# Multi-Modal Intention Prediction With Probabilistic Movement Primitives.

Oriane Dermy[1], Francois Charpillet[1], and Serena Ivaldi[1]

[1] INRIA, 615 Rue du Jardin botanique, 54600 Villers-ls-Nancy
name.surname@inria.fr

**Abstract.** This paper proposes a method for multi-modal prediction of intention based on a probabilistic description of movement primitives and goals. We target dyadic interaction between a human and a robot in a collaborative scenario. The robot acquires multi-modal models of collaborative action primitives containing gaze cues from the human partner and kinetic information about the manipulation primitives of its arm. We show that if the partner guides the robot with the gaze cue, the robot recognizes the intended action primitive even in the case of ambiguous actions. Furthermore, this prior knowledge acquired by gaze greatly improves the prediction of the future intended trajectory during a physical interaction. Results with the humanoid iCub are presented and discussed.

**Keywords:** multi-modality, probabilistic movement primitive, human robot interaction, collaboration

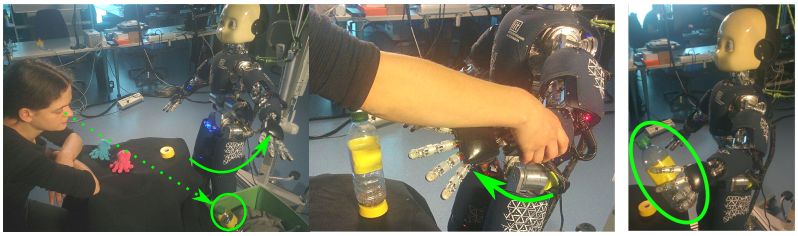
## 1 Introduction

Les humains ont dvelopps des comptences trs dveloppes en ce qui concerne la prdiction et l'adaptation de leurs actions lorsqu'ils sont en collaboration. Pour cela, ils utilisent des indices multi-modales (auditif, visuel, etc.) leur permettant de prdire l'intention de leur partenaire de manire robuste [25].

Pour collaborer efficacement avec les humains, sachant que ceux-ci exhibent des comptences anticipatives, les robots doivent aussi tre capable de prdire l'intention de leur utilisateur. La prdiction de l'intention de l'utilisateur, bas sur ses mouvements implique que ces mouvements soient la fois *lisible* et *prdictible*. En effet, ces conditions sont ncessaire pour que le robot puisse rapidement infrer le but du mouvement et la continuation du mouvement. Ici, nous dfendons l'ide qu'en utilisant des informations multimodales [8, 27], la qualit de prdictions du robot peut tre amliore.

Dans l'tude prcdente [7], nous adressions le problme de prdire le future de trajectoires du bras du robot, inities physiquement par l'utilisateur en interaction avec ce bras. Pour cela, la mthode ProMPs [21] tait utilise afin d'apprendre des primitives de mouvements partir d'un set de dmonstration ; puis de calculer la trajectoire attendue par l'utilisateur, partir de l'observation partiel d'un mouvement initi par l'utilisateur.

Dans ce papier, nous ajoutons la modalit visuelle au robot, afin qu'il prdise l'intention de son partenaire partir de mesures visuelles et cinmatique. L'intention est ici modlise comme la combinaison d'un but positionnel atteindre, et d'une trajectoire que le robot doit effectuer avec son bras.



(a) Visual Guidance. (b) Physical Guidance.

Fig. 1: The humanoid robot iCub a) recognizes the intended movement primitive using the partner's directional gaze; b) predicts the movement to perform using the partner's physical guidance at the beginning of the movement.

Cette fois ci, le robot apprend la combinaison de mesures sur la cinmatique de son bras, lorsque son utilisateur le dplace, ainsi que le mouvement du regard de ce dernier. Ces mesures permettent ainsi au robot d'apprendre des ProMPs multi-modal, qui calcul une distribution partir des dmonstrations de chaque trajectoire.

A partir de l'infrence physique, le robot est capable de reproduire le mouvement, ainsi que de continuer un mouvement initi par le partenaire, et ce mme avec peu d'observations. A partir de linfrence visuelle, le robot est capable de prdire et d'effecteur la tache, pour des tches ne ncessitant pas que l'utilisateur adapte la trajectoire en guidant le dbut du mouvement. De plus, l'infrence visuelle permet au robot de dsambiguser facilement des primitives dont le dbut de mouvement est similaire.

The paper is organized as follows. We briefly report on the literature about intention prediction and gaze as a conveyor of intention information in Section 2. Section 3 formulates the problem settled in this paper. Section 4.1 summarizes the theoretical basis of the ProMP method to learn movement primitives, applied to learning multi-modal information. Section 5 presents a multi-modal intention recognition application, where results about the action recognition improve the prediction of the future trajectory. Finally, section 6 discusses the proposed approach, its limitations and outlines our future developments.

## 2 Related Works

Pour dterminer la trajectoire effectuer, le robot doit infrer l'intention de son partenaire. Ici, nous nous intressons l'infrence effectue partir d'indices physiques et visuels. Les paragraphes suivants fournissent un rsum de la litratures des diffrentes tudes concernant la prdiction de l'*intention* et du *regard*. En ce qui concerne l'tat de l'art sur les *primitives de mouvements* et *l'infrence durant l'interaction physique homme-robot*, nous nous referons [7].

