Motivated Reinforcement Learning for Improved Head Actuation of Humanoid Robots

Jake Fountain, Josiah Walker, David Budden, Alexandre Mendes and Stephan K. Chalup

The University of Newcastle, Australia
{jake.fountain, josiah.walker, david.budden}@uon.edu.au,
{alexandre.mendes, stephan.chalup}@newcastle.edu.au

Abstract. The ability of an autonomous agent to self-localise within its environment is critically dependent on its ability to make accurate observations of static, salient features. This notion has driven considerable research into the development and improvement of feature extraction and object recognition algorithms, both within RoboCup and the robotics community at large. Instead, this paper focuses on the rarely-considered issue imposed by the limited field of view of humanoid robots; namely, determining an optimal policy for actuating a robot's head, to ensure it observes regions of the environment that will maximise the positional information provided. The complexity of this task is magnified by a number of common computational issues; specifically high dimensional state spaces and noisy environmental observations. This paper details the application of motivated reinforcement learning to partially overcome these issues, leading to an 11% improvement (relative to the null case of uniformly distributed actuation policies) in self-localisation and ball-localisation for an agent trained online for less than one hour. The method is demonstrated as a viable method for improving self-localisation in robotics, without the need for further optimisation of object recognition or tuning of probabilistic filters.

Keywords: motivated reinforcement learning, localisation, Fourier basis, head actuation, simulated curiosity

1 Introduction

Effective localisation is essential to solving tasks such as the Robocup soccer competition. The behaviour of a robot is a function strongly dependent on the robot's localisation model. Inaccurate localisation leads to ineffective behaviour and poor performance. In a game of humanoid robot soccer, successful localisation of the robot and field objects relies on the vision system to measure relative field object locations. Moreover, the behaviour of the head determines the efficacy of the vision system; limited field of view implies that all field objects cannot be measured at once. The head actuation problem involves choosing neck motor positions to optimise localisation. If the robot is completely unlocalised,

panning the head can be used to localise at least partially. The problem of interest involves achieving and maintaining a high level of localisation assuming that initially the robot is partially localised. Since there are a finite number of useful field objects, this problem can be abstracted, with inverse kinematics, to that of sequentially choosing field objects on which to focus gaze to minimise localisation uncertainty while playing soccer.

Wong, et al. [1], used humanoid robots to model human gaze behaviour in urban environments. They utilised a gaze direction model based on visually salient objects placed in a model urban environment. If a robot saw a salient object it would slow down to view it for a period of time before scanning for other objects as it continued to walk. This behaviour was then used to benchmark the performance of gaze vector computer vision techniques; that is, measuring where the robot is looking with a wide field of view camera. We postulate that this anthropomorphic behaviour is desirable in making head actuation decisions to localise humanoid robots effectively. Therefore, we utilised motivated reinforcement learning techniques [2] to implement 'curious', anthropomorphic behaviours with the goal of optimising localisation in the RoboCup KidSize soccer league [3].

There are two types of field objects important to localisation of a humanoid robot while playing soccer: *landmarks* and *objects*. Landmarks are used to localise the robot while objects must be localised in the world model by measuring position relative to a localised robot, and then transforming the information into global coordinates [4]. In our approach, the model of the soccer field world is a collection of Unscented Kalman Filters [5]; each robot maintains a filter for its own position and a filter for the ball's position. Each filter is a Gaussian probability distribution for the possible field locations of an object. This probabilistic model allows uncertainty to be managed quantitatively for each robot and object, and this allows the robot to assess its localisation performance without external reference.

The head actuation problem was phrased as a Markov decision process and solved using Q-learning [6] with a motivated agent. A motivated agent uses the concept of novelty to seek out events which are similar-yet-different to previous experiences; this *curious* behaviour was useful in partitioning head actuation decisions between measuring landmarks and objects.

