mathcal{F}, z_k)$ of labeling to C_t at time k

$$\begin{cases} P(C_\emptyset|\mathcal{F}, z_k) = 1 \text{ and } \forall t\ P(C_t|\mathcal{F}, z_k) = 0 \text{ when } \forall t\ \mathscr{L}_k^t < \tau \\ P(C_\emptyset|\mathcal{F}, z_k) = 0 \text{ and } \forall t\ P(C_t|\mathcal{F}, z_k) = \frac{\mathscr{L}_k^t}{\sum_p \mathscr{L}_k^p} \text{ otherwise .} \end{cases}$$

where τ is a threshold predefined during a learning step [5], and $C_\emptyset$ refers to the void class.

IV. PEOPLE TRACKING

A. Framework

The "search for interaction" and "guidance mission" modalities (see section II-B) involve face recognition as well as the whole/upper human body tracking. The aim of tracking is to fit a *template* relative to the tracked visitor all along the video stream, through the estimation of its image coordinates (u, v) and its scale factor s. All these parameters are accounted for in the state vector $\mathbf{x}_k$ related to the k-th frame. 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. This weak knowledge is formalized by defining the state vector as $\mathbf{x}_k = [u_k, v_k, s_k]'$ and assuming that its entries 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 covariance $\Sigma = \text{diag}(\sigma_u^2, \sigma_v^2, \sigma_s^2)$.

The following filtering strategies are then evaluated in order to check which best fulfill the requirements of the "search for interaction" and "guidance mission" tracking modalities: CONDENSATION [6], ICONDENSATION [7], hierarchical scheme [13] and Rao-Blackwellized Subspace SIR with History Sampling RBSSHSSIR [15]. Each modality is evaluated on a database of sequences acquired from the robot in a wide range of typical conditions: cluttered environments, appearance or lighting changes, sporadic disappearance of the targeted subject, jumps in his/her dynamics, etc. These evaluations, available at the URL `www.laas.fr/∼lbrethes/HRI`, emphasize the need of taking into account both the dynamics and the measurements z_k into the importance function $q(.)$ so that

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

where p_0 is the prior at initial time, and $\alpha, \beta \in [0; 1]$. Besides, the most persistent cues are used in the particles weighting stage through the measurement function $p(z_k|\mathbf{x}_k)$. The others, logically intermittent, permit an automatic initialization thanks to $\pi(.)$ and help recovery from transient tracking failures. Finally, a last requirement concerns the design of efficient trackers both in terms of selected visual cues and filtering strategies.

The current processing sampling rates range from $20Hz$ to $50Hz$ on a $3GHz$ Pentium IV personal computer, for a particles number within $[100; 200]$. These considerations motivate our choices depicted hereafter for the two people tracking modalities.

1. Upper human body tracker: From the above guidelines, we opt for the ICONDENSATION scheme. Regarding the measurement function, we consider multiple patches of distinct color distributions related to the head and the torso of the guided person (figure 3), each with its own N_{bi}-bin normalized color reference histogram in channels $\{R, G, B\}$ (resp. termed $h^c_{ref,1}$, $h^c_{ref,2}$). The color likelihood model $p(z_k^c|\mathbf{x}_k)$ is based on the Bhattacharyya distances between the two histograms pairs $\{h^c_{\mathbf{x}_k,i}, h^c_{ref,i}\}_{i=1,2}$. 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 [11]. To avoid tracker failures induced by these models updates, we also consider a shape-based likelihood $p(z_k^s|\mathbf{x}_k)$ which 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 [6]. Finally, assuming mutually independent cues, the unified measurement function comes as $p(z_k^s, z_k^c|\mathbf{x}_k) = p(z_k^s|\mathbf{x}_k).p(z_k^c|\mathbf{x}_k)$. In the considered human centered environment, more than one authorized person can be in the robot vicinity, so that the system may endlessly switch between the targeted person and other people, *e.g.* which show similar clothes appearance. From these considerations, the guidance modality must logically involve face recognition in the importance function $\pi(.)$ in (1). For the selected class C_t representing the current tutor, this becomes, with N_B the number of detected faces and $p_j = (u_j, v_j)$ the centroid coordinate of each face $\mathcal{F}_j$ – the time k being omitted for compactness reasons –

$$\pi(\mathbf{x}|z^S) \propto \sum_{j=1}^{N_B} P(C_t|\mathcal{F}_j, z).\mathcal{N}(\mathbf{x}; p_j, \text{diag}(\sigma_{u_j}^2, \sigma_{v_j}^2)).$$