*Intention* Prdire l'intention d'un humain signifie essentiellement prdire le but de son action en cours ou arrivant, ainsi que prdire le mouvement permettant d'atteindre le but vis. La comprhension de l'habilit prdire l'intention intresse diffrents domaines :

l'analyse de l'interaction entre humains [19, 5]; les tudes cherchant  rendre le comportement des robots comprhensible par les humains [16, 10]; ou encore les tudes cherchant rendre les robots capables de comprendre l'intention des humains. Notre tude se situe dans ce derniers cas, comme beaucoup d'autres applications, tels que les tudes mettant en jeu une collaboration homme-robot [11, 26], ou encore pour la navigation robotique [20]. Ici, le regard de l'utilisateur est utilis comme un indice essentiel permettant permettant de dterminer l'intention de l'utilisateur, coupl avec la direction du regard de l'humain et avec ses actions associes**reformuler**

*Gaze as a conveyor of intention information* La direction du regard les l'indice le plus fondamental utilis lors d'interaction sociale, car il permet d'avoir une attention conjointe entre les partenaires. En effet, beaucop d'etudes prennent en compte la direction de la tte/du regard humain afin d'intragir avec celui ci.

Certains utilisent cette direction afin d'estimer l'engamenet de l'utilisateur avec le robot companion [6, 1, 15]; ou d'estimer l'motion de l'utilisateur afin de corriger le comportement du robot [4]. D'autres amliore le comportement du robot en assurant la sret de l'interaction [24]; en anticipant l'action de leur partenaire [13]; ou en adaptant les actions du robot aux intentions de son partenaire [17]. Ce dernier cas correspond nos objectifs actuels.

Pour complter cet objectif, nous calculons d'abord l'orientation de la tte/du regard du partenaire.Pour ce faire, diffrentes mthodes existent telles que, Les Rseaux de Neurones [3] , le calcul de gradients [23], ou en utilisant les probabilits. Le regard est souvent utilise en tant qu' priori sur la tche voulu du partenaire (*e.g*, our work with ProMPs) to detect the object of interest (*e.g.*, [12] with Neural Networks) , or to predict the goal location (*e.g.*, [22] with dynamic models). The main differences between our study and [12, 22] is that these works are interested in the human motion prediction while we associate human gaze to the robot motions.

In some research studies, the human's gaze direction is accurately measured using eye tracker [14, 20]. In our case, we rely on visual processing of the robot's cameras, which is less invasive and it does not require to wear a device, even though it is less accurate than eye tracker.

## 3 Problem Formulation

This paper proposes a method for multi-modal prediction of intention based on a probabilistic description of movement primitives and goals. We target dyadic interaction between a human and a robot, equipped with eyes and arms, in a pick and place collaborative scenario, shown in Fig. 1. In this scenario, different objects must be sorted following different trajectories. The human partner chooses to use visual and/or physical guidance to communicate the intended movement to the robot, that should be able at some point to continue the movement on its own. During the visual guidance, the robot tracks the partner's head orientation to predict his/her intention: the gaze trajectory is recognized as belonging to one of the known action primitives. The robot predicts then the current task and the future intended movement. It completes the intended task by placing the object in the expected place, following the trajectory intended by the partner. During physical guidance, the user starts to physically move the robot to perform the action; after early observations, the robot predicts the future

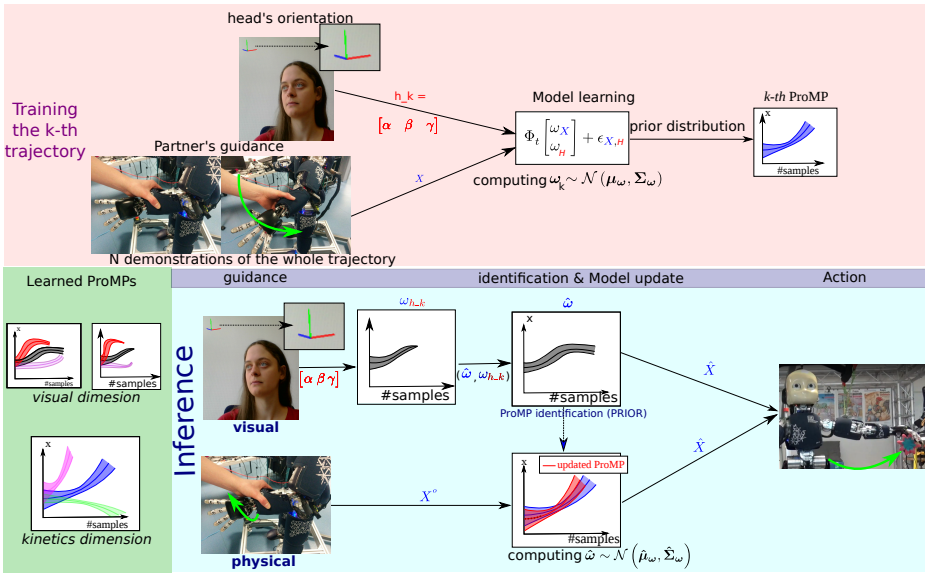



Fig. 2: Conceptual use of ProMP for predicting the desired trajectory to be performed by the robot. In the training phase (top), ProMPs are learned from several human demonstrations. In the inference phase (bottom), the robot recognizes the current ProMP using visual and/or physical information.

movement to perform. If the human partner uses both modalities, the movement primitive can be recognized from the visual guidance (prior) and physical guidance can be used to refine the predicted trajectory (posterior). To realize this scenario, we make several hypotheses. Tracking the gaze using the eyes direction is difficult because of saccadic eye movement directed towards the goal, that could cause the gaze trajectory to be inconsistent. Therefore, the partner's head orientation is used to determine his intent. We assume the user's position with respect to the robot is almost fixed during the learning and the recognition task, because the robot learning is dependent on the partner's head orientation. We assume that the partner's head orientations when he/she looks at a same goal follow a normal distribution.

