Towards safe human-robot interaction in robotic cells: an approach based on visual tracking and intention estimation

Luca Bascetta, Gianni Ferretti and Paolo Rocco
Håkan Ardö
Herman Bruyninckx, Eric Demeester and Enrico Di Lello

Abstract— **Removing the safety fences that separate humans and robots, to allow for an effective human-robot interaction, requires innovative safety control systems. An advanced functionality of a safety controller might be to detect the presence of humans entering the robotic cell and to estimate their intention, in order to enforce an effective safety reaction. This paper proposes advanced algorithms for cognitive vision, empowered by a dynamic model of human walking, for detection and tracking of humans. Intention estimation is then addressed as the problem of predicting online the trajectory of the human, given a set of trajectories of walking people learnt offline using an unsupervised classification algorithm. Results of the application of the presented approach to a large number of experiments on volunteers are also reported.**

I. INTRODUCTION

Human-robot interaction, a key feature for the innovative robotic cell, requires the elimination of the safety fences that in the traditional industrial scenario provide a rigid separation between the areas occupied by the robots and those occupied by the humans. This lack of artificially imposed safety, however, must be compensated for by new abilities of the control system. A new functionality of the safety controller might then be to detect and track the humans entering the cell and to estimate, to some extent, their intentions.

While different meanings can be associated to the concept of human intention, and consequently several ways to classify this intention can be devised, in this context we will focus on the prediction of the trajectory a human is supposed to follow, after he/she has been tracked for a sufficient amount of time. The overall goal is, in fact, to predict in the least possible time which area in the vicinity of the robot the human is heading to. Using this information, the robot control system can be aware of the actual situation and select the correct interaction mode (among those defined in its design) as soon as an intention has been reliably identified, thus increasing the chance that the safety actions conceived for the interaction mode are effective to resolve a possibly dangerous situation.

The problems of human detection and tracking, and of intention estimation, have been tackled before in the literature, from quite different points of view and with different instrumentation. Since a thorough review of the state-of-the-art of these problems in all different application domains is out of the scope of this work, we will focus on the robotic context only.

An approach based on stereo-vision, that aims at predicting the probability of an accident in working environments where human operators and robots cooperate, was presented in [1]. The methodology is based on a dynamic stochastic model of the human motion, where the human is modeled as a moving point. Another probabilistic approach, based on camera images, was illustrated in [2]. In this case, however, intentions are represented as complex manipulation operations and no motion prediction is considered.

A similar modeling approach, based on Hidden Markov Models, has been adopted in [3], but in this case a laser range finder for 3D position estimation was added to a vision system. Other approaches, that combine vision and the measurement of some psychological signals for human intent and affective state estimation during robot interaction, are presented in [4] and [5].

Finally, human intention estimation in the context of human-robot physical interaction has also been considered in [6]–[8].

With respect to the reported state-of-the-art, the approach presented in this paper has some distinctive features:

- the sensory system used: just a couple of ceiling-mounted commercial surveillance cameras, low-cost sensors easily deployable in any robotic cell;
- the particular setting of the problem, where the goal is to infer in the least possible time the area inside the robotic cell the human is heading to;
- the kinematic model adopted to describe human motion, which takes into account human orientation and does not allow for walking sideways, a situation that is rather uncommon in an industrial robotic cell.

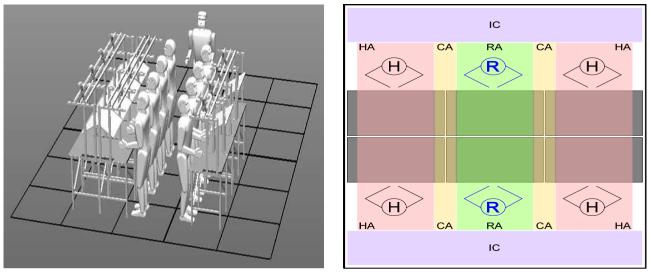
The paper is organised as follows. Section II presents the problem of estimating the human intention in an industrial robotic cell, introducing the framework that characterises the approach presented here. Sections III and IV respectively

The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 - Challenge 2 - Cognitive Systems, Interaction, Robotics - under grant agreement No. 230902-ROSETTA.

L. Bascetta, G. Ferretti and P. Rocco are with Dipartimento di Elettronica e Informazione, Politecnico di Milano, Piazza L. Da Vinci 32, 20133, Milano, Italy ({bascetta, ferretti, rocco}@elet.polimi.it).

H. Ardö is with Department of Mathematics, Lund University, Lund, 22100, Sweden. (ardo@maths.lth.se).

H. Bruyninckx, E. Demeester and E. Di Lello are with the Department of Mechanical Engineering, Katholieke Universiteit Leuven, Belgium. ({eric.demeester, Enrico.DiLello}@mech.kuleuven.be).



(a) (b)

Fig. 1. An assembly line with several manual workstations (a) and an example of area segmentation for human intention estimation (b).

describe the algorithms for human detection and tracking, and for intention estimation. Section V illustrates the results of the application of the presented approach to a large number of experiments with volunteers. Conclusions are drawn in Section VI.

