Int-HRL: Towards Intention-based Hierarchical Reinforcement Learning

Anna Penzkofer
University of Stuttgart
Stuttgart, Germany
anna.penzkofer@vis.uni-stuttgart.de

Simon Schaefer
Technical University of Munich
Munich, Germany
simon.k.schaefer@tum.de

Florian Strohm
University of Stuttgart
Stuttgart, Germany
florian.strohm@vis.uni-stuttgart.de

Mihai Bâce
University of Stuttgart
Stuttgart, Germany
mihai.bace@vis.uni-stuttgart.de

Stefan Leutenegger
Technical University of Munich
Munich, Germany
stefan.leutenegger@tum.de

Andreas Bulling
University of Stuttgart
Stuttgart, Germany
andreas.bulling@vis.uni-stuttgart.de

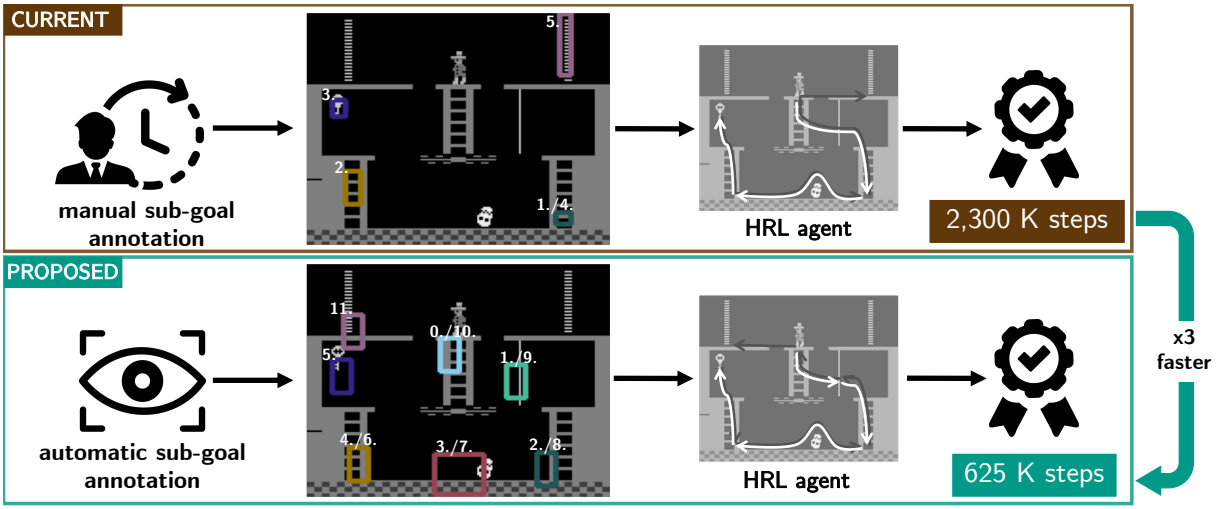

Figure 1: Previous methods require experts to manually annotate meaningful sub-goals for hierarchical reinforcement learning (HRL) agents. We propose automatic sub-goal extraction from human eye gaze, yielding a more robust and sample-efficient HRL agent that solves the first room of Montezuma's Revenge from the Atari2600 benchmark after only 625K steps.

ABSTRACT

While deep reinforcement learning (RL) agents outperform humans on an increasing number of tasks, training them requires data equivalent to decades of human gameplay. Recent hierarchical RL methods have increased sample efficiency by incorporating information inherent to the structure of the decision problem but at the cost of having to discover or use human-annotated sub-goals that guide the learning process. We show that intentions of human players, i.e. the precursor of goal-oriented decisions, can be robustly predicted from eye gaze even for the long-horizon sparse rewards task of Montezuma's Revenge – one of the most challenging RL tasks in the Atari2600 game suite. We propose *Int-HRL*: Hierarchical RL with intention-based sub-goals that are inferred from human eye gaze. Our novel sub-goal extraction pipeline is fully automatic and replaces the need for manual sub-goal annotation by human experts. Our evaluations show that replacing hand-crafted sub-goals with automatically extracted intentions leads to a HRL agent that is significantly more sample efficient than previous methods.

Proc. of the Adaptive and Learning Agents Workshop (ALA 2023), Cruz, Hayes, Wang, Yates (eds.), May 29-30, 2023, London, UK, https://alaworkshop2023.github.io/. 2023.

KEYWORDS

Hierarchical Reinforcement Learning, Intention Prediction, Eye Gaze, Montezuma's Revenge, Sub-goal Extraction

1 INTRODUCTION

Recent advances in artificial intelligence (AI) in general, and reinforcement learning (RL) in particular, have shown promising results in developing agents that can interact in complex environments and solve challenging real-world tasks, such as robotic manipulation at scale [16]. Despite these promising results, a key limitation of RL agents is that training them requires extensive exploration and training data. A large body of research [3, 6, 13, 26, 33] has thus relied on computer games and other simulated environments to develop and evaluate novel AI agents. One of the most popular testbeds are games from the Atari2600 suite implemented in the Arcade-Learning-Environment (ALE) [6]. The Atari2600 games are particularly useful to evaluate RL agents [35] as they not only have complex visuals but are also challenging for human players [21].

Research on the Atari2600 benchmark has focused on deep RL [3, 4, 26]. While deep RL agents, such as Agent57 [3], have successfully

beaten the human benchmark on all 57 Atari games, they are *sample inefficient* and, therefore, require an excessive amount of training. Moreover, deep RL methods suffer from a lack of explainability inherent to the deep neural networks used for Q-value estimation. A more promising approach is hierarchical RL (HRL) [18, 19, 32] that decomposes an RL problem into multiple sub-problems, thus also improving explainability. A key challenge with HRL is the decomposition of the task that often requires *manual and expert annotations*, which is tedious, time-consuming, and does not easily generalise to other tasks or games.

