Autonomous Learning of Smooth Pursuit and Vergence through Active Efficient Coding

T. N. Vikram\textsuperscript{1}, C. Teulière\textsuperscript{2}, C. Zhang\textsuperscript{3}, B. E. Shi\textsuperscript{3} and J. Triesch\textsuperscript{1}

\textsuperscript{1}Frankfurt Institute For Advanced Studies, Goethe University, Frankfurt am Main, Germany
\textsuperscript{2}Clermont Université, Université Blaise Pascal, Institut Pascal, Clermont-Ferrand, France
\textsuperscript{3}Department of Electronic and Computer Engineering, HK University of Science and Technology, Hong Kong, China

Abstract—We present a model for the autonomous and simultaneous learning of smooth pursuit and vergence eye movements based on principles of efficient coding. The model accounts for the joint development of visual encoding and eye movement control. Sparse coding models encode the incoming data and capture the statistics of the input in spatio-temporal basis functions while a reinforcement learner generates eye movements to optimise the efficiency of the encoding. We consider the embodiment of the approach in the iCub simulator and demonstrate the emergence of a self-calibrating smooth pursuit and vergence behaviour. Unlike standard computer vision approaches, it is driven by the interaction between sensory encoding and eye movements. Interestingly, our analysis shows that the emerging representations learned by this model are in line with results on velocity and disparity tuning properties of neurons in visual cortex.

I. INTRODUCTION

The neurobiology of vision reveals that an organism’s visual system is fine-tuned for the visual properties of its habitat. In this context the efficient coding principle postulates that the goal of early visual processing is to generate an efficient representation of the visual scene by utilizing the existing computational resources optimally. Thus the visual system attempts to represent a scene by decomposing it into statistically independent components \cite{1}. This might further result in statistical independence of responses from different neurons while processing a visual stimulus.

Visual perception is an active process, however, where the brain controls and manipulates the visual input through eye movements. During visual scene perception, high quality visual information is acquired only from a limited spatial region surrounding the center of gaze (known as the fovea). Eye movements thus reorient the fovea through the scene by saccades, smooth pursuit, vergence and other movements. Smooth pursuit eye movements keep the attended moving target stable on the fovea. On the other hand vergence eye movements align the two eyes on the same object and correct the misalignments by monitoring the difference in the positions of the two retinal images of the object \cite{2}. Active vision thus ensures that visual information is available when required, and also simplifies a variety of inherently complex computational problems \cite{3}. The common coding theory by Prinz \cite{4} hypothesises the existence of a unified representational system for perception and action, where perception leads to action and vice-versa. It is thus plausible that the emergence of active vision behaviour and optimal encoding of the visual scene are inter-related.

During smooth pursuit under dichoptic viewing conditions, the motion components perceived separately by the two eyes are coupled by the visual system into a single motion stream \cite{9}. It was further shown in \cite{5} that the visual system configures the eyes such that it minimizes the binocular disparity and retinal slip during smooth pursuit. The coupling of motion streams and reduction in disparity thus introduces redundancy of information. The authors in \cite{10} argue that the visual cortex quickly adapts to the prevailing binocular image statistics and maximize information-coding efficiency.

The efficient coding hypothesis is driven by the assumption that the sensory processing in the brain reflects the statistics of the visual stimulus. However, the behavioural response of the agent with respect to the visual stimulus also determines the nature of the statistics, as the behaviour re-orient vision to some preferred locations on the scene. The effect of behaviour on the perceptual statistics has received very little attention so far in the literature. It is plausible that behaviour and sensory processing jointly develop in an organism. This question of joint development has not been investigated much either so far.

We propose a model for the development of binocular smooth pursuit and vergence behaviour based on the principles of efficient coding. The proposed model accounts for the joint learning of scene representation and the corresponding eye motor commands. Dictionaries of sparse basis functions are learned in order to decompose the visual scene into statistically independent components. The reconstruction efficiency of the existing sparse dictionary is treated as a reward signal to a reinforcement learner which in turn generates eye velocity and vergence motor commands. After subjecting our model to a large number of perception-action training iterations we observe the emergence of Gabor-like basis functions which resemble the visual receptive fields observed in primary visual cortex. As the model explores different eye movements with the goal of improving coding efficiency, it discovers that by performing accurate vergence and smooth pursuit movements it can encode the visual inputs more efficiently. We call this “active efficient coding”.

