

## Video-based Face Recognition and Tracking from a Robot Companion

T.Germa<sup>†</sup>, F.Lerasle<sup>†</sup>, T.Simon<sup>¶</sup>

<sup>†</sup> *LAAS-CNRS, Université de Toulouse, Toulouse, FRANCE*

<sup>¶</sup> *IUT Figeac, LRPmip-Perceval, avenue de Nayrac, 46100 Figeac, France*  
 {tgerma@laas.fr, lerasle@laas.fr, thierry-simon@univ-tlse2.fr}  
<http://www.laas.fr/~tgerma/hri>

This paper deals with video-based face recognition and tracking from a camera mounted on a mobile robot companion. All persons must be logically identified before being authorized to interact with the robot while continuous tracking is compulsory in order to estimate the person's approximate position. A first contribution relates to experiments of still-image-based face recognition methods in order to check which image projection and classifier associations give the highest performance of the face database acquired from our robot. Our approach, based on Principal Component Analysis (PCA) and Support Vector Machines (SVM) improved by genetic algorithm optimization of the free-parameters, is found to outperform conventional appearance-based holistic classifiers (eigenface and Fisherface) which are used as benchmarks. Relative performances are analyzed by means of Receiver Operator Characteristics which systematically provide optimized classifier free-parameter settings. Finally, for the SVM-based classifier, we propose a non-dominated sorting genetic algorithm to obtain optimized free-parameter settings.

The second and central contribution is the design of a complete still-to-video face recognition system, dedicated to the previously identified person, which integrates face verification, as intermittent features, and shape and clothing color, as persistent cues, in a robust and probabilistically motivated way. The particle filtering framework, is well-suited to this context as it facilitates the fusion of different measurement sources. Automatic target recovery, after full occlusion or temporally disappearance from the field of view, is provided by positioning the particles according to face classification probabilities in the importance function. Moreover, the multi-cue fusion in the measurement function proves to be more reliable than any other individual cues.

Evaluations on key-sequences acquired by the robot during long-term operations in crowded and continuously changing indoor environments demonstrate the robustness of the tracker against such natural settings. Mixing all these cues makes our video-based face recognition system work under a wide range of conditions encountered by the robot during its movements. The paper concludes with a discussion of possible extensions.

### 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 in private homes (see a survey in <sup>15</sup>). This is of particular interest for elderly and disabled people given that Europe is to experience significant ageing over the next two decades. The dedicated hardware and software of such robot companions are oriented mainly towards safety, mobility in human centered environments but also towards peer-to-

2 Germa et al.

peer interaction between the robot companion and its novice human user. These unconstrained and natural interaction mechanisms will facilitate the teaching, programming and control of robot assistants and enable them to execute demanding and complex tasks under the control of and in collaboration with the current human user. The robot's interlocutor must be logically identified before being authorized to interact with the robot while his/her identity must be verified throughout the performance of any coordinated tasks. Automatic visual person recognition is therefore crucial to this process.

Person recognition based on video is preferable to using still images as motion helps in recognition. This entails the tracking of the targeted person *i.e.* the estimation of his/her image location in the video stream. Our line of investigation consists in fusing multiple visual cues, face and clothing appearance, within the well-known particle filtering formalism.

The remainder of the paper is organized as follows. Section 2 depicts the requirements imposed by our robotic application, then outlines our approach. Section 3 describes our still face image recognition system in our robotic context. To enhance recognition performances, fine-tuning of the classifier free-parameters is addressed herein. Section 4 describes our person recognition system of the previously identified person. We briefly sum up the particle filtering formalism and principles to fuse multiple cues in the tracker, especially still face image recognition probabilities. For both developed visual functions, studies are reported concerning off-line evaluations on video sequences collected by our Jido robot companion during several runs in our lab. Lastly, section 5 summarizes our contribution and discusses future extensions.

## 2. Overview

The aforementioned visual functionalities are devoted to the robot companion called Jido (figure 1). It embeds robust and efficient basic navigation and object recognition abilities. In addition, our efforts focus in this article on the design of visual functions in order to recognize individuals, verify their presence and track them in the robot's vicinity. Figure 2 illustrates a typical scenario involving peer-to-peer H/R interaction. The left and right columns show the current H/R situation as well as the video stream from the on-board camera, respectively. In this scenario, the challenge is to recognize a given

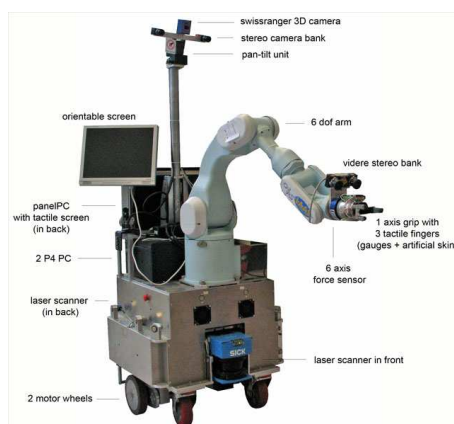


Figure 1. The Jido robot companion.

person in the video stream despite temporary occlusions by other persons, 3D rotations and out-field sight of the targeted person.

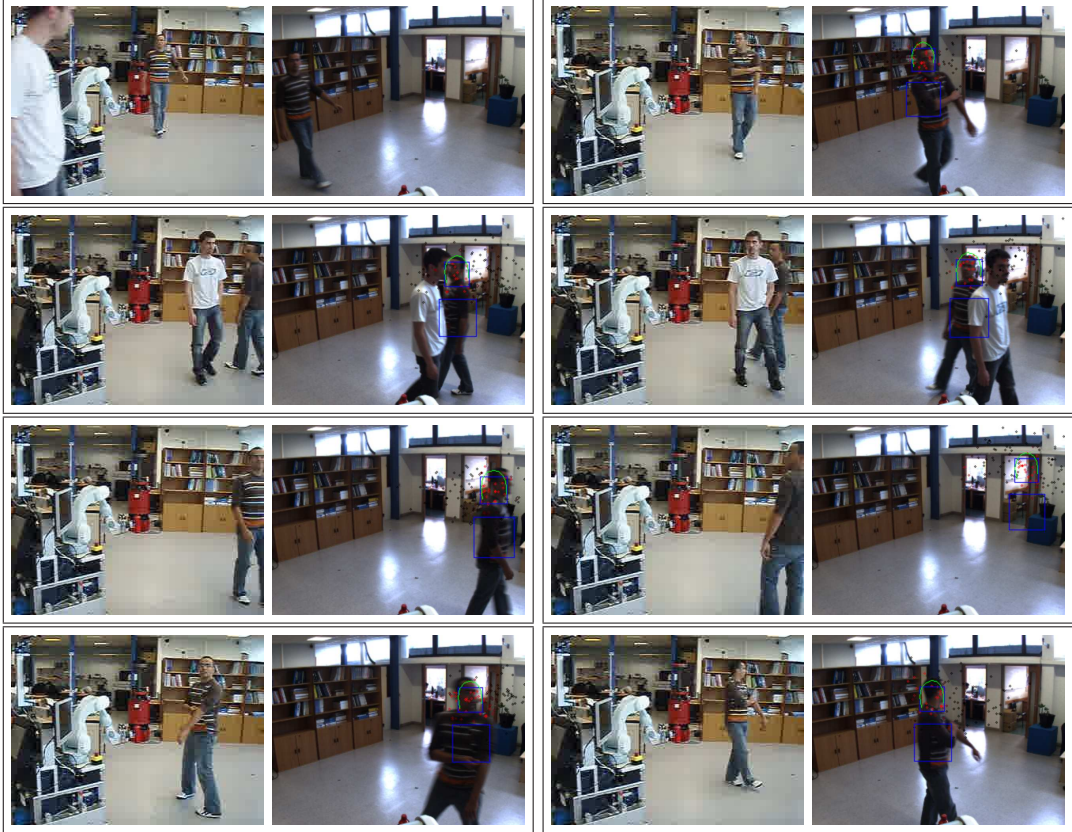


Figure 2. From top-left to bottom-right: progress of a peer-to-peer H/R interaction session. The rectangle represents the template for the targeted person.

