**Eye in hand Paper Review [Cognitive Robotics]**

Overview of Paper, for the first minute of the video.

The paper talks about how unknown the process of understanding the goal of an individual, using the input of information. Incoming flow of sensory information is used to understand, an agent’s goal.

The paper describes different sources of information, such as:

* Gaze Direction
* Arm Trajectory
* Hand Preshape

When determining the goal of a reach to grasp action.

The main unknown in human processing is understanding how much each information source contributes to the goal.

Some studies suggest, recording eye movements, while participants observed an actor reaching for and grasping one of two objects which requires two different grips.

Paper suggests some research suggests, gazes of eye movements suggest understanding on the attention of others.

Paper thinks, gaze is important in determining motor intentions.

Two experiments to understand this was given:

* Experiment 1: Which source of information participants value the most for obtaining actions?
* Experiment 2: Whether there is a difference in how these values are updated while unfolding of the action, reflecting a sensitivity for information.

Initial Idea for Implementation:

Arm Trajectory Prediction: <https://developers.google.com/ml-kit/vision/pose-detection>

Eye Cropped Eyesight: <https://medium.com/@stepanfilonov/tracking-your-eyes-with-python-3952e66194a6>

Hand Preshape: Could potentially, have a green blob on the top of someone’s hand, such that cv will crop everything within a square of it.

Paper states that experiment one, deals with two things,

* Information is progressively added, starting with gaze, and then shifting to arm trajectory and hand pre-shape.
* Both the gaze control and hand pre-shape information is unreliable, as it doesn’t complete related to final goal.

The dataset provided mentioned videos of conflicting sources to understand how reliance’s changes and what is used generally to obtain a target goal.

The conditions were:

* gaze congruent- preshape congruent.
* gaze congruent-preshape incongruent.
* gaze in-congruent-preshape congruent.
* gaze incongruent-preshape incongruent.

Experiment 1:

As the arm trajectory always did align to the actual goal needed.

In the experiment, many things were done to mitigated errors, which included, using a velocity-threshold algo to determine **eye saccades (shifts the eye gaze from one part to another)**, it also tracked two main regions, which included fixation cross (fixation AOI) and intended target (target AOI).

The experiment decided to remove trails in which eye tracked movement was not exhibiting a target-directed gaze, such as when:

* Not saccading to any of the goal areas of interest before the end of the video.
* User did not fixate on fixation AOI at the beginning of movement phase.

The paper then used the timing metric to understand how long it took to reach the goal AOI, before the action completed, calling it gaze arrival time taking a negative score, this was used to compute everyone’s ability to anticipate with the eyes the goal of the observed action.

The paper then talks about how it denoises and stabilises the pupil diameter data such as blinking etc, in page 4, which is not of relevance to us.

Experiment 2:

This is measure by understanding how the reliance on updating on unfolding actions, reflecting, here different durations were measured, like 3, 6, 9, 12, 15, or 17 frames after actor’s hand start moving, and then the participant was asked to make a prediction of the actual goal. This experiment measured if the correct goal was picked with options of small or big tomato, as well as this the mouse tracking was done to understand the complexity of the problem. It aided in understanding how different amounts of information informed the decision, with measures for removing inactive recordings that took more than 2000ms.

Results of the paper experiments:

A collage of a person holding a ball

Description automatically generated

Figure 1 Different Videos Used

Results of Experiment 1:

Using the metric of gaze arrival times mentioned in the paper metrics, (the time of getting to goal AOI subtracted to final frame ending).

It shows that based on this metric, the effect of hand preshape aided in earlier gazing of the correct target goal. The model determined that when conflicting sources was used then participants gazed at video interaction gaze for aid, however in cases where preshape was congruent, participants were faster and more accurate regardless of the information provided by the actor’s gaze.

Result of Experiment 2:

Participants accuracy increased as more information was available (more frames were available). In addition to this, participants experienced a learning effect where accuracy increased as experiment ensured.