A disparity selectivity and vergence control model was proposed in \cite{6} which is based on efficient coding. This model combines a sparse coding stage with an intrinsically motivated form of reinforcement learning where a measure

\begin{thebibliography}{9}
\bibitem{1} \textsuperscript{1}
\bibitem{2} \textsuperscript{2}
\bibitem{3} \textsuperscript{3}
\bibitem{4} \textsuperscript{4}
\bibitem{5} \textsuperscript{5}
\bibitem{6} \textsuperscript{6}
\end{thebibliography}
of the system’s coding efficiency serves as a reward signal for learning disparity perception in binocular vision. This was further extended and embodied in the iCub humanoid robot in [7]. A similar approach was proposed to learn smooth pursuit behaviour under monocular viewing conditions in [8].

The proposed model is embodied within the iCub simulator environment which unifies our previous works on learning vergence and smooth pursuit in isolation. To the best of our knowledge, this is the first attempt to model the concurrent learning of smooth pursuit and vergence eye movements on the basis of the efficient coding hypothesis.

II. PROPOSED MODEL

We propose a model where smooth pursuit and vergence behaviours emerge autonomously because they increase the efficiency in encoding of the spatio-temporal scene. The visual system takes advantage of the regularities in the stimulus, coding only what makes the stimulus unique [16], [14]. Thereby increased redundancy in the sensed spatio-temporal stimulus leads to higher information throughput. The model further discovers that the successive pairs of foveal images are encoded efficiently when they are nearly identical.

Assuming a scene where there is a fronto-parallel motion of an object at a constant depth, a high redundancy between left and right images at a given time is achieved when there is low binocular disparity. Similarly, a high redundancy between left/right images between two successive moments in time is achieved when the eye velocity is nearly equal to the object velocity. The proposed model thus attempts to generate motor commands which maximize the encoding efficiency of the two successive pairs of foveal images.

The proposed model has two distinct parts:

1) Sensory Encoding by Sparse Coding: The input images are decomposed into a set of patches, which are represented by a linear combination of basis functions. The negative of the approximation error of this representation is passed as a reward signal to the action generating agent. The residual encountered during the approximation is further used to update the basis functions.

2) Motor Action Generation by Reinforcement Learning: The basis function activations are passed to the action generation module that generates the eye commands which adjust the eye velocities and vergence. This results in the update of the images sensed by the eyes. The sparse coding modules further encode this and the approximation error is passed as the reward signal. The reward is highest when there is redundancy between the foveal images. The reinforcement module thus updates its beliefs to weight those actions which lead to higher rewards as compared to the other actions.

The diagrammatic representation of the proposed approach is given in Fig. 1.

A. Sparse Coding

The input is an RGB image of dimension 320 × 240 pixels as we use the iCub simulator environment to embody our model. We convert the input RGB image into gray scale. In order to simulate the settings of developmental vision, the system is not given access to the camera parameters. The human visual system has the highest resolution on the foveal region, and similarly a window around the centre of the input image is considered as the fovea in our model.

The foveal images were extracted from a coarse and a fine scale region. A 96 × 96 pixels region around the centre was used as the coarse scale fovea and similarly a 72 × 72 pixels region around the centre was used as the fine scale fovea. At the coarse scale, the foveal window is sub-sampled by a ratio of four in a Gaussian pyramid hierarchy. Subsequently, 81 patches of dimension 8 × 8 pixels are extracted from both foveal scales. The pixel intensity values within the patch are further converted to z-score representation assuming the data to be i.i.d samples from a Gaussian distribution.

This process is replicated for left (l) and right (r) eyes at times t and t − 1. The corresponding ith patch (p) from these four images at the coarse (c) scale are vectorized and concatenated as follows:

\[ c\mathbf{s}_t = [\begin{array}{cccc} f_p^{t_1}; & r_p^{t_1}; & l_p^{t_1-1}; & r_p^{t_1-1} \end{array}] \]  

Similarly, \( f\mathbf{s}_t \) refers to the vectorised version of a given ith patch at fine scale.

\[ f\mathbf{s}_t = [\begin{array}{cccc} f_p^{t_1}; & f_p^{t_1}; & f_p^{t_1-1}; & f_p^{t_1-1} \end{array}] \]  

Hence the fine and coarse scale encoding components each receive a vector of 256 elements. We drop the coarse (c) or fine scale (f) indices from now on. The set \( \mathbf{S}_t \) of spatio-temporal vectors (\( \mathbf{s}_t \)) thus consists of 81 elements.