To address these limitations we propose a novel approach to automatically identify sub-goals in HRL from human eye gaze behaviour. Eye gaze is particularly promising as the gaze location has been linked to human intentions and goals [5, 10, 11, 15, 29]. We hypothesise that these intentions and goals can be further linked to sub-goals so, by predicting players' intentions from their gaze, sub-goals can be identified automatically. Inspired by prior work on gaze-based intention prediction [10, 11, 15], we extract four gaze features and train a Support Vector Machine (SVM) model. We evaluated the SVM on Montezuma's Revenge (MR), a long-horizon sparse reward game from the Atari2600 benchmark, with data from Atari-HEAD [35], a data set that offers gaze data in addition to human gameplay demonstrations. Our intention prediction model achieves an average accuracy of 75%, demonstrating the relation between intention and gaze behaviour, which motivates the automatic extraction of sub-goals for HRL agents. Finding useful sub-goals, which is also known as the *option discovery problem* [8], is a major issue in HRL. However, by using user intents and gameplay demonstrations, our method is able to not only refine and extract the sub-goals, but also the sequence in which these have to be solved to complete the game level. We then integrate the predicted sub-goals into the HRL framework hg-DAgger/Q [19] and show that this approach can solve the first room of MR three times more efficiently – improving sample efficiency from around 2.3 million to around 625 thousand training steps.

In summary, our work makes two distinct contributions: (1) We propose a novel method to predict sub-goals for HRL from eye gaze and human demonstration data. Gaze information is used to predict user intentions that are linked to the sub-goal locations, while demonstration data provides the order in which these sub-goals have to be solved to complete a task. (2) We evaluate our approach on Montezuma's Revenge from the Atari2600 benchmark and demonstrate significant improvements on two key limitations: sample efficiency and the need for manual expert annotations. These first results are promising and point towards new intention-based HRL methods that leverage both hierarchical methods and additional human behavioural data, such as eye gaze, to train more efficient agents that can solve complex visual problems.

2 RELATED WORK

Hierarchical Reinforcement Learning. Deep reinforcement learning (RL) has shown great results on the Atari benchmark but still struggles to learn robust value functions from sparse feedback in long-horizon games such as MR. Specifically, current state-of-the-art methods require frame samples in the range of billions, which forces researchers to develop elaborate distributed training schemes [3, 26] that still take a considerable amount of time to train [13]. HRL, on the other hand, offers a way of exploiting the hierarchical structure of decision-making tasks, guiding the agent towards meaningful sub-goals, effectively increasing the sample efficiency of agents. Moreover, agents achieving consecutive sub-goals, are directly understandable, making HRL particularly useful in domains, where explainability is required.

Early on, even before deep RL, two ideas have emerged in HRL: the options framework [30] and feudal networks [12]. Sutton et al. [30] have proposed to temporally extend actions into *options*, which are composed of a policy, a termination condition, and a set of states in which they could be applied [30]. They have shown that Q-learning could be generalised to learning policies over options and that learning inside these options, called "intra-option" learning, allowed the agent to learn about the respective options without executing them explicitly. Feudal networks, on the other hand, define a hierarchical structure of managers and sub-managers that are only privy to the space and temporal state at their granularity, effectively hiding information from their superior and providing rewards to their sub-managers even if their superior goal was not satisfied [12]. Both hierarchical frameworks have demonstrated much faster convergence than non-hierarchical methods in their respective maze scenarios.

More recently, Kulkarni et al. have proposed a hierarchical approach to induce goal-directed behaviour that does not use separate Q-functions as in the options framework [18]. This made their method scalable and promoted shared learning between options. To this end, they proposed a two-level framework in which the top-level agent (meta-controller) was responsible for choosing sub-goals while the low-level agent was concerned with achieving these goals. Le et al. have extended this approach by integrating the interactive imitation learning approach DAgger [27] into the meta-controller [19]. This, however, introduced the need for an expert during training. Their approach is also similar to feudal networks [12, 32] in their hierarchical structure, however, needs significantly less data as it does not use standard RL on the higher level. Vezhnevets et al. have later argued that a disadvantage of[18] is the need for pre-defined sub-goals and have chosen to learn goal embeddings implicitly [32]. In this work, we take the best of both worlds and leverage the information provided by gaze data to extract sub-goals independently. This allows us to use the more sample-efficient hierarchically guided method (hg-DAgger/Q) [19].

Another work developed concurrently with ours is based on the options framework but also defines intentions as fully satisfied if a sub-goal is reached and evaluates a reduction in available actions to the ones that are affordable in a given state (affordances) via attention. Nica et al. [23] introduce these *affordance-aware sub-goal options* with a respective model-free RL algorithm and find empirically in a MiniGrid domain that this yields better sample efficiency and higher performance in long-horizon tasks. While they also incorporate visual attention, they do so by applying it to limit an agent's action choices. In our work, on the other hand, we use visual attention maps generated from eye gaze data to extract meaningful sub-goals that can be directly selected by the meta-controller of our more feudal network-like architecture.

