Facilitating Intention Prediction for Humans by Optimizing Robot Motions

Freek Stulp Jonathan Grizou Baptiste Busch Manuel Lopes

***Abstract*— Members of a team are able to coordinate their actions by anticipating the intentions of others. Achieving such implicit coordination between humans and robots requires humans to be able to quickly and robustly predict the robot's intentions, i.e. the robot should demonstrate a behavior that is *legible*. Whereas previous work has sought to explicitly optimize the legibility of behavior, we investigate legibility as a property that arises automatically from general requirements on the efficiency and robustness of joint human-robot task completion. We do so by optimizing fast and successful completion of joint human-robot tasks through policy improvement with stochastic optimization. Two experiments with human subjects show that robots are able to adapt their behavior so that humans become better at predicting the robot's intentions early on, which leads to faster and more robust overall task completion.**

I. Introduction

We humans are very good at anticipating the intentions of others from their actions, and at adapting our own actions accordingly. Many studies have shown that the capability of anticipatory action improves team work and is a natural expectation of humans [1]. If you move an open bottle in my direction, I will place my glass in an appropriate position for you to pour. If you reach for the screwdriver, I will lift the shelf to hold it in place for you. By monitoring the actions of others and inferring their intentions, a human can predict and preemptively initiate the appropriate complementary actions without the need for verbal communication [2], [3], [4]. Furthermore, it has been shown that humans unconsciously change their behavior, for instance the speed of task execution, to improve coordination [5].

Achieving such implicit, non-verbal coordination between robots and humans will greatly improve the efficiency of human-robot collaboration. The seamless integration of the complementary skills of humans (dexterity, task knowledge, flexibility) and robots (speed, strength, precision, repeatability) will have a huge impact on applications ranging from industrial manufacturing to assistance and medical robots [6]. Enabling robots to automatically recognize human intentions is an important aspect of this vision, and an active field of research [7]. But, on the other hand, how can robots make *their own behavior* easier to interpret *for humans*? How can robots improve the *legibility* of their behavior?

*The authors are with the Flowers Team, a joint lab between INRIA and ENSTA-Paristech (Unité d'Informatique et d'Ingénierie des Systèmes, ENSTA ParisTech, Université Paris-Saclay, 828 bd des Maréchaux, 91762 Palaiseau cedex France). Contact: FirstName.LastName@inria.fr

Work partially supported by INRIA, Conseil Régional d'Aquitaine, the ERC grant EXPLORERS 24007, the 3rd Hand Project (funded under 7th FWP), and the European Union's Horizon 2020 research and innovation programme under grant agreement No 640891.

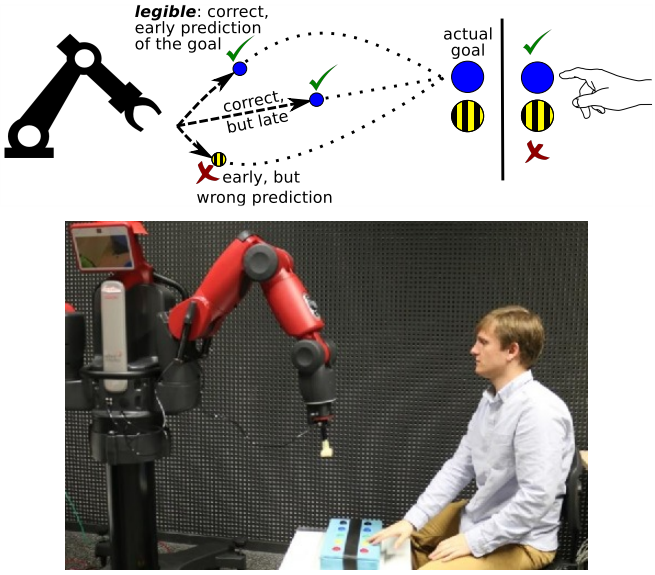


Fig. 1. Illustration of the button pressing experiment, where the robot reaches for and presses a button. The human subject predicts which button the robot will push, and is instructed to quickly press a button of the same color when sufficiently confident about this prediction. By rewarding the robot for fast and successful joint completion of the task – which indirectly rewards how quickly the human recognizes the robot's intention and thus how quickly the human can start the complementary action – the robot learns to perform more legible motion. The three example trajectories above illustrate the concept of legible behavior: it enables correct prediction of the intention early on in the trajectory.

One option is to tailor robot controllers to be legible in specific task contexts, for instance to facilitate handing over an object [8], [9], [10], [11]. This involves understanding and modelling how humans interpret actions, and implementing controllers based on this knowledge. Explicit task-specific encoding of intention prediction has also been used to coordinate robot soccer players [12], [13].

Dragan et al. [14] take a different approach, by providing a *general-purpose* definition of legibility: how probable is a goal, given a partially observed trajectory? Higher legibility implies earlier divergence of probabilities for different goals. They also clearly formalize the difference to predictability: what is the most probable trajectory, given knowledge of the goal? Although legibility and predictability are general measures, they are based on cost functions which are tailored to specific task contexts. It is a non-trivial task to adapt this cost function to novel task contexts, and especially to different (classes of) users. Robots are able to generate legible behavior by optimizing the legibility measure off-line through functional gradient optimization [15].