Each of the spatio-temporal vectors \( \mathbf{s}_t \) is encoded as a linear combination of basis functions \( \phi_j^t \) from a fixed-size dictionary of basis functions \( \mathcal{B} \). The training basis functions attempt to best represent a given \( \mathbf{s}_t \) by linear combinations of \( k \) (where \( k = 10 \)) training basis functions \( \phi_j^t \in \mathcal{B} \) and their activation co-efficients are obtained through a greedy search. The dimension of the \( \phi_j^t \) is same as the spatio-temporal vector given in Eq. 1 or 2.

The dictionary size is chosen so that \( |\mathcal{B}| > 256 \) basis functions, i.e. the dictionary is over-complete. In practice we
set $|\mathcal{B}| = 288$, but this precise value is not critical. At $t = 1$, each element of $\phi^t_j, \phi^t_n$, $n = 1...256$ is initialised as:

$$\phi^t_j \sim \mathcal{N}(0, 1) \quad (3)$$

To encode the input data, we approximate of the spatio-temporal input vectors as a sparse linear combination of basis functions from the dictionary. Formally, we determine coefficients $\alpha^t_{i,j}$ which minimize the approximation error $e^t$ defined as follows:

$$e^t = \sum_{i=1}^{S^t} (s^t_i - \sum_{j=1}^{B^t} \alpha^t_{i,j} \phi^t_j)^2 \quad (4)$$

These coefficients are computed using the matching pursuit algorithm [17] which uses a greedy search. The algorithm finds in the dictionary the basis functions to represent each input vector $s^t_i$. The sparsity of the encoding is ensured by enforcing that for each $i$ only $k$ of the $\alpha^t_{i,j}$ are non-zero.

The training basis functions $\mathcal{B}^t$ are subsequently updated using gradient descent to further minimize $e^t$. The linear coefficients of the training basis functions which were used to approximate the spatio-temporal vectors ($s^t_i$) are further used for computing the representative features $\mathcal{F}^t$ as:

$$\mathcal{F}^t = \left[|\mathcal{S}|^{-1} \sum_{i=1}^{S^t} (s^t_i)^2 ; \ldots ; |\mathcal{S}|^{-1} \sum_{i=1}^{S^t} (s^t_i|\mathcal{B}|)^2 \right] \quad (5)$$

$\mathcal{F}^t$ is computed separately for both coarse and fine scales. They are further concatenated and passed as an input to the reinforcement learning agent. The usage of $\mathcal{F}^t$ from now on implicitly means the representative vector obtained by concatenating both coarse and fine scale representative vectors. The agent subsequently learns the mapping from this feature space to the action space with the goal of minimizing the cumulated approximation error $(c e^t + f e^t)$, where $c e^t$ and $f e^t$ represent the approximation errors at coarse and fine scale respectively.

B. Reinforcement Learning

A natural actor-critic algorithm is employed where neural networks model the policy and value functions. Note that we use three different agents for manipulating the pan, tilt and vergence joints. The agents share the same state representation and receive identical reward signals but chose their respective actions independently. The weights $w_m$ for a particular joint $m$ (pan, tilt, or vergence) of the neural network are randomly initialized. The discounted cumulative reward $R^t$ with a discount factor $\gamma$ is defined as:

$$R^t = \sum_{l=0}^{\infty} -\gamma^{-l} (c e^{t+l} + f e^{t+l}) \quad (6)$$

Each agent attempts to maximize $R^t$, and henceforth aims to generate such motor actions which leads to minimal approximation error. Motor commands generated are relative increments to the current eye velocities (in degrees/sec for pan/tilt joints) or positions (in degrees for vergence joint). Let $\mathcal{A}_m$ (where $\mathcal{A}_m = \{-16, -8, -4, -2, -1, 0, 1, 2, 4, 8, 16\}$) be the set of possible actions for a particular joint $m$. The policy networks thus map the $\mathcal{F}^t$ to $a_m$ where $a_m \in \mathcal{A}_m$.

