WhiskEye: A biomimetic model of multisensory spatial memory based on sensory reconstruction

Thomas C. Knowles[0000−0002−7750−2119], Rachael Stentiford, and Martin J. Pearson[0000−0002−8642−4845]

Bristol Robotics Laboratory, University of the West of England, Bristol, UK
tom.knowles@brl.ac.uk
https://www.bristolroboticslab.com/

Abstract. We present WhiskEye, a visual tactile robot supporting a neurorobotic investigation of spatial memory as a multisensory reconstructive process. This article outlines the motivation for building WhiskEye; the technical details of the physical robot, and the publicly available simulated platform via the NeuroRobotics Platform (NRP) from the Human Brain Project; and the biomimetic control architecture. The multisensory reconstruction model of place recognition based on deep predictive coding network is presented and datasets collected from the NRP are used to train and test the network. We demonstrate that the joint latent representations inferred by this system are positively correlated to displacements in pose space suggesting it is an advantageous sensory processing front-end for our neuro-plausible model of spatial memory.

Keywords: Neurorobotics · neural networks · multisensory inference

1 Introduction

As we move through the world we see, touch, smell, taste and hear the environment around us. We use this sensory information to navigate safely and to plan routes to previously visited locations. How this multisensory information is represented, stored and recalled by the brain to aid in navigation is not fully understood. In the 19^{th} century Heinrich von Helmholtz proposed that the brain was not a passive observer of the environment through the senses, rather it was actively engaged in predicting how the world behaves [8]. This conceptual shift in understanding has become increasingly popular in contemporary neuroscience research with many works advocating and demonstrating the role of prediction in describing physiology and behaviour [6], [2], [18]. Models for how the neocortex may implement this learning have also been proposed [19] which in turn has resulted in neural network models that can be constructed and implemented using the readily available machine learning toolboxes [3]. Deep predictive coding neural networks differ from conventional deep learning neural networks in that the error correction step applied to the weight array is computed locally in each training epoch in parallel across the network, i.e., the global derivative and back propagation of error is not required. Instead each layer in the network attempts

to predict the output of the previous layer, refining its predictions by comparing them to the actual output. In other words, higher layers are trained to reconstruct the activity of lower layers but using an increasingly smaller dimensional representation space to do so. This enables a hierarchical learning of representations but with the benefit of priors that can anticipate familiar sensory inputs by generating predictions that are tested against incoming evidence.

In this paper we describe how such a network has been integrated into the processing architecture of a biomimetic multisensory robot called WhiskEye. WhiskEye has an array of active tactile whiskers and cameras for eyes that explores its environment in an ethologically plausible way. Using a model of tactile attention, it gathers visual and tactile impressions from its environment which are used to train a multimodal predictive coding implementation called MultiPredNet. The representations generated by this network show a strong correlation to pose space, and thus are useful for place recognition.

The main contributions of this paper are:

1. Overview of a novel multisensory biomimetic robot platform
2. Introduction of a publicly available simulation platform of the WhiskEye
3. A neuroplausible multimodal deep predictive coding network model that can combine vision and tactile sensory information
4. A demonstration that the network model can generate representations that are beneficial to place recognition

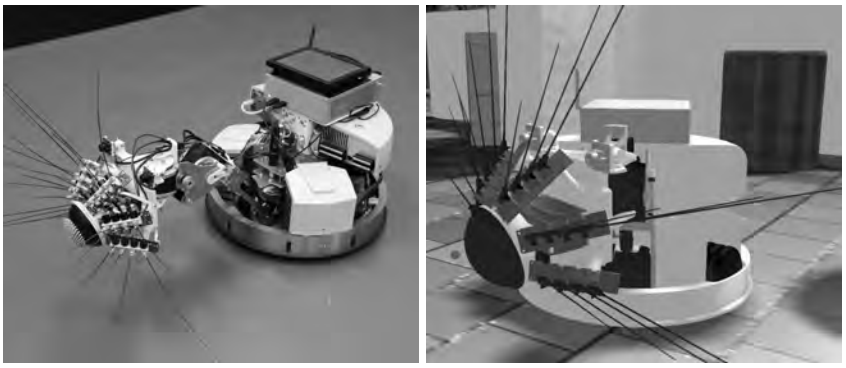
2 Related work

The brain is renowned for its ability to combine different modalities to solve problems, in artificial systems we refer to this ability as sensor fusion[10]. Model free approaches to sensor fusion include Variational AutoEncoders (VAEs) which have proven successful by being able to create joint latent spaces that encode the regularities between multiple modalities[11]. Predictive coding systems take this a step further by using bio-plausible learning rules and generating representations at each layer, whilst also showing the ability to extract disentangled latent variables[13]. To the best of our knowledge this approach has not been applied specifically to place recognition.