Rather than defining legibility as an explicit property to be optimized, we investigate legibility as an emergent adaptive property of robots who are rewarded for efficiently cooperating with humans. Our approach is based on model-free reinforcement learning, where the robot iteratively improves its legibility through trial-and-error interaction with a human. This approach has the advantage that no assumptions about the task or the human must be made, and the robot automatically adapts its legibility to the user preferences during the interaction. We evaluate our approach both in simulation and in two user studies with a humanoid robot, one of which is depicted in Fig. 1.

An important component of legibility is that human motions are highly stereotypical. For instance, Glasauer et al. [21] showed that the trapezoidal joint velocity profiles used in industrial robotics are more difficult to predict than the minimum-jerk profiles that are typical for humans. Our work, as that of Dragan et al. [14], [15] rather focuses on the question: "How can I deviate from the stereotypical behavior such that it contains information that enables an observer to quickly infer my intentions?"

The rest of this paper is structured as follows. In the next section, we describe the generic reinforcement learning algorithm used to optimize behavior. In Section III and IV we then present our simulation experiment and user study, including the relevant cost functions and results. We conclude with Section V

II. Model-Free Reinforcement Learning

The key idea in this paper is that legibility of robot behavior need not be defined and optimized explicitly, but that it arises automatically if joint task execution is penalized for not being efficient. In particular, this behavior should arise from direct interaction with users, so as to tailor the legibility to specific user preferences.

As we need to optimize a cost function, but do not have models of individual users, we formulate the problem as a model-free reinforcement learning problem. We describe the generic reinforcement learning algorithm and policy representation used in this paper.

A. Policy Improvement through Black-Box optimization

Policy improvement is a form of model-free reinforcement learning, where the parameters θ of a parameterized policy π_θ are optimized through trial-and-error interaction with the environment. The optimization algorithm we use is PI$^{\text{BB}}$, short for "Policy Improvement through Black-Box optimization" [17]. It optimizes the parameters θ with a two-step iterative procedure. The first step is to locally *explore* the policy parameter space by sampling K parameter vectors θ_k from the Gaussian distribution $\mathcal{N}(\theta, \mathbf{\Sigma})$, to execute the policy with each θ_k, and to determine the cost J_k of each execution. This exploration step is visualized in Fig. 2, where $\mathcal{N}(\theta, \mathbf{\Sigma})$ is represented as the large (blue) circle, and the samples $J_{k=1\ldots10}$ are small (blue) dots.

The second step is to *update* the policy parameters θ. Here, the costs J_k are converted into weights P_k with

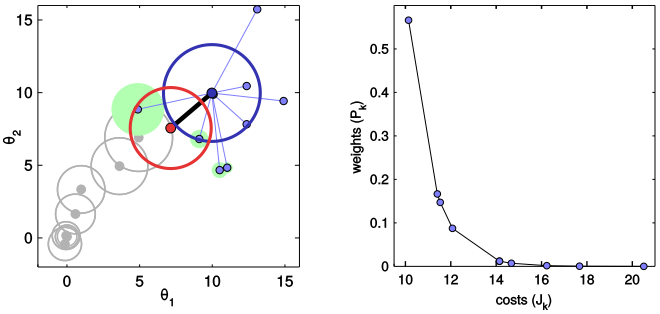


Fig. 2. Illustration of the PI$^{\text{BB}}$ algorithm on a simple cost function $J(\theta) = \|\theta\|$ (without policies). Left: iterative updating of the parameters, where the exploratory samples for the first iteration are shown. Right: mapping the costs J_k to weights P_k for the first iteration. The algorithmic parameters are $K = 10$, $\lambda = 0.7$.

$P_k = \exp\left(\frac{-h(J_k - \min(\mathbf{J}))}{\max(\mathbf{J}) - \min(\mathbf{J})}\right)$, where low-cost samples thus have higher weights. For the samples in Fig. 2, this mapping is visualized to the right. The weights are also represented in the left figure as filled (green) circles, where a larger circle implies a higher weights. The parameters θ are then updated with reward-weighted averaging $\theta \leftarrow \sum_{k=1}^{K} P_k \theta_k$.

Exploration is decreased after each iteration $\mathbf{\Sigma} \leftarrow \lambda \mathbf{\Sigma}$ with a decay factor $0 < \lambda \leq 1$. The updated policy and exploration parameters (red circle in Fig. 2) are then used for the next exploration/update step in the iteration. Despite its simplicity, PI$^{\text{BB}}$ is able to learn robot skills efficiently and robustly [17]. Alternatively, algorithms such as PI^2, PoWER, NES, PGPE, or CMA-ES could be used, see [16], [18] for an overview and comparisons.