2 Motivated Reinforcement Learning

Reinforcement learning is a form of machine learning used to solve problems involving a series of decisions made by an *agent* based on perceptions of its *environment*, with a metric indicating performance after every decision [7]. This type of problem is called a Markov decision process, and is described fully by: a state space S; a set of actions A and a function $\Lambda : S \to 2^A$ where $\Lambda(s)$ is the subset of actions available in state $s \in S$; a transition function $T : S \times A \times S \to [0, 1]$ describing the probability of state transitions; and a reward function $R : S \times A \to \mathbb{R}$ [6]. Here $2^A = \{U : U \subseteq A\}$ is the power set of A, or the set of

all subsets of A. We additionally simplify the problem for the purpose of head actuation by including the assumption that $\Lambda(s)$ is finite and discrete for each $s \in S$. Q-learning is a method of learning the optimal value function Q^* and can be performed as an *online* learning method; that is, the model of Q^* is updated between actions to reflect experiences. Actions are chosen from successive states to affect the environment and thus explore the state action space $S \times A$. After each action, a function $Q : S \times A \to \mathbb{R}$ is updated to approximate Q^* using the rule

$$Q(s,a) \leftarrow Q(s,a) + \alpha[R(s,a) - Q(s,a) + \gamma \max_{a' \in \Lambda(s')} Q(s',a')]$$

where s' is sampled according to the transition function T, and α and γ are constants known as the learning rate and the discount factor. It has been shown that this method converges to the optimal value function Q^* provided the agent explores all states and each action from each state sufficiently [6]. A further assumption for convergence is that the process is *Markov*. A process satisfies the Markov property if the reward and transition functions depend only on the current state, and not the history of the system. The state space in the head actuation problem was constructed with with the aim of satisfying this property (Sec. 4). Once Q^* is known, the agent can make optimal decisions given each state s in an effort to maximise long term reward by choosing a to maximise $Q(s,a)$.

Motivation theory attempts to describe the behaviour of biological intelligent agents by giving motivational reasons for behaviour. This includes describing the behaviour of agents removed from stimuli. When reward is a constant function in some region of the state action space, the choice of actions becomes ambiguous as the environment no longer discriminates. Typically, an animal given no environmental stimuli will seek out novel experiences, rather than taking no action [8]. Reinforcement learning infrastructure has been used to model motivated behaviour for application in generating complex, exploratory behaviours in unsupervised intelligent agents for non-player characters in online multi-player games [2]. Such motivated reinforcement learning agents differ from standard agents in that they generate their own reward, independent of the environmental reward, based on state perceptions and their own actions. It has been established that natural agents will seek out a middle ground in terms of novelty of sensation, resulting in an aversion to experiences too familiar or too unfamiliar. Saunders and Gero [9] implemented motivated reinforcement learning agents to study the progression of architectural designs, with successive designs similar-yet-different to previous designs. Merrick, et al. [10], have used motivated reinforcement learning agents to create game content procedurally, conforming to the similar-but-different concept of motivation, to simulate creativity. Further research by Merrick [11] involves agents which generate goals based on motivation. The agent then seeks to learn these goals with standard reinforcement learning techniques.

Given the novelty $N = N(s,a)$ of a state action pair, the Wundt curve is used to model the motivation reward function $M : S \times A \to \mathbb{R}$

$$M(s,a) = M_0 + \frac{M_1}{(1 + e^{-\rho_1(N-N_1)})} - \frac{M_2}{(1 + e^{-\rho_2(N-N_2)})}$$

where $\rho_1, \rho_2, N_1, N_2, M_0, M_1$ and M_2 are real parameters which define the function's behaviour; Fig.1 illustrates the Wundt curve utilised in solving the head actuation problem. The novelty $N(s,a)$ can be calculated using methods such as a Habituated Self-Organising Map [9]. In contrast, the novelty detection method utilised for the head actuation agents involved maintaining a model $T' : S \times A \to S$ of the expected transition function $E\{T\} : S \times A \to S$. $T'(s,a)$ modeled the most likely next state after taking action a from state s. After each action a from state s, the expected next state, $T'(s,a)$, is compared to the actual next state, s', using the Euclidean norm on $S \subseteq \mathbb{R}^m$. That is, the novelty is given by

$$N(s,a) = \|T'(s,a) - s'\|^2 \tag{1}$$

After each novelty calculation, the function T' was updated to agree more closely with $E\{T\}(s,a) = s'$, simulating habituation, with novelty declining over multiple similar experiences. This method measures the agent's ability to predict environmental reaction and ascribes high prediction accuracy with low novelty values. The magnitude of the novelty depends on numerous unpredictable factors, and so the motivation function required extensive tuning to achieve desirable behaviour (see Sec. 5 for more detail).

