Inverse Eye Tracking for Intention Inference and Symbol Grounding in Human-Robot Collaboration

Svetlin Penkov
School of Informatics
The University of Edinburgh
Email: sv.penkov@ed.ac.uk

Alejandro Bordallo
School of Informatics
The University of Edinburgh
Email: alex.bordallo@ed.ac.uk

Subramanian Ramamoorthy
School of Informatics
The University of Edinburgh
Email: s.ramamoorthy@ed.ac.uk

***Abstract*—People and robots are required to cooperatively perform tasks which neither one could complete independently. Such collaboration requires efficient and intuitive human-robot interfaces which impose minimal overhead. We propose a human-robot interface based on the use of eye tracking as a signal for intention inference. We achieve this by learning a probabilistic generative model of fixations conditioned on the task which the person is executing. Intention inference is then achieved through inversion of this model. Importantly, fixations depend on the location of objects or regions of interest in the environment. Thus we use the model to ground plan symbols to their representation in the environment. We report on early experimental results using mobile eye tracking glasses in a human-robot interaction setting, validating the usefulness of our model. We conclude with a discussion of how this model improves collaborative human-robot assembly operations by enabling intuitive interactions.**

I. Introduction

Despite significant recent advances in the autonomous capabilities of humanoid robots, much remains to be done before robots are able to function effectively in complex human environments. This is especially the case when robots require understanding of contextualized information within cluttered and dynamic environments. However, it is clear that people and robots may cooperatively perform tasks which neither one could complete independently. Such collaboration requires intuitive and efficient human-robot interfaces that impose minimal overhead on the agents involved. We propose a human-robot interface for collaborative tasks based on real time eye tracking for intention prediction.

Studies related to eye tracking during the execution of natural tasks report that most of the fixations are located on objects or locations in the environment which are relevant to the task [3]. This suggests that it is possible to infer the intention of the person (in other words the task they are performing or the plan they are executing) by learning a probabilistic generative model of fixations and then inverting this model in the Bayesian sense. Achieving such inference of motion intent allows the robot to take suitable actions such as to assist the human operator.

A generative model enables inferring latent variables that may be otherwise hard to measure, but still have significant influence on the behaviour of the human operator. For example, consider a person and a robot collaboratively constructing a multi-part assembly (see Figure 1). A first question in this setting is to estimate the extent of the effective workspace within which the person wishes to operate. At a higher level, the robot may need to recognize new objects (e.g. a screwdriver) that the person has begun to use. The robot knows the plan which the person is executing and it is aware that a new tool is to be used. Since it is natural for the person to look at the tool while using it (much more so, on average, than at other task-unrelated parts of the visual scene), fixations provide a noisy supervisory signal and enable the robot to better learn the appearance of the new tool. Thus, through the use of measurements offered by eye tracking, we are able to address questions related to the grounding of symbols.

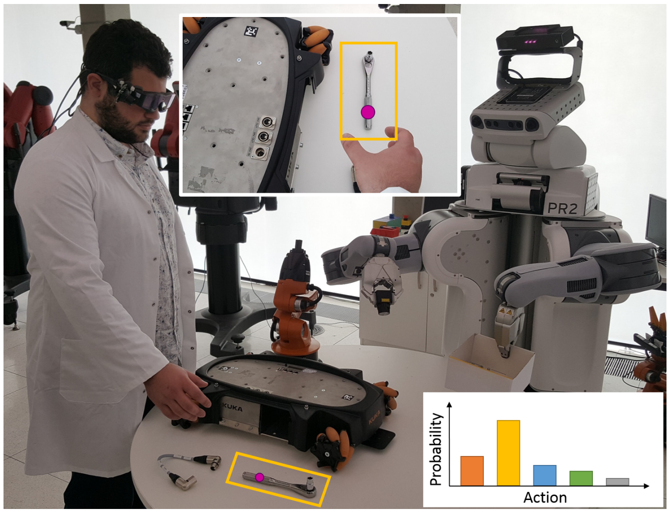
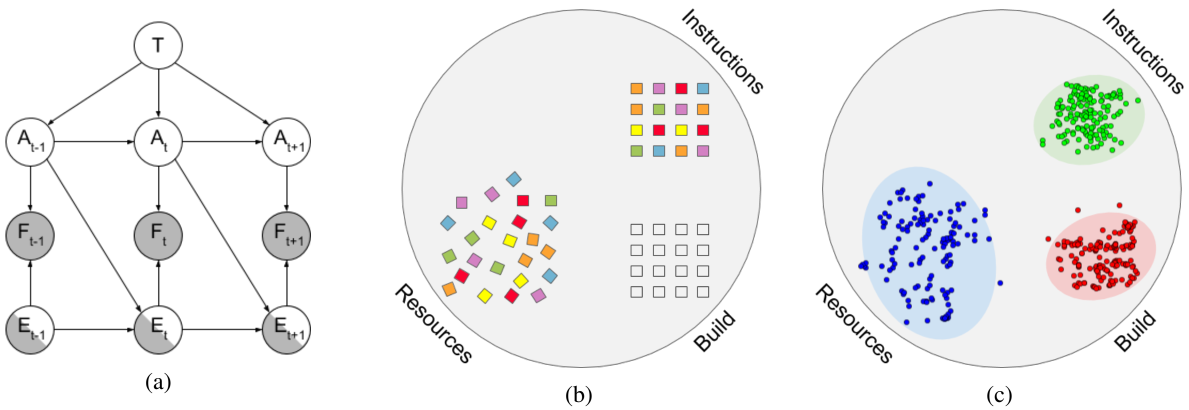