B. Policy Representation

The policy π_θ itself is implemented as a dynamical movement primitive (DMP) [19]. DMPs combine a closed loop controller (a spring-damper system with rest point g) with an open loop controller (a function approximator) to generate goal-directed movements. DMPs are convenient for our experiments, as they ensure convergence towards a goal g, whilst allowing the trajectory towards this goal to be adapted by changing the parameters θ of the function approximator used inside the DMP. The function approximator we use is a radial basis function network with 3 basis functions. The policy parameters θ thus correspond to the weights of the basis functions. The output of the DMP is the angular position of the joints of the robot. Our approach does not hinge on the use of DMPs as a policy representation, and we refer to [19] for details.

III. Simulated Experiments

Before turning to the user studies, we first verify that trajectory legibility can emerge by optimizing time efficiency in joint tasks in simple simulated environments. We do so with two experiments: 1) We simulate an 'intention recognizer' – based on the legibility measure defined by Dragan et al. [14] – and give higher rewards (lower penalties) to the robot if the intention recognizer recognizes the intention earlier.

2) We use another intention recognizer, but it does not use the legibility measure.

We are not interested in quantitative differences between the solutions the methods generate, but rather the fact that they are able to generate the same qualitative results in the first place. These experiments are a proof-of-concept that legible behavior may emerge even if the legibility measure is not explicitly encoded in the cost function. Thus, our approach can also – as we show in the user studies – be used by robots to (implicitly) optimize their legibility for humans, who are 'intention recognizers' whose cost function is not objectively accessible.

Experimental Design. The design is the same for all experiments. The start position of the movement is always the same, and there are two targets, cf. Fig. 3. We run two optimizations, in which one target is the goal and the other the distractor, and vice versa. The 2D trajectories are generated by a 2D DMP with 6 basis functions, and the initial weights are all set to zero, leading to a straight line towards the goal. The algorithmic parameters of PI$^{\text{BB}}$ are $\mathbf{\Sigma} = 5\mathbf{I}$ $K = 10$ and $\lambda = 0.95$.

A. Experiment 1: Optimizing Recognition Time (Based on Legibility)

In the work of Dragan et al. [15], a functional gradient algorithm is used to directly optimize the legibility measure along the trajectory. We have reproduced one of their example results (Fig. 1 from [15]) in Fig. 3. In Experiment 1, we provide a less informative measure: the time at which this probability exceeds a certain threshold. Why use a less informative measure? Because we can objectively measure it in human subjects, whereas explicitly modeling legibility measures for humans is difficult, and needs to be adapted to each individual.

Cost Function. We compute a predictor using the Eq. 4 from [15] which estimates the probability of a goal given part of a trajectory. To evaluate the decision time, we compute the evolution of the probability of each goal through time.

$$p(G_h|\xi_{0\ldots t}) \propto p(\xi_{0\ldots t}|G_h)p(G_h) \tag{1}$$

with $p(\xi_{0\ldots t}|G_h)$ from Eq. 8 of [14] and $\xi_{0\ldots t}$ being the trajectory up to time t.

We then normalize the probabilities to 1 across all goal hypotheses. A goal is predicted as being the targeted one when its probability exceeds a threshold β. The response time of the human is the first time t that $p_t(G_h)$ exceeds β.

The cost function is defined as the response time of the user. The earlier the reply, the better the legibility. If an incorrect goal is predicted, the cost is set to a high value, here 20. We also add a term that penalizes high jerk trajectory.

Results. The resulting trajectories (Fig. 3) confirm that optimizing a proxy cost function – the time for an 'intention recognizer' to estimate the intended goal – can lead to similar trajectories as when optimizing directly for legibility.

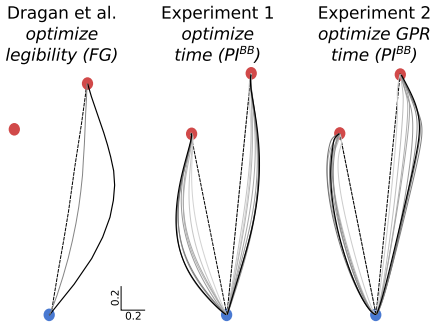


Fig. 3. Results of the simulation experiments. The left figure is adapted from Fig. 1 in [15]. The initial trajectories are the straight dashed lines between the start and the two goals, the thick black lines is the result after optimization. Intermediate policies during optimization are plotted in gray.

B. Experiment 2: Optimizing Recognition Time (Not Based on Legibility)

Although in Experiment 1 the cost only considered the time at which the goal intention was recognized confidently, it still used the measure of legibility to perform this recognition. In this experiment, we use an ad-hoc cost function that does not use the legibility measure from [15]. Our aim is to show that if ad-hoc cost functions (which mimic the prediction time of a human) lead to similar trajectories than previous works considering an explicit definition of legibility (e.g. [15]), then our approach is applicable to human subjects.

Cost function. As in Experiment 1, this cost function observes the trajectory time step by time step. A (partial) trajectory is modeled as a Gaussian process between the x and y coordinate [20]. The Gaussian process uses a Gaussian kernel with maximum variance σ^2=0.25 and length l=1.0. Using Gaussian process regression, we estimate the mean μ and variance σ of x for the y-coordinates of the goal and the distractor. Given these means and variances, we compute the conditional probability that the trajectory will pass through the goal. If this probability > 0.9, the intention is correctly recognized. The time at which this recognition occurs is the cost. The main aim of this implementation is to demonstrate that an 'ad hoc' intention recognizer, not based on the formal concept of legibility, enables legible behavior to be generated. Although this recognizer models and approximates human intent recognition, the optimization algorithm does not have access to this model.

