The Hybrid Integration of Perceptual Symbol Systems and Interactive Reinforcement Learning

Michael J. Knowles and Stefan Wermter
Centre for Hybrid Intelligent Systems
University of Sunderland
www.his.sunderland.ac.uk
Michael.Knowles@Sunderland.ac.uk Stefan.Wermter@Sunderland.ac.uk

Abstract

In order to produce robots which can interact more effectively with humans we propose that it is necessary for their cognitive processes to be grounded in the same perceptual elements as humans deal with. Perceptual Symbol Systems offer an attractive mechanism for capturing the symbolic properties of the senses and for integrating them into higher level cognitive processes. We have designed a Perceptual Symbol System where the robot learns about objects through interaction and reinforcement and have carried out experiments to assess the merits of this approach. We show that the use of human perceptual elements combined with interactive reinforcement leads to intuitive learning and interpretable knowledge structures.

1. Introduction

In many restricted or industrial scenarios robots are able to outperform humans but in terms of natural interactivity robots lag behind even young children. Even sophisticated robots rely heavily on pre-programmed behaviour and the performance of such robots is dependent on the limitations of the pre-programming [1].

The interactive abilities of robots can be enhanced by developing their ability to learn from both the environment around them and the individuals they interact with. This can trigger inferences about the nature of the environment, learning actions or behaviours under human instruction. Both activities require a robot to learn from various modalities of sensory data which may be presented simultaneously or perhaps more likely, sequentially [2].

In order to learn, interpret and use acquired sensory data in some form of cognitive process it is necessary to form associations between modalities at a low level of abstractions, and to form hierarchies of these linked representations to form higher level, more abstract concepts [3]. Once these concepts are learnt they can be utilized to construct behaviour when the robot is faced with a particular scenario.

Some work exists on robot systems which have the desired ability to learn and adapt and which are grounded in the components of human perception. Roy et al [4][5] use high level visual information to aid speech recognition, but this is not based on human perceptual features, but instead on features such as colour and shape histograms.

One attractive mechanism for capturing sensory features, forming conceptual representations and using them is the concept of Perceptual Symbol Systems (PSS) as originally proposed by Baralou [6] and further described by Niedenthal et al [3]. Limited work has been undertaken on implementing PSSs. Cangelosi et al [7] developed a system which captures the symbolic behaviour of neural activity during a simple vision task. Pezzulo and Calvi [8] pursued an approach where the symbols took the form of visual and motor schemas which were combined to form multimodal frames that describe the execution of a particular task.

Rather than building a system which is based on neural architectures for direct processing of low level sensory input, we begin by using well established machine vision techniques to produce a simple percept from an incoming scene. This is combined with linguistic input and a simple implementation of a PSS to allow an agent to learn about the presented objects. We then use interaction with a human user combined with reinforcement learning to refine these learned perceptual representations. Both reinforcement-based learning and interaction are essential elements for learning and have been identified by Cognitive Scientists as key factors in the development of human infants and children [9].

The remainder of the paper is structured as follows. The vision system used is outlined in section 2, the language system in section 3 and the implementation of the PSS in section 4. Sections 5 and 6 describe the application of reinforcement learning and section 7 concludes the paper with discussion of the results.

2. Vision system

The primary objective of the vision system is to provide a symbolic representation of an incoming scene. This is achieved by first detecting simple shapes in the image. Once a shape has been determined its colour is estimated by averaging the RGB values within its boundary.

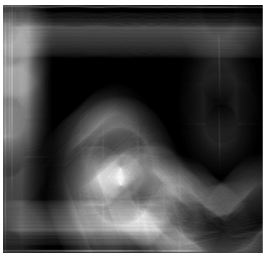
Incoming scenes can be analysed from either live video, recorded video or from a still image. The first stage of processing is the generation of an edge image using a Canny edge detector [10]. The resultant edge image is then subjected to two parallel Hough transforms [11], one to detect circles in the image and one to detect squares.

The result of the Hough transforms is a pair of three dimensional images which indicate the likelihood of detecting a circle or square of a given size at a given position within the image. A two stage approach is used to determine the most likely location. Firstly the peak position is determined, which is then refined by examining the overall likelihood based on position alone accumulating the evidence for all feature sizes.

This results in the most likely location for a circle and a square within the frame along with a likelihood of each. If the likelihood of either exceeds a preset threshold then they are labelled as such and the average RGB values are computed within the features. These are compared with the RGB values for the three available colours: Red (255,0,0), Yellow (255,255,0) and Blue (0,0,255). Each detected circle or square is assigned a colour based on the closest match.