RatSLAM[14] is a successful Simultaneous Localisation and Mapping (SLAM) approach inspired, like WhiskEye, by rat behaviour. Unlike RatSLAM, this paper does not purport to solve the full SLAM problem, instead focusing on representation learning for place recognition. This is equivalent to the sensory front-end of RatSLAM, processing raw sensor data into a form suitable for a future downstream mapping system. The use of whisker-based touch has been successfully incorporated into a SLAM system before[5] and is promising in terms of the redundancy and robustness it offers. WhiskEye builds on prior works using whisker based tactile sensing for mobile robots[17], [16] by introducing the head-mounted cameras to coarsely approximate rat vision and allowing us to capture rich multisensory datasets during mobile exploration.

3 Materials and methods

3.1 WhiskEye platform



(a) Physical WhiskEye in the BRL test arena

(b) Simulated WhiskEye in an NRP virtual arena

Fig. 1: Both incarnations of WhiskEye. Note the differences in whisker shape and simplified structure of the simulated model, with extraneous detail like wires and the onboard display omitted.

Hardware The main physical components of WhiskEye are the head, neck and body. The body is a Robotino^{TM} chassis from Festo Didactic, with an onboard Intel computer running the robot control software, including ROS. This computer communicates to a head mounted master SPI bus that controls much of the robot's behaviour. Within the Robotino^{TM} is an ARM microcomputer that itself runs ROS, interacting as a ROS device with the onboard computer. Logs and data are sent via wi-fi to a remote desktop. Three omni-wheels allow for arbitrary motion in x, y and θ.

The neck is custom-built, attaching to the front of the Robotino chassis with a USB connection to the onboard computer. This USB is set up as a ROS device, allowing for data to be read from sensors and commands to be sent to neck and head actuators.

The head is also custom-built, mounting the aforementioned head SPI master. This controls the 6 whisker arrays and neck via 7 slave SPIs. Each whisker array consists of 4 whisker complexes, each with its own motor, ARM processor, a 2-axis Hall Effect sensor and the whisker proper; a flexible, tapering plastic rod that mimics small mammal whiskers. Each ARM processor coordinates its whiskers to generate 'whisks' of synchronised movement across the array, but allows each to respond individually to impingement for whisker-specific retraction.

Neurorobotics Platform The NeuroRobotics Platform (NRP) [4] is a web based robotics and neuroscience research tool for neuroscience based robotics experiments, particularly through time sensitive coupling between Gazebo and spiking neural network simulators such as NEST. For very large network models it also provides an API to deploy on the SpiNNaker neuromorphic supercomputer [7]. A CAD model of WhiskEye has been instantiated into the NRP with a Gazebo-ROS plugin deployed to mirror the interface of the physical platform described above. To accommodate the flexible whiskers within Gazebo's rigid body physics, whisker collisions were disabled; instead, surface penetration depth was used to calculate the corresponding force experienced at the base of each whisker in the 2 orthogonal planes (x_{whisk}, y_{whisk}). Crucially, the NRP hosts the same ROS control architectures as the physical robot, ensuring parity between simulated and physical behaviour.

Control architecture WhiskEye's movements are initiated and coordinated through a model of whisker based tactile attention derived from prior work [15]. It is composed of an interconnected network of functional models of mid brain structures of the rat that have been modelled using Python and compatible with ROS. Each module encapsulates a specific set of functions necessary for control, with many modules implementing neuro-plausible functional models.