The initializations of the histograms $h^c_{ref,1}, h^c_{ref,2}$ are achieved during the "proximal interaction" phase from these frames which lead to $P(C_t|\mathcal{F}_j, z)$ probabilities equal to one. In the tracking loop, the histogram model $h^c_{ref,2}$ (torso) is re-initialized with the current values when the user verification is highly confident, typically $P(C_t|\mathcal{F}_j, z) = 1$.

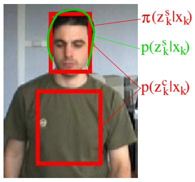


Fig. 3. The body tracking template.

2. Whole human body tracker: Evaluations have been performed in the same way as before so as to characterize the trackers associated with this modality. The two filters ICONDENSATION and RBSSHSSIR strategies ar well suited. Importance and measurement functions are based on the motion and color N_{bi}-bin normalized histograms of ROIs including the whole human body (Figure 4). The importance function $\pi(\mathbf{x}_k|z_k^m)$ involves a motion detector based on the Bhattacharyya distance between a uniform motion histogram h_{ref}^M and histograms of regions located on the nodes of a regular grid overlaid on the difference of two successive images [13]. This cue is also used in the motion likelihood model $p(z_k^m|\mathbf{x}_k)$. From the detected regions, a N_{bi}- bin normalized histogram in channels $\{R, G, B\}$ is defined (annoted h_{ref}^c). As previously, the color likelikood model $p(z_k^c|\mathbf{x}_k)$ favors candidate histograms $h_{\mathbf{x}_k}^c$ which are close to this reference histogram h_{ref}^c. These cues are assumed mutually independent conditioned on the state, *i.e.* weak correlation exists between the color, and motion of the tracked objets. Consequently, the unified measurement function thus factorizes as: $p(z_k^c, z_k^m|\mathbf{x}_k) = p(z_k^c|\mathbf{x}_k).p(z_k^m|\mathbf{x}_k)$.

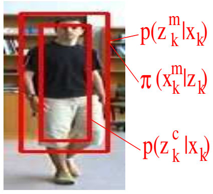


Fig. 4. The "search for interaction" template.

V. Commanding gestures interpretation

A. Framework

The last modality concerns communicative gestures. These fall into two main categories, namely acts or symbols. Interpreting act-based gestures is not trivial in our context, as both the targeted person and the robot are moving during the "guidance mission". We thus focus on symbolic gestures which are expressed by hand postures and/or canonical displacements. Due to space reasons, only static hand postures are depicted here. The reader is referred to videos available at the URL `www.laas.fr/∼lbrethes/HRI` for a handling of such gestures but also for a similar handling of dynamic gestures.

Many studies have been undertaken in order to interpret hand gestures with a single camera [12]. Conventional approaches involve two sequential stages, namely the tracking stage and the recognition stage

Our approach does not distinguish so clearly these tasks. Indeed, the aim is to recognize, in the tracking loop, a number of well-defined hand configurations which represent a limited set of commands that the user can communicate to the robot. We opt for the mixed-state CONDENSATION [8], an extension of CONDENSATION to state vectors which gather continuous-valued pose parameters (denoted $\mathbf{x}_k$) and discrete indexes c_k encoding the hand configurations. The state vector becomes $\mathbf{X}_k = (\mathbf{x}_k', r_k')'$, where the entry θ_k of the continuous part $\mathbf{x}_k = (u_k, v_k, \theta_k, s_k)'$ encodes the template situation. The continuous state components are assumed to evolve according to mutually independent Gaussian random walk models. The discrete state entry r_k evolves according to predefined transition probabilities $p(r_k|r_{k-1})$. Besides, the weighting stage relies on the evaluation of the likelihood $p(z_k|\mathbf{X}_k) = p_{r_k}(z_k|\mathbf{x}_k)$.