II. ESTIMATING THE HUMAN INTENTION IN AN INDUSTRIAL ROBOTIC CELL

Consider an assembly line (Fig. 1(a)) where humans and robots coexist side-by-side, and even cooperate, without any safety fence providing a rigid separation between the areas occupied by the robots and those occupied by the humans. Three different modes of human-robot interaction can be introduced: coexistence, cooperation and interference. These three modes are defined as follows:

- coexistence: a robot and a human are working side-by-side in two contiguous workstations, or a human is passing in the inspection corridor while a robot is operating;
- cooperation: a robot and a human, working in two contiguous workstations, are exchanging workpieces;
- interference: a human is entering the robot workspace.

Following the idea of stating the human intention estimation problem as a prediction of the area to which the human is heading, the robotic cell is segmented into four different areas (Fig. 1(b)) as follows:

- Human worker Area (HA) and Robot Area (RA): rectangular regions spanning the space in front of a workstation occupied by a human worker or a robot, including the space needed by the worker or the robot to accomplish the task and a portion of the space behind them, as well.
- Inspection Corridor (IC): a rectangular region located behind the human worker and robot areas (a corridor that goes through the robotic cell, used, for example, by human workers to reach their workstations).
- Cooperation Area (CA): a rectangular aisle located between the human worker and robot areas, representing the space where a cooperation between a human worker and a robot can take place.

On the basis of the robotic cell segmentation in Fig. 1(b), the three interaction modes can be characterised with respect to the human's position/velocity, the position of the human arms and the robot Tool Centre Point (TCP) position/velocity as shown in Table I. As suggested by this table, to perform intention estimation during close cooperation the human arms' position have to be tracked as well. A ceiling mounted surveillance camera, however, is not suitable to track small features such as human arms, moving at rather high speed, with the required precision. For this reason, the present paper will focus on intention estimation of a walking human. Possible extensions of the proposed methodology to intention estimation during close human-robot interaction will be considered in the future.

III. HUMAN DETECTION AND TRACKING

The human detection and tracking algorithm relies on a simplified model of the world, including some prior knowledge about walking humans (e.g. the fact that humans walk on the floor, enter the scene from its borders, etc.) as will be explained further on.

This simplified model is designed as a state machine, where each state represents a configuration of objects in the scene. The transitions between different states represent object movements, as well as the event of objects entering or leaving the scene.

The task of human detection and tracking can then be formalised as follows: given a set of observed images of the scene, deduce a likely state sequence that could have produced those images.

In this paper, we propose an approach where the images observed by several ceiling mounted cameras are first processed using a foreground/background segmentation, and then interpreted in terms of a state sequence in the state machine model of the world.

A. Foreground/background segmentation

A static camera viewing the scene produces one image for each time step, where most of the pixels belong to a static or pseudo-static background. The task of the foreground/background segmentation algorithm is to segment out the foreground, i.e. moving objects that represent the features of interest in the scene, from the background.

The foreground/background segmentation is an important step for many video surveillance applications, being thus a well studied problem in the context of computer vision [9]. For the present application the segmentation algorithm, besides being robust, should be causal and computationally efficient, in order to be implementable on a realtime system. Two OpenCV [10] advanced algorithms, fulfilling the aforementioned requirements, were thus selected: one based on colour and colour co-occurrence features [11], another using a Gaussian mixture model with shadow detection [12] that allows the background model to be multi-modal and to take object colours into account in the segmentation process.

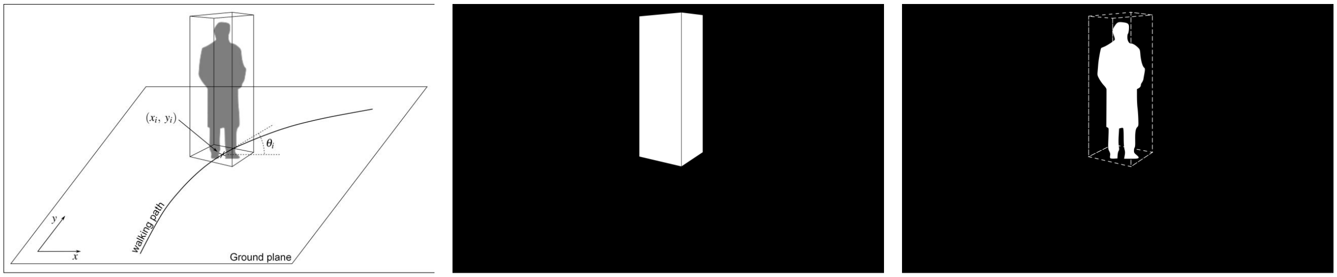
B. The observation model

To explain the segmented images observed in the camera frames, a simplified model of the world is used, assuming that the scene consists of a flat ground plane on which

Interaction modes	Human position	Human velocity	Arm position	TCP position	TCP velocity
Co-existence	belongs to IC	towards HA			
	belongs to IC	parallel to IC			
	belongs to HA	zero	belongs to HA		
Co-operation	belongs to IC	zero	belongs to CA	belongs to CA	
				belongs to RA	towards CA
	belongs to HA	zero	belongs to CA	belongs to CA	
				belongs to RA	towards CA
Interference	belongs to IC	towards RA			
	belongs to RA				
	belongs to HA	towards RA			
	belongs to HA	zero	belongs to RA		

TABLE I

CHARACTERISATION OF INTERACTION MODES.