A conceptual representation of the problem is shown in Fig. 2. To learn the movement primitives (top), two partners run several demonstrations: one moves the robot's arm while another moves his head, following the trajectories to learn. From these demonstrations, the robot collects the Cartesian position of its arm and the partner's gaze (head orientation). The trajectories make the base for learning the primitives (prior distribution). The bottom of the figure represents the inference step. The partner follows with his/her head the robot's movement and/or he/she physically initiates the robot's hand movement. When the prediction is done, the robot finishes autonomously the movement (i.e., drop the hand-held object). To show the improvement with respect to our previous work, the learned trajectories of the dropping phase have identical initial and final positions (making the prediction from early observations harder, and possible here only thanks to the multimodal primitive).

# 4 Methods

This section presents the ProMP method used to learn the motion primitives and to predict the trajectory of the ProMP given one modality. See [7] for further information.

## 4.1 Learning Motion Primitives With ProMP

A ProMP is a Bayesian parametric model of demonstrated trajectories in the form:

$$\xi(t) = \Phi_t \boldsymbol{\omega} + \epsilon_\xi \tag{1}$$

where $\xi(t)$ is the vector containing all the multi-modal variables to be learned at time $t$ (*e.g.*, $A(t)$ for visual modality or $X(t)$ for physical modality); $\boldsymbol{\omega} \in R^M$ is a time-independent parameter vector weighting the $\Phi$ matrix; $\epsilon_\xi \sim \mathcal{N}(0, \beta)$ is the trajectory noise; and $\Phi_t$ is a matrix of $M$ Radial Basis Functions (RBFs) evaluated at time $t$: $\Phi_t = [\psi_1(t), \psi_2(t), \ldots, \psi_M(t)]$. Note that all the $\psi$ functions are scattered across time. The robot first records a set of $n_1$ trajectories $\{\Xi_1, \ldots, \Xi_{n_1}\}$, where the $i$-th trajectory is $\Xi_i = \{\xi(1), \ldots, \xi(t_{f_i})\}$. The duration $t_{f_i}$ of each recorded trajectory varies, following the user demonstrations. To find a common representation (in terms of primitives), a time modulation is applied to all trajectories, such that they have the same number of samples $\bar{s}$. To do so, we consider "$\Phi_{\alpha t}$" instead of "$\Phi_t$", to rescale the RBFs to each trajectory, with the time modulation parameter "$\alpha = \frac{\bar{s}}{t_{fi}}$" . Such modulated trajectories are then used to learn a ProMP.

For each $\Xi_i$ trajectory, we compute the $\omega_i$ parameter vector that minimizes the error between the observed $\xi_i(t)$ trajectory and its model $\Phi_{\alpha t}\omega_i + \epsilon_\xi$. This is done using the Regularized Least Mean Square algorithm.

Thus, we obtain a set of parameters upon which a normal distribution is computed:

$$p(\boldsymbol{\omega}) \sim \mathcal{N}(\mu_{\boldsymbol{\omega}}, \Sigma_{\boldsymbol{\omega}}) \tag{2}$$

$$\text{with } \mu_{\boldsymbol{\omega}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\omega_i} \tag{3}$$

$$\text{and } \Sigma_{\boldsymbol{\omega}} = \frac{1}{n-1}\sum_{i=1}^{n} (\boldsymbol{\omega_i} - \mu_{\boldsymbol{\omega}})^\top (\boldsymbol{\omega_i} - \mu_{\boldsymbol{\omega}}) \tag{4}$$

## 4.2 Predicting the Trajectory of the ProMP

The learned ProMPs corresponds to several skills or action primitives. They are used as a prior knowledge by the robot to predict the current action and its future trajectory, so that it can continue the movement autonomously. Here, early observations of the trajectory are a subset of the variables to learn:

$$\Xi^o = [\Xi_1 \ldots \Xi_{n_o}]^\top = \{X^o || A^o || \begin{bmatrix} X^o \\ A^o \end{bmatrix}\} \tag{5}$$

Where $X^o$ is the haptic measurement and $A^o$, the visual measurement.

The first step of the recognition process is to recognize the current ProMP $\hat{k} \in [1:2]$ , and the temporal modulation parameter $\hat{\alpha}$ from this partial observation $\Xi^o$. This is

done by computing the most likely couple of temporal modulation parameter and ProMP type $(\hat{\alpha}_{\hat{k}}, \hat{k})$ corresponding to the early trajectory. We use two methods to perform this computation.

- The first called "*maximum likelihood*" (*ML*) is computed by:

$$(\hat{\alpha}_{\hat{k}}, \hat{k}) = \text{argmax}_{(\alpha \in S_{\alpha_{\hat{k}}}, \hat{k} \in [1:2])} \{\text{loglikelihood}(\Xi^o, \mu_{\boldsymbol{\omega}_{\hat{k}}}, \sigma_{\boldsymbol{\omega}_{\hat{k}}}, \alpha_{\hat{k}})\}. \quad (6)$$

  , where $S_{\alpha_{\hat{k}}} = \{\alpha_{\hat{k}1}, \ldots, \alpha_{\hat{k}n}\}$ is the set of all the $\alpha$ parameters computed during the learning for each observation of the ProMP $\hat{k}$.
- The second called "*model*" is based on the assumption there is a correlation between the time modulation $\alpha$ and the variation of the trajectory $\delta_{n_o}$ from the beginning until the instant $n_o$. Indeed, we assume that the time modulation parameter $\alpha$ is linked to the movement speed, which can be roughly approximated by "$\dot{\Xi} = \frac{\delta \Xi}{t_f}$". For the physical inference, the "variation" of the hand position is computed by "$\delta_{n_o} = X(n_o) - X(1)$", whereas for the visual inference, the variation of the partner's head orientation is computed by "$\delta_{n_o} = A(n_o) - A(1)$". We model the mapping between $\delta_{n_o}$ and $\alpha$ by:

$$\alpha = \Psi(\delta_{n_o})^\top \boldsymbol{\omega}_\alpha + \epsilon_\alpha, \quad (7)$$

  where $\Psi$ are RBFs, and $\epsilon_\alpha$ is a zero-mean Gaussian noise. During learning, we compute the $\boldsymbol{\omega}_\alpha$ parameter, using the same method as in Equation 1 and during the inference, we compute $\hat{\alpha} = \Psi(\delta_{n_o})^\top \boldsymbol{\omega}_\alpha$. Finally, we compute the maximum likelihood in the set of $\{\hat{\alpha_1}, \hat{\alpha_2}\}$