The MAP estimate $[\hat{r}_k]_{\text{MAP}} = \arg\max_{r_k} p(r_k|z_{1:k})$ of r_k can be approximated by

$$\hat{r}_k = \arg\max_l \sum_{i \in \Upsilon_l} w_k^{(i)} \; ; \; \Upsilon_l = \{i : \mathbf{X}_k^{(i)} = (l, \mathbf{x}_k^{(i)})\},$$

where i indexes the i-th particle $\mathbf{X}_k^{(i)}$ with probability –or "weight"– $w_k^{(i)}$. It then follows

$$\hat{\mathbf{x}}_k = \frac{\sum_{i \in \Upsilon_{\hat{r}_k}} w_k^{(i)} \mathbf{x}_k^{(i)}}{\sum_{i \in \Upsilon_{\hat{r}_k}} w_k^{(i)}} \; ; \; \Upsilon_{\hat{r}_k} = \{i : \mathbf{X}_k^{(i)} = (\hat{r}_k, \mathbf{x}_k^{(i)})\}.$$

B. Implementation and evaluations

The discrete index switching probabilities –related to the seven configuration types (Table II)– are defined manually, so as to reflect the lexicon associated with commands.

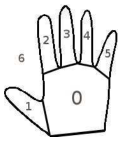


Fig. 5. The template with its seven ROIs.

Hand configurations are represented by coarse 2D rigid models, such as their silhouette contours, by means of splines. We suggest to classify static hand gestures as either direction-oriented (e.g. "turn-left", "turn-right", "move-forward", "move-backward") or motion-oriented ("move", "stop").

As was done for people tracking, the unified measurement function fuses color and shape cues. Further, defining the color likelihood on multiple patches proves efficient to discriminate between hand configurations. This is achieved within our color model by splitting the tracked region into ROIs corresponding to the palm and fingers (Figure 5).

Two reference histograms h_{ref}^c and $h_{ref}^{\neg c}$ are considered in the likelihood $p(z_k^c|\mathbf{x}_k)$. The histogram h_{ref}^c is related to a human skin color distribution trained from an images database [9], while the histogram $h_{ref}^{\neg c}$ is selected to be uniform in order to accommodate to the background variations. Local Bhattacharyya distances on the ROIs can exhibit the presence or absence of open fingers, thus improving the discriminative power between templates associated with configurations. Assuming pixel-wise independence, the color-based likelihood $p(z_k^c|\mathbf{x}_k)$ factorizes as

$$p(z_0^c, \ldots, z_6^c|\mathbf{x}) = \prod_{i \in \{0\} \cup \mathcal{O}} p^{h^c}(z_i^c|\mathbf{x}) \prod_{j \in \mathcal{C}} p^{h^{\neg c}}(z_j^c|\mathbf{x})$$

where $\mathcal{O}$ (resp. $\mathcal{C}$) gathers the indexes of the ROIs corresponding to open (resp. closed) fingers, $i = 0$ indexes the palm, and subscripts/superscripts k and ref have been omitted for compactness reason. Practically, the smaller is the color discrepancy between a given ROI and h_{ref}^c or $h_{ref}^{\neg c}$ (depending on the open fingers of the tested configuration), the higher is its associated probability. The tracker initialization logically involves skin-blobs detection.

Evaluations have been performed for this modality. Table II shows the results of a quantitative comparison with or without cues fusion for heavy cluttered background. It can be noticed that fusing shape and color seldom leads to a posture misclassification. Figure 6 shows a recognition run for such a modality.