The value function network maps the representative feature $\mathcal{F}^t$ to a corresponding value ($V^t_m$) as:

$$V^t_m = \langle w_m, \mathcal{F}^t \rangle \quad (7)$$

The number of neurons in the output layer of the policy network thereby equals $|\mathcal{A}_m|$. The activation $z^t_{a_m}$ of the output neuron corresponding to the action $a_m$ is given by:

$$z^t_{a_m} = \langle \theta^t_{a_m}, \mathcal{F}^t \rangle \quad (8)$$

where $\theta^t_{a_m}$ are the regression coefficients which map $\mathcal{F}^t$ to action $a_m$. We employ a softmax function to compute the probability $\pi^t_{a_m}$ of selecting an action $a_m$ as:

$$\pi^t_{a_m} = \frac{\exp(z^t_{a_m} T^{-1})}{\sum_{m=1}^{\mathcal{A}_m} \exp(z^t_{a_m} T^{-1})} \quad (9)$$

The temperature $T$ where $(T = 1)$ controls the exploration versus exploitation behaviour of the learning agent.

III. Experiments

We first evaluate the model’s performance while learning pan, tilt and vergence eye movements. Subsequently we analyse the structure of the learned basis functions for their sensitivity to velocity and disparity and compare them to biological findings.

A. Performance Evaluation of the Model

We operate on the pan ($m_1$), tilt ($m_2$) and vergence ($m_3$) joints of the iCub simulator environment to train and test our model. Flat textured objects which moved in a fronto-parallel plane at constant depth were considered for experiments. The textured objects were changed once in 200 iterations. Pan and tilt velocities are measured in deg/s while object depth is measured in meters. The pan velocity ($V_{m_1}$), tilt velocity ($V_{m_2}$) and the object depth ($D_{m_3}$) were sampled from uniform distribution as follows: $V_{m_1} \in [-25, 25]$, $V_{m_2} \in [-15, 15]$, $D_{m_3} \in [0.5, 2]$. Each iteration took approx. 0.15s to complete, and a 4.5 pixel shift reflected a change of 1 degree at the fine scale. The $V_{m_1}, V_{m_2}, D_{m_3}$ were changed at every 20th iteration, and at every 10th iteration the signs of $V_{m_1}$, and $V_{m_2}$ are inverted. A trial consists of 20 iterations where eye velocities and disparity were recorded at the 9th and 19th iteration, and compared with that of the object velocity and depth. Each experiment consisted of 10^5 trials, and a total of four experiments was performed to compute the mean average errors (MAE) of pan, tilt and vergence actions.

In Fig. 2 we present the MAE of pan, tilt and vergence actions. It can be observed from Fig. 2a that the MAE of pan has reduced from 17 deg/s to approximately 8 deg/s. Similarly, the tilt MAE (Fig. 2b) which was at 11.5 deg/s in the beginning...
reduces to 6 deg/s in the end. In Fig. 2c the MAE for vergence actions is presented. It can be observed that the vergence error which was 3.5 deg in the beginning reduces to 0.5 deg in the end. The results in Fig. 2 show that the proposed objective of coding efficiency is sufficient to account for the development of vergence and smooth pursuit behavior.

The modeling of vergence action is apparently less difficult as compared to pan and tilt in our case change of depth occurs only at every 20th iteration while the velocity changes every 10th iteration. It should be noted that both pan and tilt velocities have a high reduction in MAE until $4 \times 10^4$ trials and they subsequently saturate. The reduction in MAEs is directly related to the convergence of the respective reinforcement learning agents.

Another possibility is to have a single learning agent which generates all the three actions. However, this might lead to computational intractability due to the large number of state-action combinations. Taking this into account we employed three reinforcement learning agents which are independent and cater to a single motor degree of freedom. This is similar to the naive Bayes formulation where the joint probability between events assumes their conditional independence. Many models in vision research hypothesize that the brain attempts to solve the vision problem by efficient approximations of computationally intractable functions.

### B. Analysis of basis functions properties

We investigated whether the learned basis functions have a Gabor-like structure. It has been shown in [12], [20] that orthogonal decomposition of natural scenes leads to basis functions which are best characterized by Gabor-like functions. From a biological perspective, receptive fields of the so-called simple cells in the primary visual cortex have Gabor-like structures and are sensitive to orientation, contrast, luminance, speed and direction of motion [21], [13].

We consider the following Gabor function:

$$g(x, y) \Delta, \psi = \exp(-\frac{x^2 + \beta^2 y^2}{2\sigma^2}) \cos(2\pi \frac{x'}{\lambda} + \psi)$$

$$\Delta = \{\lambda, \theta, \sigma, \beta, x_c, y_c\}$$

$$x' = (x - x_c) \cos \theta + (y - y_c) \sin \theta$$

We fit Gabor functions to the four sub-fields of the basis functions. The fitted Gabor functions have identical $\Delta$ but differ only in their phase offsets $\psi$.