Intention Prediction. Intentions are goals and desires associated with a concrete plan, i.e. an intention causes a sequence of actions that lead to achieving a certain goal [2]. In other words,

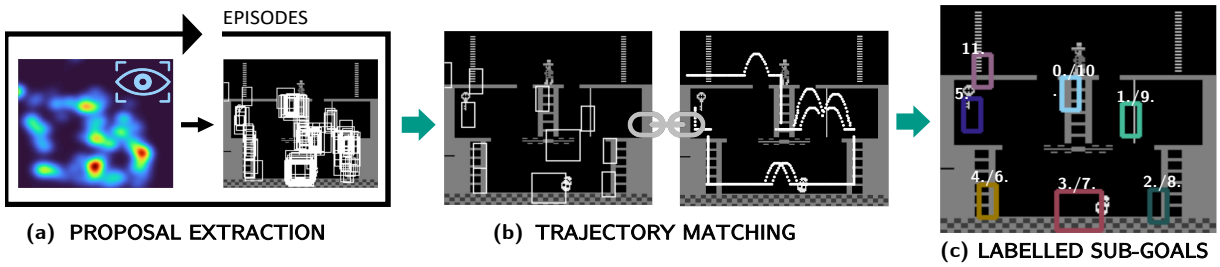

Figure 2: Sub-goal extraction pipeline: (a) proposal extraction is performed from human attention maps for each episode and resulting proposals are merged via non-maximum supression (NMS), then final proposals for one room are matched with human agents' trajectories (b), yielding labeled sub-goals and visitation order (c).

intentions are the precursor of actions, which poses the question whether human intentions are able to pose as sub-goals for HRL agents, where the hardest problem is to discover suitable sub-goals [8]. However, human intention prediction has never been done in this context before. Therefore, before we can replace hand-crafted sub-goals with human intention, we verify whether intention prediction is feasible on MR.

Model-free intention prediction models rely on eye gaze as the most important feature [5, 10, 11, 15, 29]. For a tabular summary of intention and activity recognition using eye gaze in Virtual Reality (VR), PC, table-top, and real-world environments we refer the reader to Chen and Hou [10]. The work of Huang et al. [15] is the most relevant to ours, as they consider intention prediction as a multi-class classification in a real-world scenario. They achieved 89% accuracy in their collaborative ingredient prediction task, where a customer instructs a worker to add displayed ingredients to a sandwich, and 76% accuracy with gaze features alone. The gaze features used in their SVM model were: *total duration of looks*, *most recently looked at*, *number of glances*, *duration of first glance*. We successfully test their model on MR with gaze data from the demonstration data set Atari-HEAD [35].

Belardinelli [5] offers a more general review on gaze-based intention estimation, identifying application areas of intention prediction as human-computer interfaces, human-robot interaction, and Advanced Driving Assistance Systems (ADAS) with relevant works from the last decade of research. However, the application of intentions to solve the *option discovery problem* in HRL, or in our case the *sub-goal discovery problem*, is to the best of our knowledge a novel idea and constitutes the main contribution of our work.

3 SUB-GOAL DISCOVERY

Prerequisites. MR is one of the most challenging games in the Atari2600 suite because of its long planning horizon and sparse rewards [19]. A RL agent only receives feedback sparingly, requiring many actions to achieve a small reward. Unlike similar long-horizon planning tasks, e.g. artificial grids [8, 23], MR is more challenging because it features different rooms that change according to the current level and collecting items allows for different actions in them. Therefore, to identify a specific state of the game it is necessary to know the position of the agent, room ID, level number, and the number of keys held [9]. The room ID is particularly important for our method because gaze data should be evaluated separately for each room so that gaze points can be mapped to the specific areas of interest.

The required state information can be extracted directly from the ALE via an environment wrapper called Atari Annotated RAM Interface (AtariARI) [1]. The wrapper parses information from the state variables in the ALE and makes it available for each environment step. However, the AtariARI wrapper was not used in the collection process of the data set Atari-HEAD [35]. To acquire the necessary labels subsequently, we simulated the episodes played by humans. This was possible as the original collection was done in a frame-by-frame mode, labeling each consecutive action.

Sub-goal Extraction. Our method for sub-goal extraction is inspired by previous research that showed that visual attention is a predictor of human intentions [5, 10, 11, 15, 24] and is further validated by successfully performing intention prediction on the extracted sub-goals in room one. The novel extraction pipeline is visualised in Figure 2: separate visual saliency maps are calculated for each episode and further isolated to only include the gaze data from the first room in the first level. These saliency maps are generated by adding each gaze point to the frame and passing a Gaussian filter over the generated fixation map with variance σ being one visual degree (pixel / visual degrees of the screen). Finally, the saliency map is normalised into the range of $[0, 1]$. An example saliency map can be seen in Figure 2 (a), where hot areas (red/yellow) indicate a high focus of attention across the selected time frame, and cold areas (blue) were not gazed upon at all.

After generating saliency maps for all episodes, each saliency map was thresholded , i.e. only values above 0.4 were kept. This threshold was fine-tuned to yield the best results over the entire sub-goal extraction pipeline, a qualitative assessment of other thresholds can be seen in Figure 3. Then, sub-goal proposals were generated, by drawing an agent-sized bounding box around each remaining saliency map point. These proposals were then processed with a custom implementation of the non-maximum suppression algorithm (NMS) [22]. In general, NMS is applied to suppress overlapping bounding boxes if they exceed an Intersection-over-Union (IoU) threshold, and, in our case, boxes with higher saliency values were favoured. Then, the remaining overlapping boxes were merged into one. After the number of sub-goal proposals for each episode was greatly reduced in this manner, the process was repeated to combine proposals across all episodes, yielding a final number of 11 possible sub-goals for room one, as shown in Figure 2 (b).