(a) (b) (c)

Fig. 2. Left: illustration of the world state in the case of a single human located at pose (x_i, y_i, θ_i). Middle: the expected observation given the current world state. Right: the true noise-free observation.

humans walk around.

A walking human is represented by a rectangular box (Fig. 2(a)), that can translate and rotate around an axis parallel to the vertical dimension of the box, and crossing the base in its centre. The pose of the i-th human in the scene is thus completely described by the coordinate of the rectangular box base centre (x_i, y_i), with respect to a reference frame fixed on the ground plane, and by the angle θ_i formed between the tangent to the walking path and the x-axis. To produce complete trajectories with many people in the scene, however, an integer number l_i, that univocally identifies the i-th human, is added as well.

To keep the world model simple and efficient, only a few different types of humans, each one characterised by different (but fixed and a priori known) box dimensions, are considered. Summarising, the state of this simplified world model is represented by a configuration of humans, and is specified by the number N of humans present in the scene and, for each of those humans, by a pose $s_i = (x_i, y_i, \theta_i)$ and an identifier l_i, for $i = 1, \ldots, N$.

Given a set of possible states, an expected segmentation image can be generated, by rendering the box model, and comparing it with the observed foreground (Fig. 2). How similar the rendered and measured foreground are depends on how precisely the box model approximates a human, and how many small objects, that will be considered noise, appear in the scene foreground (more details can be found in [13]).

The previous statement, however, is based on the assumption that humans are the unique moving objects in the camera field of view. If this is not true for some parts of the image, such as where robots are operating, those parts can be masked out, by rendering a 3D model of the robot in its current pose. A simpler and more efficient approach is to place a static box covering the entire working area of each robot. The system will be blind in the masked area but, by using several cameras viewing the scene from different angles, it will still be able to detect and track humans within the entire scene.

Whatever the level of detail of the selected observation model is, the idea of generating a virtual world representation, based on the actual estimate of the world state, and comparing this representation with the measured foreground to update the state estimate, is the key aspect of the human detection and tracking algorithm.

C. Human tracking with a particle filter

The presence of noise and imprecisions in the segmented image, the roughness of the observation model, the possibility that different sets of states yield similar virtual foreground images, do not allow to deterministically evaluate the exact state of the world at each time instant. Instead, a probability distribution over all the possible states is maintained, in the form of a set of weighted particles that are propagated forward in time – namely, a particle filter. This forward propagation can be based on a stochastic motion model that describes how the states typically evolve or, as it will be explained in Section III-D, on a simplified (but deterministic) kinematic model of a walking human.

The distribution of the multi-object state q_t at time t, given the observations up to that time, $O_{0,\ldots,t}$, is represented by a set $S_t = \left\{q_t^{(1)}, q_t^{(2)}, \ldots, q_t^{(p)}\right\}$ of p samples called particles. Each particle $q_t^{(j)}$, in turn, represents a set of $Q_t^{(j)}$ single object states

$$q_t^{(j)} = \left\{(x_i, y_i, \theta_i, l_i) \,\middle|\, i = 1, \ldots, Q_t^{(j)}\right\}$$

Then, the initial distribution S_0 is assumed to be known, and equivalent to a scene that is empty with probability 1. The effect of this initial assumption, however, is only local in time, since the background model is updated. Even if there were humans in the scene at time $t = 0$, the algorithm will converge to the correct distribution once they move out of the scene or close to the borders.

The algorithm propagates the distribution forward in time, from frame to frame. In particular, the distribution of the previous frame, S_{t-1}, is propagated by sampling some stochastic state transfer function h that models the motion of the humans (see Section III-D), for each of the particles in S_{t-1}. The set of transformed particles forms a prediction distribution

$$S_t^- = \left\{q_t^{-(j)} = h\left(q_t^{(j)}\right) \,\middle|\, j = 1, \ldots, p\right\}$$

representing the distribution of the predicted state q_t^- at time t, given the observations up to time $t-1$. Each of the elements of S_{t-1} is then compared to the measured foreground, as described in Section III-B, yielding a weight $\alpha_t^{(j)}$ that describes how likely the state $q_t^{-(j)}$ is, given the current foreground.

Finally, a new distribution S_t is generated by randomly choosing a single sample from S_t^- p times, where the probability of choosing a particle corresponds to its weight, i.e. the particle $q_t^{-(j)}$ is chosen with probability $\alpha_t^{(j)}/\bar{\alpha}_t^{(j)}$ where

$$\bar{\alpha}_t^{(j)} = \sum_{j=1}^{p} \alpha_t^{(j)}$$

is a normalisation factor.

Note that the same sample might be chosen several times, in which case S_t will contain several identical particles.

To generate single object measurements, the multi-object distribution S_t is simplified, extracting, for each identifier l, the state of the corresponding human and the probability that he/she is present in the current frame.