Visual person recognition from a mobile platform operating in a human-centered scene is a challenging task which imposes several requirements. First, on-board processing power must enable the concurrent execution of other non-visual functionalities as well as decisional routines in the robot's architecture. Thus, care must be taken to design efficient vision algorithms. Contrary to conventional biometric systems, the embedded visual sensor moves in uncooperative human centered settings where people stand at a few meters - approximately at social and intimate distances - when interacting with the robot. Because of this context dependence, we cannot use well-known public face still image and video galleries for our evaluations.

Given this framework, our FR system must be capable of handling: (i) poor video quality and low image resolution which is computationally faster, (ii) heavier

lighting changes, (iii) larger pose variations in the face images *i.e.* 2D (image plane) but also 3D rotations, (iv) occlusion or background clutter. These requirements have led to interest in the design of a system that can fuse other cues in addition to face appearance and recognize these faces from video sequences instead of still images. This requires solving tracking (estimation of the targeted person image location) with automatic re-initialization capabilities, apart from the recognition task. However, the robot does not deal solely with still images. By considering subsequent frames and, as a result, spatiotemporal relationships, it is possible to make the FR problem more tractable.

Historically, video FR originated from still-image-based techniques. In other words, the system automatically detects and segments the face from the video, and then applies still-image FR techniques. Though a detailed description of the state of the art related to still-image FR falls outside the scope of this paper, the interested reader is referred to the comprehensive surveys<sup>1,49</sup> for more details. Briefly, they can be classified into two broad categories: holistic and analytic strategies although the two are sometimes combined to form a complete FR system<sup>26</sup>. We focus on the former as analytic or feature-based approaches<sup>32</sup> are not really suited to our robotic context. In fact, possible small face images (depending on the H/R distance) and low image quality of the faces captured by the onboard camera increase the difficulty in extracting local facial features. Other hand, holistic or appearance-based approaches<sup>7,38,40</sup> consider the face as a whole and operate directly on pixel intensity array representation of faces without the detection of facial features.

Besides still-image-based techniques devoted to mug-shot matching applications, approaches exploiting spatiotemporal information have recently been suggested for access control or video surveillance applications. These approaches can also be divided in two broad categories: still-to-video FR systems (*e.g.*<sup>12</sup>) based on a gallery of still face images and a probe set of videos, and video-to-video FR systems (*e.g.*<sup>3</sup>). Latter category is unsuitable in our robotic context which considers scalable face motions in videos in terms of relative distance, 2D and 3D (out-of-plane) rotations. Spatiotemporal analysis, namely tracking, is in this case based on Monte Carlo simulation methods, also known as particle filters (PF)<sup>13</sup>. The key idea is to represent at each time the posterior over the state space by a set of samples -or particles- with associated importance weights. The principle follows two steps. The particle set is first sampled (predicted) from the state vector initial probability distribution and a proposal distribution which constitutes a mathematical way of targeting the search in the state vector space. The particles are then weighted by their likelihoods *w.r.t* the measurements. PF constitutes a powerful probabilistic framework for tracking and fulfils the above robotic requirements due to the easy probabilistic based fusion of multiple measurements and the non-parametric distribution of probability distributions.

Data fusion for PF has been discussed extensively by Pérez *et al.* in<sup>34</sup>. They highlight the fact that intermittent cues make them excellent candidates for the

construction of detection modules and efficient proposal distributions. Besides, the likelihood is computed by means of measurement functions according to cues which must be persistent. Unlike <sup>51,50</sup>, the FR output is an intermittent cue in this robotic application as it is necessary to handle probe set of videos acquired from the robot in a wider range of conditions during key runs: occlusion, out-field of view or non frontal face. Consequently, our proposal distribution combines the dynamics (like CONDENSATION) but also measurements, namely the FR output. Such data-driven distribution has been surprisingly rarely exploited for tracking purposes <sup>23,34</sup>. Besides motion and face appearance cues, we consider other visual cues for person recognition, namely head silhouette, head and clothing color distributions. Clothe appearance is known to significantly facilitate person recognition, especially in low resolution when fine facial features cannot be seen. In our view, using such multiple cues simultaneously, both in the importance and measurement functions of the underlying estimation scheme, makes it possible not only to use complementary and redundant information but also enables a more robust person recognition and automatic targeted person recovery.

Prior to their fusion in the overall software robot architecture, the scalable systems for FR and head tracking are evaluated individually in order to identify their associated strengths and weakness. Fusion of both systems is investigated with special emphasis on real-time capabilities and robustness against the aforementioned H/R situations.

### 3. Face recognition from still face images

#### 3.1. *Related work*

Since the 1990s, appearance-based methods have been dominant approaches in still face image recognition systems. They involve two sequential processes: (1) image projection into subspaces to construct lower dimensional image representation, (2) final decision rule for classification purposes. Adini *et al.* in <sup>2</sup> point out that there is no image representation that can be completely invariant to lighting conditions and image-preprocessing is usually necessary.

Besides non-linear techniques <sup>33</sup>, principal component analysis (PCA), linear discriminant analysis (LDA), and independent component analysis have been popular linear techniques used for image projection. PCA uses image projection into PC (eigenface) to determine basis vectors that capture maximum image variance per class <sup>38</sup> or for the overall classes <sup>40</sup>. LDA determines a set of optimal discriminant basis vectors so that the ratio of the between-class and within-class scatters is maximized. LDA finds the best projection direction in which training samples of different classes are best separated. The LDA is either operated on the raw image to extract the Fisherface <sup>7,25</sup> or on the eigenface to obtain the discriminant eigenfeatures <sup>48</sup>. ICA provides a set of basis vectors that possess maximum statistical independence <sup>6</sup>. We design experiments in which faces are represented in both PC and LD subspaces parameterized by the information ratio (noted  $\eta$ ) they

6 *Germa et al.*

encompass.

The decision rule differs from the classification algorithms. Euclidean distance or normalized correlation<sup>25</sup>, Hausdorff distance<sup>30</sup>, distance from face space (DFFS)<sup>40</sup> showed successful results. In earlier studies<sup>17</sup>, we proposed an error norm distance which is highlighted to outperform the well-known DFFS. These rules require a decision threshold hereinafter referred to as  $\tau$ . Inspired from<sup>25</sup>, the evaluations are extended to support vector machines (SVM) in combination with PCA or LDA for dimensionality reduction. SVMs map the observations from input space into a higher dimensional feature space using a non-linear transformation, then find a hyperplane in this space which maximizes the margin of separation in order to minimize the risk of misclassification between faces. An RBF kernel is usually used for this transformation<sup>21,25</sup> where the width free-parameter (herein annotated  $\gamma$ ) controls the width of the Gaussian kernel. Another important free-parameter to tune is  $C$  the upper bound of Lagrangian multipliers required for the minimization under constraints. SVM shows significantly different performance according to kernel functions but especially the SVM free-parameters  $\gamma$ ,  $C$  and also  $\tau$ .

The issue of automatic optimization of the aforementioned free-parameters is either ad-hoc, or based on receiver operator characteristics (ROC) curves<sup>16,35</sup> or on numerical methods dealing with the minimization of non-linear objective functions. In this vein, local gradient descent methods<sup>11</sup> or global optimization<sup>10,46</sup> are proposed to maximize the generalization performance. Genetic algorithms (GA) are also well-known techniques for optimization problems, and have proved to be effective for selecting SVM parameters<sup>27,45</sup>. Our primary motivation for the study referred to below is to fine-tune properly the free-parameters produced by each classifier model in order to highlight its optimal performance for detailed classifier performance comparison.

### 3.2. *Our approach*

From these reminders, recognition experiments are performed for histogram equalization-based preprocessing, two different representations (PC and LD basis), and three decision rules (error norm, Mahalanobis distance and SVM). Note that the final goal is to classify facial regions  $\mathcal{F}$ , segmented from the input image, into either one class  $C_t$  out of the set  $\{C_l\}_{l=1}^M$  of  $M$  subject faces using training algorithms. For detecting faces, we apply the well known window scanning technique introduced by Viola *et al.*<sup>41</sup>, and improved in<sup>42,43</sup>, which covers a range of  $\pm 45^\circ$  out-of plane rotation. The bounding boxes of faces segmented by the Viola's detector are then fed into the FR systems referred to below.

#### 3.2.1. *Face recognition systems*

We enumerate hereafter the developed and evaluated classifiers as well as their free-parameters subject to optimization and which mainly influence classification performances.