Fig. 1: A human and a robot collaboratively assemble another robot. The human wears eye tracking glasses enabling the robot to infer the actions which the person is executing. Fixation data is also used to teach the robot about the particular instance of the abstract symbols in the assembly plan.

II. Generative Model

Computational models of human fixations have been extensively researched in various domains such as reading or free image viewing. It is a well known fact that salient features such as high contrast edges or motion attracts eye fixations [5]. However, it is less clear how the task influences eye motions [6]. Yarbus first showed in 1967 that fixation patterns



(a) (b) (c)

Fig. 2: **Left:** The proposed graphical model expanded for 3 consecutive time steps. The action, A_t, executed by the person, depends on the task, T, they are working on. The resulting fixations, F_t, depend on the action and the state of the environment, E_t, which is composed of the locations of objects and regions related to the task. Shaded variables are observed and semi-shaded variables are partially observed. **Centre:** Diagram of the experimental setup in which a person rebuilds the pattern shown in the instructions area by using cubes from the available resources. **Right:** Fixations recorded during 1 trial, color coded for each area. The ellipses represent the learnt distribution $p(F_t|A_t)$ where green is $p(F_t|A_t = check)$, blue is $p(F_t|A_t = find)$, and red is $p(F_t|A_t = build)$

differed significantly when people observed an image and were asked to answer different questions about that image [7]. Borji et al. [2] demonstrated that it is possible to determine the task which the person is solving while observing the image by using aggregate features of the eye tracking data such as average fixation duration and mean distance between fixations. Following these results, an interesting question is whether it is possible to infer the intention of a person while executing a natural task instead of observing an image.

There have been numerous studies related to eye tracking in natural tasks such as driving, cooking, navigation and playing sports [3, 6]. The results from all those studies conclude that the majority of fixations are placed on objects or regions of interests relevant to the task. Thus, if a robot has access to eye tracking data it should be able to infer the task which the person is executing and assist accordingly. In order to achieve this we describe a probabilistic graphical model based on a hidden Markov model, which is shown in Figure 2a.

We denote the task which a person could be executing as $T \in \{T_1, T_2, \ldots, T_M\}$ where M is the number of possible tasks. Each task is completed through the execution of a sequence of actions. The action which a person could perform at time t is $A_t \in \{a_1, a_2, \ldots, a_N\}$ where N is the total number of possible actions. We assume that each action is associated to an object or a region which we will refer to as an item of interest. For example, the action "pick" is associated to an object which is to be picked and the action "put" is associated to a region where the object is to be put. Thus when action A_t is executed we expect to observe fixations F_t on the item of interest associated to the action. In comparison to previous probabilistic fixation models [4], we do not assume that A_{t-1} and A_t are independent given the task T. This enables us to work with more complex plans often found in natural tasks. If we observe the environment E_t at time t and have access to the physical position of the item of interest then we can define $F_t \in \mathbb{R}^3$ as the physical coordinates in 3D space of where the person will fixate. We now present the set of generative processes which are needed to complete the model definition.

A. Generating Fixations

One of the benefits of using eye tracking data is that even though it is locally noisy, people tend to move their eyes with substantial regularity when they perform a particular task [6]. Therefore, we could pre-define where a person would look if they were executing action A_t. A more flexible approach is to learn these models from empirical data. Starting with an initial understanding of objects in the scene E_t during the training stage, we model $p(F_t|A_t, E_t)$ as a normal distribution centred at the position of the item of interest of A_t with variance proportional to the size of the item. The state of the environment, E_t, may be encoded as prior knowledge or extracted from analysing the visual feed through standard methods for recognising objects. A low cost intermediate solution between assuming very strong prior knowledge and fully open ended object recognition is to utilise AR tags to index into a database of potential objects of interest.