Figure 1 shows the stages of the vision system, the edge map (a), the overall positional likelihood for a circle derived from the Hough transform (b), the inferred position (c) and the perceptual output (d). The vision system is implemented using the OpenCV library.

a) Edge map



b) Hough transform

c) Detected circle

d) Percept

Figure 1. Vision System

3. Language system

We propose that a simple "language system" is essential to provide some feedback for learning. We believe that simple language input still allows the user to interact with the agent in a rich and complex manner. Language input is currently simulated using a selection of buttons which the user clicks to trigger spoken utterances. A limited vocabulary consisting of the following words is used:

- 'This is'
- 'Circle'
- 'Square'
- 'Red'
- 'Yellow'
- 'Blue'

In order to allow testing of the system the following utterances are also available:

- 'What is'
- 'Colour'
- 'Shape'

4. Perceptual symbol system design

The perceptual symbol system used consists of two categories of symbol, visual symbol and linguistic symbol. The visual system can currently deal with two shapes and three colours as described above. The linguistic symbols consist of those utterances described above. There is no predefined link between the visual symbols and the related linguistic symbols.

Each symbol carries an activation value α. When a symbol is activated by the appropriate perceptual system i.e. vision or language, its activation is set to 1. When a perceptual symbol is activated, its activation spreads to adjacent symbols using the connections described later.

Barsalou identifies working memory as the forum in which symbols are combined and in which perceptual simulations are run [6]. The working memory architecture proposed by Barsalou and which will be adopted here is that proposed by Baddeley and Hitch [12] and extended by Baddeley [13]. The working memory is modelled as a set of limited capacity buffers. The Baddeley model contains the following elements:

- A phonological loop which stores auditory and linguistic information. This element also involves a rehearsal loop which revives the auditory information in the loop.
- A visual sketchpad which is involved in the storage and processing of visual and spatial information.
- An episodic buffer which handles the integration of information from different modalities into coherent temporal episodes.
- A central executive which manages the other systems within working memory and regulates attention.

One of the key aspects of working memory is that it is a limited resource. Various theories have been put forward regarding the limitations of resources. Case et al [14] proposed that working memory is a limited resource shared between processing and storage. Towse and Hitch [15] suggested an alternative view that items stored in working memory decay in time. Both the resource sharing and time decay hypotheses are contained in the model proposed by Barrouillet et al [16]. In this model attention is the limited resource. Items within working memory decay unless they are refreshed. This refreshing can only occur when attention is not focused on some processing task.

In our implementation, working memory is a buffer which can hold up to seven symbols in line with established estimates of working memory capacity [17]. Each symbol in the working memory has an associated weight denoted as β. This is used solely to represent the duration the object has been held in working memory. Two independent threads of execution are used to process the contents of the working memory. A decay thread is used to decay the activation of the symbols within the buffer exponentially with an approximate decay time of 30 seconds. Simultaneously an attention thread will process any new symbols, and if none are present it will refresh the linguistic symbols in working memory using the following update rule:

$$\beta = 1 - (0.05k) \tag{1}$$

k is the number of times that the linguistic symbol has been refreshed.

Categories of objects are represented by a frame which consists of associations between symbols. Symbols can not only be associated with the frame itself, e.g. the visual symbol for a round object is associated with the circle frame, but symbols can be associated with each other to represent patterns within the instances of the objects which are perceived – e.g. if balls are either large and blue, or small and red, this association between colour and size will be captured. Associations can be excitory or inhibitory such as to further capture combinational classes within a frame.

When an instance of a known object is presented to the frame representing the appropriate class, the frame's representation of the class in question is refined. This occurs by reinforcing all the appropriate associations within the frame. If a symbol is attached to either end of any of the associations that make up the frame, then that association is updated. If the association is between contradictory symbols of the same class, red and blue for example, then the association strength c is weakened:

$$c^{t+1} = c^{t-1} - \left| \alpha c^{t-1} \right| \tag{2}$$

If the association is not between contradictory symbols then the association is strengthened:

$$c^{t+1} = c^{t-1} + \left| c s^{t-1} \right| \tag{3}$$

In addition to updating the strengths of the associations, the instance of the object is added to a list of experienced instances of the frame.

While this system functions and learns the representations through demonstration, a large number of training examples are needed. We propose that even limited human interaction offers more effective learning if used in a targeted manner.