**A. FSS+EN system: Face-Specific Subspace and error norm** - As described in <sup>38</sup>, for each class  $C_t$ , eigenface  $W_{pca,t}$  basis is deduced by solving

$$S_{T,t}.W_{pca,t} - W_{pca,t}.\Lambda_t = 0, \quad (1)$$

where  $S_{T,t}$  is the scatter matrix, and  $\Lambda_t$  the ordered eigenvalue vector for class  $C_t$ . We keep the first  $N_{v,t}$  eigenvectors as the eigenface basis such that

$$\frac{\sum_{i=0}^{N_{v,t}} \Lambda_{i,t}}{\sum \Lambda_{i,t}} \leq \eta, \quad (2)$$

accounting for a predefined ratio  $\eta$  of the total class  $C_t$  variance, given that  $\Lambda_t$  is the ordered eigenvalue vector. The decision rule is based on the error norm introduced in <sup>17</sup>. Given an unknown test face  $\mathcal{F} = \{\mathcal{F}(i), i \in \{1, \dots, nm\}\}$  and  $\mathcal{F}_{r,t}$  the reconstructed face onto face specific subspace of the class  $C_t$ , this error norm is given by

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

and the associated likelihood follows  $\mathcal{L}(C_t|\mathcal{F}) = \mathcal{N}(\mathcal{D}(C_t, \mathcal{F}); 0, \sigma_t)$ , where  $\mathcal{F} - \mathcal{F}_{r,t}$  is the difference image of mean  $\mu$ ,  $\sigma_t$  terms the standard deviation of the error norms within the  $C_t$  training set, and  $\mathcal{N}(\cdot; m, \sigma)$  is the Gaussian distribution with moments  $m$  and covariance  $\sigma$ . This error norm has been shown in <sup>17</sup> to outperform both the Euclidian distance and the DFFS. The last issue concerns the appropriate selection of the threshold in the decision rule. From a set of  $M$  learnt subjects/classes noted  $\{C_l\}_{l=1}^M$  and a detected face  $\mathcal{F}$ , we can define for each class  $C_t$  the likelihood  $\mathcal{L}^t = \mathcal{L}(C_t|\mathcal{F})$  for the detected face  $\mathcal{F}$  and the posterior probability  $P(C_t|\mathcal{F}, z)$  of labeling to  $C_t$  as

$$\begin{cases} \forall t \ P(C_t|\mathcal{F}, z) = 0 \text{ and } P(C_\emptyset|\mathcal{F}, z) = 1 \text{ when } \forall t \ \mathcal{L}^t < \tau \\ \forall t \ P(C_t|\mathcal{F}, z) = \frac{\mathcal{L}^t}{\sum_p \mathcal{L}^p} \text{ and } P(C_\emptyset|\mathcal{F}, z) = 0 \text{ otherwise,} \end{cases} \quad (3)$$

where  $C_\emptyset$  refers to the void class, and  $\tau$  is a predefined threshold. This classifier depends on the free-parameters  $\eta$  and  $\tau$ .

**B. GPCA+MD system: global PCA and Mahanalobis distance** - Here a single PC basis is estimated given equation (1) and the total scatter matrix  $S_T$ . The decision rule is based on the Mahanalobis distance. This classifier depends also on the free-parameters  $\eta$  and  $\tau$ .

**C. LDA+MD system: Fisherface and Mahanalobis distance** - Fisherface  $W_{lda}$  basis is deduced by solving  $S_B.W_{lda} - S_W.W_{lda}.\Lambda = 0$ , where  $S_B$ , and  $S_W$  are the between-class, and within-class scatter matrices while the eigenvectors also follow equation (2). The decision rule is based on the Mahanalobis distance. The free-parameters are also  $\eta$  and  $\tau$ .

**D. GPCA+SVM system: global PCA and SVM** - This system performs global PCA and SVM delivers probability estimates. The associated theory and implementation details are described in <sup>44</sup>. This classifier model produces the free-parameters  $\eta$ ,  $C$ ,  $\gamma$  and  $\tau$ .

### 3.2.2. Fine-tuning strategy based on ROC curves and classifiers comparison

We conducted FR experiments using the proposed framework on the face dataset composed of 6600 examples including 8 possible human users and 3 impostors corresponding to unknown individuals for the robot. In this dataset, the subjects arbitrarily move their heads, possibly change their expressions while the ambient lighting, the background, and the relative distance might change. A few sample images from this dataset are shown in figure 3 while the entire face gallery is available at [www.laas.fr/~tgerma/hri](http://www.laas.fr/~tgerma/hri).



Figure 3. Examples of samples for a given class.

The evaluation protocol specifies a partitioning of the face database into four disjoint sets: (1) a training set #1 (8 users, 30 images per class), (2) a training set #2 (8 users, 30 images per class), (3) an evaluation set (8 users and 3 impostors, 40 images per class), (4) a test set (8 users and 3 impostors, 500 images per class). The training sets #1 and #2 are used to learn the users' face representations and the support vectors. The evaluation set allows us to estimate the aforementioned free-parameters, and the test set to characterize the optimal performances for each classifier on independent data.

These performances of the above classifier are analyzed by means of ROCs when varying the free-parameter vector  $\mathbf{q}$  subject to optimization for each classifier. The idea, pioneered by Provost *et al.* in <sup>35</sup>, is outlined as follows. We search over a set of free-parameters by computing a ROC point *i.e.* the false positive and true positive (or hit) rates, namely FPR and TPR. For a given classifier, the set  $\mathcal{Q}$  of all admissible parameter vectors  $\mathbf{q}$  generates a set of ROC points, of which we seek the dominant, or optimal Pareto points along the ROC convex hull. More formally, we seek for the subset  $\mathcal{Q}^* \subset \mathcal{Q}$  of parameter vectors  $\mathbf{q}$  for which there is no other parameter vector that outperforms both FPR and TPR:

$$\mathcal{Q}^* = \{\mathbf{q} \in \mathcal{Q} | \forall \mathbf{q}' \in \mathcal{Q}, FPR(\mathbf{q}') \geq FPR(\mathbf{q}) \wedge TPR(\mathbf{q}') \leq TPR(\mathbf{q})\}. \quad (4)$$

Clearly,  $\mathcal{Q}^*$  identifies the subset of parameter vectors that are potentially optimal for a given classifier. Figure 4 shows ROC points and the Pareto front when varying the free-parameters over their ranges. The subfigures, when plotting TPR and FPR on the Y- and X-axis, allows an informal visual comparison of the four classifiers. System D clearly dominates the other classifiers as its Pareto front lies in the northwest corner of the ROC space (TPR is higher, FPR is lower). Considering the equal error rate (EER) leads to the same analysis. This rate is derived by isocost lines where the cost represents the simple sum of the cost of misclassifying positive and negative examples. Assuming equal numbers of positive and negative