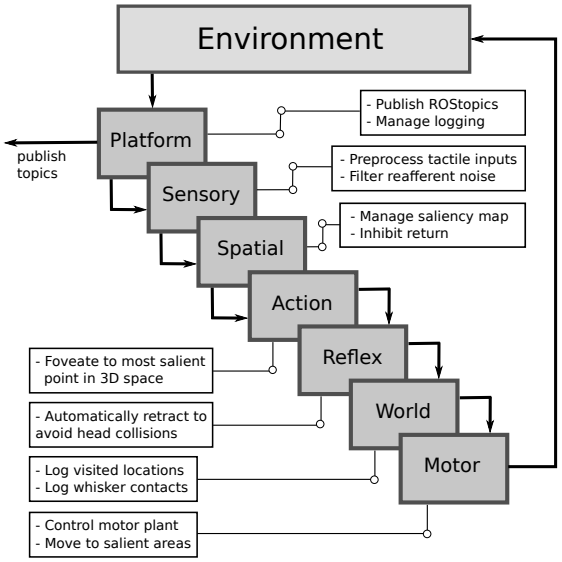


Fig. 2: Cascading view of control functions. Each function is called sequentially and contributes to the final, salience-guided foveation, sampling the environment in an ethologically plausible way. ROStopics of cameras and whiskers are published, permitting collection of datasets for MultiPredNet training (3.2)

Figure 2 shows WhiskEye's cascade of controller functions that each contribute to the final behaviour of the robot:-

- Platform - creates publishers for all relevant ROStopics that can be subscribed to both internally (such as whisker inputs for tactile attention) and externally (for monitoring and data collection).
- Sensory - preprocesses incoming sensory data, reshaping and removing reafferent noise with a high-pass filter, preserving deflections caused by impingement; loosely analogous to a proposed cerebellar role for re-afferent sensory prediction [1].
- Spatial - manages a Superior Colliculus-inspired salience map fed by tactile data. This determines where the robot will orient to. Local space is mapped as head-centric (x_h, y_h, z_h) and the most salient location identified. If its salience exceeds a threshold, the coordinates pass to the Action module. If not, structured noise is applied that raises salience around the fovea until a candidate location is found.
- Action - inspired by the Basal Ganglia - deciding how to act, and how much - the desired position in head space is transformed into world space (x_w, y_w, z_w). The difference between the current and desired position forms the movement vector describing the orient required.
- Reflex - responds via callbacks to any potential collisions that a movement can cause; since obstacles can be interesting features themselves, this is a common occurrence. Proportional retraction ensures collisions are minimised.
- World - logs visited locations, implementing Inhibition of Return (IoR) by temporarily masking their coordinates in the salience map. This avoids incessant exploration of a single location, encouraging orienting to novel areas.
- Motor - translates the Action module transformations to motor commands. Orienting is head-led, only moving the neck and body if head movement alone cannot reach the destination. Once the salient location is reached, a whisking bout is induced, repeating the cycle.

3.2 Multisensory integration and reconstruction using multimodal predictive coding network

To generate multisensory inferences, a MultiPredNet architecture is used[1]. Based on principles of predictive coding[2][6][19] and building on prior work[3], this network flips the conventional Deep Learning information flow on its head. Rather than being led by the sensory data filtering through weight matrices, the MultiPredNet instead leads by predictions. Hypothesised 'causes', high-level predictions of what the world contains, are passed in a top-down fashion and compared with the sensory input at each level. The remainder of the signal - that not predicted by the causes - will continue to propagate upwards.

[1] Code and data can be found at:
https://github.com/TomKnowles1994/MultiPredNet/releases/tag/1.3.2

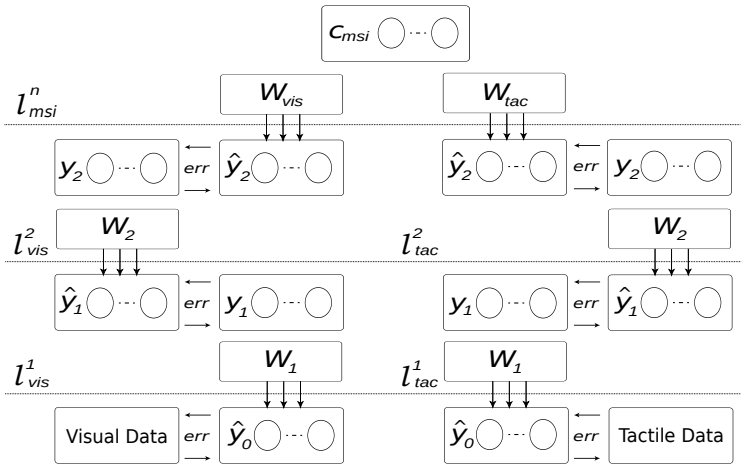


Fig. 3: The MultiPredNet architecture. Each layer contains a filter of learned weights (W_x) and receives top-down, hypothesised causes (c_x) of the input at the preceding $(l-1)$ layer. Causes pass through these weights, generating predictions of lower layer cause values. Discrepancy between predictions and causes propagate to higher layers as error gradients. The topmost layer integrates both modalities, learning a single set of causes that, filtered through modality-specific weights, reconstruct each unimodal data input.