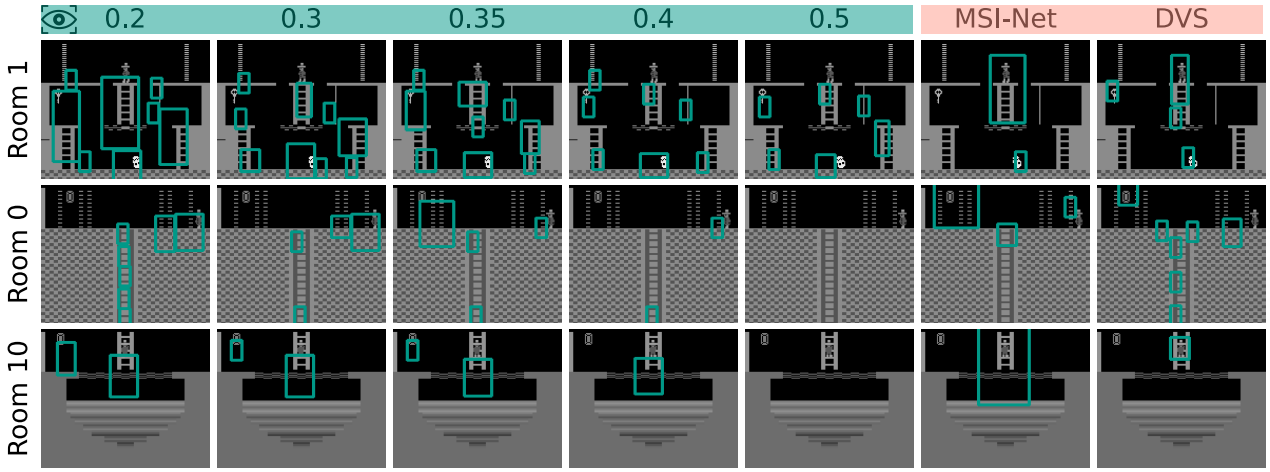

Figure 3: Sub-goal extraction examples on three rooms of MR with different saliency map thresholds from 0.2 to 0.5 on human gaze data, as well as with saliency maps generated by the MSI-Net [17] and DVS [20] saliency models with fixed threshold of 0.4.

With the definition of intention in mind, where intention directly leads to goal-directed behaviour [2], it is intuitive to only include sub-goals as labels for intentions that are visited during gameplay. Therefore, we ran another simulation of the game data from human players to find the sub-goals that were visited and in which order (trajectory matching). As MR is considered to be an almost deterministic game, i.e. there exists a best sequence of sub-goals, this resulted in almost identical orders across episodes. The remaining discrepancies were rectified by implementing a majority vote.

Overall, this extraction procedure resulted in 7 remaining sub-goals, labeled in order as shown in Figure 2 (c): moving from the middle ladder (0) to rope (1) and bottom right ladder (2), to crossing the middle area with the dangerous and dynamic skull (3), to the bottom left ladder (4), climbing up to collect the key (5), and then reversing this order to get to the left door (11). Interestingly, in all the episodes collected of human gameplay, only the left door was used, most likely because this is the best route suggested in MR solution guides.

In comparison to the four hand-crafted sub-goals selected by the HRL framework of Le et al. [19] (top row of Figure 1), which they hand-picked from the six sub-goals manually selected by Kulkarni et al. [18], our automatic pipeline extracted the same goals and added more areas of interest. In detail, Kulkarni et al. originally performed object detection on the image of room one and then chose the two doors, the three ladders, and the key as entities to define relational goals in the form of *agent reaches goal*.

Sub-goal Analysis. We further analysed our sub-goal extraction pipeline (Figure 2) qualitatively by generating proposals and final sub-goals for additional rooms of MR, with different saliency map thresholds, but also with artificial saliency maps generated by saliency models [17, 20], the results of which can be seen in Figure 3. We showcase *Room 1* as the starting point of the game, *Room 0* as the second room reached when choosing the left door, and *Room 10* as it features a special room layout. The saliency map threshold is a hyperparameter that needs to be finetuned on the overall extraction, where we have chosen 0.4 as it includes all hand-crafted sub-goals proposed by prior work [18, 19] with the meaningful addition of the area around the skull, without adding insignificant goals as with lower thresholds, but still including the door, which would be left out by a higher threshold. While there are no hand-crafted sub-goals to compare to for other rooms of MR, we can see that the pipeline also selects meaningful sub-goals, e.g. the bottom pathway in *Room 0*, the disappearing floor in *Room 10*, or the diamonds that give an external reward in both. In contrast, artificially generated saliency maps by MSI-Net [17], a standard saliency model with state-of-the-art results on the saliency benchmark CAT2000 [7], and DVS [20], a saliency model optimised for data visualisations, have a predominant focus on the agent itself and otherwise fail to find important steps like the doors, even though they are highlighted in the same colour as the key. Note here that saliency map prediction was done on the RGB images.

Intention Prediction. For testing the intention prediction model of Huang et al. [15] on our extracted sub-goals, we preprocessed the gaze data following prior work [10, 11], extracting saccade and fixation events and calculating the four features: *total duration of looks*, *most recently looked at*, *number of glances*, *duration of first glance*. We then implemented intention prediction as a multi-class classification for the 7 sub-goals of room one with a SVM. We achieve an average prediction accuracy of 75% in a 10-fold cross-validation, which is significantly better than a random model and also outperforms results reported on other data sets [5, 15]. We argue that this corroborates the efficacy of using human gaze data as an indicator of intention and motivates the extraction of sub-goals for HRL from human intention.