5. Applying reinforcement learning to a perceptual symbol system

In order to improve the learning of the system we extend a reinforcement learning (RL) algorithm. As previously stated, reinforcement involves increasing the likelihood of a behaviour, based on some feedback

presented in response to the execution of the said behaviour. In terms of simple single stage tasks such as object recognition, this can be achieved using simple learning techniques such as Hebbian learning. However for more complex sequential learning tasks involving multiple decisions, the machine learning paradigm of reinforcement learning becomes relevant. Here the difficulty is assigning the credit to the correct decisions and not to any sub-optimal stages.

Various algorithms exist to tackle this problem. Typically reinforcement learning algorithms function by learning a reward function which attempts to capture the amount of reward available at each stage of the behaviour. We apply the Q-Learning algorithm proposed by Watkins [18]. In Q-Learning an action-value function *Q(s,a)* is used to select the correct action *a* based on the current state *s*.

In order to apply Q-Learning it is necessary to define the system state, the available actions and the *Q* function. The objective is to present the system with a stimulus and ask it a question, before providing reward based on the answer. Our state consists of two elements. Firstly the request that has been made by the user which will either be the shape or the colour of the presented object. The second element of the system state is the encoding of the concepts in the associations within the frames.

The available actions are simply the available linguistic utterances. In order to obtain the available actions the system determines which frame is most active following presentation of the object. The linguistic symbols within the frame form the candidate actions. For each of these a Q value is inferred as the product of the association between the candidate utterance and the request, and the activation of the candidate utterance:

$$Q(s,a) = c(request - utterance) \times \alpha(utterance) \quad (4)$$

The inference of Q values from more complex systems is a common problem in reinforcement learning since it is often undesirable to tabulate all combinations of state and action.

Once all candidate utterances have been examined the one with the highest Q value is presented to the user as the answer to the question. The user then responds with either 'Correct' or 'Wrong'. If the answer is correct then a reward of 1 is delivered. If the answer is wrong then a reward of -0.1 occurs. This bias towards positive learning is designed to allow the agent to move on quickly once the correct answer is established since an incorrect answer does not necessarily specify where the error lies and which components that produced the Q value are at fault. The update rule for Q learning is:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha(s_t,a_t)\left[r_t + \gamma \max_a Q(s_{t+1},a_t) - Q(s_t,a_t)\right] \quad (5)$$

This equation is used to derive the new Q value. The challenge is to update the system in such a way as to achieve this new Q value. As stated in equation 4 the Q value for the chosen action is derived from the association between request and utterance and the activation of the utterance. In order to update these two components an update factor, *j*, is determined from the ratio of the previous and new Q values:

$$j = \sqrt{\frac{Q_{t+1}}{Q_t}} \quad (6)$$

The association between the utterances and the request simply has its strength multiplied by the above factor. Modifying the activation of the utterance is, however, more problematic, since the goal is to modify the activation that would result if the same stimulus was presented again.

The activation of a symbol is determined by the sum of the products of the activation of associated symbols and the association strengths. Thus the association strengths are the correct target for modification. Since we are aiming to update connection strengths based on the current activation we bias the update to these associations by the strength of the activation of the symbol on the other end of the association. We do not wish to update associations with symbols which are not active since they have not participated in the current, incorrect, activation and may actually be correct. The update on the connection strength is:

$$c = c + (c \times \alpha \times \omega) \quad (7)$$

ω is set such that:

$$\frac{\sum c + (c \times \alpha \times \omega)}{\sum c} = j \quad (8)$$

Thus ω scales the update to produce the desired corrected activation value.

The system was tested by first training it with three objects: a blue circle, a red square and a yellow square. The system was then refined by showing the system the objects and making queries.

In order to assess the learning, two metrics are employed. The Request Connection Strength, RCS, is the sum of the associations between the request symbols, for colour and shape, and the linguistic symbols with which each should be associated. The Linguistic Symbol Grounding, LSG, is the sum of the associations between the visual symbols and the appropriate linguistic symbols. The value of these two metrics through time during a series of trials is shown in Figures 2 and 3.