The MultiPredNet begins with randomly initialised weights and arbitrary cause values (0.1 by default). Each layer of causes ($\mathbf{y}^{(l)}$) is updated in parallel with a Hebbian-like learning rule:

$$\Delta\mathbf{y}^{(l)} = \eta_y \left(\mathbf{W}^{l(l-1)} \phi'(\hat{\mathbf{y}}^{(l-1)}) \left(\left(\mathbf{y}^{(l-1)} - \hat{\mathbf{y}}^{(l-1)} \right) + \left(\mathbf{y}^{(l)} - \hat{\mathbf{y}}^{(l)} \right) \right) \right) \quad (1)$$

where η_y is the learning rate and ϕ' is the derivative of the activation function. Error component $(\mathbf{y}^{(l-1)} - \hat{\mathbf{y}}^{(l-1)})$ is the bottom-up error, comparing the prediction derived from the upper cause to the actual value of the causes. This penalises causes that cause poor predictions of lower layer causes. Error component $(\mathbf{y}^{(l)} - \hat{\mathbf{y}}^{(l)})$ is the top-down error, comparing the current value of the cause to what it was predicted to be by $\mathbf{y}^{(l+1)}$. This penalises causes that are difficult to predict by higher layers. Note that $\mathbf{y}^{(l+1)}$ is not a component of this learning rule, as its own value is not required to update $\mathbf{y}^{(l)}$, only its prediction $(\hat{\mathbf{y}}^{(l)})$. Note that for the uppermost layer, there is no higher layer to predict causes, and thus top-down error is treated as 0.

Each layer has a threshold defines the margin of error (10^{-3} to 10^{-4}) between a cause (or data item) and its prediction. Once all layers are within their error criteria (or after a maximum number of iterations), inference stops and the final causes values compared to the predictions. Further discrepancy between final causes and predictions leads to a weight update as per:

$$\Delta \mathbf{W}_{l(l-1)} = \eta_w \mathbf{y}^{(l)} \phi'(\hat{\mathbf{y}}^{(l-1)}) \left(\mathbf{y}^{(l-1)} - \hat{\mathbf{y}}^{(l-1)} \right)^T \quad (2)$$

with η_w being the learning rate for the weights. This iterative adjustment of causes occurs both during training and when generating inferences. Inferences do not invoke weight updates - the filters are 'fixed' - and rely on adjustment of causes to match predictions to the data presented. These predictions should therefore not be considered a direct window into the latent representations of the network, nor a decoded reconstruction of such, instead being a live hypothesis of the network as to the causes of the l_x^0 sensory impingement.

4 Results

Datasets were collected from WhiskEye exploring a virtual ovoid arena populated with coloured cubes and cylinders. Visual data consisted of 3-channel RGB images from the left camera, downsized to 80x45x3 pixels and flattened into a 1-D array of 10,800 elements. Tactile data consisted of 24 whisker protractions (θ_{whisk}) and 24 x 2 values of deflection data (x_{whisk} and y_{whisk}) concatenated into a 1-D, 72 element array. Sampling was driven by the rat-inspired whisking behaviour described in Section 3.1, with 'views' in both modalities captured at the moment of whisker peak protraction; whether the whiskers reached their desired theta angle or not (due to obstacles and/or IOR).

Note the relationship between visual and tactile data (Figure 4); a visual scene displaying largely wall implies proximity to the wall (f), and thus many whiskers colliding with the surface. The tactile data reflects this, with greater and more numerous deflections (d) in comparison to a clear visual scene (c, e). Relationships like these can be learned by MultiPredNet's multisensory layer, inferring that denser tactile input implies a more occluded visual scene and vice-versa.

The MultiPredNet was initialised with random filter weights and causes set to 0.1; two visual layers of 1000 (L_{vis}^1) and 300 (L_{vis}^2) neurons; two tactile layers of 50 (L_{tac}^1) and 20 (L_{tac}^2) neurons; and a single L_{msi} layer of 100. Causes were allowed to infer for 50 cause epochs before weight updates took place. 1900 samples of training data were divided into minibatches of 10 and the network trained for 200 training epochs. During some inferences, modalities were masked to test robustness to sensory dropout.