From the experiment, incongruent actor’s gaze led to the lowest accuracy, meaning the most reliance to this specifically.

The paper states that interesting, these incongruent actor’s gaze videos became less determinantal as the experiment continued, meaning participants relied less on information provided completely from the actor’s gaze due to learning.

It states although complicated, that preshape hand aided in improving accuracy up until the longest duration in which hand trajectory then informed the target user of the action.

It also states that when both gaze and preshape was used, it would be much accurate if both were accurate and vice versa.

But interestingly, when gaze information was only given, and no preshape, accuracy was hit, when only the first part of the video was shown.

And that it was only when duration was the largest that performance changes was considering on the preshape.

The analyses also revealed a significant main effect of duration, indicating that RTs were faster as more visual detail of the actor’s action became available.

Understanding the results of the experiments mentioned.

**In the first experiment**, we tried to find out which sources of information participants value the most in the goals, it concluded that the most viewed thing was looking at the videos gaze.

**In the second experiment**, it was understanding whether these source of information weightings were constant or not.

This is supported by the evidence showing that when the hand preshape correctly cues the actor’s goal and/or the video duration increases, the information provided by the gaze decreases.

Experiment 2 show that, for longer videos, the importance lowers for gaze and rises for arm trajectory.

Response from paper, which is better than rewriting:

Our results also show that the actor’s hand preshape had widespread effects on participants’ prediction ability, affecting both their predictive eye movements and their mouse responses. Indeed, our results show that participants were much more accurate and faster in gazing at the object to be manipulated by the other’s hand when the actor’s hand preshape was congruent with the intended target of the actor’s hand movement, regardless of the information provided by the actor’s gaze.

Final Points:

Interestingly, when the preshape was incongruent instead of when gaze mislead cued her goal.

In addition, as time progressed, the participants relied more on arm movements, such as arm trajectory.

It states that the participants relied on the preshape even though it was only reliable 0.5 of the time. And that participants behaved on their prior beliefs about precision of a source, and failed to update these, making us assume that they do not update these priors during the experiment.

Also, that over time of the experiment, actor’s gaze was less reliantly used for accuracy, and was modulated by learning.

In sum, our results suggest that gaze information can affect the ability to predict the outcome of others’ actions, but only when no other information about their behavioural intention is provided.

In an addition, participants experience motor cues when preshaped hand with specific grip is available, even when unreliable, and choosing it even when gaze is contradicting.

A graph of a number of indicators

Description automatically generated with medium confidence

Now from the conclusion, eye gaze is dependent on when no other information is given, and then prehandle even when unreliable, but as time goes by, the arm trajectory gives more indication, however most important for gaze initially.

Possible issues with the paper:

It doesn’t consider that perhaps the reason why performance increases as the experiment continues is that even though there are 480 trails, they were limited combinations and could mean, speaking from experience of these psychology trails lack of learning and just plain remembrance due to fatigue of experiment as it ensues.

The paper mentions something called **Linear Mixed-Effects Modelling** which I am not aware of currently, looks like a method of modelling, linear and random variables within models.

Part 2: Computational Model

I would like to make a project that could obtain the same type of conclusions. In this project, I would like to understand the same two questions, which are:

Given 3 different inputs, which source of information does the model classify as most valuable while observing the movements.

And whether these values change over time from one source to another.

Here the probabilities that we want to figure out is the following:

This model will have only 2 goals, so either one or another option can be obtained.

The model may have 3 inputs of information and 2 outputs, with a cross entropy loss function and SoftMax for specifically the output of say Object A or Object B.

P(Object A | Eye Position, Arm Trajectory, Preshape of Hand)

The current scenarios that can occur (at different timesteps & both goals) are:

|  |  |
| --- | --- |
| **Not or Congruent Eye Position** | **Not or Congruent Pre-Shape** |
| Not | Not |
| Congruent | Not |
| Not | Congruent |
| Congruent | Congruent |

With pose estimation:

We can try to obtain a better accuracy, at the different stages of the model.

Part 3: Results & Conclusion