4 INTENTION-BASED LEARNING

Baseline. One approach for solving long-horizon decision-making tasks is HRL, where two popular frameworks emerged in the past: the options framework [23, 30] and feudal networks [12, 32]. Building upon a feudal architecture, by combining deep HRL with predefined sub-goals, Kulkarni et al. [18] are able to outperform naïve deep Q-learning. Their h-DQN model was tested on two delayed-reward domains, including the first room of MR, where their approach is able to reach the door after 2.5 M samples. Taking the idea

further, Le et al. [19] combined imitation learning (IL) with HRL showing that their hierarchical guidance model (hg-DAgger/Q) significantly reduces expert effort compared to other interactive IL approaches and is also able to learn faster and more robustly than h-DQN, solving the first room of MR after 2.3 M steps.

Our baseline, the hg-DAgger/Q model by Le et al., consists of two levels, the meta-controller level, and the agents level. They use the data aggregation method DAgger [27] on the top level, which trains the meta-controller policy with iteratively aggregated data sets. The meta-controller is used to predict one of the four hand-crafted sub-goals, initiating the corresponding agent.

On the low level, Le et al. used a double deep Q-Network (DDQN) [31] with prioritised experience replay [28]. A major difference between their approach and h-DQN [18] is that separate agents are trained for each sub-goal instead of passing the goal vector as a feature into a single policy network. This ensures the mitigation of the issue of catastrophic forgetting and also has the advantage of separate exploration schedules. However, maintaining a separate network for each sub-goal is not scalable across different rooms of MR.

Combining the low-level RL agents with hierarchical guidance from the meta-controller ensures that the experience buffer for the DDQN only contains valuable samples for the next sub-goal, as wrong meta-controller choices terminate the episode. Le et al. argue that this is the main reason for their higher robustness in training. However, Le et al. also report that their architecture only learned all sub-goals successfully in 50 out of 100 trials. This high variability is most likely due to different implementations and random seeding, an issue common in RL [14], which would also explain, why we were unable to reproduce their results. Consequently, we will compare our results to the ones reported in their paper.

In summary, hg-DAgger/Q [19] is significantly more sample efficient than other methods [18, 32]. However, it requires an expert at training time to select hand-crafted sub-goals and only implements a rudimentary sub-goal check.

Model. Similar to [19], we use a hierarchical reinforcement learning approach, with 8 possible actions (no action, cardinal moving directions, jumping up, left, and right). Starting with a custom implementation of hg-DAgger/Q [19], we tested different approaches to making training more stable. By using a Dueling deep Q-Network (DQN) architecture [34] in addition to the DDQN [31] and including the lower left ladder (see Fig. 1 top row goal 2) as an additional hand-crafted goal, we were able to train a model to reach the first external reward, the key. To further improve the model performance, we replace the hand-crafted sub-goals with the fields of interest derived from human gaze data, as previously described, thereby expanding the set of sub-goals used in [18, 19]. Sub-goals are a way of providing pseudo rewards [18, 19, 30] to populate the sparse reward map in MR. Next to the sub-goal reached reward $R_{\text{sub-goal}}$, we introduce a dense reward signal R_{dir}, R_{dist}, and R_{step} to further stabilise training.

$$R = R_{\text{sub-goal}} + \alpha R_{\text{dir}} + \beta R_{\text{dist}} + \gamma R_{\text{step}} \tag{1}$$

We define a direction reward R_{dir} to steer the agent in the direction of the next sub-goal. It is computed as the scalar product of the selected action's direction vector $\vec{a}$ and the vector between the next and previous goal $\vec{g} = G_{prev} - G_{next}$:

$$R_{\text{dir}} =< \vec{a}, \vec{g} > \tag{2}$$

The distance reward R_{dist} guides the agent to the next sub-goal by minimising the euclidean distance between the agent and current goal d_{ac}, as well as previous goal d_{ap}, and the distance between previous and current goal d_{pc}:

$$R_{\text{dist}} = \frac{\sqrt{d_{ap}} - \sqrt{d_{ac}}}{\sqrt{d_{pc}}} \tag{3}$$

The step reward R_{step} penalises each time-step used to reach the next goal with a constant $\tau = 0.001$, which is also used to scale R_{dist} and R_{dir}.

The distance reward R_{dist} requires the agent location at each step, we propose to either use a specifically trained object detector, which was tested with a pre-trained FasterRCNN model [25] fine-tuned on 100 manually labelled training examples, or to use the RAM state labels provided via the AtariARI Wrapper [1]. Both of these approaches are much more robust than the rudimentary approach used in [19] and, additionally, are able to track non-static regions of interest.

Results. We evaluated the full-sequence hierarchical model without any dense reward (*fullmodel*), a single-agent model with negative step reward and goal feature: $\gamma = 1$ (*singlegoal*), and the single-agent model with a distance reward: $\beta = 1$ (*singledist*) in comparison to the direction reward: $\alpha = 1$ (*singledir*). To further help the single-agent models, we passed the sub-goals as additional feature vectors. Throughout all experiments, we use an ϵ-greedy exploration policy with a linear scheduler ranging from 1 to 0.02 in 200 K steps, a prioritized replay buffer [28], and a learning rate of 0.0001. Results are shown in Figure 4. Each sub-goal is considered learned when the trailing performance is above 0.9, i.e. when the agent reaches the goal 90 out of 100 trials. Directly compared models are trained with the same random seed to ensure comparability [14].