Results. The same general trend of exaggerated movements away from the distractor are visible. Thus, legible behavior can also be generated with cost functions involving only time, without any underlying concept of legibility.

C. Discussion

These results demonstrate that A) legible behavior may arise if only recognition time is used as a cost function. B) model-free stochastic optimization is able to optimize this cost function, without requiring a model of the 'intention recognizer'. This makes our approach applicable to human subjects, as we demonstrate in the next section.

IV. Robot Experiments with Human Subjects

We perform two experiments with human subjects. In Experiment A, the task consists in pressing two corresponding buttons. First the robot decides on a button to press and then the subject needs to press the corresponding button as soon as possible; subjects are to press a button of the same color. In Experiment B, the task consists in selecting a button based on the location a robot will deposit a bottle. First the robot decides on a target location to put a bottle after grasping it, the subject needs to press the button corresponding to the location of the target location before the robot grasps the bottle.

In both experiments the time taken by the user to press the button is used as the main cost of the collaborative task. In practice if the motion of the robot is more informative then the subject will be able to predict sooner the target of the robot.

Both experiments follow the same protocol. The subject is instructed to press the corresponding button as soon as he understand the intention of the robot. All experiments start with an *habituation phase* of 32 trials where the robot performs always the same trajectory for the same task. This phase allows the subject to get used to the robotic motions, and practice the prediction and button pressing. Preliminary results indicate that 32 trials are sufficient for habituation. The recognition time at the end of this phase is a baseline on which we will evaluate the improvement that the robotic optimization gave for the shared task.

After habituation, we start the *optimization phase* of 96 trials with the PI$^{\text{BB}}$ algorithm as presented in Sec. II, during which the robot executes variations of the movement by exploring policy parameter space. The two DMPs are optimized in two separate optimization processes. The parameters of PI$^{\text{BB}}$ are $K = 8$ (trials per update), $\mathbf{\Sigma} = 5\mathbf{I}$ (initial exploration magnitude) and $\lambda = 0.9$ (exploration decay).

We recruited 20 volunteers among INRIA staff, PhD students in computer science, and under-grad students of cognitive science. Among them 13 participated to Experiment A (9 starting from initial trajectories and 4 starting from previously optimized trajectories) and 7 to Experiment B.

A. Experiment A: Joint button pressing

In this task, the robot reaches for and presses one of two buttons. The subject is instructed to press a button of the same color as early as possible, whilst avoiding mistakes, i.e. pressing another button than the robot intended to. We used the set-up presented in Fig. 1, where the subject sits in front of the robot.

The two initial policies, i.e. one for each button, have been recorded through programming by demonstration, as described in [19]. The starting position is the same and the path to each button is straight. Thus trajectories are hard to differentiate in the first part of the movement.

Cost function. The cost function consists of three components: *Efficiency*: The time taken by the human and the robot to press their respective buttons from the onset of the robot's movements, $T_{\text{robot}} + T_{\text{subject}}$. *Robustness*: Whether the subject pressed the same button as the robot (δ_{buttons}=0) or not (δ_{buttons}=1). γ is an arbitrary high cost, it was set to 20 in this experiment. *Energy*: The sum over the jerk (third derivative of the joint position, $\alpha\dddot{\mathbf{q}}_{t_i}$) at each time step i in the trajectory. α is a scaling factor chosen such that the cost of the jerk is about $1/20$ of the total cost in the initial trajectories.

$$J = \underbrace{T_{\text{robot}} + T_{\text{subject}}}_{\text{efficiency}} + \underbrace{\gamma\delta_{\text{buttons}}}_{\text{robustness}} + \underbrace{\alpha|\dddot{\mathbf{q}}_{1\ldots N,1\ldots T}|}_{\text{energy}} \quad (2)$$

Note that this cost function does not contain a direct measure of legibility. Our hypothesis is that legibility can arise by penalizing efficiency and robustness only, which are only indirect measures of legibility.

Results. For illustration purposes, the top graph of Fig. 4 shows a learning session for one of the subjects. The x-axis corresponds to the number of trials, and the y-axis to the time between the onset of the robot's movements and the pushing of the button by the human, which we denote "prediction time". Individual trials are depicted as circles or crosses corresponding to successful predictions of the robot's intentions or "prediction errors" respectively. For this subject, we see that during the initial habituation phase, prediction times drop from $\approx 3s$ to $\approx 2.5s$. After the optimization phase the prediction time is reduced further. This reduction is not due to the subject quickly pushing random buttons, because the error rate does not increase (crosses indicate prediction errors, i.e. pressing the wrong button).