B. Action Transitions

The quantity $p(A_t|T, A_{t-1})$ can be calculated by generating plans which fulfil the task and recording the frequency of transitions from A_{t-1} to A_t. This strategy is effective for tasks with well defined plans and low branching factor. However, we are interested in recognizing informative actions such as "looking at instructions", "inspecting a detail" or "searching for a tool". Those actions depend on many latent variables and as such it is not clear when they should occur in the plan. This makes it infeasible to generate all possible plans. Therefore, during the training stage, we learn the transition probabilities by calculating the maximum likelihood estimate from the training data. We also model the duration of each action as a normal distribution which is also learnt from data.

C. Environment Evolution

We describe how the environment changes given that action A_{t-1} was executed while the environment was in state E_{t-1} as $p(E_t|E_{t-1}, A_{t-1})$. Some actions such as moving objects have an observable effect on the environment, unlike others such as "look for a tool" which have no impact on the environment.

III. Task Inference and Symbol Grounding

Let us assume that we can observe all necessary features of the environment such as the position of objects and regions of interest. Once we have learnt the parameters of the probabilistic model in Figure 2a, we can run importance sampling in order to infer the latent variables of interest by generating potential fixation paths and comparing them with the evidence. We keep recursive estimates of the probability distribution over what action the person is executing $p(A_t|F_t, E_t)$, and secondly what task they are working on $p(T|F_t, E_t)$.

If we relax our assumption that the environment is fully observable, then we face the larger problem of grounding the abstract symbols in the plan to their particular instances in the environment. However, we have learnt an initial model of the structure of the task, so we can infer $p(E_t|A_t, F_t)$ given the observed fixations. We may thus locate objects or regions in the environment without explicitly detecting them. Or, the input image around the fixations may be cropped and used to train a detector. In the experimental scenario shown in Figure 1, the most probable action is to pick the wrench. However, the specific instance of the wrench is unknown, which is resolved by using fixation data evidence to obtain labelling information over the image.

IV. Experiments and Preliminary Results

In our experiments, the above mentioned model is applied to joint collaborative assembly. We use SMI Eye Tracking Glasses that we have augmented through the addition of an 120FPS camera in order to be able to run MonoSLAM for estimation of the head pose in addition to the raw fixation data. Thus, we can project the fixations onto the 3D environment instead of simply using the 2D fixation positions in the image.

A. Theoretical Investigation

Based on initial successes with simpler tasks, we designed and conducted an experiment which involves longer plans with high branching factor and actions which do not have an observable impact on the environment (e.g. "searching", "looking at instructions"). Similar to [1], we asked people to stand in front of a table and copy an example pattern by picking cubes from the resources area and placing them in the building area. The layout of the table is shown in Figure 2b. There are 10 different predefined patterns which a person could be building resulting in 10 distinct tasks. The actions that we take into account are *"check instructions"*, *"find cube"*, *"put cube"*. The environment variable E_t contains information about the position of each area and the position of each cube. Our aim is to predict online which pattern the person is building by observing only the resources area and the fixations data. Figure 2c shows part of the recorded fixations during the execution of the task and the learnt $p(F_t|A_t)$ distribution which is composed of 3 normal distributions corresponding to each action (item of interest).

B. Collaborative Assembly

Our main goal is to deploy the model in a collaborative assembly scenario where the human and the robot share the same workspace and work together in order to achieve a task. Mobile 3D eye tracking provides information from an additional modality which enables intention inference and localisation of objects which may be ambiguous otherwise. An earlier version of the proposed model, implemented with the Baxter robot to assist a human operator in performing assembly operations can be seen in this video demonstration: `https://youtu.be/2At2k2Gzd58`.

Our current work, involving implementation of the entire stack on the PR2 robot, is focussed on allowing the robot to more flexibly acquire models of the environment in the process of grounding instructions, utilising modalities including eye tracking to better disambiguate what would have been otherwise been ill-posed commands.

V. Conclusion

We report on the design and preliminary experiments with a human-robot interface that utilises evidence from a novel 3D mobile eye tracking system to infer motion intent. We show that our model is capable of symbol grounding in interactive tasks leading to improved context-aware collaboration.

References

[1] D Ballard, M Hayhoe, and J Pelz. Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, 1995.

[2] Ali Borji and Laurent Itti. Defending Yarbus: Eye movements reveal observers' task. *Journal of Vision*, 14(3):29, 2014.

[3] Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in cognitive sciences*, 9(4):188–94, 2005.

[4] Mary Hayhoe and Dana Ballard. Modeling task control of eye movements. *Current biology : CB*, 24(13):R622–8, 2014.

[5] Laurent Itti. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Visual Cognition*, 12(6):1093–1123, 2005.

[6] Michael Land and Benjamin Tatler. *Looking and Acting: Vision and Eye Movements in Natural Behaviour*. OUP Oxford, 2009.

[7] Benjamin W Tatler, Nicholas J Wade, Hoi Kwan, John M Findlay, and Boris M Velichkovsky. Yarbus, Eye Movements, and Vision. *i-Perception*, 1(1):7–27, 2010.