The full-sequence model is the first to successfully learn to solve the first room of MR by reaching the left door. The final goal (11) is consistently reached by the *fullmodel* after only 625K steps. Hence, our model is more than three times more sample efficient than our baseline [19], which needs 2.3M samples to complete the first room. Furthermore, our new framework works without an expert, as we extracted the chosen sub-goals previously from gaze data and perform a true goal reached check via RAM state mapping. Overall, given labelled gaze demonstration data, the pipeline can be trained end-to-end with no manual effort.

In comparison to the full-sequence model, the single-agent model with the goal feature and small negative rewards per step does not succeed at all. As expected, the single Dueling DDQN agent suffers from a lack of *exploration* [19], i.e. as Figure 4 shows, the *singlegoal* model fails to learn anything new and gets stuck at the first sub-goal. Resetting the schedule when a new sub-goal is explored was tested, but did not succeed as it resulted in ϵ being too high which also prevents learning. Further ideas for solving insufficient exploration are included in the discussion and are left for future work.

Providing a more dense reward structure by adding a distance or direction reward improves the single-agent model as it increases sample efficiency so that the model learns to reach sub-goal 4, i.e.

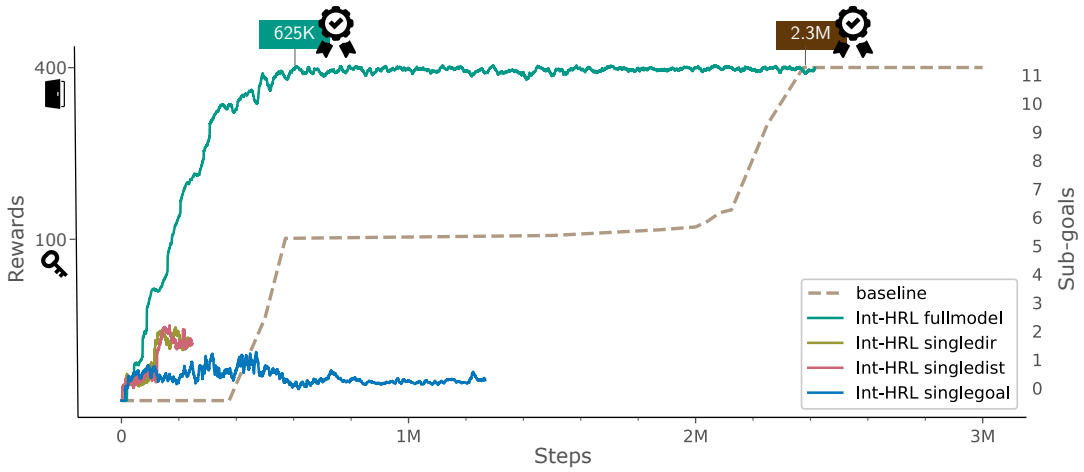

Figure 4: Step-wise sub-goal trailing performance of Int-HRLand baseline with sub-goal 5 as a first external reward from the key and sub-goal 11 the right/left door, which completes the first room.

passing the skull and reaching the lower left ladder (see the *singledist* and *singledir* model in Figure 4). In our trial, the direction reward even facilitates learning to reach the key sub-goal (5), which provides the first external reward. Interestingly, the direction reward is simpler to implement as it does not require knowledge of the agent's location. Both models still suffer from the issues encountered by the single-agent model approach and performance deteriorates after 200K steps when the ϵ-exploration schedule reaches its final value. However, this confirms that more intrinsic rewards improve performance and should therefore be incorporated into future models.

5 DISCUSSION AND OUTLOOK

We have shown that gaze features are indicative of intentions by successfully training a simple intention prediction model on the first room of the Atari game MR. The model predicts the next intention of an agent with 75% accuracy, thus validating the relation between eye gaze and intentions, and could be easily extended to a more powerful classification model by using more detailed fixation and saccade dynamics [11], or with neural networks and temporal information [5]. Further, we developed a novel sub-goal extraction pipeline from gaze data. To this end, we labelled an available demonstration data set via simulation, analysed the visual attention heatmaps for each room, and aligned the proposals with the agent trajectories. This process yields sub-goals that are on par with hand-crafted ones from prior work [18, 19]. We demonstrate the efficacy of our sub-goal extraction pipeline by using the extracted sub-goals to train an HRL agent that can solve the first room of MR significantly more sample efficiently than any previous method. Moreover, our pipeline is fully automatic and allows for a transparent explanation of agent behaviour. In comparison to previous methods, where sub-goals have been chosen manually without further analysis [18, 19].

Generalisability to other games. One requirement for our approach is a fixed layout of rooms for extracting meaningful information from gaze data. Areas of interest need to be stationary enough for a high duration of attention and depict isolated or unique objects. We have chosen the long-horizon sparse reward game MR of the Atari2600 suite because it can be structured into sub-tasks across different static rooms and standard RL still struggles with solving it efficiently. Other games available in the Atari-HEAD [35] data set are not suitable for this analysis because agents and sprites can move across all lanes, because objects of interest scroll too fast across the game screen, or because areas of interest are trivial, as in shooter games where all sprites are at the top of the screen and agent movement is restricted to be horizontal across the bottom. Other games similar to MR are *Hero* and *Venture*, where different rooms need to be navigated, which include static sprites or objects and clear goals. While *Hero* is structured like a search tree, iteratively expanding the depth of exploration for solving a level by finding people lost in the caverns, *Venture* is like a Maze with an overview screen from which the agent can reach different rooms to find treasure. They were not selected because they are not considered particularly difficult for RL; however, it will be interesting to see whether the intention predictor can add explainability to agents for these games in future work.