Once identified the $(\hat{\alpha}_{\hat{k}}, \hat{k})$ couple, the recognized distribution (called the "prior") can be updated by:

$$\begin{cases} \hat{\mu}_{\boldsymbol{\omega}_{\hat{k}}} = \mu_{\boldsymbol{\omega}_{\hat{k}}} + K(\Xi^o - \Phi_{\hat{\alpha}_{\hat{k}}[1:n_o]} \mu_{\boldsymbol{\omega}_{\hat{k}}}) \\ \hat{\Sigma}_{\boldsymbol{\omega}_{\hat{k}}} = \Sigma_{\boldsymbol{\omega}_{\hat{k}}} - K(\Phi_{\hat{\alpha}_{\hat{k}}[1:n_o]} \Sigma_{\boldsymbol{\omega}_{\hat{k}}}) \\ K = \Sigma_{\boldsymbol{\omega}_{\hat{k}}} \Phi^\top_{\hat{\alpha}_{\hat{k}}[1:n_o]} (\Sigma_{\xi^o} + \Phi_{\hat{\alpha}_{\hat{k}}[1:n_o]} \Sigma_{\boldsymbol{\omega}_{\hat{k}}} \Phi^\top_{\hat{\alpha}_{\hat{k}}[1:n_o]})^{-1} \end{cases} \quad (8)$$

with $\hat{\alpha}_{\hat{k}}[1:n_o] = \hat{\alpha}_{\hat{k}}\, t$ (in matrix form), with $t \in [1:n_o]$.

Finally, the inferred trajectory is given by:

$$\forall t \in [1:\hat{t}_f], \hat{\xi}(t) = \Phi_t\, \hat{\mu}_{\boldsymbol{\omega}_{\hat{k}}}$$

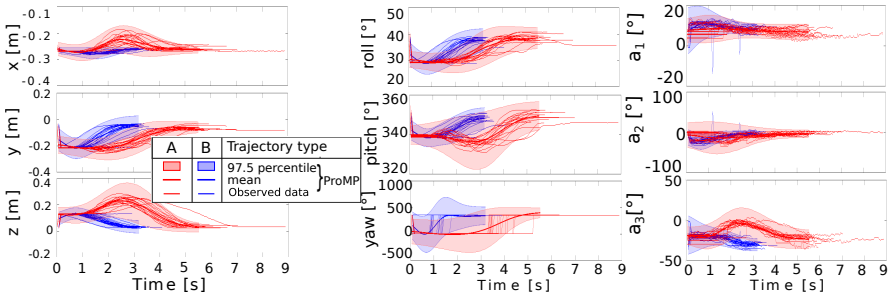
with the expected duration of the trajectory $\hat{t}_f = \frac{\bar{s}}{\hat{\alpha}_k}$. The robot is now able to finish the movement executing the most-likely "future" trajectory $\hat{X} = [\hat{X}_{n_o+1} \ldots \hat{X}_{\hat{t_f}}]^\top$.

## 5 Experiments

### 5.1 Experimental Setup

We carried out experiments with the humanoid robot iCub. To retrieve the approximated gaze direction, we use the roll/pitch/yaw angles of the user's head orientation, extracted from the camera image of the iCubs eyes by Intraface [28], To retrieve the Cartesian information, we use an iCub module that computes the Cartesian position and orientation (iKinCartesianSolver). The experimental procedure is outlined in Fig. 2. The training phase requires a robot operator (performing kinesthetic teaching) and a human partner (guiding the robot via gaze), for a total of two people. In the inference phase, only the partner interacts with the robot.

### 5.2 Teaching iCub the Action Primitives




(a) End-effector position. (b) End-effector orientation. (c) Head orientation.

Fig. 3: Demonstrations (trajectories) and primitives. In red (ProMP A) the "curved" trajectory, and in blue (ProMP B) the "direct" trajectory.

We taught the robot two multi-modal movement primitives that make it drop an object inside a target bin (roughly at the same position) but following two different type of trajectories coupled with the corresponding trajectories of the human partner. These primitives contain the Cartesian position and orientation of the robot's left hand (guided by the robot operator), and the head orientation of the human partner that visually guides the robot: $\xi(t) = [X(t), A(t)]^\top$, with $X(t) \in \Re^6$ the Cartesian pose and $A(t)$ the roll-pitch-yaw orientation angles of the partner's head.

We performed 20 trajectory demonstrations per primitive action. Fig. 3 shows the demonstrations and the learned-distribution for the two ProMPs.

### 5.3 Activating Primitives With Gaze

The gaze cue is used to identify the current action. This procedure has two advantages. First, it does not require physical interaction, which could ease interacting with the robot for some people. Second, it enables to improve the prediction of intended trajectory, especially in case of ambiguous primitives that overlap and could make it difficult to obtain a good prediction with few early observations. An intuitive case is shown in Fig. 5.

From [7], we retain two methods to compute the time modulation: "maximum likelihood" (*ML*) and "*model*", where the latter consists on estimating the trajectory duration according to the global partner's head orientation variation: "$\delta_{n_o} = A(n_o) - A(1)$".