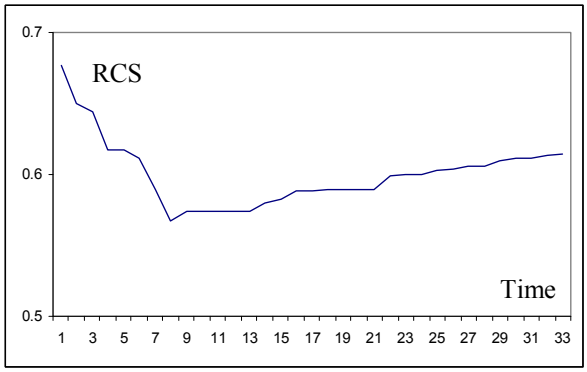


Figure 2 RCS for standard RL

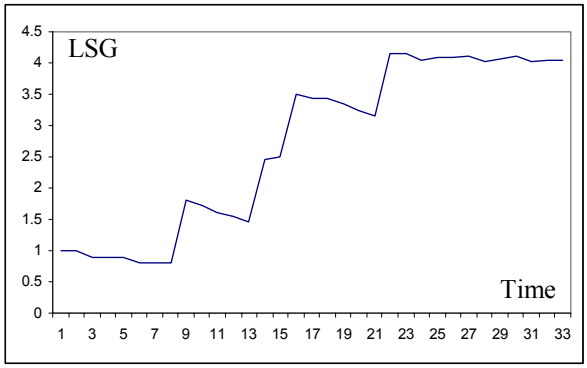


Figure 3 LSG for standard RL

It can be seen that the RCS falls initially quite heavily. This occurs when the agent selects the correct instance from the wrong category, e.g. when asked for the colour of a red square, the agent replies 'square'. This deficiency is addressed in the next section.

6. Interactive reinforcement learning

The use of targeted feedback for reward has been explored by Thomaz [19] in a simulated environment where the timing of feedback is used to infer correct assignment of the reward function. We go beyond this by using verbal feedback from the user for inferring the desired assignment of the reward. Furthermore we incorporate a means for our agent to refine its knowledge under restricted guidance from human users rather than simply punishing or rewarding the agent.

In order to achieve guided interaction, two further feedback options are added, 'wrong concept' or 'wrong instance'. Thus if the request is 'what colour' and the answer is 'round' then the reward can be targeted as such – i.e. at the link between the utterance and the request. Alternatively, if the answer is 'blue' but the object is red then the reward is targeted at the associations that led to the activation of the 'blue' utterance. In such a case the update factor used is j^2 since only one factor in the Q value is being updated

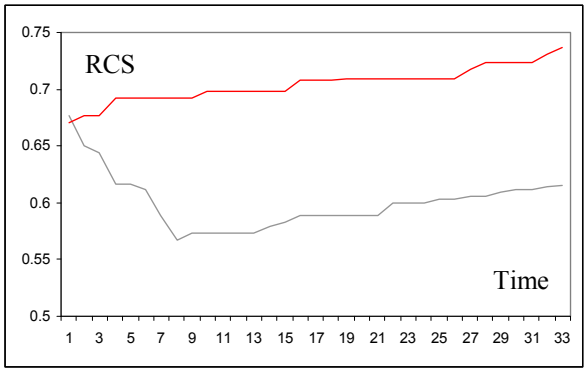


Figure 4 RCS for targetted RL (red) and standard RL (grey)

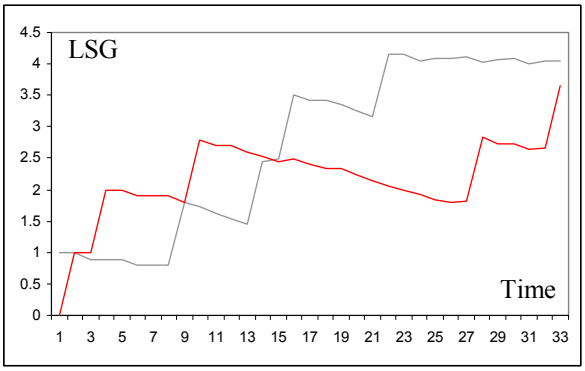


Figure 5 LSG for targetted RL (red) and standard RL (grey)

The results of a trial of the guided interaction following the same procedure as for the unguided interaction are shown in Figures 4 and 5, along with the results for the unguided interaction for comparison. It can be seen that although the LSG has a dip, due in this case to a repeated wrong guess, it learns at the same rate otherwise. It can also be seen RCS learning is greatly improved since correct associations are not targeted by reward application.

7. Conclusions

We have demonstrated that grounding human-robot interaction at a perceptual level leads to agents which begin to learn high level concepts immediately without having to learn first to process sensory data.