The center graph in Fig. 4 shows the prediction times averaged over all 9 subjects. To allow comparison between subjects without introducing variance due to the natural overall differences in their prediction time, we normalized their results by their intrinsic prediction times, computed as the average of their last 8 prediction times in the habituation phase. Finally, the bottom graph in Fig. 4 shows the number of prediction errors per block of 8, averaged over all subjects. During optimization the execution time drops by 20%, without leading to significantly more prediction errors, i.e. execution time does not decrease because humans are simply guessing.

However, the number of errors slightly increases at the beginning of the optimization phase. This is probably due to the sudden variability in the robot trajectories. Indeed, after 32 trials, the robot starts exploring the parameter space. As a result, two successive trajectories might look different while aiming at the same target and the error rate is thus likely to increase. Soon enough, the co-adaptation between the human and the robot takes place and the error rate decreases.

The box-plots at the top of Fig. 5 highlight the statistics of the normalized prediction times at important transitions during learning: the start (trial 1 to 8), the last trials of the habituation phase (25-32), the first trials of the optimization phase (33-40), and the final trials (121-128). We observe a substantial (20%) and significant ($p = 5.3e^{-8}$, Wilcoxon signed-rank test) drop in prediction time between the end of the habituation phase (25-32) and the end of the optimization

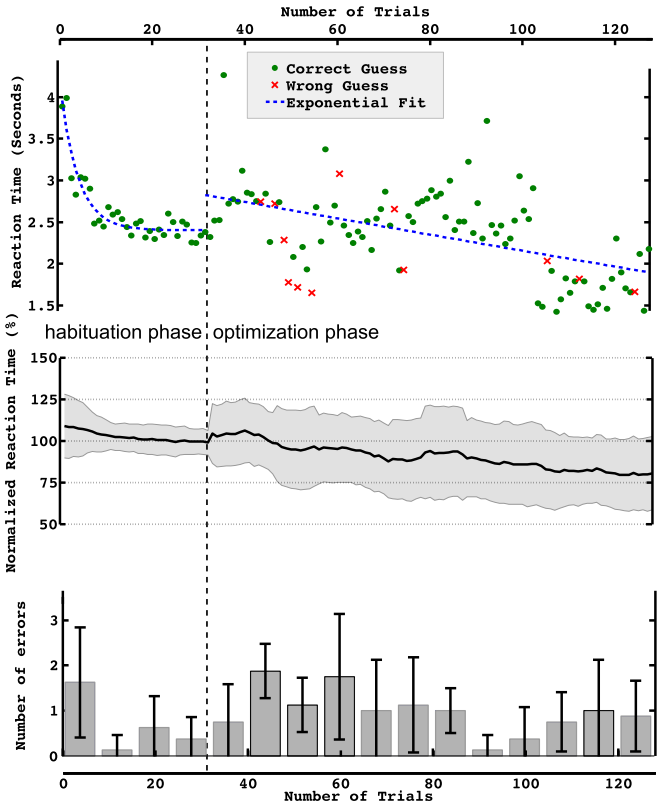


Fig. 4. Results for Experiment A. Top) Prediction times of an example subject during learning. Middle) Average and standard deviation ($\mu \pm \sigma$) over the normalized prediction times of the 9 subjects. Bottom) Number times the incorrect button was pushed, averaged over 9 subjects. The start of the optimization phase is indicated by the vertical dashed line.

(121-128). The small decrease during the habituation phase is also significant ($p = 0.001$), indicating that initial trajectories are different. Yet, the prediction time is further improved by 20% after the optimization showing that initial trajectories are not differentiable as early as optimized ones.

Do subjects learn quicker when starting with policies that have been optimized previously with another subject? To analyze this, we ran the same experiment as above with another 4 more subjects, but started the habituation phase with policies resulting from optimization with other subjects. The results are plotted in the lower graph in Fig. 5. Please note that prediction times are again normalized with respect to trials 25-32, i.e. the second box plot in each row. In contrast to the previous experiment, the decrease in prediction time is still significant during habituation ($p = 0.0003$), but not during optimization ($p = 0.427$). This suggesting the trajectories optimized with another subject are sufficiently separable to be differentiate by a new human in only a few iterations, and cannot be optimized further.

The (unnormalized) prediction times at the end of the habituation phase is significantly lower ($p < 0.001$, Mann-Whitney U-test) when starting with pre-optimized trajectories than using the initial straight approach. Interestingly, at the end of the optimization phase the prediction times are not significantly different ($p = 0.93$).

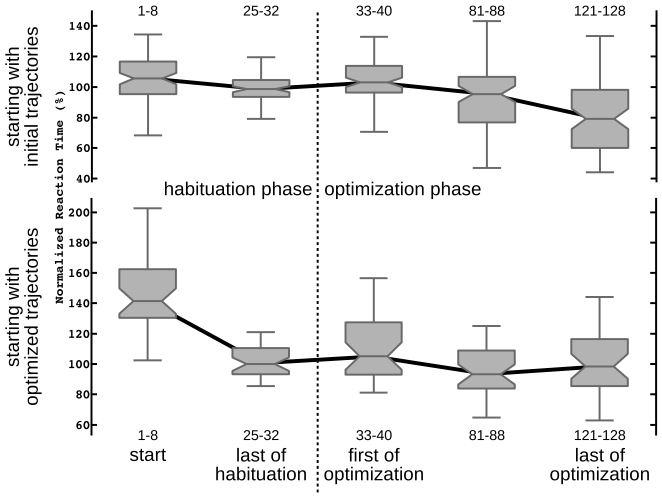


Fig. 5. Box plots for the normalized prediction times, averaged over all subjects, and blocks of 8 trials. Top) Experiment with 9 subjects, with initial trajectories, as in Fig. 4. Bottom) Experiment with 4 subjects, when using previously optimized trajectories from the beginning.