We tested off-line the gaze prediction of the trajectories on the acquired data

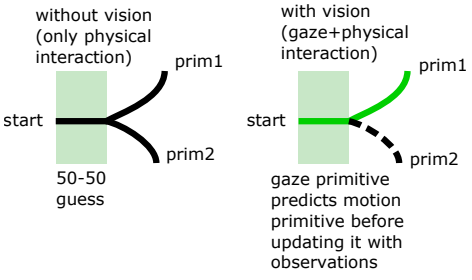



Fig. 5: Gaze helps disambiguate two overlapping primitives.

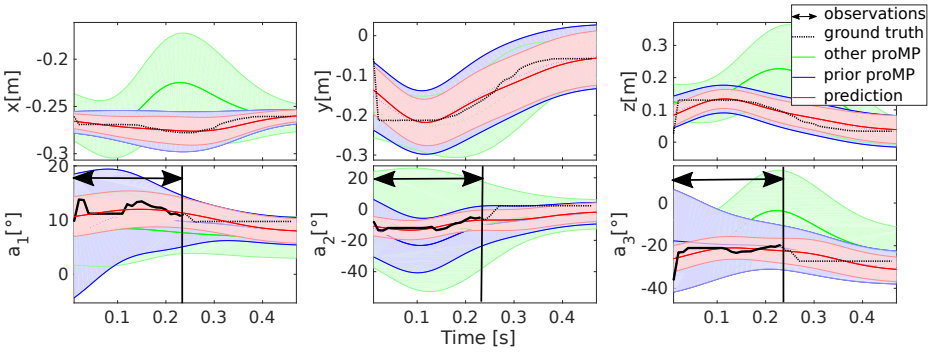



Fig. 4: Example of position inference from 50% of the head orientation trajectory. The dots represent the trajectory the robot has to perform (ground truth). The black curves represent the measurements done by the robot. The blue distribution represents the recognized ProMP and the green distribution the other ProMP. The red distribution represents the posterior of the blue distribution, computed from the measured data.

set using cross-validation. Fig. 4 shows a prediction example after having observed 50% of the trajectory. The inferred trajectory is the mean trajectory of the red posterior distribution. Note that this posterior distribution is included in the prior distribution and pass by the observed data with some *flexibility*, that correspond to the expected measurement noise fixed a-priori. Even though the partner's head orientation observations are not accurate, the prediction is good enough to allow the robot to complete the task correctly.

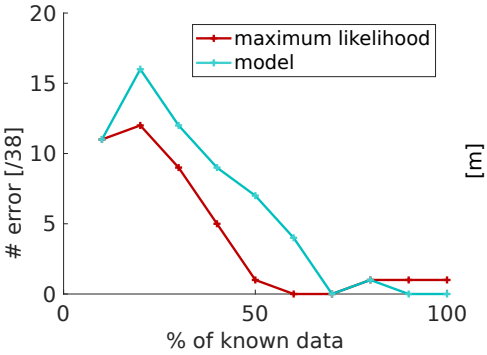
Fig. 6a represents the error of ProMP recognition according to the percentage of observations of the test trajectory. The longer the head trajectory is observed, the smaller is the prediction error, for both methods for computing the time modulation. This figure also shows that the *model* is less accurate than the *ML* method when the robot observes less than 70% of the whole trajectory, while with more observation the *model* method is a slightly more accurate. Since head movements are fast, the robot can use the whole head movement trajectory and still react quickly. So, we can use the *model* method to allow the robot to recognize which ProMP to follow for the visual guidance. With 70% observation of a trajectory, there is no ProMP type recognition error, thus, the robot can roughly infer the trajectory to perform (which corresponds to 3 seconds).

We represent in Fig. 6b the average error of the Cartesian position of the inferred trajectory. It shows that the error of the predicted trajectory goes from 4*cm* (10% of the trajectory) to 2cm (from 80%). Thus, the more the robot observes its partner's head trajectory, the more it is able to achieve it own movement intended by its partner.

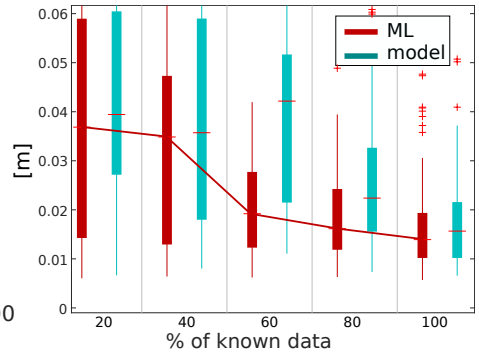
However, we can wonder if the posterior distribution is more accurate than the prior. It would be the case if the partner's head orientation was totally correlated to the robot's hand position and the measurement accurate enough to infer exactly the end-trajectory. Fig. 6c represents the difference of the Normalized Root Mean Square

Error (NRMSE) between the prior and the posterior distribution. From 40% of the trajectory observation, this difference is inferior to zero, meaning that by updating the distribution, the robot improves the trajectory inference. Thus, the visual guidance can be used to determine which ProMP the robot has to follow, but also to adapt the ProMP distribution from the user's head guidance in an accurate way.

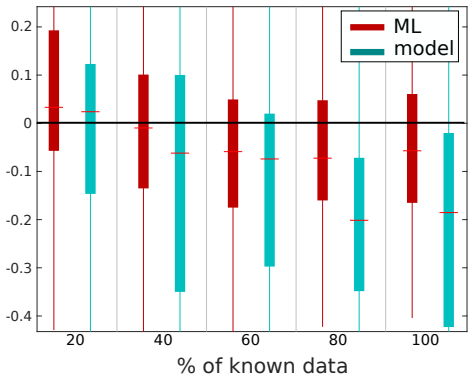
To achieve a better accuracy, we assume the physical interaction will more indicated. To verify this assumption, the next session presents the physical guidance experiment.