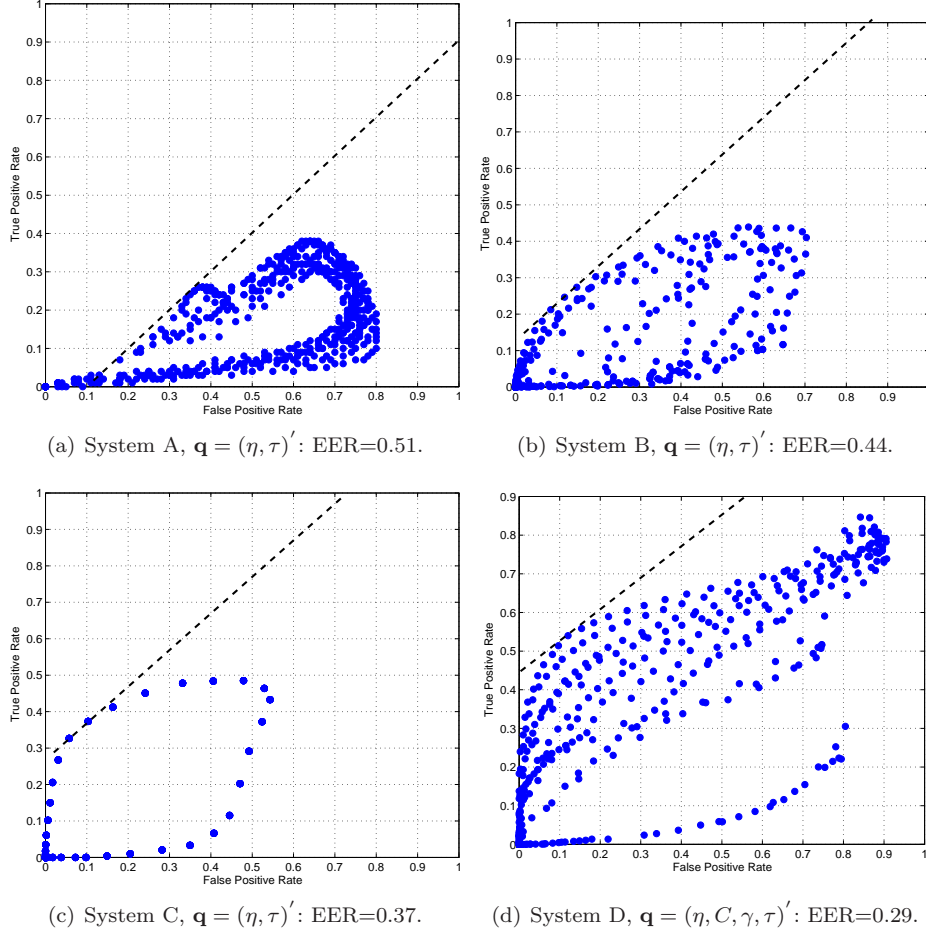


Figure 4. ROC points for each classifier and the associated isocost line for EER. Free-parameter vector  $\mathbf{q}$  for optimization are listed under the corresponding classifier.

examples for a given classifier, the EER is the point on its ROC which lies on a  $45^\circ$  line closest to the north-west corner of the ROC plot. Thus, the best system, namely D, provides a Pareto front with a lower EER, namely 0.29. Finally, note that its computational cost is 0.5 ms against 0.3 ms per image for systems B-C. Unfortunately, an exhaustive search for the selection of all parameters, especially for model D which produces more free-parameters, is computationally intractable on an autonomous robot as the finality is to learn human faces on-the-fly when interacting with new persons. Consequently, we propose a genetic algorithm (GA) to discover optimal free-parameter vectors of system D more quickly due to its multi-objective optimization framework. By limiting the number of ROC points to be considered, GA renders the optimization procedure computationally feasible.

### 3.2.3. Fine-tuning strategy based on genetic algorithm

Conventional methods using GA are single-objective optimization problems<sup>27</sup>. Non-dominated sorting GA (NSGA-II) has proved to be suited to multi-objective optimization problem<sup>45</sup>. The algorithm aims at minimizing the distance of the generated solutions to the Pareto front (4) and maximizing the diversity of the achieved Pareto front approximation. Figure 5 shows the evolution of the Pareto front when varying the population size in the range [16, 20] and the preset generation count in the range [1, 30].

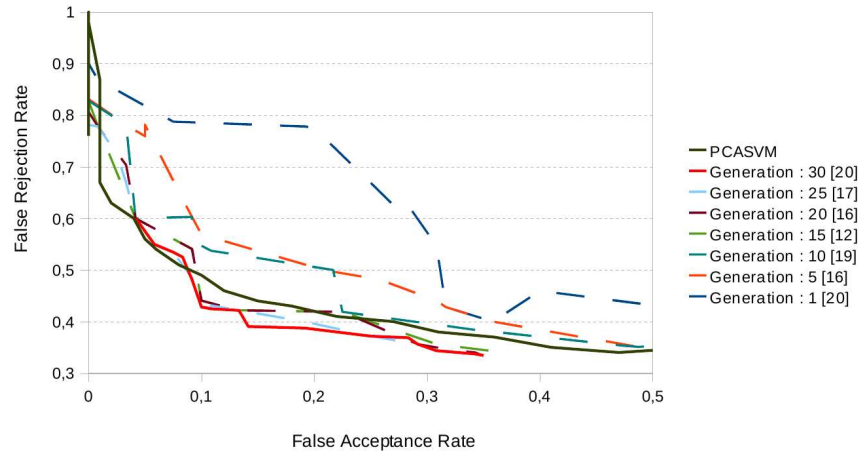


Figure 5. NSGA-II Pareto front evolution *vs.* PCA+SVM based system D.

Note that the generated solutions “move” so as to reduce both FPR and FRR objectives. This optimization strategy is no longer guaranteed to find the Pareto front optimum but there is an experimental evidence that the solution is close to optimal when increasing the preset generation count. Given a population initialized randomly (first generation in figure 5), we can see that after the first 10 generations, there is already one solution that outperforms the one without optimization while 30 generations increase the performance compared to ROC means slightly. Therefore, the minimum EER for 30 generations becomes 0.26 against 0.29 in subfigure 4(d). Informally, this non-exhaustive search strategy, parameterized by the generation count and the population size, makes it possible to control the tradeoff between the computational cost and the classification performance.

## 4. Person recognition and tracking from videos

### 4.1. Related work

Lasting recent years, many video-based FR systems have been developed<sup>47</sup>. Note that such systems can be classified into two categories introduced in § 2: still-to-

Symbol	Meaning	Value
$\eta$	Ratio of the total class variance for PCA	0.99
$C$	Upper bound of Lagrangian multipliers	80391
$\gamma$	Parameter in the RBF kernel	0.002526
$\tau$	Threshold for decision rule (3)	0.71

Table 1. Free-parameter values used in the system D.

video and video-to-video systems. This latter category addresses the problem of FR from spatial and temporal information learnt in video gallery. Hidden Markov models (HMM) have been widely applied to model temporal information and perform FR<sup>20,31</sup>. HMMs are trained to learn both the statistics and the temporal dynamics of each individual. FR in<sup>3</sup> is performed using the concept of subspace angles to compute distances between probe and gallery video sequences. Biuk *et al.* in<sup>9</sup> build a trajectory in eigenspace where each trajectory belongs to one face sequence (profile to profile). Similarly, Lee *et al.* in<sup>29</sup> use appearance manifolds approximated by piecewise linear subspace coupled with a transition matrix representing the dynamics. This method seems to be at most capable of handling large 2D and 3D head rotations. Although all these systems increase the recognition rates significantly as compared with still-to-video FR systems, these FR systems can be applied only to a subset of canonical sequences, namely prototype face trajectories<sup>9</sup>, video sequences involving specific individual *vs.* camera situations<sup>31</sup>, and so on. All these assumptions are clearly unsuitable for our robotic application or would require an excessive amount of training sequences to capture all the H/R situations the onboard FR system would have to handle.

In fact, our approach belongs to the first category which is reviewed and discussed hereafter. Despite the fact that both static and dynamic information is available, preliminary research reported in<sup>28,49</sup> has limited the scope of the problem to the use of still image-based methods to some selected frames. Spatiotemporal analysis was initially considered by<sup>12</sup> even if the two tasks were split: the person-specific estimated dynamic characteristics helped the FR system and reciprocally. Lastly, solving these two tasks simultaneously by probabilistic reasoning<sup>51,50</sup> has been proven to enhance recognition performances significantly. To the best of our knowledge, this strategy is the most similar to ours even if important differences exist. Zhou *et al.* in<sup>51</sup> consider the well-known PCA and the CONDENSATION strategy to estimate for each frame the face kinematic and FR in a joint posterior distribution. In<sup>50</sup>, the authors improved their previous approach by incorporating two models, respectively for the interframe appearance changes and the appearance changes between probe videos and gallery images. Unfortunately, the FR cue is logically persistent as the probe set of videos involves people gazing at the camera (namely near frontal face view) while this is intermittent in our application context (see § 2). The ICONDENSATION scheme is applied in this case to permit

12 *Germa et al.*


---

$\{x_k^{(i)}, w_k^{(i)}\}_{i=1}^N = \text{SIR}(\{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}_{i=1}^N, z_k)$	
1:	<b>IF</b> $k = 0$ , <b>THEN</b> Draw $x_0^{(1)}, \dots, x_0^{(i)}, \dots, x_0^{(N)}$ i.i.d. according to $p(x_0)$ , and set $w_0^{(i)} = \frac{1}{N}$ <b>END</b>
2:	<b>IF</b> $k \geq 1$ <b>THEN</b> $\{x_{k-1}^{(i)}, w_{k-1}^{(i)}\}_{i=1}^N$ being a particle description of $p(x_{k-1} z_{1:k-1})$
3:	<b>FOR</b> $i = 1, \dots, N$ , <b>DO</b>
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 $w_k^{(i)} \propto w_{k-1}^{(i)} \frac{p(z_k x_k^{(i)})p(x_k^{(i)} x_{k-1}^{(i)})}{q(x_k^{(i)} x_{k-1}^{(i)}, z_k)}$ , prior to a normalization step so that $\sum_i w_k^{(i)} = 1$
6:	<b>END FOR</b>
7:	Compute the conditional mean of any function of $x_k$ , <i>e.g.</i> the MMSE estimate $E_{p(x_k z_{1:k})}[x_k]$ , 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 $\{1, \dots, N\}$ the indexes $s^{(1)}, \dots, s^{(N)}$ 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:	<b>END IF</b>

---

Table 2. Generic particle filtering algorithm (SIR).

intermittent cue fusion in the importance function and persistent cue data fusion in the measurement function and thus automatic re-initialization after target loss. Other visual cues like the appearance of clothes and skin color are also considered.