Sacalability of HRL method. While the full-sequence model has outperformed all other baselines tested on the first room of MR in terms of sample efficiency [18, 19, 32], it needs to be more scalable to solve the entire game. The HRL approach based on Le et al. [19] requires the separate handling of 12 agents in the first room of MR but an additional 23 rooms need to be explored to solve the first level. While both [18, 32] only use a single low-level agent, the former's successful trials were not reproducible [19] and the latter requires 200M samples for the first room alone, most likely getting close to the 10 billion samples needed by standard deep RL approaches [3]. We have tested single agents with more dense reward structures (R_{step}, R_{dist}, R_{dir}), however, were unable to circumvent the issue of *insufficient exploration* in single agents. In future work, we would like to address this by adding per-episode and full-game novelty values as intrinsic rewards, which succeeded in deep RL methods [3, 4].

ACKNOWLEDGMENTS

Anna Penzkofer was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2075 – 390740016. Simon Schaefer was supported by TUM AGENDA 2030, funded by the Federal Ministry of Education and Research (BMBF) and the Free State of Bavaria under the Excellence Strategy of the Federal Government and the Länder as well as by the Hightech Agenda Bavaria. Mihai Bâce was funded by a Swiss National Science Foundation (SNSF) Postdoc.Mobility Fellowship (grant number 214434). Florian Strohm and Andreas Bulling were funded by the European Research Council (ERC) under the grant agreement 801708.

REFERENCES

[1] Ankesh Anand, Evan Racah, Sherjil Ozair, Yoshua Bengio, Marc-Alexandre Côté, and R. Devon Hjelm. 2020. Unsupervised State Representation Learning in Atari. https://doi.org/10.48550/arXiv.1906.08226 arXiv:1906.08226 [cs, stat].

[2] Janet Wilde Astington. 1994. *The Child's Discovery of the Mind:*. Harvard University Press, Cambridge, MA.

[3] Adrià Puigdomènech Badia, Bilal Piot, Steven Kapturowski, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, and Charles Blundell. 2020. Agent57: Outperforming the Atari Human Benchmark. https://doi.org/10.48550/arXiv.2003.13350 arXiv:2003.13350 [cs, stat].

[4] Adrià Puigdomènech Badia, Pablo Sprechmann, Alex Vitvitskyi, Daniel Guo, Bilal Piot, Steven Kapturowski, Olivier Tieleman, Martín Arjovsky, Alexander Pritzel, Andew Bolt, and Charles Blundell. 2020. Never Give Up: Learning Directed Exploration Strategies. https://doi.org/10.48550/arXiv.2002.06038 arXiv:2002.06038 [cs, stat].

[5] Anna Belardinelli. 2023. Gaze-based intention estimation: principles, methodologies, and applications in HRI. https://doi.org/10.48550/arXiv.2302.04530 arXiv:2302.04530 [cs].

[6] Marc G. Bellemare, Yavar Naddaf, Joel Veness, and Michael Bowling. 2013. The Arcade Learning Environment: An Evaluation Platform for General Agents. *Journal of Artificial Intelligence Research* 47 (June 2013), 253–279. https://doi.org/10.1613/jair.3912 arXiv:1207.4708 [cs].

[7] Ali Borji and Laurent Itti. 2015. CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research. https://doi.org/10.48550/arXiv.1505.03581 arXiv:1505.03581 [cs].

[8] Matthew Botvinick and Ari Weinstein. 2014. Model-based hierarchical reinforcement learning and human action control. *Philosophical Transactions of the Royal Society B: Biological Sciences* 369, 1655 (Nov. 2014), 20130480. https://doi.org/10.1098/rstb.2013.0480 Publisher: Royal Society.

[9] Yuri Burda, Harrison Edwards, Amos Storkey, and Oleg Klimov. 2018. Exploration by Random Network Distillation. https://doi.org/10.48550/arXiv.1810.12894 arXiv:1810.12894 [cs, stat].

[10] Xiao-Lin Chen and Wen-Jun Hou. 2022. Gaze-Based Interaction Intention Recognition in Virtual Reality. *Electronics* 11, 10 (Jan. 2022), 1647. https://doi.org/10.3390/electronics11101647 Number: 10 Publisher: Multidisciplinary Digital Publishing Institute.

[11] Brendan David-John, Candace Peacock, Ting Zhang, T. Scott Murdison, Hrvoje Benko, and Tanya R. Jonker. 2021. Towards gaze-based prediction of the intent to interact in virtual reality. In *ACM Symposium on Eye Tracking Research and Applications (ETRA '21 Short Papers)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3448018.3458008

[12] Peter Dayan and Geoffrey E Hinton. 1992. Feudal Reinforcement Learning. In *Advances in Neural Information Processing Systems*, Vol. 5. Morgan-Kaufmann. https://papers.nips.cc/paper/1992/hash/d14220ee66aeec73c49038385428ec4c-Abstract.html

[13] Jiajun Fan. 2022. A Review for Deep Reinforcement Learning in Atari:Benchmarks, Challenges, and Solutions. https://doi.org/10.48550/arXiv.2112.04145 arXiv:2112.04145 [cs].

[14] Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. 2019. Deep Reinforcement Learning that Matters. https://doi.org/10.48550/arXiv.1709.06560 arXiv:1709.06560 [cs, stat].

[15] Chien-Ming Huang, Sean Andrist, Allison Sauppé, and Bilge Mutlu. 2015. Using gaze patterns to predict task intent in collaboration. *Frontiers in Psychology* 6 (2015). https://www.frontiersin.org/articles/10.3389/fpsyg.2015.01049

[16] Dmitry Kalashnkov, Jake Varley, Yevgen Chebotar, Ben Swanson, Rico Jonschkowski, Chelsea Finn, Sergey Levine, and Karol Hausman. 2021. MT-OPT: Continuous Multi-Task Robotic Reinforcement Learning at Scale. *arXiv* (2021).