For qualitative evaluation we provide the coarse and fine scales of three different basis functions obtained at the end of training in Fig. 3 and Fig. 4. It can be observed from the first and middle column of Fig. 3 that the basis functions are tuned to zero velocity but non-zero disparity. The middle column in Fig. 4 shows a basis function which is tuned to zero velocity and zero disparity.

The residual norms between the actual basis functions and the fitted Gabor functions are computed to assess the quality...
Fig. 5: Error in terms of residual norms when Gabor filters were fit to basis functions.

Fig. 6: Histogram of Basis Orientations

Fig. 7: Histogram of Disparity Preferences

Fig. 8: Histogram of Velocity Preferences

Fig. 9: Scatter Plot of Disparity Preferences

Fig. 10: Scatter Plot of Velocity Preferences

of the fits. In Fig. 5, we provide the histograms of the residual norms for coarse and fine scale Gabor fits. Only those coarse and fine scale basis functions whose residual norms were less than 0.3 were considered for further analysis. The histograms indicate that we obtain a good quality of Gabor fits over the basis functions with approximately 100 basis functions at both coarse and fine scales have a residual norm of less than 0.15.

In Fig. 6, we visualize the histograms of basis function orientations. It should be noted that the most dominant preferred orientation is vertical. This is in line with the research from [15] which reports that more receptive fields have vertical or horizontal orientation preference than oblique ones. Both coarse and fine scale basis functions have a similar orientation preference and hence orientation preference is agnostic to scale of the foveal image. The bias towards horizontal and vertical orientation is also in parts due to the presence of horizontal and vertical edges in the set of images on which we operated.

We further analyse the disparity preference of the basis function pairs at time \( t \) and \( t - 1 \) and the results are given in Fig. 7. It is apparent that the most preferred disparity is zero and the spread of the distribution increases at the coarse scale. The velocity preferences of the basis functions for left and right eyes are computed and the respective histograms for coarse and fine scale basis functions are given in Fig. 8. Most of the basis functions have preference for low velocity at both coarse and fine scales.

We further examine the change in disparity preference of the basis functions with respect to time. Fig. 9 shows that there exists a linear correlation between the disparity of a given basis
with respect to \( t \) and \( t - 1 \). It can observed that at there are a few outliers which do not have identical disparity preferences at both \( t \) and \( t - 1 \). This can be attributed to the experimental settings where in we change the depth of the object at every 20\(^{th}\) iteration.

We plot the velocity preference of Left and Right eyes in Fig. 11. The majority of the basis functions have a preference for zero velocity. We further plot the average velocity preference of both the eyes versus the average disparity preference for \( t \) and \( t - 1 \) in Fig. 11. This shows that majority of the basis functions have a low variance in velocity preference but have a large variance in disparity preference. This is similar to the biological findings from [18], which reported that most monkey VI cells are tuned to disparity but only a small fraction of them are tuned to motion.

IV. CONCLUSION

In this research, we have explored a model which autonomously learns smooth pursuit and vergence through active efficient coding. The model exploits the mutual relationship between motor response and sensory representation. This aspect of developmental vision has not been explored by the vision sciences community to the best of our knowledge. Our model does not require camera calibration parameters unlike most engineering models. Research from the vision community has claimed that the problem of active binocular vision is efficiently solved in the brain by reorienting the eyes such that the visual streams are redundant. This paradigm was utilized in [6, 7] which addressed the problem of disparity preference and also in [8] which addressed the problem of monocular smooth pursuit.

In our work, we concurrently address the problem of binocular smooth pursuit and vergence. To crudely mimic human vision, we only process a region around the centre of the input image at coarse and fine scale which are analogous to peripheral and foveal vision. The proposed model allows a simultaneous representation of velocity and disparity among a population of learned basis functions. This is consistent with the fact that many complex cells in the V1 area are tuned to orientation, velocity and disparity [19].

It has been well established in the literature that the receptive fields in certain parts of visual cortex in mammals have a Gabor-like structure. Our experiments have confirmed that the majority of the basis functions which emerged through active efficient coding have Gabor-like structure. Overall, we explored the emergence of smooth pursuit and vergence eye movements through active efficient coding.

V. ACKNOWLEDGEMENTS

Funded by the German Ministry of Education and Research (Bernstein Focus: Neurotechnology – FKZ 01GQ0840) and by the Hong Kong RGC (619111). Thanks to N. C. Pramod and L. Lonini for the many useful discussions.

REFERENCES