Figure 5 shows sample inferences generated from testsets 1 and 4 as per Section 3.2. Representational Similarity Analysis [12] was used to compare the distances between samples within each space. Assuming the robot can only rotate its head around the z-axis and is bound to a flat plane, its position and orientation is represented completely by a pose vector (x_{pose}, y_{pose}, θ_{pose}); a high quality reference representation useful for localisation. Therefore, if dissimilarity within pose space correlates well with that of MultiPredNet inference space, the inference will be of good quality proportional to that correlation. Results from all test sets under all conditions show a mean correlation well above significance, thus the representations generated are useful for localisation.

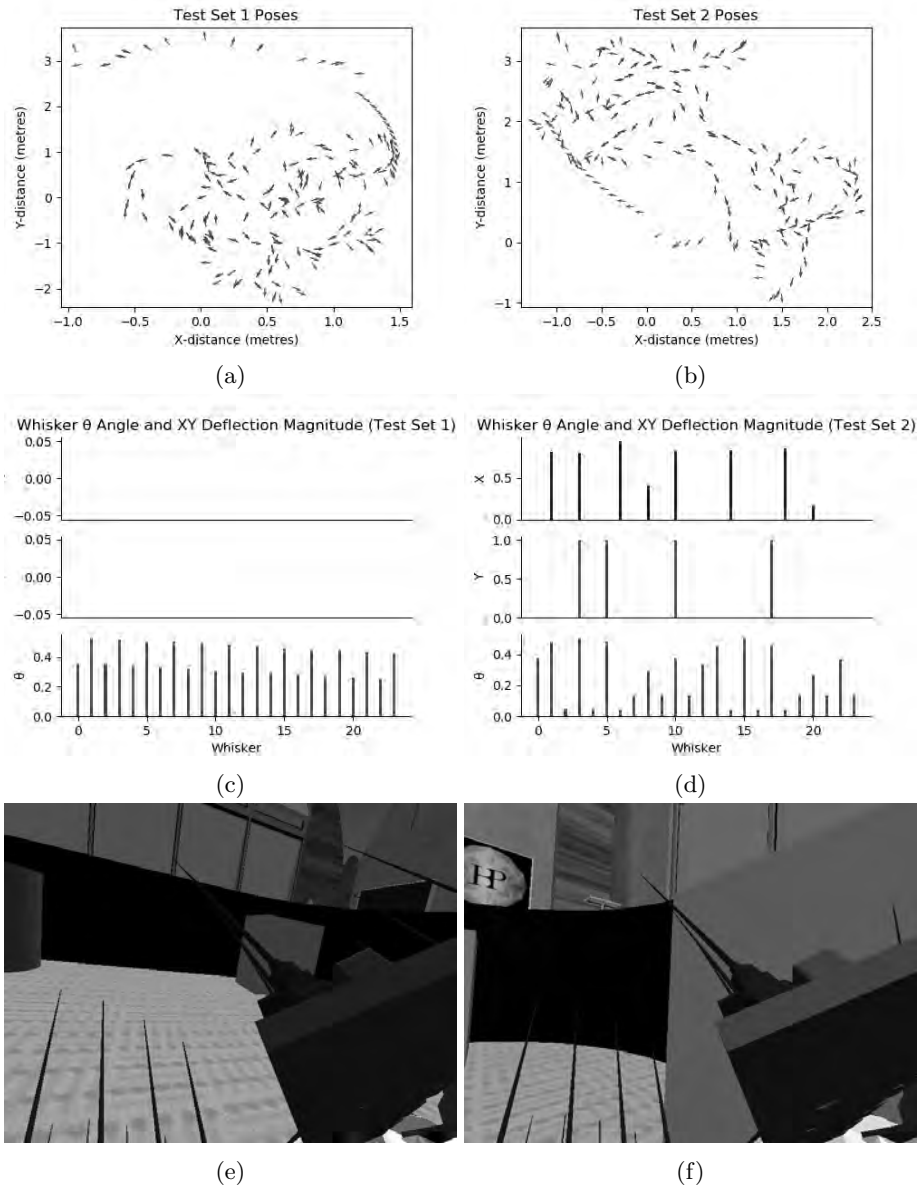


Fig. 4: A sample of MultiPredNet data from testsets 1 and 4. *a* and *b*: Quiver plot of poses (x_{whisk}, y_{whisk}, θ_{whisk}) *c* and *d*: Sample instances of whisker θ_{whisk} angle alongside the resulting magnitude of whisker deflection in x_{whisk} and y_{whisk} axes. *e* and *f*: Sample instances of camera visual input.