(a) Number of prediction errors.




(b) Inference error of the Cartesian position: average$|X_{des} - \hat{X}|$.




(c) RNMSE difference between the prior and posterior distribution.

Fig. 6: Visual guidance analysis.

### 5.4 Inference of Intended Trajectories With Physical Guidance

The same prediction experiment from early-demonstrations than the previous section is presented here with haptic signals. Fig. 7 presents an example of such prediction. If

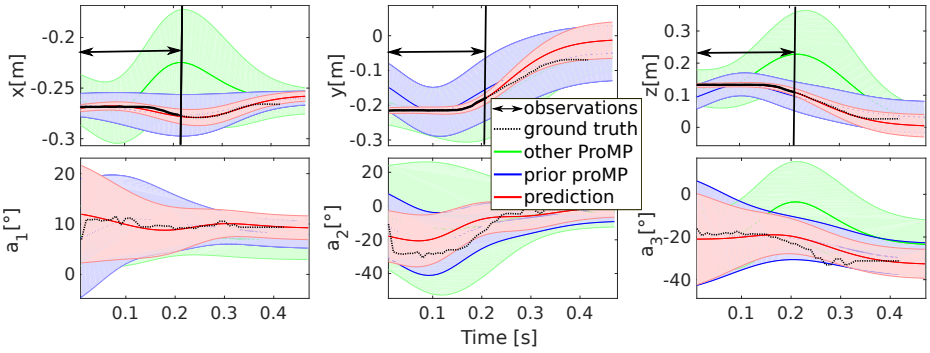
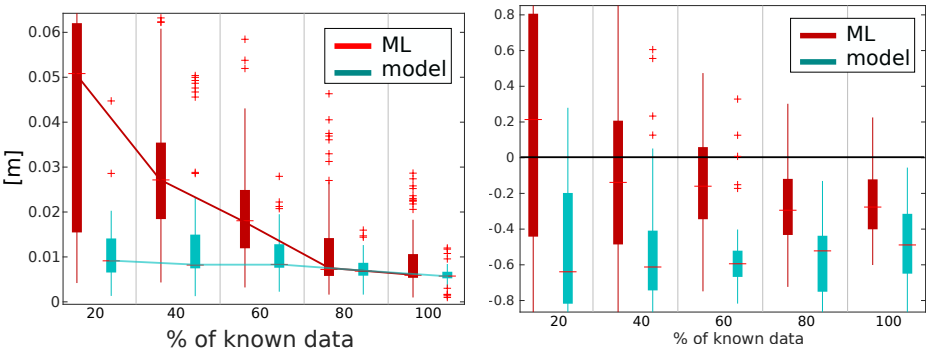



Fig. 7: Example of trajectory inference from physical guidance.




(a) Inference error of the Cartesian position: average$|X_{des} - \hat{X}|$.

(b) average $|nrmse_{post} - nrmse_{prior}|$.

Fig. 8: Physical guidance analysis.

we compare to the visual experiment, we can note that the inferred trajectory (mean of the red posterior distribution) is closer to the ground truth. Fig. 8a verifies this idea. It represents the average distance between the inferred trajectory ($\hat{X}$) and the ground truth ($X_{des}$), and the results show that the trajectory prediction using physical estimation is more accurate than the visual estimation, whether with the *model* or the *ML* method, with an average of less than $1cm$ of distance error for the *model* and from $3cm$ (40% of known data) to $1cm$ (80%) for the *ML*. Moreover, Fig. 8b shows that the posterior distribution of the ProMP improves the accuracy of the trajectory, mainly for the *model* method which explains why the distance error using this method is short in the previous figure.

Now, we can wonder if using the two modalities could improve the performance of this inference ability. Thus, the next section is the multi-modal experiment on the same data set.

### 5.5 Inference of Intended Trajectories With Multi-modal guidance

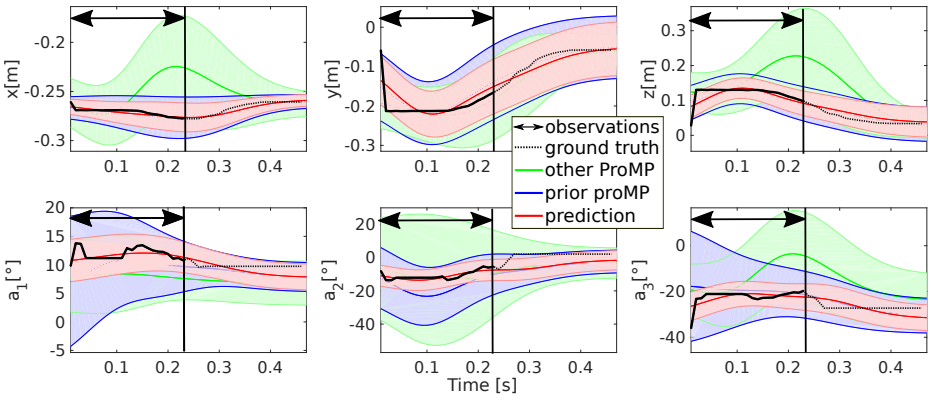
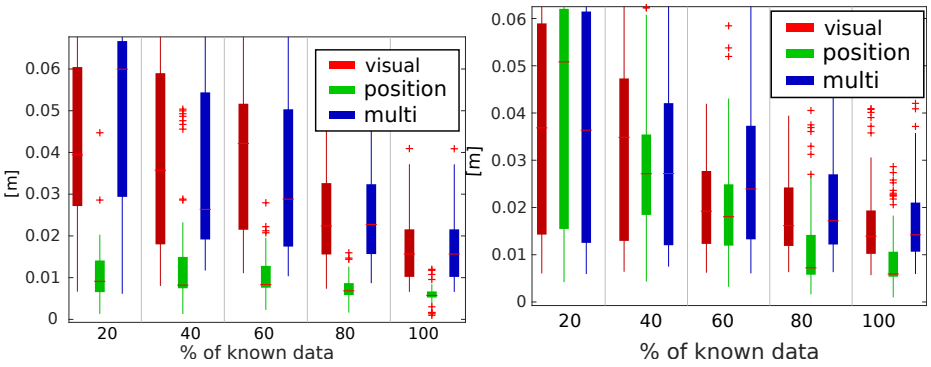



Fig. 9: Example of position inference from 50% of the head orientation and the Cartesian position trajectories.