Given the set $\mathcal{Q}_l$ of the single object states with id l, found within the multi-object particles $q_t^{(j)}$, i.e.

$$\mathcal{Q}_l = \left\{(x_i, y_i, \theta_i) \,\middle|\, (x_i, y_i, \theta_i, l_i) \in \cup_j q_t^{(j)} \text{ and } l_i = l\right\}$$

the expected human state $\hat{s}_l$ is given by

$$\hat{s}_l = \frac{1}{|\mathcal{Q}_l|} \sum_{s_i \in \mathcal{Q}_l} s_i$$

and the probability he/she is in the current frame by $\mathcal{Q}_l/p$.

D. A simplified kinematic model for walking humans

To propagate the particles a simplified kinematic model, describing a walking human, can be adopted. In particular, in this paper the extended unicycle model proposed by [14] is considered

$$\begin{cases} \dot{x}_i = v_i \cos(\theta_i) \\ \dot{y}_i = v_i \sin(\theta_i) \\ \dot{\theta}_i = \kappa_i v_i \\ \dot{\kappa}_i = \phi_i \end{cases}$$

where $(x_i,\, y_i,\, \theta_i)$ represents the pose of the i-th human, v_i is the linear (nonholonomic) velocity along the direction of motion, κ_i is the curvature and ϕ_i is the derivative of the curvature, respectively.

Consider now the forward velocity v_i. In principle it varies with time along the path and depends on a large number of factors (see [15] and [16]), but it has been also demonstrated that for short paths it can be considered constant (see e.g. [14]). Holding this assumption, the previous model can be rewritten in terms of the natural coordinate as follows

$$\begin{cases} x'_i = \cos(\theta_i) \\ y'_i = \sin(\theta_i) \\ \theta'_i = \kappa_i \\ \kappa'_i = \phi_i / v_i \end{cases} \tag{1}$$

where the notation $'$ represents the derivative with respect to the natural coordinate.

IV. INTENTION ESTIMATION

Our intention estimation algorithm for walking humans consists of two phases: offline trajectory classification and learning, and online interaction area prediction. The former is based on an unsupervised classification algorithm, which uses the Expectation Maximisation (EM) technique (see [17] for further details). The latter concerns the realtime prediction of the area to which each human is heading, of the time at which the area will be reached, and of the probability associated to this prediction.

A. Offline unsupervised human path classification

To classify the experimental trajectories recorded by the human tracking algorithm into a small set of motion patterns, the technique presented in [17] was used. The recorded data, represented by a set of paths (i.e. sequences of Cartesian coordinates of the base centre of the box representing a human) of different length in time and space, are classified into a set $\Theta = \{\theta_1, \ldots, \theta_M\}$ of M patterns. Each motion pattern θ_m is represented by a sequence of Gaussian distributions, i.e. a set $\theta_m = \left\{\theta_m^1, \ldots, \theta_m^K\right\}$ of K Gaussian distributions, where the k-th Gaussian θ_m^k is represented by a Cartesian coordinate μ_m^k and its covariance Σ_m^k.

The covariance matrices Σ_m^k and the number K of Gaussian distributions are manually chosen, making a compromise between the desired prediction accuracy of where people will be at a certain future time instant and the inherent noise on the walking trajectories. The same standard deviation σ is chosen for all M motion patterns: $\Sigma_m^k = \sigma^2 I_2 \, \forall m, k$. The lower the standard deviation is chosen to be, the higher the number of model trajectories will be.

The classification algorithm, based on the EM technique, learns in an unsupervised way the number of model trajectories M and the means μ_m^k of the Gaussian distributions.

B. Online path and interaction area prediction

The area prediction is performed online in two steps. First, the trajectory that the human is following is estimated based on the learnt motion patterns Θ. Second, the area that the human is headed to is predicted based on the intersection between the predicted path and the interaction areas (co-existence, cooperation, or interference) described in Section II.

In the first step, for each observed human, online trajectory prediction is performed, using a Hidden Markov Model (HMM) with a number of K discretised states[1] μ_m^k located along the M motion patterns. The human forward velocity is also estimated through a suitable Kalman filter. Together with knowledge about the followed motion pattern, this allows to compute when the person will enter a certain area.

The probability distribution over the discrete HMM states along the motion patterns is updated as follows. In a prediction step, the probability of each discretised state is propagated along the motion pattern, taking into account the estimated walking velocity of the person. This results in a probability distribution with probabilities over states $\mathbf{p}^{-}\left(\mu_m^k\right)$. In an observation step, the probability of each discretised state is updated based on the distance between the state and the position of the observed human.

Given a HMM state located at μ_m^k and the observed position of the walking human $\pi_i = (x_i, y_i)$, the probability of this state is updated as

$$\mathbf{p}\left(\mu_m^k|\pi_i\right) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-\dfrac{1}{2\sigma^2}\left\|\pi_i-\mu_m^k\right\|^2} \cdot \mathbf{p}^{-}\left(\mu_m^k\right).$$

After this, the probabilities of all HMM states are normalised such that the sum of all probabilities equals one.

Each time a human is detected for the first time, the Hidden Markov Model is initialised with a uniform distribution.