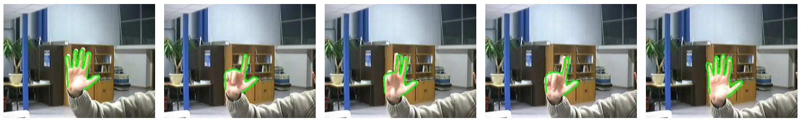
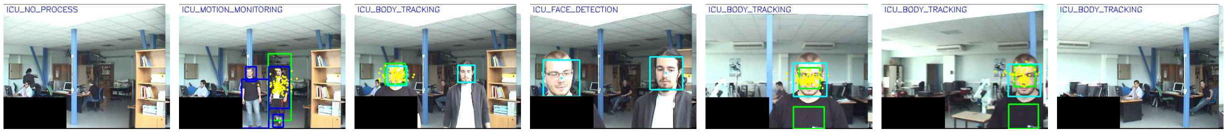


Fig. 6. “Interaction through commanding gestures”: hand configurations tracking on a sequence involving cluttered background when fusing color and shape cues in the particles likelihood.



(a) (b) (c) (d) (e) (f) (g)

Fig. 7. Switch between modalities. (a) and (g) INIT, (b) Search for interaction, (c) Body tracking, (d) Proximal interaction, (e) and (f) Guidance mission.

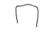
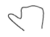
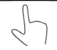
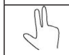
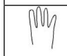
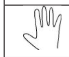
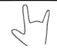
	Shape cue			Shape and color cues		
N=	100	200	400	100	200	400
	61%	83%	83%	94%	94%	94%
	0%	0%	0%	100%	100%	100%
	8%	30%	17%	75%	80%	83%
	41%	43%	43%	70%	96%	96%
	100%	100%	100%	100%	100%	94%
	1%	0%	7%	95%	95%	96%
	0%	0%	0%	85%	97%	97%
Total	**13%**	**18%**	**19%**	**89%**	**93%**	**94%**

TABLE II

AVERAGE RECOGNITION RATE PER CONFIGURATION *vs* PARTICLES NUMBER ON SEQUENCES INCLUDING CLUTTERED BACKGROUND WITH OR WITHOUT MULTIPLE CUES FUSION.

VI. DESCRIPTION OF OUR VISION-BASED MODULES

The module ICU – for “I see you” – encapsulates the aforementioned person recognition/tracking modalities while the module GEST – for “Gestures tracking” – relates to the gestures recognition system. Subsection VI-A enumerates all the visual functions provided by the module ICU. Subsection VI-B details the way how the modules ICU and GEST are entailed in the tour-guide scenario, and discusses the automatic switching between trackers.

A. Visual functions provided by the module ICU

These can be organized into three broad categories.

a) Functions related to human body/limbs detection: Independently from the tracking loop, the Viola’s face detector can be invoked depending on the current H/R distance and the scenario status.

b) Functions related to user face recognition: The face recognition process underlies the following functions

- a *face learning function* based on the face-based detector in order to train the classifier;
- a *face classification function* based on these training examples and eigenfaces representation;
- a *user presence function* which updates a presence table of the robot’s users thanks to (2). The probability of the presence of the class/person C_t at time k is updated by applying the following recursive Bayesian scheme from the classifier ouputs in the p previous frames, i.e.

$$P(C_t|z_{k-p}^k) = \left[1 + \frac{1 - P(C_t|z_k)}{P(C_t|z_k)} \cdot \frac{1 - P(C_t|z_{k-p}^{k-1})}{P(C_t|z_{k-p}^{k-1})} \cdot \frac{p(C_t)}{1 - p(C_t)}\right]^{-1}, \quad (2)$$

where

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

with N_B the number of detected faces $\mathcal{F}$ at time k. During the execution of the mission, the robot can decide to switch from a targeted person to another one depending on both: (i) the classification probabilities $\{P(C_l|\mathcal{F}_j), l \in \{1,..,M\}\}$ for each detected face $\mathcal{F}_j$, $j = 1, \dots, N_B$ at time k, (ii) the classes with the highest presence probabilities $\{P(C_l|z_{k-p}^k), l \in \{1,..,M\}\}$ in the p previous frames.

c) Functions related to user tracking: These are

- the *two tracking functions* characterized and evaluated in section IV. Recall that they have been designed so as to best suit to the interaction modalities;
- an *estimator of the H/R distance* of the targeted person from the scale s_k of the updated template during the tracking loop.