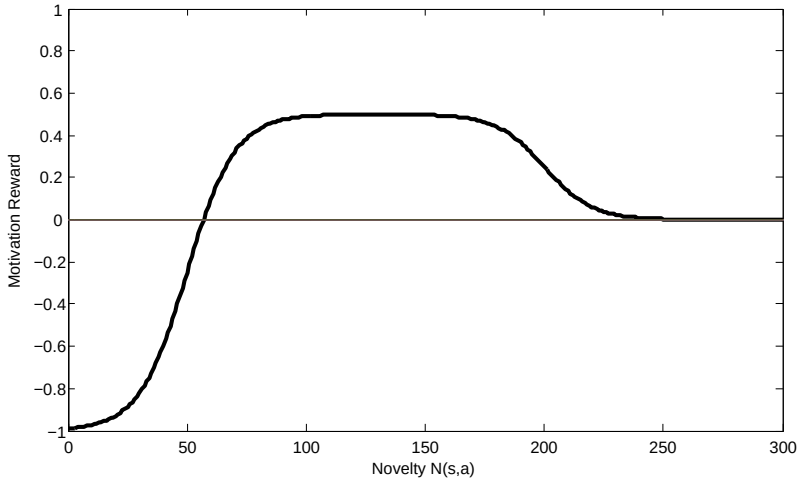

Fig. 1. The Wundt function [8] used for calculating the motivation reward $M(s,a)$ is visualised. A reinforcement learning agent interprets motivation as reward and this drives exploration of the state action space.

3 Approximating Continuous Value Functions

A method for storing a continuous value function, Q, in reinforcement learning involves a weighted sum of basis functions and learning of a set of scalar weights using gradient descent. Based on the Fourier series expansion of periodic functions, a Fourier basis can be used to approximate the value functions in a given domain [12]. The Fourier basis linear approximator is given by the cosine part of a truncated Fourier series and is updated with a sampled point, gradient descent update rule. By using only the cosine terms of the series, the number of required terms for a given accuracy is halved. This comes at the cost of the approximator being even; we overcome this limitation by restricting the state space to a non-periodic domain of the function, namely $S = [0, \tau]^m$ for some $\tau > 0$. In the domain $[0, \tau]^m$ the function is neither periodic nor even, and thus arbitrary continuous functions $f : [0, \tau]^m \to \mathbb{R}$ can be approximated. The performance of the Fourier basis linear approximator at value function approximation has been shown to compete with other leading methods such as radial basis estimation and popular learned basis approximation architectures [12]. For an approximator $F : [0, \tau]^m \to \mathbb{R}$ of order $k \in \mathbb{N}$, the value of the approximation at $\mathbf{x} \in [0, \tau]^m$ is given by

$$F(\mathbf{x}) = \langle \mathbf{w}, \overrightarrow{\phi}(\mathbf{x}) \rangle = \sum_{\mathbf{c} \in C} w_{\mathbf{c}} \cos\left[\frac{\pi}{\tau} \langle \mathbf{c}, \mathbf{x} \rangle\right]$$