In the second step, the probability over interaction areas is computed. Suppose that we want to predict the probability that interaction area A_j is visited by the i-th person at a certain time t in the (near) future. This is expressed as $\mathbf{p}(A_j, t|\pi_i)$. By applying the total probability theorem and Bayes' rule, probability $\mathbf{p}\left(A_j, t|\pi_i\right)$ can be expressed as

$$\begin{aligned}\mathbf{p}(A_j, t|\pi_i) &= \sum_{m,k} \mathbf{p}(A_j, t, \mu_m^k|\pi_i) \\ &= \sum_{m,k} \mathbf{p}(A_j, t|\mu_m^k, \pi_i)\mathbf{p}(\mu_m^k|\pi_i)\end{aligned}$$

In this expression, probability $\mathbf{p}(\mu_m^k|\pi_i)$ is known, since it is the HMM state that is updated each time a new observation is available. Probability $\mathbf{p}(A_j, t|\mu_m^k, \pi_i)$ is modeled to equal 1 if the person, starting from μ_m^k, is predicted to be inside area A_j at time t, using the same prediction model that is used to update the HMM states. If the person is predicted not to be inside area A_j at time t, probability $\mathbf{p}(A_j, t|\mu_k^m, \pi_i)$ equals 0.

Notice that, the probability distribution $\mathbf{p}(A_j, t|\pi_i)$ can at certain times be a multi-modal distribution over the interaction areas.

V. EXPERIMENTAL RESULTS

As a proof of concept of the feasibility of the proposed methodology, an experimental scenario was set up (Fig. 3), that resembles to some extent the one described in Section II[2]. An environment of approximately $3 \times 2.5\,m$, delimited by walls and fences and with a single entrance, was selected. An ABB IRB140 robot is placed in the centre of this space, surrounded by an interference area (red area marked with number 3 in Fig. 3) that covers the workspace of the selected task. Two tables, representing workstations, were added at the left and right side of the robot, each one with a corresponding cooperation area (blue areas marked with number 2 in Fig. 3).

Finally, two more areas, that are far from the robot and are intended to be coexistence areas (blue areas crossed with a red line and marked with number 1 in Fig. 3), were added as well.

The robotic cell was equipped with two AXIS 212 ceiling mounted surveillance cameras. The two cameras were suspended at about $3\,m$ and located at a distance that ensures a complete overlap on the interference area. The zoom factor and the pan/tilt settings were selected according to this requirement, as well. The acquisition rate was fixed at $30\,fps$, prioritising the frame-rate in case of low light. A multi-threaded software architecture optimised for an 8-cores Intel® i7 processor allows to execute the algorithm with a cycle time that ensures to exploit the maximum camera acquisition rate.

A set of five volunteers was selected to perform five different experiments of human detection and intention estimation in the robotic cell previously described. Each experiment is structured according to the following steps:

1) the volunteer enters the robotic cell from a door located in the bottom-right corner;
2) the volunteer steps towards a preconceived destination (one among five of the interaction areas previously described);
3) the volunteer stops at the destination and performs a simple task;
4) the volunteer comes back to the entrance door;
5) the volunteer leaves the robotic cell.

An experimental protocol, including a thorough description of each experiment and a detailed set of instructions for the

[1] K may be chosen to be different from the number of Gaussian distributions learnt in the offline classification phase.

[2] Though the size of the cell and the layout of the interaction areas are only similar to the environment depicted in Fig. 1, the experimental scenario here considered (Fig. 3) is even more general and complex, and thus more suitable to demonstrate the effectiveness of the proposed approach.

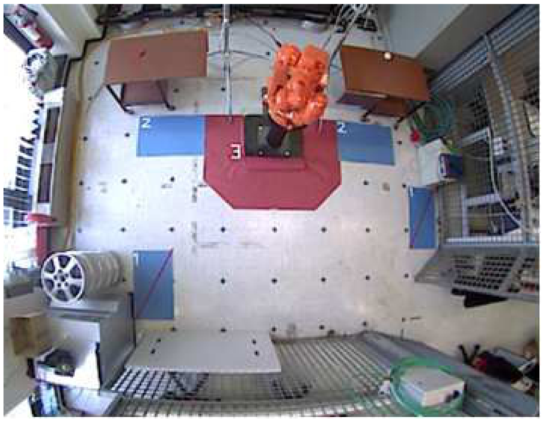


Fig. 3. A top view of the robotic cell environment.

volunteers, was also prepared (the protocol is not reported here due to space limitations). According to this protocol, five healthy volunteers were selected, characterised by different heights and wearing different but common clothes.

Fifty experiments, two for each volunteer and interaction area, were run to test the functionality of the human detection and intention estimation algorithms, following the protocol described above. Snapshots taken during three such experiments are shown in Fig. 4, for an approach towards a coexistence area, a cooperation area, and an interference area, respectively. Notice that, throughout all the experiments the subject is correctly detected and tracked by the system (the bounding box correctly follows the subject). Also notice that, though the robot was moving during the experiments, its motion is masked by the system in order to avoid any misinterpretation of this motion as coming from a moving person.