Finally, we must ensure that the global task is executed more efficiently, and that the robot is not reducing human prediction time by, for instance, dramatically increasing energy consumption. Fig. 6 shows the evolution of each component of the cost function for Exp. A and for both initial conditions. It confirms that the global cost is effectively decreased and that the human prediction time has the most impact during the optimization process. Interestingly, the robot time also decreases slightly. The jerk component slightly increases over time, indicating that trajectories with more pronounced features are generated.

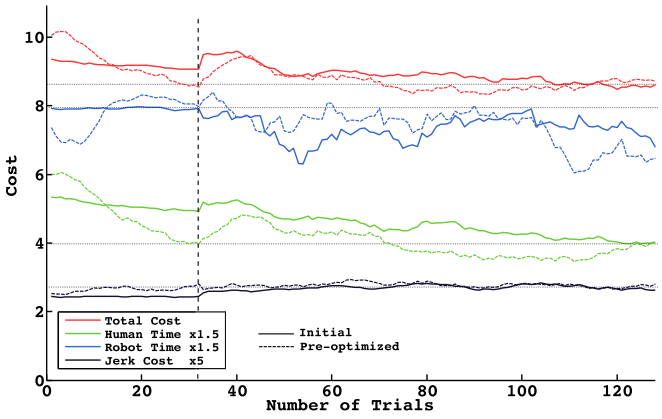


Fig. 6. Components of the cost function for the two conditions of Exp. A averaged across all subjects.

Discussion. The algorithm presented improves human-robot collaboration by producing motions that subjects find easier to predict. By penalizing errors and the joint execution time, the robot learns policies that enable the human to distinguish the robot's intentions earlier without more errors. New subjects are able to quickly read intentions from robot behavior that has been optimized for another subject. Optimized behavior is thus transferable to other subjects.

B. Experiment B: Pick-and-place

In the second task, the robot reaches for and grasps a single object. After grasping the object, the robot has to place it inside one of two boxes located on each side (see Fig. 7). The 7 subjects are instructed to press the button corresponding to the aimed box as early as possible while avoiding mistakes.

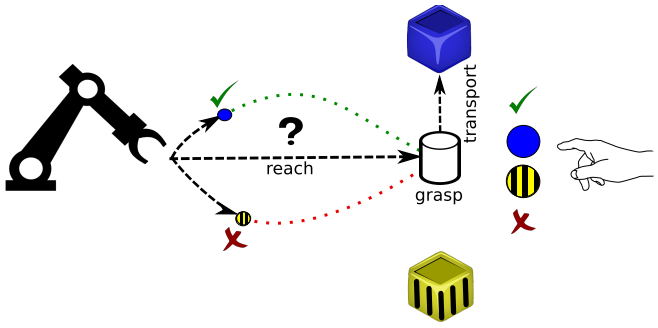


Fig. 7. Illustration of the pick-and-place experiment

To improve the task, the subject must predict which side the object will be moved to and press the corresponding button. This experiment differs from the joint button task because the motions to approach the object are initially identical for both outcomes, guessing before that point results in 50% chances of success. As the robot aims at eliciting an early response from the subject, differentiating the approach motion is necessary to improve joint coordination. Thus, our hypothesis is that at the end of the optimization phase, the subject should be able to predict the robot's intention before it even grasps the object.

Cost function. We use the same cost function as in the previous experiment, with γ again set to 20, and α chosen as in previous experiment.

Results. As in the previous experiment, Fig. 8 summarizes the results by showing the results for one example subject (top), the average prediction times over all 7 subjects (middle), and the number of errors (bottom).

The results for the example subject show that during the habituation phase, this subject waits for the robot to actually start moving the object (approx. $14s$) towards the box to predict the ultimate goal. Because the initial trajectories for each box are identical during habituation, apparent guessing before that point results in an error rate of 50%. When the optimization starts however, the two trajectories start distinguishing themselves and there is co-adaptation between the robot and the human on the intent of each trajectory. After some trial and error, the prediction time of the human drops to a consistent $1.5s$, which implies that the trajectories for the left or right box already differ early on. Despite such early decisions, this subject has almost no prediction errors (only one in the last 70 trials). In comparison to the previous task, the improvements due to optimizing are more pronounced.