where C is a subset of $(\mathbb{Z}_{k+1})^m$. We say $\phi_{\mathbf{c}}(\mathbf{x}) = \cos(\frac{\pi}{\tau}\langle \mathbf{c}, \mathbf{x} \rangle)$ is the basis function corresponding to $\mathbf{c} \in C$ and $w_{\mathbf{c}} \in \mathbb{R}$ is the weight of the basis function corresponding to $\mathbf{c} \in C$. $\mathbb{Z}_j$ denotes the set of integers modulo j, thus C is a collection of $m-$dimensional vectors of integers less than or equal to k. $\overrightarrow{\phi}(\mathbf{x})$ is the vector of basis functions. If $f : \mathbb{R}^m \to \mathbb{R}$ is to be approximated by F with the form above, then sampling f at $\mathbf{x} \in \mathbb{R}^m$ allows F to be updated according to a gradient descent update rule to shift the value of $F(\mathbf{x})$ to agree more closely with $f(\mathbf{x})$. To approximate a function $g : \mathbb{R}^m \to \mathbb{R}^n$, n Fourier basis approximators are used, one for each component of the output.

4 Experiment and Results

The head actuation problem required a solution which accounted for the subtleties involved with maintaining field models with high levels of noise in measurements. This involves making appropriate decisions based on what is most likely to be seen given the estimated locations of objects, which objects give the most information and several other factors. Thus, the problem of choosing head actuation decisions was framed as a Markov decision process and solved with online reinforcement learning methods.

The Markov decision process was constructed by sampling data from the robot infrastructure. The states were constructed as vectors of useful data. Each entry in the state vector was scaled using sigmoid ($f(x) = \frac{1}{1+e^{-a(x-b)}}$) or decaying growth ($f(x) = 1 - e^{-ax}$) functions to confine the ranges within $[0, \tau]$ to

ensure convergence of the Fourier basis approximator which was used to store the value function Q. The action space was the collection of landmarks and objects which were on the field; namely the four goal posts and the ball. For a given action, the robot scans the region in which the object is likely to be found, given by the Kalman filter, for a set period of time, or until it finds the object. An unlocalised robot would thus scan its full field of view, giving a good initial level of localisation. The state space had 10 dimensions and the action space contained 5 actions. The state variables were selected to be approximately independent for inclusion in an uncoupled Fourier basis linear approximator. The state variables were chosen to be the distances to each goal, the total uncertainty in the robot's location filter, the distance to the ball from the robot, the total uncertainty in the ball's location filter, and five *object priorities* corresponding to the action space. The priority of an object is given by the time since the object was last seen, t, and the head movement cost to look at the object, measured in radians, c. If $c \neq 0$ then the priority is given by

$$\tau(1 - e^{-c_p \frac{t}{c}})$$

where c_p is a constant which accounts for the different units and ranges of t and c. If $c = 0$ then the priority is zero, corresponding to when the object is in the field of view of the robot. Objects which have not been seen for longer times or which require little movement to view have a high priority. Thus it can be predicted that high priority objects should be chosen more frequently. This combination of two state variables with complementary expected optimal policies minimises the dimensionality of the state space and reduces computation times. Distances were calculated based on data stored on the robot or measured by the robot. Measured location was used when the object was in the field of view of the robot, and filtered location was used otherwise. The robot's (x, y) location was not included in the state variables because sufficient information is encoded in the distance to the goals. Two reflectively symmetric possibilities exist for a given pair of goal distances. This information is sufficient for head movement decisions as the best action will not depend on which wing of the field the robot is positioned.

The action function $\Lambda(s)$ was defined as the set of objects the robot could look at in state s without exceeding the restrictions on head rotation imposed by the KidSize League RoboCup rules ($-135^o \leq \theta \leq 135^o$, $-90^o \leq \varphi \leq 90^o$; where θ and φ are the angles measured from forward facing direction in the horizontal and vertical planes respectively [13]). The environmental reward function was constructed as a decreasing function of total localisation variance of the robot model and field objects model. The reward was calculated according to

$$R(s, a) = \frac{1}{2}[e^{-c_b(\sigma_{bx}^2 + \sigma_{by}^2)} + e^{-c_r(\sigma_{rx}^2 + \sigma_{ry}^2) - c_\phi(\sigma_{r\phi}^2)}] \tag{2}$$