A subset of the walking trajectories obtained from these experiments (2D plots of the estimated trajectories, for the same experiments as in Fig. 4, are shown in Fig. 8) was then used to learn, in an unsupervised way, the different motion patterns for the environment in Fig. 3. For this, the algorithm described in Section IV-A was used, adopting a standard deviation of $0.5\,m$ and representing each motion pattern with 12 Gaussian distributions.

During the experiment the intention estimation algorithm computes, at each sampling time, the probability that each interaction area is reached by the tracked person for a number of discrete future time instants (up to $10\,s$ in the future). For example, a 1-step ahead prediction, corresponding to $0.5\,s$ look ahead into the future, gives, at each time instant and for each of the interaction areas, the probability that the tracked human will be in that area $0.5\,s$ later.

Plots of the interaction area probabilities, generated with a 4-step ahead[3] prediction ($2\,s$ look ahead into the future), are shown in Figs. 5, 6, and 7, for each of the experiments in Fig. 4. Consider, for example, the coexistence experiment (first row of Fig. 4, and Fig. 5), i.e. the one in which the human walks towards the interaction area at the bottom left corner of the robotic cell. From Fig. 5 it follows that, when the human has walked for approximately $1.17\,m$ along his trajectory, corresponding to the $43\,\%$ of the trajectory length, the safety system is able to predict that the coexistence area at the bottom left corner of the robotic cell will be reached in $2\,s$.

We can thus conclude that in the experiments of coexistence, cooperation and interference, shown in Figs. 5, 6, and 7, the intention estimation algorithm is successful in predicting the correct interaction area before the first half of the path has been covered (the correct interaction area is predicted when the human has walked $43\,\%$, $25\,\%$ and $28\,\%$ of the path for the coexistence, cooperation and interference experiment, respectively). This aspect is made clear in Fig. 8, where a point is highlighted on each estimated path, corresponding to the position at which the probability associated to the correct interaction area is larger than 0.65.

From Figs. 5, 6, and 7, it is also evident that when a human is close to the entrance, trajectories heading to different interaction areas are so similar that cannot be easily distinguished. For this reason, it might happen that at the beginning of a trajectory more than one interaction probability is significantly greater than zero[4] (see e.g. Figs. 5 and 6). As a consequence, a decision threshold of 0.65 is adequate to predict the interaction area sufficiently in advance, without affecting the robustness of the approach.

Results have then been generalised to a large number of experiments on volunteers, in order to give statistical evidence of the reliability of the method. Out of 50 experiments performed with 6 volunteers, in 46 of them, corresponding to 92% of the total, the intention of the human is correctly recognised, meaning that the first interaction area whose probability reaches 0.65 is the correct one.

Finally, in the accompanying video seven different clips are presented, showing the results of the human detection and tracking and of the intention estimation algorithms. In the video, the predicted interaction area is marked, as soon as the associated probability becomes larger than the threshold, by a red tag.

VI. CONCLUSIONS

This work fits into the broader task of developing the part of a robot controller in charge of ensuring a certain degree of safety in the interaction between humans and industrial robots, even in the absence of protective fences. Estimation of human intention is an important ingredient of the safety strategy, as it allows the system to enter the correct interaction mode and then to anticipate the enforcement of the appropriate safety behaviour.

This paper has presented an approach to human intention estimation, where cognitive vision algorithms are used in

[3] Adopting a 4-step ahead prediction is a good compromise between performance and robustness. In a real application, however, the safety controller will select the best look ahead time to resolve a possibly dangerous situation on the basis of the interpretation of the situation.

[4] This issue is strongly related to the size and geometry of the considered robotic cell, making the intention estimation in this environment particularly challenging.