The next section recalls some basics on particle filtering (PF) algorithms for data fusion. Then, section 4.3 details our tracking-and FR approach by resolving uncertainties in tracking and recognition simultaneously in the ICONDENSATION framework. Finally, both quantitative and qualitative evaluations on videos shot using the Jido robot are presented in section 4.4.

#### 4.2. Basics on particle filtering algorithms for data fusion

Particle filters are sequential Monte Carlo simulation methods of the state vector estimation of any Markovian dynamic system<sup>5,13</sup>. Their aim is to recursively approximate the posterior probability density function (pdf)  $p(x_k|z_{1:k})$  of the state vector  $x_k$  at time  $k$  conditioned on the set of measurements  $z_{1:k} = z_1, \dots, z_k$ . A linear point-mass combination

$$p(x_k|z_{1:k}) \approx \sum_{i=1}^N w_k^{(i)} \delta(x_k - x_k^{(i)}), \quad \sum_{i=1}^N w_k^{(i)} = 1, \quad (5)$$

is determined – where  $\delta(\cdot)$  is the Dirac distribution – which expresses the selection of a value – or “particle” –  $x_k^{(i)}$  with probability – or “weight” –  $w_k^{(i)}$ ,  $i = 1, \dots, N$ . An approximation of the conditional expectation of any function of  $x_k$ , such as the MMSE estimate  $E_{p(x_k|z_{1:k})}[x_k]$ , then follows.

The generic particle filtering algorithm – or “Sampling Importance Resampling” (SIR) –, shown on Table 2, is fully described by the prior  $p(x_0)$ , the dynamics pdf  $p(x_k|x_{k-1})$  and the observation pdf  $p(z_k|x_k)$ . After initialization of independent identically distributed (i.i.d.) sequence drawn from  $p(x_0)$ , the particles evolve

stochastically, being sampled from an importance function  $q(x_k|x_{k-1}^{(i)}, z_k)$ . They are then suitably weighted so as to guarantee the consistency of the approximation (5). To this end, step 5 assigns each particle  $x_k^{(i)}$  a weight  $w_k^{(i)}$  involving its *likelihood*  $p(z_k|x_k^{(i)})$  w.r.t. the measurement  $z_k$  as well as the values of the dynamics pdf and importance function at  $x_k^{(i)}$ .

In order to limit the degeneracy phenomenon, as it is well known in the literature<sup>(5, 14)</sup>, step 8 inserts a resampling stage introduced by Gordon *et al.* in<sup>19</sup> so that the particles associated with high weights are duplicated while the others collapse and the resulting sequence  $x_k^{(s^{(1)})}, \dots, x_k^{(s^{(N)})}$  is i.i.d. according to (5). Note that this resampling stage should rather be performed only when the filter efficiency – related to the number of “useful” particles – falls below a predefined threshold<sup>14</sup>.

The CONDENSATION – for “Conditional Density Propagation”<sup>22</sup> – is the instance of the SIR algorithm such that the particles are drawn according to the system dynamics, viz. when  $q(x_k|x_{k-1}^{(i)}, z_k) = p(x_k|x_{k-1}^{(i)})$ . Then, in visual tracking, the original algorithm<sup>22</sup> defines the particles likelihoods from contour primitives, yet other visual cues have also been exploited<sup>34,39</sup>. On this point, resampling may lead to a loss of diversity in the state space exploration. The importance function must thus be defined with special care. As CONDENSATION draws the particles  $x_k^{(i)}$  from the system dynamics but “blindly” w.r.t. the measurement  $z_k$ , many of these may well be assigned a low likelihood  $p(z_k|x_k^{(i)})$  and thus a low weight in step 5, significantly worsening the overall filter performance.

An alternative, henceforth labeled “Measurement-based SIR” (MSIR), merely consists in sampling the particles – or just some of their entries – at time  $k$  according to an importance function  $\pi(x_k|z_k)$  defined from the current image. The first MSIR strategy was ICONDENSATION<sup>23</sup>, which guided the state space exploration by a color blob detector. Other visual detection functionalities can be used as well, *e.g.* face detection/recognition (see here below), or any other intermittent primitive which, despite its sporadicity, is very discriminant when present<sup>34</sup>. Thus, the classical importance function  $\pi(\cdot)$  based on a single detector can be extended to consider the outputs from  $L$  detection modules, *i.e.*

$$\pi(x_k^{(i)}|z_k^1, \dots, z_k^L) = \sum_{l=1}^L \kappa_l \pi(x_k^{(i)}|z_k^l), \text{ with } \sum \kappa_l = 1. \quad (6)$$

In an MSIR scheme, if a particle  $x_k^{(i)}$  drawn exclusively from the image (namely  $\pi(\cdot)$ ) is inconsistent with its predecessor  $x_{k-1}^{(i)}$  from the point of view of the state dynamics, the update formula leads to a small weight  $w_k^{(i)}$ . One solution to this problem, as proposed in the genuine ICONDENSATION algorithm, consists in also sampling some particles from the dynamics and some w.r.t. the prior so that, with  $\alpha, \beta \in [0; 1]$

$$q(x_k^{(i)}|x_{k-1}^{(i)}, z_k) = \alpha \pi(x_k^{(i)}|z_k) + \beta p(x_k|x_{k-1}^{(i)}) + (1 - \alpha - \beta) p_0(x_k). \quad (7)$$

Besides the importance function, the measurement function involves visual cues which must be persistent but are however more prone to ambiguity for cluttered scenes. An alternative is to consider multi-cue fusion in the weighting stage. Given  $L$  measurement sources  $(z_k^1, \dots, z_k^L)$  and assuming the latter are mutually independent conditioned on the state, the unified measurement function can then be factorized as

$$p(z_k^1, \dots, z_k^L | x_k^{(i)}) \propto \prod_{l=1}^L p(z_k^l | x_k^{(i)}). \quad (8)$$

It can be argued that data fusion using particle filtering schemes has been fairly seldom exploited within this tracking context<sup>34,39</sup>. In our view, using multiple cues simultaneously, both into the importance and measurement functions of the underlying ICONDENSATION scheme, makes it possible to increase the tracker versatility to variable environments encountered by the mobile robot. This strategy also enables automatic initialization when the robot user appears or re-appears in the scene and improve the recovery of deadlocks induced by target loss due for instance to occlusions<sup>18,23</sup>.

#### 4.3. *Our approach*

In a populated environment, more than the current engaged person might be in the robot vicinity. Consequently, the mobile robot maintains visual contact (thanks to its on-board camera) with this particular interlocutor during any continuous peer-to-peer H/R communicative or interactive act. This logically requires a merge of the face verification probabilities in a tracking loop of the targeted person. The aim of our image-based tracker is classically to fit the template relative to the tracked person throughout the video stream, through the estimation of the state vector  $\mathbf{x}_k$  related to the  $k$ -th frame which is composed of image coordinates  $(u_k, v_k)$  and scale  $s_k$  of the template. 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 thus 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  $\mathcal{N}(\cdot; \mu, \Sigma)$  is a Gaussian distribution with mean  $\mu$  and covariance  $\Sigma = \text{diag}(\sigma_u^2, \sigma_v^2, \sigma_s^2)$  being determined *a priori*. Regarding the filtering strategy, we opt for the ICONDENSATION algorithm as it enjoys the aforementioned nice properties and a low time consumption. Both importance and measurement functions involved in the tracker are characterized below.