The prediction times have a bimodal distribution. Subjects either wait until the robot starts transporting the object, or make a prediction early on during the reaching phase. Rather than averaging over this bimodal distribution data,

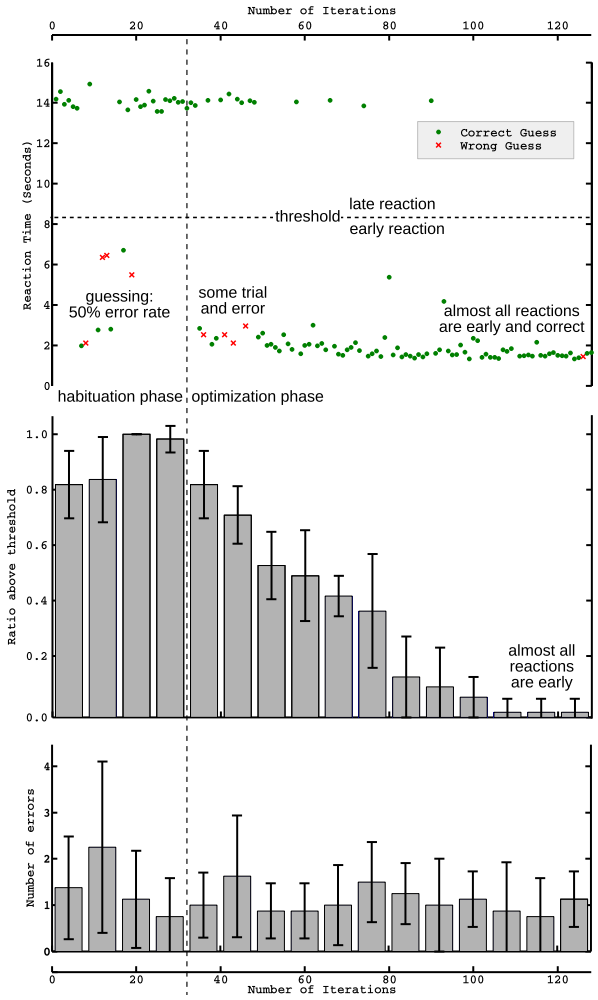


Fig. 8. Results for Experiment B. Top) Prediction times of an example subject during learning. Middle) Ratio of late prediction times (see threshold in top plot), averaged over the 7 subjects. Bottom) Number times the incorrect button was pushed, averaged over 7 subjects. The start of the optimization phase is indicated by the vertical dashed line.

we compute the ratio of early/late prediction times, averaged over blocks of 8 trials and all 7 subjects. The threshold is the average over all the prediction times for one subject. In the middle graph, we see that the late prediction ratio decreases from [0.8-1.0] during the habituation phase to $<$0.05 at the end of the optimization phase. Furthermore, this early prediction is not accompanied by an increase in the number of errors, as the bottom graph shows. Thus, the robot learned behaviors that enabled subjects to predict the correct box before the grasp was even performed.

Because the robot's reaching behavior is the same for both boxes during the habituation phase, subjects can expect a 50% error rate when pressing the button when the robot is still reaching. The increase in the ratio of late predictions and the decrease of errors during the habituation phase may indicate that subjects learn that early guesses lead to errors, and should thus be avoided.

Finally, Fig. 9 confirms that the global cost is effectively reduced and that the human prediction time is dramatically

reduced by the optimization process. We note that the robot only grasps the object once the trajectory is finished, hence the constant cost (DMPs are of fixed duration). The jerk component slightly increases over time, indicating more pronounced trajectories are performed.

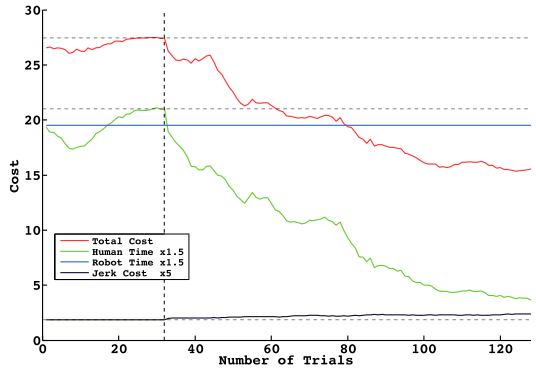


Fig. 9. Components of the cost function for Exp. B averaged across all subjects.

Discussion. During habituation, subjects must wait for the robot to transport the object before being able to robustly predict its intention. After optimization however, all subjects are able to predict the intention of the robot very early on during the reaching phase, when the robot has not yet grasped the object. Experiment B thus confirms the observations in Experiment A, that earlier intention recognition is achieved, but without an increase in errors.

V. Conclusion

The main contribution of this paper is to demonstrate that robots are able to autonomously adapt their behavior so that humans recognize the robot's intentions early and robustly, and that robots are able to do so without a model of how humans predict motion intentions, or knowledge of the concept of legibility. Achieving this with our approach hinges on two steps. The first is to define a cost function that penalizes efficiency (joint execution time), robustness (task errors), and energy (jerk); these are generic measures, as illustrated by the fact that we used the exact same cost function for the two very different tasks in Section IV. The second is to use a model-free optimization algorithm, such as PI^{BB}, to efficiently optimize this cost function through trial-and-error interaction of the robot with the human. We thus provide a generic approach for improving human-robot collaboration, where legible behavior is a side-effect of generic, easy to specify cost measures.