where $\sigma_{b\zeta}^2$ and $\sigma_{r\zeta}^2$ are the variances in the variable ζ in the next state s' for the ball and the robot respectively. The variable ϕ denotes the robot's heading on the field and the terms of the form c_ζ are constants which were adjusted to weight

each variance equally, accounting for differing units and ranges. The reward is normalised to the interval $[0, 1]$, with perfect localisation indicated by $R = 1$. The reward function is a function of data also directly included in the state vector, namely the total ball error, $c_b(\sigma_{bx}^2 + \sigma_{by}^2)$, and the total self localisation error, $c_r(\sigma_{rx}^2 + \sigma_{ry}^2) + c_\phi(\sigma_{r\phi}^2)$. Thus, the reward depended on the current state only and did not compromise the Markov property. All data was sourced from the world model of the robot, or from the robot's measurements. Data could have been generated by overhead cameras and other external sensors, particularly the absolute localisation error. However, by restricting data to that stored and computed on the robot, the process could be solved online, even during a RoboCup match. Online learning is necessary for the motivated reinforcement learning agent, as the behaviour relies on the changing novelty and thus changing reward function. Standard reinforcement learning agents should also benefit from the adaptability of online learning. Additionally, by using the localisation uncertainty for the reward, and not the absolute error, the performance of the agents can be measured with noise from other systems, like vision, excluded.

Three agents used Q-learning to solve the Markov decision process with reward functions differing between agents. Training sessions were performed with off-policy (RL Agent) or on-policy (others) action selection. During games all agents operate by making decisions on-policy and learning simultaneously. One

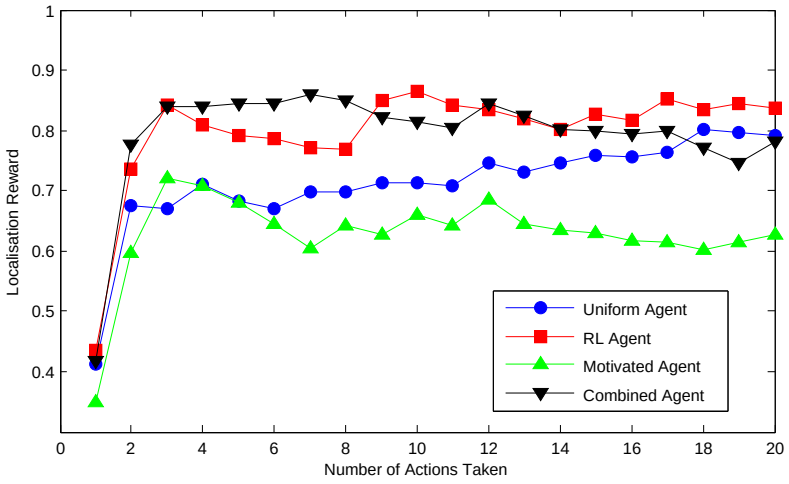

Fig. 2. Results of on-policy, kidnapped robot testing of the head actuation agents. The performance of each agent was assessed by placing the robot in fifteen field positions and allowing the head actuation agent to make twenty decisions while learning. The localisation reward, $R(s, a)$, was recorded after each action was taken and averaged over the 15 positions to provide a metric of performance over the whole field. A localisation reward of 1 corresponds to complete certainty in localisation and the relative localisation uncertainty is given by simply inverting this graph.