[17] Alexander Kroner, Mario Senden, Kurt Driessens, and Rainer Goebel. 2020. Contextual encoder–decoder network for visual saliency prediction. *Neural Networks* 129 (Sept. 2020), 261–270. https://doi.org/10.1016/j.neunet.2020.05.004

[18] Tejas D. Kulkarni, Karthik R. Narasimhan, Ardavan Saeedi, and Joshua B. Tenenbaum. 2016. Hierarchical Deep Reinforcement Learning: Integrating Temporal Abstraction and Intrinsic Motivation. https://doi.org/10.48550/arXiv.1604.06057 arXiv:1604.06057 [cs, stat].

[19] Hoang M. Le, Nan Jiang, Alekh Agarwal, Miroslav Dudík, Yisong Yue, and Hal Daumé III. 2018. Hierarchical Imitation and Reinforcement Learning. https://doi.org/10.48550/arXiv.1803.00590 arXiv:1803.00590 [cs, stat].

[20] Laura E. Matzen, Michael J. Haass, Kristin M. Divis, Zhiyuan Wang, and Andrew T. Wilson. 2018. Data Visualization Saliency Model: A Tool for Evaluating Abstract Data Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 24, 1 (Jan. 2018), 563–573. https://doi.org/10.1109/TVCG.2017.2743939 Conference Name: IEEE Transactions on Visualization and Computer Graphics.

[21] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin A. Riedmiller. 2013. Playing Atari with Deep Reinforcement Learning. *ArXiv* abs/1312.5602 (2013).

[22] A. Neubeck and L. Van Gool. 2006. Efficient Non-Maximum Suppression. In *18th International Conference on Pattern Recognition (ICPR'06)*, Vol. 3. 850–855. https://doi.org/10.1109/ICPR.2006.479 ISSN: 1051-4651.

[23] Andrei Nica, Khimya Khetarpal, and Doina Precup. 2022. The Paradox of Choice: Using Attention in Hierarchical Reinforcement Learning. https://doi.org/10.48550/arXiv.2201.09653 arXiv:2201.09653 [cs].

[24] Jella Pfeiffer, Thies Pfeiffer, Martin Meißner, and Elisa Weiß. 2020. Eye-Tracking-Based Classification of Information Search Behavior Using Machine Learning: Evidence from Experiments in Physical Shops and Virtual Reality Shopping Environments. *Information Systems Research* 31, 3 (Sept. 2020), 675–691. https://doi.org/10.1287/isre.2019.0907 Publisher: INFORMS.

[25] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2016. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. https://doi.org/10.48550/arXiv.1506.01497 arXiv:1506.01497 [cs].

[26] Jerome Revaud, Philippe Weinzaepfel, César De Souza, Noe Pion, Gabriela Csurka, Yohann Cabon, and Martin Humenberger. 2019. R2D2: Repeatable and Reliable Detector and Descriptor. https://doi.org/10.48550/arXiv.1906.06195 arXiv:1906.06195 [cs].

[27] Stephane Ross, Geoffrey Gordon, and Drew Bagnell. 2011. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 627–635. https://proceedings.mlr.press/v15/ross11a.html ISSN: 1938-7228.

[28] Tom Schaul, John Quan, Ioannis Antonoglou, and David Silver. 2016. Prioritized Experience Replay. https://doi.org/10.48550/arXiv.1511.05952 arXiv:1511.05952 [cs].

[29] Ronal Singh, Tim Miller, Joshua Newn, Eduardo Velloso, Frank Vetere, and Liz Sonenberg. 2020. Combining gaze and AI planning for online human intention recognition. *Artificial Intelligence* 284 (July 2020), 103275. https://doi.org/10.1016/j.artint.2020.103275

[30] Richard S. Sutton, Doina Precup, and Satinder P. Singh. 1998. Intra-Option Learning about Temporally Abstract Actions. In *Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 556–564.

[31] Hado van Hasselt, Arthur Guez, and David Silver. 2015. Deep Reinforcement Learning with Double Q-learning. https://doi.org/10.48550/arXiv.1509.06461 arXiv:1509.06461 [cs].

[32] Alexander Sasha Vezhnevets, Simon Osindero, Tom Schaul, Nicolas Heess, Max Jaderberg, David Silver, and Koray Kavukcuoglu. 2017. FeUdal Networks for Hierarchical Reinforcement Learning. https://doi.org/10.48550/arXiv.1703.01161 arXiv:1703.01161 [cs].

[33] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. 2019. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature* 575, 7782 (2019), 350–354.

[34] Ziyu Wang, Tom Schaul, Matteo Hessel, Hado van Hasselt, Marc Lanctot, and Nando de Freitas. 2016. Dueling Network Architectures for Deep Reinforcement Learning. https://doi.org/10.48550/arXiv.1511.06581 arXiv:1511.06581 [cs].

[35] Ruohan Zhang, Calen Walshe, Zhuode Liu, Lin Guan, Karl Muller, Jake Whritner, Luxin Zhang, Mary Hayhoe, and Dana Ballard. 2020. Atari-HEAD: Atari Human Eye-Tracking and Demonstration Dataset. *Proceedings of the AAAI Conference on Artificial Intelligence* 34, 04 (April 2020), 6811–6820. https://doi.org/10.1609/aaai.v34i04.6161 Number: 04.