The robot activates these functions depending on the current H/R distance, user identification and scenario status. The next subsection details the way how they are scheduled.

B. Heuristic-based switching between trackers

A finite-state automaton can be defined from the tour-guide scenario outlined in section II, as illustrated in Figure 8. Its four states are respectively associated to the INIT mode and to the three aforementioned interaction modalities. Two heuristics relying on the current H/R distance and the presence table status allow to characterize most of the transitions in the graph. The robot in INIT mode invokes the motion-based detector thanks to $\pi(\mathbf{x}_k|z_k^m)$, so that any visitor entering the exhibition initializes the whole body tracking

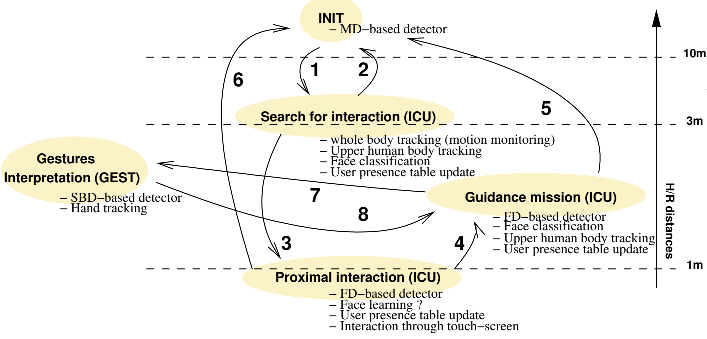

INIT
– MD–based detector
1
2
6
5
Search for interaction (ICU)
– whole body tracking (motion monitoring)
– Upper human body tracking
– Face classification
– User presence table update
Gestures Interpretation (GEST)
– SBD–based detector
– Hand tracking
7
8
Guidance mission (ICU)
– FD–based detector
– Face classification
– Upper human body tracking
– User presence table update
3
4
Proximal interaction (ICU)
– FD–based detector
– Face learning ?
– User presence table update
– Interaction through touch–screen
10m
3m
1m
H/R distances

Fig. 8. Transitions between tracking modalities.

(arrow 1). The robot assumes that the visitors are willing to interact when they have come closer and their frontal faces are frequently detected. If so, a "proximal interaction" begins (arrow 3). The face learning function and the human presence table update function are possibly invoked if no visitor is known in the robot surroundings. When starting the "guidance mission", the robot launches the upper human body tracker (arrow 4). During its execution, the robot can involve multiple persons into interaction but does remain visually in contact with only one of them, especially when the targeted person suddenly moves away. The robot displacements can be controlled without any contact thanks to the module GEST. Finally, the robot returns in INIT mode when: (i) no moving blobs are detected (arrow 2), (ii) all the presence probabilities go below a certain threshold (arrow 5), (iii) the end mission is signified by the user (arrow 6).

Thanks to an efficient modular implementation, all the ICU and GEST functions can be executed in real time on our robot. Experiments show their complementary and efficiency in cluttered scenes (Figure 7).

VII. Conclusion

This paper has presented the development of a set of visual functions dedicated to H/R interaction for our tour-guide robot. We introduced mechanisms for data fusion within particle filtering to develop trackers combining/fusing visual cues, including face recognition, in order to track people or gestures.

A first contribution relates to visual data fusion with respect to the considered robotics scenarii. Data fusion using particle filtering schemes has been extensively tackled, typically by Pérez *et al.* in [13]. The authors propose a hierarchical particle filtering algorithm, which successively takes into account the measurements so as to efficiently draw the particles. To our belief, using multiple cues simultaneously, both into importance and measurement functions, enables a more robust failures detection and recovery. More globally, other existing particle filtering strategies have been evaluated in order to check which people trackers best fulfill the requirements for the envisaged modalities. From this guiding principle, an extension for understanding hand configurations is also proposed.