Fig. 9 represents the inference of the Cartesian position trajectory when the robot knows 50% of the trajectory data to achieve and when it uses both visual and physical measurements (black curves). In this example, the inferred trajectory (mean of the red posterior distribution) is close to the trajectory expected by the partner (black dots). To compare this multi-modal prediction with visual or physical prediction only, Fig. 10 and 11 represent all the statistics for each prediction type. Fig. 10 represents the distance error between the Cartesian position of the expected and the inferred trajectory. Whether with the *model* (in Fig. 10a) or the *ML* method (in Fig 10b), the inference using the Cartesian position measurement only is more accurate than using the multi-modal or the visual-only measurement. The performance of this physical guidance is mainly visible with the *model* method, where the distance error is really short. Thus, the multi-modality guidance did not improve the inference ability of the robot.

From Fig. 11, we can see the number of ProMP recognition error according to the type of modality used to perform the inference. An interesting result is that by using the *model* method (in Fig. 11a), the robot is entirely able to recognize the initiated movement from 70% of know data, and with the *ML* method, the robot has only done one error from the 38 trials (which corresponds to 2%). Thus, the multi-modal clearly improves the ProMP recognition step of the inference, even though it did not improve the final inferred trajectory precision.

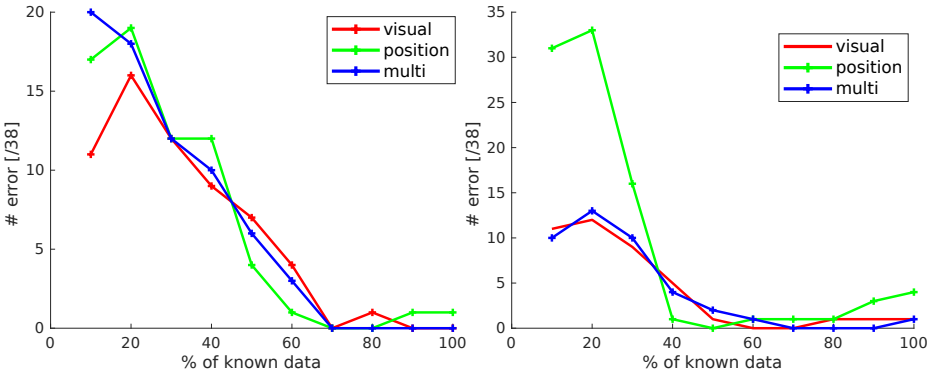



(a) With the model. (b) With the Maximum Likelihood.

Fig. 10: Inference error of the Cartesian position: average$|X_{des} - \hat{X}|$ according to modality used.

## 6 Conclusions

This paper presents a multi-modal method for robots to predict the partner's intended trajectory during HRI using haptic and/or gaze cues. We tested our system with the humanoid iCub collaborating with a human partner in a task where the robot has to grasp an object using different trajectories. The human physically interacts with the robot's arm to start an action and/or uses his directional gaze to guide the robot. We build on our previous work [7], where elementary actions are represented by Probabilistic Movement Primitives that enable prediction of goals from early observations. During physical guidance, the robot uses the haptic information to recognize the current action, then it is able to accurately predict the goal, the future intended trajectory and its duration. A limitation of previous inference method is that the robot is not able to determine which movement primitive to follow when the early-observations are ambiguous, *i.e.*, identical to more than one primitive. In that case, the visual guidance is used to identify the correct movement primitive. While during the visual guidance, the same prediction is done using the directional gaze, approximated here by the head orientation. The association between gaze cues and robot primitives is done by a multi-modal learning phase. The visual modality has two main advantages: first, it does not




(a) With model. (b) With Maximum Likelihood.

Fig. 11: Prediction error according to modality used.

require the partner to physically touch the robot to start his intended movement; second, it provides a faster recognition of the action primitive if compared with physical signals. However, results show that by using the visual instead of the physical guidance, the performance of the inference decreases slightly (around $1.5cm$). A limit of this modality is the accuracy of the gaze estimation. To improve it, we have many possibilities: use the Kinect to have more relevant data; use another head recognition software instead of Intraface; or use the Xsens 3D tracking. It is also possible to add another *"no-human"* modality to even surpass human inference skills, by guiding the robot from a watch that contains sensors to detect the human partner's arm pose and to use this pose to learn and recognize ProMPs.

Regarding the inference using multi-modal measurements, results show that by adding the visual recognition in addition to the physical recognition, it did not improve the accuracy of the inferred trajectory (*i.e.*, it did not improve the posterior distribution computation), but it improves the ProMP recognition (*i.e.*, it improves the first step of the inference that consists on recognizing which movement the robot has to execute among the one it has learned). Thus, to have the better inference skills, we should use the multi-modal guidance to allow robots to recognize the movement/action to perform, and then we should use the haptic guidance to improve the movement precision according to the early measurements. However, the multi-modal guidance currently requires to use two human partners (one in front of the robot to guide it with his/her head and the other one to guide it physically) or to perform the guidance type one after the other. The utilization of the Xsens is a good way to improve this study because one partner will be able to guide physically and visually the partner at the same time, hence in a more natural way.