Furthermore we have demonstrated the importance of using reinforcement combined with human interaction in guiding machine learning to speed the process by correctly targeting the update of the agent knowledge.

This constitutes an important first step in the development of perceptual grounded agents and future work must focus on further refining the perceptual basis of learning guided by studies of human perception as well as scaling up these concepts.

8. Acknowledgement

This work has been supported by European Community as part of the NESTCOM project (NEST-043374).

9. References

[1] S. Wermter, G. Palm, C. Weber and M. Elshaw, "Towards Biomimetic Neural Learning for Intelligent Robots" in *Biomimetic Neural Learning for Intelligent Robots,* S. Wermter, G. Palm, and M. Elshaw (eds), Springer, Heidelberg, 2005.

[2] S. Oviatt, R. Coulston, S. Tomko, B. Xiao, R. Lunsford, M. Wesson and L. Carmichael, "Toward a Theory of Organised Multimodal Integration Patterns during Human-Computer Interaction" *Proceedings of the 5th international conference on Multimodal interfaces*, ACM, Vancouver, November 2003, pp. 44-51.

[3] P. M. Niedenthal, L. Barsalou, P. Winkielman, S. Krauth-Gruber and F. Ric, "Embodiment in Attitudes, Social Perception, and Emotion", *Personality and Social Psychology Review,* Sage Publications, Vol 9, No 3, 2005, pp 184-211.

[4] D.K. Roy and A. Pentland, "Learning words from sights and sounds: a computational model", *Cognitive Science*, Elsevier, Vol. 26, 2002, pp. 113-146.

[5] D.Roy, "Learning Visually Grounded Words and Syntax of Natural Spoken Language", *Evolution of Communication,* Benjamins, Vol 4, 2000.

[6] L.W. Barsalou, "Perceptual Symbol Systems", *Behavioural and Brain Sciences*, Cambridge University Press, No 22, 1999, pp. 577-660.

[7] D. Joyce, L. Richards, A. Cangelosi and K.R. Coventry, "On the foundations of perceptual symbol systems: Specifying embodied representations via connectionism", *The Logic of Cognitive Systems. Proceedings of the Fifth International Conference on Cognitive Modeling*, pp147-152, Universitätsverlag Bamber, 2003.

[8] G. Pezullo and G. Calvi, "Towards a Perceptual Symbol System", *Proceedings of the Sixth International Conference on Epigenetic Robotics: Modeling Cognitive Development in Robotic Systems.* 2006.

[9] D. Mareschal, M.H. Johnson, S. Sirois, M. Spratling, M. Thomas and G. Westermann, Neuroconstuctivism Vol. 1 and 2. Oxford University Press, UK, 2007.

[10] J. Canny, "A Computational Approach To Edge Detection", *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol 8, 1986, pp. 679-714.

[11] P.V.C. Hough, "Machine Analysis of Bubble Chamber Pictures"*, Proc. Int. Conf. High Energy Accelerators and Instrumentation*, 1959.

[12] A.D. Baddeley and G.J. Hitch, "Working Memory", *The psychology of learning and motivation: advances in research and theory,* New York: Academic Press, Vol. 8, 1974, pp. 47-89.

[13] A.D. Baddeley. "The episodic buffer: A new component of working memory?" *Trends in Cognitive Science*, Cell Press, Vol 4, pp. 417-423.

[14] R. Case, M.D. Kurland and J. Goldberg, "Operational efficiency and the growth of short-term memory span", *Journal of Experimental Child Psychology,* Vol. 33*,* pp. 386-404.

[15] J.N. Towse and G.J. Hitch, "Is there a Relationship between Task Demand and Storage Space in Tests of Working Memory Capacity?" *The Quarterly Journal of Experimental Psychology Section A*, Vol. 48, 1995, pp. 108 – 124.

[16] P. Barrouillet, S. Bernardin, and V. Camos, "Time constraints and resource sharing in adults' working memory spans", *Journal of Experimental Psychology: General*, Vol 133, 2004, pp. 83-100.

[17] G.A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information", *Psychological Review*, Vol 63, 1956, pp. 81-97.

[18] C.J.C.H. Watkins, "Learning from Delayed Rewards" PhD thesis, Cambridge University, Cambridge, England.

[19] A.L. Thomaz and C. Breazeal, "Teachable robots: Understanding human teaching behavior to build more effective robot learners" *Artificial Intelligence,* Vol. 172, 2008, pp 716-737.