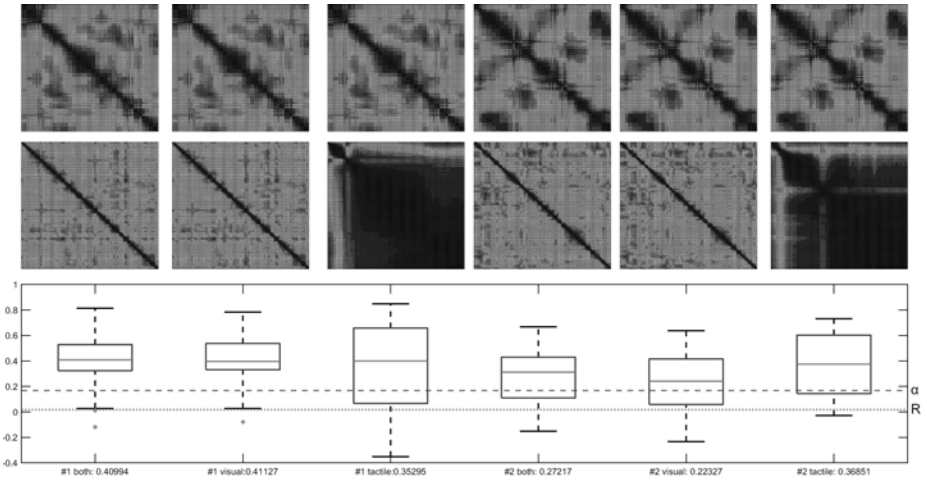


Fig. 5: RDM plots and Spearman's rank correlation coefficient scores for inferences on 100 samples from test sets 1 and 2 with visual, tactile, or both inputs unmasked. The top row of heatmaps show Euclidean distance between visited locations in pose space (x_{pose}, y_{pose}, θ_{pose}). The bottom rows of heatmaps show the 1-Pearson correlation distance between samples in MultiPredNet inference space; [12] shows this to be a more suitable metric for high-dimensional representation spaces. Below the heatmaps are boxplots of the correlation between spaces, with dotted line R marking correlation with a uniform random RDM, and dashed line α showing the threshold for significance (0.167). Significance is determined by $p < 0.05$ for $N = 100$ samples.

5 Discussion

In this paper we have described a novel multisensory robot which investigates salient environmental features in an ethological manner. The datasets from these investigations have then been used to train a multisensory predictive coding network that can generate inferences useful for place recognition. Furthermore, generated inferences remain useful even when modalities are obscured, a trait useful for real-world situations where vision is poor or whiskers are damaged.

Though fit for purpose, the datasets gathered have several areas of improvement. A prominent feature in every camera frame is the whisker array itself; with no benefit to localisation, this is an irrelevant feature that will be removed in future work to allow the network to learn more about the external environment, rather than itself. The unused right camera feed (with its own whisker array portion removed) can be used to make up the difference without altering the input shape of the network; important both for comparing results and re-using trained weights.

The results show a clear correlation between inference space and pose space, showing that *something* useful for place recognition is captured by the network,

and the correlation with pose specifically suggests that the MultiPredNet's is able to extract latent features relating to the observer; namely position and orientation. However, unlike in some other generative models such as β-VAEs[9], these latent features are highly entangled and not human-legible; there are no explicit 'x_{pose}', 'y_{pose}' or 'θ_{pose}' variables in the representation, and to the extent these are represented, it is as a high-dimensional mix of other variables. In a similar vein, MultiPredNet current stores representations as single, discrete numbers, rather than as a distribution (as VAEs in general do); as VAE disentanglement factors e.g. Kullbach-Leiber Divergence require distributed representations, this makes disentangling MultiPredNet's representations by these methods intractable in their current form.

To address both these issues, future work will look towards creating a 'Variational MultiPredNet' to learn disentangled representations at each layer. We will then use this as the sensory front end of a full localisation system, using the multimodal inferences produced by the MultiPredNet as a prediction of the current pose. This inferred pose will then be used to correct for inherent drift in internal representations of self motion modelled as spiking neural networks inspired by mammalian spatial cells, a task made easier by the NRP's integration with both SpiNNaker and NEST.