agent was trained to maximise the reward given by the environment, R(s,a); this agent will be denoted the *RL Agent*. The second agent was trained to ignore the reward given by the environment and instead use a motivation reward, M(s,a); this agent will be denoted the *Motivated Agent*. A third agent, the *Combined Agent*, used a sum of the environmental and the motivation reward, biased additively by -0.5 (Table 1). The Wundt parameters used for motivation were $N_1 = 50, N_2 = 200, M_0 = -1, M_1 = 1.5, M_2 = 0.5$ and $\rho_1 = \rho_2 = 0.1$ (Fig. 1). These parameters were chosen to suit the range of the novelty by experimental tuning, and are specific to the novelty calculation method and desired behaviour. It was found that, for the purpose of the head actuation, it was more effective to penalise highly novel states minimally, hence the parameters chosen. This choice of parameters gave a motivation reward function $M(s, a)$ ranging from -1 to 0.5 (Fig. 1). The combination of environmental and motivation rewards for the Combined Agent was selected by tuning to balance the importance of exploration and exploitation. The agents utilised Fourier basis linear approximators of order $k = 30$ with gradient descent update rule to store and learn the value function Q and the expected transition function T'. The set $C \subseteq (\mathbb{Z}_{k+1})^m$ for the approximators was chosen to be $C = \{\mathbf{c} \in (\mathbb{Z}_{k+1})^m : c_i \neq 0$ for at most one $i \in \mathbb{N}$ s.t. $1 \leq i \leq m\}$. This is called a *decoupled* Fourier basis, as only single variable cosines form the basis functions. The state variables were chosen carefully because of this; variables which may have been correlated were re-formulated to be independent variables where possible. For example, the robot's position was not included in the state variables, as the goal distances and position would not be independent.

Table 1. The reward schemes for learning agents are summarised.

Agent	Reward
RL Agent	$R(s, a)$
Motivated Agent	$M(s, a)$
Combined Agent	$R(s, a) + M(s, a) - 0.5$

The agents were trained for approximately 30 minutes each. Fifteen minutes while playing soccer and fifteen minutes of *fixed position training*. Fixed position training involves placing the robot on the field in a series of positions and allowing the agent to make decisions and learn while the robot remains stationary. Fixed position training was used to guide the robot to explore states that rarely occur during soccer playing but which are still important. During training, off-policy soft-max action selection was used to train the RL Agent, whereas the motivated agents made all decisions on-policy. Off-policy training was required only for the RL Agent as the others had inbuilt explorative tendencies in the form of motivation reward. Another technique for off-policy training, ϵ-greedy, was also tested, but found to produce no better results. The purpose of this training was to provide the agents with a base value function and policy upon

which they can build during game play by learning online.

Obtaining quantitative results from a soccer playing robot proved too difficult due to the stochastic elements introduced by the interaction of several other robot systems, so the agents were assessed by measuring on-policy performance during fixed position, *kidnapped robot* localisation problems [14]. That is, each agent was moved to a random location on the field and was allowed to localise by making twenty head movement decisions, on-policy, while learning. This simulates the scenario where the robot loses localisation, perhaps due to falling down, and must re-localise before re-engaging with the game. This was done for 15 positions for each agent and the reward after a given number of actions was averaged over the 15 positions. The environmental reward for each action was used to measure performance as it directly indicates the relative localisation accuracy of the robot (Equation 2). A third agent was used as a control: the *Uniform Agent*. The Uniform Agent simply chose head movement with uniform probability from the available objects, as indicated by Λ. The Uniform Agent performed well and, as it was simple to implement, the Uniform Agent became the baseline to which the other agents were compared.

The results of the experiment are shown in Fig. 2 and Table 2. The mean environmental localisation reward over the 15 field positions was used as a metric for agent performance after each action. The environmental reward after a given number of actions varied on average 23% from the displayed mean value. Although this seems large, it must be recognised that between different field locations, during kidnapped robot testing, the bounds on the localisation uncertainty vary due to the differing amounts of information available. For example, the magnitudes of σ^2_{bx} and σ^2_{by} will be bounded below by some value determined by the distance from the ball to the robot and by the parameters involved with the Kalman filter. Therefore a spread of values for the localisation uncertainty is expected. Learning was fast enough to be performed online without affecting the other robot systems. However, this result was sensitive with respect to the size of the Fourier linear approximators. This is why a decoupled approximator was a necessity.