In future work, we will also study the human preference for the use between the haptic and visual guidance modes.

**Acknowledgments.** The authors wish to thank Olivier Rochel, Alexandros Paraschos, Marco Ewerton, Waldez Azevedo Gomes Junior and Pauline Maurice for their help and feedbacks.

## References

1. Anzalone, S.M., Boucenna, S., Ivaldi, S., Chetouani, M.: Evaluating the engagement with social robots. I.J. of Social Robotics 7(4), 465–478 (2015)
2. Bader, T., Vogelgesang, M., Klaus, E.: Multimodal integration of natural gaze behavior for intention recognition during object manipulation. In: PIC on Multimodal interfaces. pp. 199–206. ACM (2009)
3. Baluja, S., Pomerleau, D.: Non-intrusive gaze tracking using artificial neural networks. In: Advances in NIPS. pp. 753–760 (1994)
4. Boucenna, S., Gaussier, P., Andry, P., Hafemeister, L.: A robot learns the facial expressions recognition and face/non-face discrimination through an imitation game. International Journal of Social Robotics 6(4), 633–652 (2014)
5. Bretherton, I.: Intentional communication and the development of an understanding of mind. Children's theories of mind: Mental states and social understanding pp. 49–75 (1991)
6. Castellano, G., Pereira, A., Leite, I., Paiva, A., McOwan, P.W.: Detecting user engagement with a robot companion using task and social interaction-based features. In: PIC on Multimodal interfaces. pp. 119–126. ACM (2009)
7. Dermy, O., Paraschos, A., Ewerton, M., Peters, J., Charpillet, F., Ivaldi, S.: Prediction of intention during interaction with icub with probabilistic movement primitives, Frontiers in robotics and AI (2017)
8. Dillmann, R., Becher, R., Steinhaus, P.: ARMAR II-a learning and cooperative multimodal humanoid robot system. International Journal of Humanoid Robotics 1(01), 143–155 (2004)
9. Dragan, A., Srinivasa, S.: Generating legible motion. In: Proceedings of Robotics: Science and Systems. Berlin, Germany (June 2013)
10. Dragan, A., Srinivasa, S.: Integrating human observer inferences into robot motion planning. Autonomous Robots 37(4), 351–368 (2014)
11. Ferrer, G., Sanfeliu, A.: Bayesian human motion intentionality prediction in urban environments. Pattern Recognition Letters 44, 134–140 (2014)
12. Hoffman, M.W., Grimes, D.B., Shon, A.P., Rao, R.P.: A probabilistic model of gaze imitation and shared attention. Neural Networks 19(3), 299 – 310 (2006)
13. Huang, C.M., Mutlu, B.: Anticipatory robot control for efficient human-robot collaboration. In: HRI, 2016 pp. 83–90
14. Ishii, R., Shinohara, Y., Nakano, T., Nishida, T.: Combining multiple types of eye-gaze information to predict user's conversational engagement. In: 2nd workshop on eye gaze on intelligent human machine interaction (2011)
15. Ivaldi, S., Lefort, S., Peters, J., Chetouani, M., Provasi, J., Zibetti, E.: Towards engagement models that consider individual factors in HRI. Int. J. of Social Robotics 9, 63–86 (2017)
16. Kim, J., Banks, C.J., Shah, J.A.: Collaborative planning with encoding of users' high-level strategies. In: AAAI (2017)
17. Kozima, H., Yano, H.: A robot that learns to communicate with human caregivers. In: Proceedings of the First International Workshop on Epigenetic Robotics. pp. 47–52 (2001)

18. Ma, C., Prendinger, H., Ishizuka, M.: Eye movement as an indicator of users' involvement with embodied interfaces at the low level. In: Proc. AISB pp. 136–143 (2005)
19. Meltzoff, A.N., Brooks, R.: Eyes wide shut: The importance of eyes in infant gaze following and understanding other minds. Gaze following: Its development and significance, ed. R. Flom, K. Lee & D. Muir. Erlbaum.[EVH] (2007)
20. Mitsugami, I., Ukita, N., Kidode, M.: Robot navigation by eye pointing. Lecture notes in computer science 3711, 256 (2005)
21. Paraschos, A., Daniel, C., Peters, J.R., Neumann, G.: Probabilistic movement primitives. In: NIPS pp. 2616–2624 (2013)
22. H.C. Ravichandar, H., Kumar, A., Dani, A.: Bayesian human intention inference through multiple model filtering with gaze-based priors. In: Information Fusion (FUSION) pp. 2296–2302. IEEE (2016)
23. Timm, F., Barth, E.: Accurate eye centre localisation by means of gradients. Visapp 11, 125–130 (2011)
24. Traver, V.J., del Pobil, A.P., Pérez-Francisco, M.: Making service robots human-safe. In: Proceedings.(IROS 2000) on. vol. 1, pp. 696–701. IEEE (2000)
25. Walker-Andrews, A.S.: Infants' perception of expressive behaviors: differentiation of multimodal information. Psychological bulletin 121(3), 437 (1997)
26. Wang, Z., Deisenroth, M.P., Amor, H.B., Vogt, D., Schölkopf, B., Peters, J.: Probabilistic modeling of human movements for intention inference. In: Robotics: Science and Systems. (2012)
27. Weser, M., Westhoff, D., Huser, M., Zhang, J.: Multimodal people tracking and trajectory prediction based on learned generalized motion patterns. In: Int. Conf. Multisensor Fusion and Integration for Intelligent Systems, pp. 541–546 (2006).
28. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: IEEE CVPR (2013)