Recall that the unified importance function  $\pi(\cdot)$  in equation (6) offers a mathematical way of directing search according to several visual detectors. This function combines two importance functions  $\pi(\mathbf{x}_k | z_k^c)$  and  $\pi(\mathbf{x}_k | z_k^s)$  respectively based on a skin-blob detector and the aforementioned face classification process. Human skin colors have a specific distribution in color space. Training images from the



database<sup>24</sup> are used to construct a reference  $(R, G, B)$  color histogram model. Blob detection is performed by subsampling the input image prior to grouping the classified skin-like pixels. Let  $N_B$  be the number of detected faces and  $\mathbf{p}_j = (u_j, v_j), j = 1, \dots, N_B$  the centroid coordinate of each such region. The importance function  $\pi(\mathbf{x}_k | z_k^c)$  at location  $\mathbf{x} = (u, v)$  follows, as the Gaussian mixture proposal

$$\pi(\mathbf{x} | z^c) = \sum_{j=1}^{N_B} \mathcal{N}(\mathbf{x}; \mathbf{p}_j, (\sigma_{u_j}^2, \sigma_{v_j}^2)),$$

where the time index  $k$  has been omitted for compactness reasons. The importance function  $\pi(\mathbf{x}_k | z_k^s)$  is defined by a similar Gaussian mixture. Given the selected class  $C_l$  *i.e.* the current tracked face and the associated probabilities  $P(C_l | \mathcal{F}_j)$  for each detected face  $\mathcal{F}_j, j = 1, \dots, N_B$  at time  $k$ , the importance function becomes

$$\pi(\mathbf{x} | z^s) \propto \sum_{j=1}^{N_B} P(C_l | \mathcal{F}_j, z) \cdot \mathcal{N}(\mathbf{x}; \mathbf{p}_j, \text{diag}(\sigma_{u_j}^2, \sigma_{v_j}^2)), \quad (9)$$

where  $\mathbf{p}_j$  is the centroid of each detected face  $\mathcal{F}_j$ .

The influence of fusing intermittent visual cues in the importance function, as well as the influence of data fusion in the measurement function, is shown in Table 3. Regarding that the unified measurement function (8), which aims at fusing persistent cues, must handle varying light conditions and head poses as well as occlusions, we opt for a template based both on color histograms and head silhouette, similar to the data fusion presented in<sup>8,37</sup>. Multi-patches of distinct color distribution related to the head and the clothing appearance of the targeted person (figure 6) are here considered. Each specific  $N_{bi}$ -bin normalized reference histograms model in channel  $c$  is hereinafter annotated  $h_{ref,1}^c, h_{ref,2}^c$  respectively. Let the union  $B_{\mathbf{x}} = \bigcup_{p=1}^2 B_{\mathbf{x},p}$  for any state  $\mathbf{x}_k$  be associated with the set of reference histograms  $\{h_{ref,p}^c : c \in \{R, G, B\}, p = 1, 2\}$ . By assuming conditional independence of the color measurements, the likelihood  $p(z_k^c | \mathbf{x}_k)$  becomes

$$p(z_k^c | \mathbf{x}_k) \propto \exp \left( - \sum_c \sum_{p=1}^2 \frac{D^2(h_{\mathbf{x},p}^c, h_{ref,p}^c)}{2\sigma_c^2} \right),$$

provided that  $\sigma_c$  terms a standard deviation being determined *a priori* and  $D$  the Bhattacharyya distance<sup>4</sup> between the two histograms  $h_{ref,p}^c$  and  $h_{\mathbf{x},p}^c$  *i.e.* for a channel  $c$

$$D(h_{\mathbf{x}}^c, h_{ref}^c) = (1 - \sum_{j=1}^{N_{bi}} \sqrt{h_{\mathbf{x},j}^c \cdot h_{ref,j}^c})^{1/2},$$



Figure 6.  
Color template.

where the index  $p$  has been omitted for compactness reasons. This multi-part extension is more accurate thus avoiding the drift, and possible subsequent loss, experienced sometimes by the single-part version<sup>34</sup>. The initialization of  $(h_{ref,1}^c, h_{ref,2}^c)$  is achieved according to frames which lead to high probabilities in terms of face recognition, typically  $P(C_l|\mathcal{F}) \sim 1$ . To overcome the ROI 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.*

$$h_{ref,k}^c = (1 - \kappa) \cdot h_{ref,k-1}^c + \kappa \cdot h_{E[\mathbf{x}_k]}^c, \quad (10)$$

where  $\kappa$  weights the contribution of the mean state histogram  $h_{E[\mathbf{x}_k]}^c$  to the target model  $h_{ref,k-1}^c$  and index  $p$  has been omitted for compactness reasons. This model update process can lead to drifts with the consequent loss of the target. To avoid such tracker failures, we also consider a shape-based likelihood  $p(z_k^s|\mathbf{x}_k)$  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 (figure 7) *i.e.* the shape-based likelihood is given by<sup>22</sup>

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



Figure 7.  
Shape cue.

where  $l$  indexes the  $N_p$  template point  $x(l)$  and associated closest edge  $z(l)$  in the image. Finally, the unified measurement function in step 5 of Table 2 can then be formulated as  $p(z_k^s, z_k^c|\mathbf{x}_k) = p(z_k^s|\mathbf{x}_k) \cdot p(z_k^c|\mathbf{x}_k)$ . The examples in Table 3(a) and 3(b) show for the above example the likelihood function  $p(z_k^s|\mathbf{x}_k)$  and the more discriminant unified likelihood function  $p(z_k^c, z_k^s|\mathbf{x}_k)$ .

The runs presented in Table 3 show the efficiency of the strategy of data fusion in both the importance and measurement function. These results are discussed below. The template corresponding to the estimate of the position of the target is represented by the blue rectangles (color template) and the green curve (shape template) while the dots materialize the hypotheses and their weight after normalization (black is 0 and red is 1).

The first run (Table 3(a)) shows the execution of a simple CONDENSATION strategy based on both aforementioned random walk dynamic for particles sampling and multi-patch color measurement. After some iteration, we can observe a drift of the tracker as the histogram model update corrupts the reference histogram due to the cluttered background. The second run (Table 3(b)) better matches to the targeted person because the measurement function considers the shape-based likelihood in addition to the color measurement. Even if the template fits a person, fusing cues in the measurement function is not enough to remain robust to occlusion between persons (in this instance between  $t = 15$  and  $t = 81$ ). The run (c) in Table 3 combines face and skin color detection with the random walk dynamic in

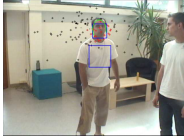
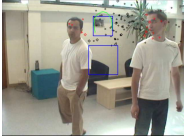
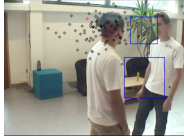
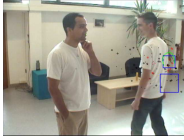
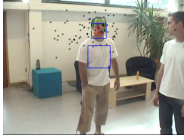
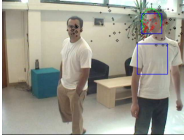
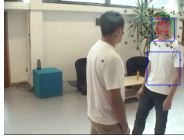
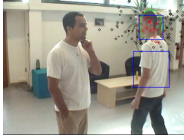
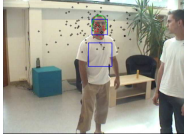
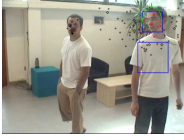
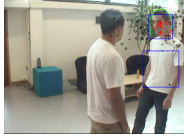

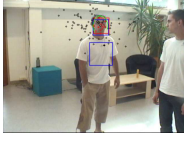
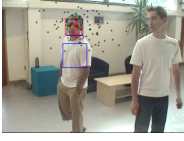
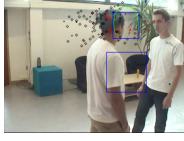
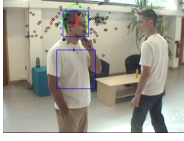
Data fusion strategy	$t = 15.$	$t = 81$	$t = 126$	$t = 284$
(a) $q(x_k x_{k-1}, z_k) = p(\cdot)$ $p(z_k \mathbf{x}_k) = p(z_k^c \mathbf{x}_k)$				
(b) $q(x_k x_{k-1}, z_k) = p(\cdot)$ $p(z_k \mathbf{x}_k) = p(z_k^s \mathbf{x}_k).p(z_k^c \mathbf{x}_k)$				
(c) $q(x_k x_{k-1}, z_k) = \alpha\pi(\cdot) + \beta p(\cdot)$ with face detection $p(z_k \mathbf{x}_k) = p(z_k^s \mathbf{x}_k).p(z_k^c \mathbf{x}_k)$				
(d) $q(x_k x_{k-1}, z_k) = \alpha\pi(\cdot) + \beta p(\cdot)$ with face classification $p(z_k \mathbf{x}_k) = p(z_k^s \mathbf{x}_k).p(z_k^c \mathbf{x}_k)$				

Table 3. Four different data fusion strategies involved in importance sampling and measurement function.

the importance function in order to guide the particle sampling on specific additional areas of the current image (mainly on detected faces). We can see that this strategy is not sufficient to distinguish whether the template is on the right targeted person or not. The last run in Table 3(d) shows the complete system used in our experiments involving the face classification process in the importance function as described in (9). We can see, at time  $t = 81$ , that after a sporadic occlusion of the target by another person (with the black trousers), the face classification helps to direct the particle sampling only on the desired person and so helps the template to recover the target.