Acknowledgments

This research has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 945539 (Human Brain Project SGA3).

References

1. Anderson, S.R., Porrill, J., Pearson, M.J., Pipe, A.G., Prescott, T.J., Dean, P.: An internal model architecture for novelty detection: Implications for cerebellar and collicular roles in sensory processing. PLOS ONE **7**(9), 1–17 (09 2012)
2. Clark, A.: A nice surprise? predictive processing and the active pursuit of novelty. Phenomenology and the Cognitive Sciences **17**(3), 1572–8676 (2018)
3. Dora, S., Pennartz, C., Bohte, S.: A deep predictive coding network for inferring hierarchical causes underlying sensory inputs. In: Artificial Neural Networks and Machine Learning – ICANN 2018. vol. 11141. Springer (2018)
4. Falotico, E., Vannucci, L., Ambrosano, A., Albanese, U., Ulbrich, S., Vasquez Tieck, J.C., Hinkel, G., Kaiser, J., Peric, I., Denninger, O., Cauli, N., Kirtay, M., Roennau, A., Klinker, G., Von Arnim, A., Guyot, L., Peppicelli, D., Martínez-Cañada, P., Ros, E., Maier, P., Weber, S., Huber, M., Plecher, D., Röhrbein, F., Deser, S., Roitberg, A., van der Smagt, P., Dillman, R., Levi, P., Laschi, C., Knoll, A.C., Gewaltig, M.O.: Connecting artificial brains to robots in a comprehensive simulation framework: The neurorobotics platform. Frontiers in Neurorobotics **11**, 2 (2017)
5. Fox, C., Evans, M., Pearson, M., Prescott, T.: Tactile slam with a biomimetic whiskered robot. In: 2012 IEEE International Conference on Robotics and Automation. pp. 4925–4930 (2012)

6. Friston, K.: The free-energy principle: a unified brain theory? Nature Reviews Neuroscience **11**, 127–138 (2010)
7. Furber, S., Bogdan, P.: SpiNNaker: A Spiking Neural Network Architecture. Boston-Delft: now publishers (2020)
8. von Helmholtz, H.: Treatise on Physiological Optics Vol. III. Dover Publications (1867)
9. Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., Lerchner, A.: beta-vae: Learning basic visual concepts with a constrained variational framework. ICLR 2017 (2016)
10. Khaleghi, B., Khamis, A., Karray, F.O., Razavi, S.N.: Multisensor data fusion: A review of the state-of-the-art. Information Fusion **14**(1), 28–44 (2013)
11. Korthals, T., Hesse, M., Leitner, J., Melnik, A., Rückert, U.: Jointly trained variational autoencoder for multi-modal sensor fusion. 2019 22th International Conference on Information Fusion (FUSION) pp. 1–8 (2019)
12. Kriegeskorte, N., M., M., Bandettini, P.: Representational similarity analysis – connecting the branches of systems neuroscience. Frontiers in Systems Neuroscience **2** (2008)
13. Lotter, W., Kreiman, G., Cox, D.: Deep predictive coding networks for video prediction and unsupervised learning. ArXiv **abs/1605.08104** (2017)
14. Milford, M., Wyeth, G., Prasser, D.: Ratslam: a hippocampal model for simultaneous localization and mapping. In: IEEE International Conference on Robotics and Automation, 2004. Proceedings. ICRA '04. 2004. vol. 1, pp. 403–408 Vol.1 (2004)
15. Mitchinson, B., Prescott, T.J.: Whisker movements reveal spatial attention: A unified computational model of active sensing control in the rat. PLOS Computational Biology **9**(9), 1–16 (09 2013)
16. Pearson, M.J., Fox, C., Sullivan, J.C., Prescott, T.J., Pipe, T., Mitchinson, B.: Simultaneous localisation and mapping on a multi-degree of freedom biomimetic whiskered robot. In: 2013 IEEE International Conference on Robotics and Automation. pp. 586–592 (2013)
17. Pearson, M.J., Pipe, A.G., Melhuish, C., Mitchinson, B., Prescott, T.J.: Whiskerbot: A robotic active touch system modeled on the rat whisker sensory system. Adaptive Behavior **15**(3), 223–240 (2007)
18. Pennartz, C.M.: The Brain's Representational Power: On Consciousness and the Integration of Modalities. Cambridge: The MIT Press (2015)
19. Rao, R., Ballard, D.: Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature Neuroscience **2**, 79–87 (1999)