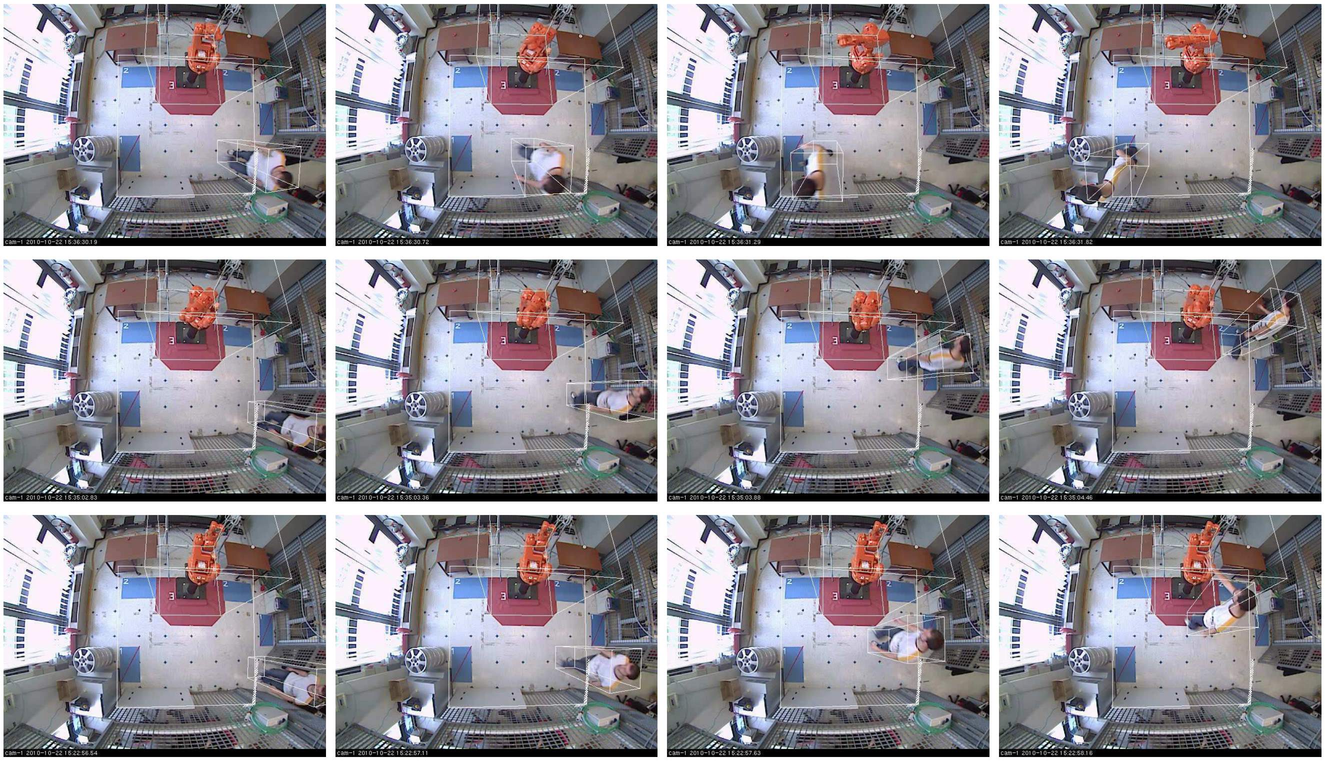

cam-1 2010-10-22 15:36:30.19
cam-1 2010-10-22 15:36:30.72
cam-1 2010-10-22 15:36:31.29
cam-1 2010-10-22 15:36:31.82
cam-1 2010-10-22 15:35:02.83
cam-1 2010-10-22 15:35:03.36
cam-1 2010-10-22 15:35:03.88
cam-1 2010-10-22 15:35:04.46
cam-1 2010-10-22 15:22:56.54
cam-1 2010-10-22 15:22:57.11
cam-1 2010-10-22 15:22:57.63
cam-1 2010-10-22 15:22:58.16

Fig. 4. A sequence of frames extracted from a coexistence experiment (first line), a cooperation experiment (second line), and an interference experiment (third line).

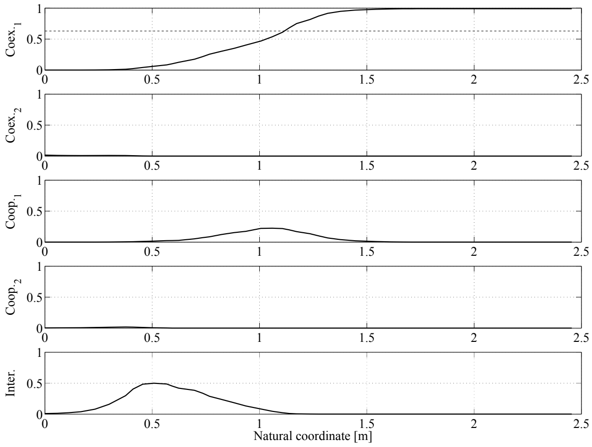


Fig. 5. Probabilities of the 5 interaction areas with respect to the natural coordinate for the coexistence experiment (the dashed line shows the threshold adopted for the prediction of the interaction area).

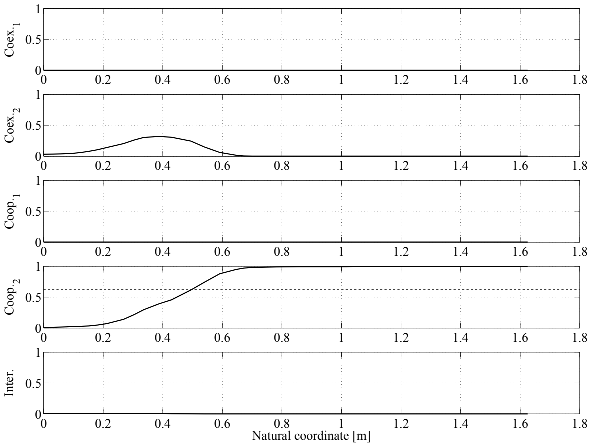


Fig. 6. Probabilities of the 5 interaction areas with respect to the natural coordinate for the cooperation experiment (the dashed line shows the threshold adopted for the prediction of the interaction area).

combination with statistical methods to estimate the probability of occupancy of a few areas in the robotic cell in future times. An experimental validation shows the practical validity of the method.

References

[1] R. Asaula, D. Fontanelli, and L. Palopoli, “Safety provisions for human/robot interactions using stochastic discrete abstractions,” in *Proc. IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 2175–2180.

[2] M. Awais and D. Henrich, “Human-robot collaboration by intention recognition using probabilistic state machines,” in *Proc. 19th IEEE International Workshop on Robotics in Alpe-Adria-Danube Region*, 2010, pp. 75–80.