#### 4.4. Evaluations and results

The above tracker has been prototyped on a 1.8GHz Pentium Dual Core using Linux and the OpenCV library. Both quantitative and qualitative off-line evaluations on sequences are reported below. This database of two different sequences (800 images) acquired from our Jido mobile robot in a wide range of realistic conditions allows us to: (i) determine the optimal parameter values of the tracker, (ii) identify its strengths and weaknesses, and in particular characterize its robustness to environmental artifacts: clutter, occlusion or out-field of sight, lighting changes. Several filter runs per sequence are performed and analyzed.

Quantitative performance evaluations summarized below have been carried out on the sequence database. Since the main concern of tracking is the accuracy of the tracker results, location as well as face label, we compare the tracking performance quantitatively by defining the False Position Rate (FPR) and the False Label Rate (FLR). If the tracker locks onto none of the observed person, this is considered as a position failure while a tracker lock onto the non-desired person is considered as a label failure.

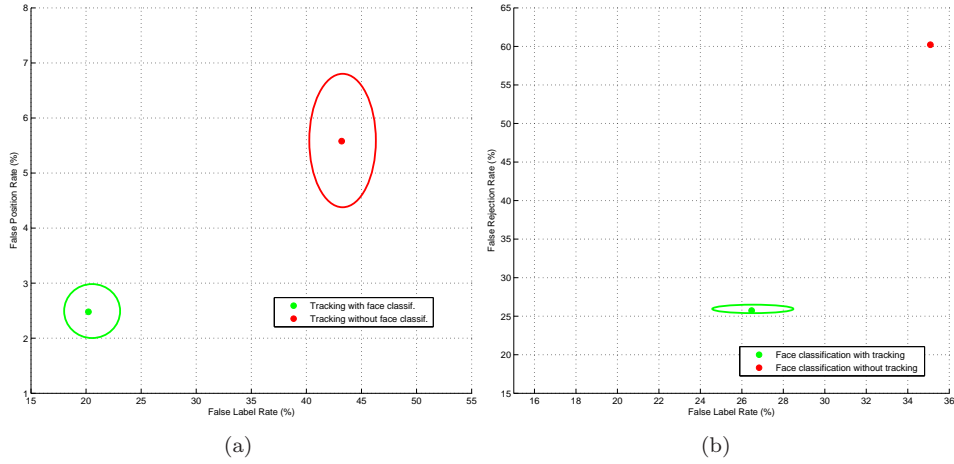


Figure 8. (a) Face tracker performance for the whole sequence database. (b) Face classification performance for the database image subset involving detected frontal faces.

Figure 8(a) presents the performance considering or not the FR in the tracking loop whereas Figure 8(b) considers the FR performance with or without tracking. Our advanced tracker is shown to outperform the conventional tracker (without FR) with much lower false position and label rates for slight additional time consumption. In Figure 8(a), we note that the estimate of the position of the targeted person is more precise when the tracking loop is fed by the FR results. In this case, the average FPR is reduced from 5.58% to 2.47% and the average FLR falls from 43.20% to 20.21%. These results have been processed on the basis of 10 runs of our tracker on each sequence due to stochastic context. The standard deviations of these results are represented by the ellipses on the graph. In the same vein, Figure 8(b) presents the classification results. For each sequence, these results are compared to tracking results in terms of FLR (or False Acceptance Rate) and FRR (False Rejection Rate). To be more consistent, the only images involving face detection have been taken into account. We note that the runs involving tracking are more robust to environmental changes, mainly due to spatiotemporal effects. More precisely, the FLR is decreased from 35.09% to 26.47% (with a standard deviation

of 1.97%) while the FRR is divided by more than 2 (60.22% against 25.73%). These evaluations prove that when combining all the above cues, both FR and tracking performances are considerably increased.

These results have been obtained for the empirically designed tracker parameter values listed in Table 4.

Symbol	Meaning	Value
$(\alpha, \beta)$	coeff. in the importance function $q(\mathbf{x}_k \mathbf{x}_{k-1}, z_k)$	(0.4, 0.6)
$(\sigma_u, \sigma_v, \sigma_s)$	standard deviation in random walk models	(40, 20, 0.2)
$(\kappa_{face}, \kappa_{skin})$	coeff. in the weighted sum $\pi(x_k^{(i)} z_k^1, \dots, z_k^L)$ defined in (6)	(0.8, 0.2)
$\kappa$	coeff. for reference histograms $h_{ref,1}^c, h_{ref,2}^c$ update in (10)	0.1
$\sigma_s$	standard deviation in shape-based likelihood $p(z^s \mathbf{x}_k)$	20
$\sigma_c$	standard deviation in color-based likelihood $p(z^c \mathbf{x}_k)$	0.2
$N_{bi}$	number of color bins per channel involved in $p(z^c \mathbf{x}_k)$	32

Table 4. Parameter values used in our face tracker.

## 5. Conclusion and future works

This paper has presented the development of a set of visual functions dedicated to Human/Robot interaction in a household framework used for face recognition. First, we propose a non-dominated sorting genetic algorithm to find the optimal free-parameters of a SVM-based face classifier in an optimized fashion. Besides, the second and main contribution is the design of a video-based face recognition process integrated through a particle filtering framework combining both intermittent features (face and skin blob detection, face recognition) and multiple persistent visual cues (shape and color) in a principled, robust and probabilistically motivated way.

Off-line evaluations on sequences acquired from the robot show that the overall system enjoys the valuable capabilities: (1) remaining locked on to the targeted person in populated and continuously changing environments, (2) recovering this person automatically after full occlusion or temporary disappearance from the field-of view. Eigenface subspace and SVM makes it possible to improve the face recognition process while the multi-cue fusion in the tracking loop is proven to be more robust than any of the individual cues. Clothing color and also face classification probabilities increase tracker reliability in presence of several persons in the vicinity of the robot. Finally, we have integrated this advanced tracker into a mobile robot companion called Jido. The visual-based tracker was then successfully tested in Jido's long-term operations in natural settings. To the best of our knowledge, quite few mature robotic systems enjoy such scalable human perception capabilities.

Several directions are studied regarding our video-based face recognition. A first line of investigation concerns the fusion of heterogeneous information such as RFID

or sound cues. Detection of an RFID tag worn by individuals will allow us to direct the camera using a pan-tilt unit and thus trigger tracker initialization, and will contribute as another measurement in the tracking loop. The sound cue will endow the tracker with the ability to switch its focus between speakers. Then, we aim to adapt our tracker in order to be able to recognize and track multiple persons simultaneously. In the same vein as our previous developments, we will consider distributed Bayesian multiple-target trackers based on particle filtering <sup>36</sup>.

### Acknowledgment

The work described in this paper was partially conducted within the EU STREP Project CommRob funded by the European Commission Division FP6 under Contract FP6-045441.