A second contribution relates to the integration of the developped visual functions on our robot to highlight their relevance and complementarity. To our knowledge, quite few mature robotic systems enjoy such advanced capabilities of human and gestures perception. To illustrate our tour-guide scenario, the reader is referred to the URL `www.laas.fr/∼tgerma/rackham` for videos or more images.

Several directions are currently studied regarding our trackers. First, we study how to fuse other information such as stereo or sound cues. The sound cue won't just contribute to the localization in the image plane, but will also endow the tracker with the ability to switch its focus between speakers. Second, our tracking modalities will be made much more active.

Acknowledgement: The work described in this paper was partially conducted within the EU Integrated Projects COGNIRON ("The Cognitive Companion") and funded by the European Commission Division FP6-IST Future and Emerging Technologies under Contract $FP6 - 002020$.

References

[1] R. Alami, R. Chatila, S. Fleury, and F. Ingrand. An architecture for autonomy. *Int. Journal of Robotic Research*, 17(4):315–337, 1998.

[2] W. Burgard, A.B. Cremers, D. Fox, D. Hahnel, G. Lakemeyer, D. Schulz, W. Steiner, and S Thrun. Experiences with an interactive museum tour-guide robot. *Artificial Intelligence*, 114(1):3–55, 1999.

[3] W. Burgard, D. Fox, D. Hahnel, G. Lakemeyer, D. Schulz, W. Steiner, S. Thrun, and A.B. Cremers. Real robots for the real world – the RHINO museum tour guide project. In *National Conf. on Artificial Intelligence (AAAI'98)*, Stanford, CA, 1998.

[4] A. Doucet, N. De Freitas, and N. J. Gordon. *Sequential Monte Carlo Methods in Practice*. Series Statistics For Engineering and Information Science. Springer-Verlag, New York, 2001.

[5] T. Germa, L. Brèthes, F. Lerasle, and T. Simon. Data fusion and eigenface based tracking dedicated to a tour-guide robot. In *Int. Conf. on Vision Systems (ICVS'07)*, Bielefeld, Germany, March 2007.

[6] M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *Int. Journal on Computer Vision*, 29(1):5–28, 1998.

[7] M. Isard and A. Blake. Icondensation: Unifying low-level and high-level tracking in a stochastic framework. In *European Conf. On Computer Vision (ECCV'98)*, pages 893–908, London, UK, 1998. Springer-Verlag.

[8] M. Isard and A. Blake. A mixed-state condensation tracker with automatic model-switching. In *International Conference on Computer Vision*, page 107, Washington, DC, USA, 1998. IEEE Computer Society.

[9] M.. Jones and J. Rehg. Color detection. Technical report, Compaq Cambridge Research Lab, 11 1998.

[10] I. Nourbakhsh, C. Kunz, and Willeke. The Mobot museum robot installations: A five year experiment. In *Int. Conf. on Intelligent Robots and Systems (IROS'03)*, 2003.

[11] K. Nummiaro, E. Koller-Meier, and L. Van Gool. Object tracking with an adaptative color-based particle filter. In *Symp. For Pattern Recognition of the DAGM*, pages 353–360, 2002.

[12] V. Pavlovic, R. Sharma, and T. S. Huang. Visual interpretation of hand gestures for human-computer interaction : A review. *IEEE Trans. On Pattern Analysisand Machine Intelligence*, 19(7):677–695, 1997.

[13] P. Pérez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proc. IEEE*, 92(3):495–513, 2004.

[14] S. Thrun, M. Beetz, M. Bennewitz, W. Burgard, A.B. Cremers, F. Dellaert, D. Fox, D. Halnel, C. Rosenberg, N. Roy, J. Schulte, and D. Schulz. Probabilistic algorithms and the interactive museum tour-guide robot MINERVA. *Int. Journal of Robotics Research (IJRR'00)*, July 2000.

[15] P. Torma and C. Szepesvári. Sequential importance sampling for visual tracking reconsidered. In *AI and Statistics*, pages 198–205, 2003.

[16] 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.