5 Discussion

The Uniform Agent performed better than previous methods, such as the simple hard coded logic statement method which was used in previous competitions, with a steady increase in mean reward over the twenty actions after the initial localisation. The Uniform Agent performed well because it was obtaining information regardless of its choices. The RL Agent generally performed better than the Uniform Agent at playing soccer. It was able to balance looking at the goals and the ball in most situations, and this enabled goal scoring to increase and defending to be more effective. Moreover, the Combined Agent performed better than the RL Agent for the first 8 actions due to the interaction of *boredom* with the environmental reward. Boredom is the state where, due to low novelty, the motivation reward is low. By choosing to penalise boredom heavily (Fig. 1) rela-

Table 2. The localisation performance of reinforcement leaning agents during kidnapped robot testing is summarised (corresponding to Fig. 2). The RL Agent and Combined Agent outperformed the Uniform Agent both at localising quickly and in long-term performance. The expression 'localised over x' means achieving a localisation reward greater than x.

Agent Type	Uniform	Motivated	RL	Combined
Number of cases(out of 15) localised over 0.8 within 4 actions	10	10	13	13
Actions to achieve over 0.8 mean localisation reward	18	>20	3	3
Average % improvement over Uniform Agent over 20 actions	-	-11%	12%	11%

tive to the environmental reward in the Combined Agent, a curious but focused head actuation agent was trained. Overall the Combined Agent was able to make head actuation decisions to localise robustly. The worsening performance of the Combined Agent for more than 8 actions is likely due to the agent becoming bored after being in the one field position for extended periods and revisiting the same subset of the state space. This would cause worse decisions due to the motivation reward dominating, and this is reflected in the reward curve trending similarly to the Motivated Agent's curve after 12 actions. This would not seriously affect soccer playing as the field location changes quickly during a soccer match, so the robot will likely never make more than a few decisions from one field position at a time. The motivated agent performed poorly as it could not distinguish which actions were better and it did not become bored quickly enough due to the Wundt function. A better agent may have used a *hyperactive* Wundt function, with high novelty required for positive reward, causing the agent to choose the next action based on the least frequently chosen action. However, the Wundt function shown in Fig. 1 was effective for the Combined Agent, so it was also examined independently.

The Combined Agent is only random in its initial state, before training. After the first learning iteration, it uses no random variables to make decisions but rather a complex motivation system that balances exploration and exploitation. It should be noted that the Combined Agent requires no off-policy action selection during training as the motivation reward induces a natural exploration. This exploration may provide a viable alternative to using off-policy action selection techniques, such as soft-max or ϵ-greedy, during training. Due to its balanced exploration, motivation may prove more effective than these techniques; more research would be required to measure its effectiveness.

As with any machine learning task, finding an acceptable set of training parameters was difficult. The key parameters for the Fourier basis approximators

was the domain size τ, order k and the basis function set C; k and C must be chosen to balance computation time and resolution of the function. The dimensionality costs of choosing $C = (\mathbb{Z}_{k+1})^m$ were too large for a high dimensional state space (m=10) as $|(\mathbb{Z}_{k+1})^m| = (k+1)^m$. The use of the uncoupled basis function set, with size $|C| = m(k+1)$, allowed for liberal choice for the order k at the cost of requiring uncoupled state variables. The calculation of the novelty proved effective for the purposes of motivating the head actuation agents to explore new actions and states. It was found that the expected transition function T' does not have to approximate $E\{T\}$ with arbitrary accuracy. However, due to the way in which the novelty was calculated with a norm on $\mathbb{R}^m$, it was impossible to predict the range of values which the novelty would take or where the Wundt function should be most sensitive. An upper bound for the novelty can be estimated to be $m\tau^2$, based on the Equation 1 and the restriction of the range of the state vectors to $[0, \tau]$. However, T' is not bounded by these limits and this value says little about the useful range of novelty once the function approximator has been trained. For example, this experiment had a state space dimension of $m = 10$ and a function approximator range $\tau = 10$ giving a theoretical maximum novelty of 1000. However, the most useful set of Wundt parameters only distinguished between novelties in the range of 50 to 300 (Fig. 1). These Wundt parameters were obtained by, on the first instance of training an agent on a given state space, re-adjusting the parameters between training sets. Training sets involved between 20 and 100 actions with full training involving about 500 actions over about an hour. This training was not automated; specifically the fixed position training required manual repositioning the robot and ball after each training set. After one agent was trained, and the Wundt function parameters found, other agents could be trained without adjustment of the parameters provided the state space and function approximator parameters did not change.