### Bibliography

1. A.F. Abata, M. Nappi, D. Riccio, and G. Sabatino. 2D and 3D face recognition: a survey. *Pattern Recognition Letters*, 28(14):1885–1906, 2007.
2. Y. Adini, Y. Moses, and S. Ullman. Face recognition: the problem of compensating for changes in illumination direction. *Trans. on Pattern Analysis Machine Intelligence (PAMI'97)*, 19(7):721–732, 1997.
3. G. Aggarwal, A. Roy-Chowdhury, and R. Chepalla. A system identification approach for video-based face recognition. In *Int. Conf. on Pattern Recognition (ICPR'04)*, Cambridge, UK, August 2004.
4. F. Aherne, N. Thacker, and P. Rockett. The bhattacharyya metric as an absolute similarity measure for frequency coded data. *Kybernetika*, 32(4):1–7, 1997.
5. 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.
6. M. Bartlett, J. Movellan, and T. Sejnowski. Face recognition by independent component analysis. *Trans. on Neural Networks*, 13(6):1450–1464, 2002.
7. P. Belhumeur, J. Hespanha, and D. Kriegman. Eigenfaces vs. fisherfaces. In *European Conf. on Computer Vision (ECCV'96)*, pages 45–58, 1996.
8. S. Birchfield. Elliptical head tracking using intensity gradients and color histograms. In *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'98)*, pages 232–237, Santa Barbara, USA, 1998 1998.
9. Z. Biuk and S. Loncaric. Face recognition from multi-pose image sequence. In *Int. Symp. on Image and Signal Processing and Analysis (ISPA'01)*, pages 319–324, 2001.
10. M. Boardman and T. Trappenberg. A heuristic for free parameter optimization with SVM. In *Int. Joint Conf. on Neural Networks (IJCNN'06)*, pages 610–617, Pula, Croatie, June 2006.
11. O. Chapelle, V. Vapnik, O. Bousquet, and S. Mukherjee. Choosing multiple parameters for support vector machines. *Machine Learning*, 46(1):131–159, 2002.
12. T. Choudhury, B. Clarkson, T. Jebara, and A. Pentland. Multimodal person recognition using unconstrained audio and video. In *Int. Conf. on Audio- and Video-based Person Authentication*, pages 176–180, 1999.
13. 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.



14. 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.
15. T. Fong, I. Nourbakhsh, and K. Dautenhahn. A survey of socially interactive robots. *Robotics and Autonomous Systems*, 42:143–166, 2003.
16. D.M. Gavrila and S. Munder. Multi-cue pedestrian detection and tracking from a moving vehicle. *Int. Journal of Computer Vision (IJCV'07)*, 73(1):41–59, 2007.
17. 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, November 2007.
18. T. Germa, F. Lerasle, P. Danès, and L. Brèthes. Human/robot visual interaction for a tour-guide robot. In *Int. Conf. on Intelligent Robots and Systems (IROS'07)*, San Diego, USA, November 2007.
19. N.J. Gordon, D.J. Salmond, and A.F.M. Smith. Novel approach to nonlinear/non-gaussian bayesian state estimation. *Radar and Signal Processing, IEE Proceedings F*, 140(2):107–113, Avril 1993.
20. A. Hadid and Pietikäinen. From stil image to video-based face recognition: an experimental analysis. In *Int. Conf. on Face and Gesture Recognition (FGR'04)*, pages 813–818, Seoul, Korea 2004.
21. B. Heisele, P. Ho, and T. Poggio. Face recognition with support vector machines. In *Int. Conf. on Computer Vision (ICCV'01)*, pages 688–694, July 2001.
22. M. Isard and A. Blake. CONDENSATION – conditional density propagation for visual tracking. *Int. Journal on Computer Vision*, 29(1):5–28, 1998.
23. 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.
24. M. J. Jones and J. M. Rehg. Statistical color models with application to skin detection. In *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'00)*, pages 274–280, 1999.
25. Jonsson K., Matas J., Kittler J., and Li. Y. Learning support vectors for face verification and recognition. In *Int. Conf. on Face and Gesture Recognition (FGR'00)*, pages 208–213, Grenoble, France, March 2000.
26. Lam K. and Yan H. An analytic-to-holistic approach fo face recognition based on a single frontal view. *Trans. on Pattern Analysis Machine Intelligence (PAMI'98)*, 7(20):673–686, 98.
27. Seo K. A GA-based feature subset selection and parameter optimization of SVM for content-based image retrieval. In *Int. Conf. on Advanced Data Mining and Applications (ADMA'07)*, pages 594–604, Harbin, China, August 2007.
28. S. Kong, J. Heo, B. Abidi, J. Paik, and M. Abidi. Recent advances in visual and infrared face recognition - a review. *Computer Vision and Image Understanding (CVIU'05)*, 97(1):103–135, 2005.
29. K. Lee, J. Ho, M. Yang, and D. Kriegman. Video-based face recognition using probabilistic appearance manifolds. *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, 1:I–313–I–320 vol.1, June 2003.
30. K.H Lin, Lam K.M., and Siu W. Spatially eigen-weighted Hausdorff distances for human face recognition. *Pattern Recognition (PR'03)*, 36(8):1827–1834, 2003.
31. X. Liu and T. Chen. Video-based face recognition using adaptative hidden markov models. In *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'03)*, Madison, USA, June 2003.
32. Quintiliano P., Santa-Rosa A., and Guadagnin R. Face recognition based on eigen-features. In *SPIE: Image extraction, segmentation and recognition*, pages 140–145,

22 Germa et al.

- 2001.
33. Y. Pang, Z. Liu, and N. Yu. A new nonlinear extraction method for face recognition. *Neurocomputing*, (69):949–953, 2006.
34. P. Pérez, J. Vermaak, and A. Blake. Data fusion for visual tracking with particles. *Proc. IEEE*, 92(3):495–513, 2004.
35. F. Provost and T. Fawcett. Robust classification for imprecise environments. *Machine Learning*, 42(3):203–231, 2001.
36. W. Qu, D. Schonfeld, and M. Mohamed. Distributed bayesian multiple-target tracking in crowded environments using multiple collaborative cameras. *EURASIP Journal on Advances in Signal Processing*, 2007.
37. McKenna S. and H. Nait-Charif. Tracking of human motion using auxiliary particle filters and iterated likelihood weighting. *Image and Vision Computing (IVC'07)*, 25(6):852–862, 2007.
38. S. Shan, W. Gao, and D. Zhao. Face recognition based on face-specific subspace. *Int. Journal of Imaging Systems and Technology*, 13(1):23–32, 2003.
39. M. Spengler and B. Schiele. Towards robust multi-cue integration for visual tracking. *Machine Vision and Applications (MVA'03)*, 14:50–58, 2003.
40. M.A. Turk and A.P. Pentland. Face recognition using eigenfaces. In *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'91)*, pages 586–591, 1991.
41. 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.
42. P. Viola and M. Jones. Fast multi-view face detection. In *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'03)*, 2003.
43. Yan Wang, Yanghua Liu, Linmi Tao, and Guangyou Xu. Real-time multi-view face detection and pose estimation in video stream. *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, 4:354–357, 0-0 2006.
44. T. Wu, C. Lin, and R. Weng. Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research*, 5:975–1005, 2004.
45. L. Xu and C. Li. Multi-objective parameters selection for SVM classification using NSGA-II. In *Industrial Conference on Data Mining (ICDM'06)*, pages 365–376, 2006.
46. H. Yang, X. Jiao, L. Zhang, and F. Li. Parameter optimization for SVM using sequential number theoretic for optimization. In *Int. Conf. on Machine Learning and Cybernetics*, Dalian, August 2006.
47. Y. Zhang and A. Martinez. A weighted probabilistic approach to face recognition from multiple images and videos sequences. *Image and Vision Computing (IVC'03)*, 24(6):626–638, 2006.
48. W. Zhao and R. Chellappa. Discriminant analysis of principal components for face recognition. In *Int. Conf. on Computer Vision and Pattern Recognition (CVPR'98)*, pages 336–341, Santa Barbara, USA, 1998 1998.
49. W. Zhao, R. Chellappa, P.J. Phillips, and A. Rosenfeld. Face recognition: a literature survey. *ACM Computing Surveys*, 35(4):399–458, 2000.
50. S. Zhou, R. Chellappa, and B. Moghaddam. Visual tracking and recognition using appearance-adaptive models in particle filters. *Trans. on Image Processing*, 13(11):1491–1506, November 2004.
51. S. Zhou, V. Krueger, and R. Chellappa. Probabilistic recognition of human faces from videos. *Computer Vision and Image Understanding (CVIU'03)*, 91:214–245, 2003.