[3] R. Kelley, A. Tavakkoli, C. King, M. Nicolescu, M. Nicolescu, and G. Bebis, “Understanding human intentions via hidden markov models in autonomous mobile robots,” in *Proc. 3rd ACM/IEEE International Conference on Human-Robot Interaction*, 2008, pp. 367–374.

[4] D. Kulic and E. Croft, “Pre-collision safety strategies for human-robot interaction,” *Autonomous Robots*, vol. 22, pp. 149–164, 2007.

[5] ——, “Affective state estimation for human–robot interaction,” *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 991–1000, 2007.

[6] J.-Y. Kuan, T.-H. Huang, and H.-P. Huang, “Human intention estima-

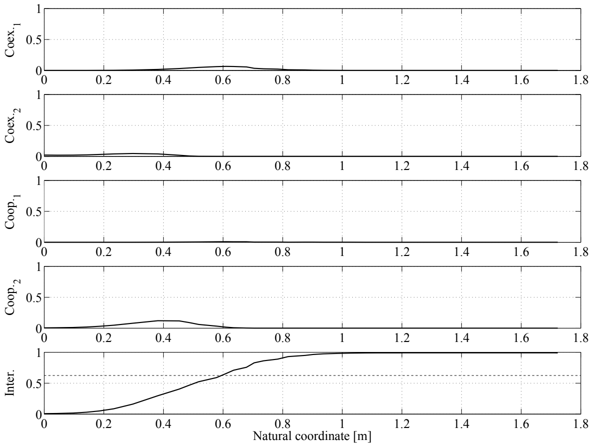


Fig. 7. Probabilities of the 5 interaction areas with respect to the natural coordinate for the interference experiment (the dashed line shows the threshold adopted for the prediction of the interaction area).

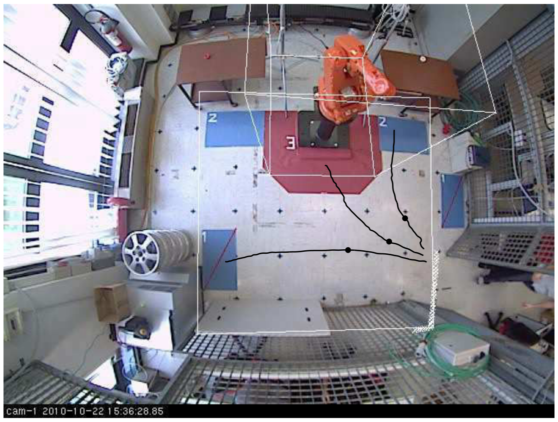


Fig. 8. The estimated 2D paths reported into the robotic cell environment with indication of the point corresponding to the guess of the correct interaction area.

tion method for a new compliant rehabilitation and assistive robot," in *Proc. SICE Annual Conference 2010*, 2010, pp. 2348–2353.

[7] T. Takeda, Y. Hirata, and K. Kosuge, "Dance step estimation method based on HMM for dance partner robot," *IEEE Transactions on Industrial Electronics*, vol. 54, no. 2, pp. 699–706, 2007.

[8] K. Kosuge, "Dance partner robot: An engineering approach to human-robot interaction," in *Proc. 5th ACM/IEEE International Conference on Human-Robot Interaction*, 2010, p. 201.

[9] H. Zhang, E. Fritts, and A. Goldman, "Image segmentation evaluation: A survey of unsupervised methods," *Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 260–280, 2008.

[10] OpenCV: http://opencv.willowgarage.com/wiki/.

[11] L. Li, W. Huang, I. Gu, and Q. Tian, "Foreground object detection from videos containing complex background," in *Proc. 11th ACM International Conference on Multimedia*, 2003, pp. 2–10.

[12] P. KadewTraKuPong and R. Bowden, "An improved adaptive background mixture model for real-time tracking with shadow detection," in *Proc. 2nd European Workshp on Advanced Video-Based Surveillance Systems*, 2001.

[13] H. Ardö, "Multi-target tracking using on-line viterbi optimisation and stochastic modelling," Ph.D. dissertation, Centre for Mathematical Sciences LTH, Lund University, Sweden, 2009.

[14] G. Arechavaleta, J.-P. Laumond, H. Hicheur, and A. Berthoz, "An optimality principle governing human walking," *IEEE Transactions on Robotics*, vol. 24, no. 1, pp. 5–14, 2008.

[15] T. Öberg, A. Karsznia, and K. Öberg, "Joint angle parameters in gait: reference data for normal subjects, 10-79 years of age," *Journal of Rehabilitation Research and Development*, vol. 31, no. 3, pp. 199–213, 1994.

[16] R. Knoblauch, M. Pietrucha, and M. Nitzburg, "Field studies of pedestrian walking speed and start-up time," *Transportation Research Record, No. 1538*, pp. 27–38, 1995, TRB, National Research Council, Washington, DC.

[17] M. Bennewitz, W. Burgard, G. Cielniak, and S. Thrun, "Learning motion patterns of people for compliant robot motion," *The International Journal of Robotics Research*, vol. 24, no. 1, pp. 31–48, 2005.