6 Conclusion

A motivated reinforcement learning framework was successfully developed and implemented for the optimisation of head actuation policies. Self-localisation was demonstrated as improving by over 11% relative to the null case of uniformly distributed actuation policies, for agents trained online for no more than one hour. Within 4 actions, the reinforcement learning agents were able to localise accurately for 13 out of 15 cases. In contrast, the uniform agent localised accurately in 10 from 15 cases within 4 actions. It was observed that, by exhibiting some level of artificial 'boredom' and 'curiosity', the motivated reinforcement learning agent is able to modestly improve its self-localisation by observing its environment in a more intelligent manner; a viable method for improving localisation performance without the need for improved object recognition algorithms or the tuning of probabilistic filters.

Acknowledgements. This project was supported by the Australian Mathematical Sciences Institute (AMSI) summer research scholarship program. Thanks

to the University of Newcastle RoboCup team, the NUbots, for providing hardware resources and support.

References

1. Wong, A.S.W., Chalup, S.K., Bhatia, S., Jalalian, A., Kulk, J., Nicklin, S., Ostwald, M.J.: Visual gaze analysis of robotic pedestrians moving in urban space. Architectural Science Review **55**(3) (2012) 213–223
2. Merrick, E.K., Maher, M.L.: Motivated Reinforcement Learning: Curious Characters for Multuser Games. Springer, Dordrecht (2009)
3. Kitano, H., Asada, M., Kuniyoshi, Y., Noda, I., Osawa, E., Matsubara, H.: Robocup: A challenge problem for ai. AI Magazine **18**(1) (1991)
4. Budden, D., Fenn, S., Walker, J., Mendes, A.: A novel approach to ball detection for humanoid robot soccer. In Thielscher, M., Zhang, D., eds.: Advances in Artificial Intelligence (LNAI 7691), Springer (2012)
5. Wan, E., van der Merwe, R.: The unscented kalman filter for nonlinear estimation. In: Adaptive Systems for Signal Processing, Communications, and Control Symposium 2000. AS-SPCC. The IEEE 2000. (2000) 153 –158
6. Watkins, C.: Learning from Delayed Rewards. PhD thesis, Cambridge University (1989)
7. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA (1998)
8. Wundt, W.: Principles of Physiology and Psychology. Macmillan, New York (1910)
9. Saunders, R., Gero, J.S.: Designing for interest and novelty - motivating design agents. In de Vries, B., van Leeuwen, J., Achten, H., eds.: Proceedings of the ninth international conference on computer aided architectural design futures, Kluwer Academic Publishers (2001) 725–738
10. Merrick, K.E., Isaacs, A., Barlow, M., Gu, N.: A shape grammar approach to computational creativity and procedural content generation in massively multiplayer online role playing games. Entertainment Computing **4**(2) (2013) 115 – 130
11. Merrick, K.: Intrinsic motivation and introspection in reinforcement learning. Autonomous Mental Development, IEEE Transactions on **4**(4) (dec. 2012) 315 –329
12. Konidaris, G., Osentoski, S., Thomas, P.S.: Value function approximation in reinforcement learning using the Fourier basis. In Burgard, W., Roth, D., eds.: Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, AAAI Press (2011) 380–385
13. the RoboCup Institution: RoboCup Soccer Humanoid League Rules and Setup for the 2013 Competition in Eindhoven (DRAFT), http://www.tzi.de/humanoid /bin/view/Website/Downloads. (2012)
14. Majdik, A., Popa, M., Tamas, L., Szoke, I., Lazea, G.: New approach in solving the kidnapped robot problem. In: Robotics (ISR), 2010 41st International Symposium on and 2010 6th German Conference on Robotics (ROBOTIK). (2010) 1–6