A limitation of our work, as for all model-free optimization approaches, is that novel tasks or substantial changes to a task require re-optimization of the policies.

An open question, also raised by Dragan et al. [15], [14], is: which aspects of legibility are universal, and which are specific to individual humans? Although we have shown that optimized policies are sufficiently different to elicit earlier responses from subjects which never interacted with the robot before (Fig. 5), we clearly cannot claim that the behavior generated by the robot are universally legible. Also, to which extent can legibility be separated from co-adaptation? To which extent did the robot's behavior truly become generally legible, and to which extent did the user learn to recognize idiosyncratic robot behavior that arose from exploration? In our future work, we will investigate these questions, for instance by having different subjects alternatively interacting with the robot during the same optimization process. Nevertheless the importance of co-adaptation cannot be dismissed in any complex learning robot that interacts with humans.

References

[1] W. Erlhagen, A. Mukovskiy, F. Chersi, and E. Bicho, "On the development of intention understanding for joint action tasks," in *ICDL*, London, UK, 2007.

[2] N. Sebanz, H. Bekkering, and G. Knoblich, "Joint action: bodies and minds moving together," *Trends in Cogn. Sciences*, 10(2):70–76, 2006.

[3] L. Sartori, C. Becchio, and U. Castiello, "Cues to intention: the role of movement information," *Cognition*, 119(2):242–252, 2011.

[4] C. Becchio, V. Manera, L. Sartori, A. Cavallo, and U. Castiello, "Grasping intentions: from thought experiments to empirical evidence," *Frontiers in human neuroscience*, vol. 6, 2012.

[5] C. Vesper, R. van der Wel, G. Knoblich, and N. Sebanz, "Making oneself predictable: reduced temporal variability facilitates joint action coordination," *Experimental brain research*, pp. 1–14, 2011.

[6] A. Bauer, D. Wollherr, and M. Buss, "Human–robot collaboration: a survey," *International Journal of Humanoid Robotics*, vol. 5, no. 01, pp. 47–66, 2008.

[7] J. Aggarwal and M. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 16:1–16:43, Apr. 2011.

[8] M. Cakmak, S. Srinivasa, M. Lee, S. Kiesler, and J. Forlizzi, "Using spatial and temporal contrast for fluent robot-human hand-overs," in *Proceedings of the 6th international conference on Human-robot interaction*. ACM, 2011, pp. 489–496.

[9] M. Lee, J. Forlizzi, S. Kiesler, M. Cakmak, and S. Srinivasa, "Predictability or adaptivity?: designing robot handoffs modeled from trained dogs and people," in *Proceedings of the 6th international conference on Human-robot interaction*. ACM, 2011, pp. 179–180.

[10] K. Strabala, M. K. Lee, A. Dragan, J. Forlizzi, and S. S. Srinivasa, "Learning the communication of intent prior to physical collaboration," in *RO-MAN, 2012 IEEE*. IEEE, 2012, pp. 968–973.

[11] K. W. Strabala, M. K. Lee, A. D. Dragan, J. L. Forlizzi, S. Srinivasa, M. Cakmak, and V. Micelli, "Towards seamless human-robot handovers," *Journal of Human-Robot Interaction*, vol. 2, no. 1, pp. 112–132, 2013.

[12] F. Stulp, M. Isik, and M. Beetz, "Implicit coordination in robotic teams using learned prediction models," in *ICRA*, 2006, pp. 1330–1335.

[13] E. Pagello, A. D'Angelo, F. Montesello, F. Garelli, and C. Ferrari, "Cooperative behaviors in multi-robot systems through implicit communication," *Robotics and Autonomous Systems*, vol. 29, no. 1, pp. 65–77, 1999.

[14] A. D. Dragan, K. C. Lee, and S. S. Srinivasa, "Legibility and predictability of robot motion," in *Int'l Conference on Human-Robot Interaction (HRI)*, 2013, pp. 301–308.

[15] A. Dragan and S. Srinivasa, "Generating legible motion," in *Robotics: Science and Systems*, June 2013.

[16] F. Stulp and O. Sigaud, "Policy improvement methods: Between black-box optimization and episodic reinforcement learning," 2012, hal-00738463.

[17] F. Stulp, L. Herlant, A. Hoarau, and G. Raiola, "Simultaneous on-line discovery and improvement of robotic skill options," in *IROS*, 2014.

[18] J. Kober and J. Peters, "Policy search for motor primitives in robotics," *Machine Learning*, vol. 84, no. 1, pp. 171–203, 2011.

[19] A. Ijspeert, J. Nakanishi, P. Pastor, H. Hoffmann, and S. Schaal, "Dynamical Movement Primitives: Learning attractor models for motor behaviors," *Neural Computation*, vol. 25, no. 2, pp. 328–373, 2013.

[20] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*. MIT Press, 2006.

[21] S. Glasauer, M. Huber, P. Basili, A. Knoll, and T. Brandt, "Interacting in time and space: Investigating human-human and human-robot joint action," in *IEEE International Workshop on Robot and Human Interactive Communication*, 2010.