To appear in *Advanced Robotics*
Vol. 00, No. 00, Month 20XX, 1–17

ARTICLE

A Neural Model for Eye-Head-Arm Coordination

Wasif Muhammad[a]* and Michael W. Spratling[b]

[a]*Department of Electrical Engineering, University of Gujrat, Gujrat, Pakistan;* [b]*Department of Informatics, King's College London, London, UK*

(*Submitted: 03 October 2016*)

The coordinated movement of the eyes, the head and the arm is an important ability in both animals and humanoid robots. To achieve this the brain and the robot control system need to be able to perform complex non-linear sensory-motor transformations in the forward and inverse directions between many degrees of freedom. In this article, we apply an omni-directional basis function neural network to this task. The proposed network can perform 3-D coordinated gaze shifts and 3-D arm reaching movements to a visual target. Particularly, it can perform direct sensory-motor transformations to shift gaze and to execute arm reach movements and can also perform inverse sensory-motor transformations in order to shift gaze to view the hand.

Keywords: basis function network; sensory-motor transformation; spatial auto-encoder; direct visuo-motor transformation; inverse visuo-motor transformation

1. Introduction

In primates coordinated sensory-motor activity is a very common behaviour for visually guided reaching and manipulation tasks *e.g.*, writing, picking and holding *etc.*. All vision guided arm movements involve a series of sensory-motor transformations starting from visual sensory space and ending in arm joint space involving motor spaces and proprioceptive signals from the eyes and the head [1–3]. In primates this sensory-motor transformation is bi-directional. Hence, visual sensory information can be used to drive eyes-head motor spaces and the arm joint angles which is known as the "direct visuo-motor transformation" [1, 2]. Alternatively the arm joint angles can be used as a driving signal to shift gaze to view the hand which is known as the "inverse visuo-motor transformation" [4]. Such sensory-motor transformations in the brain are believed to be performed using basis functions [4–12].

Basis function networks are very popular in robotics for complex non-linear sensory-motor transformations [13–23]. For non-linear and complex sensory-motor transformations setting of number of basis function neurons, their receptive fields (RFs) sizes and peak locations is a non-trivial task which cannot be pre-defined or hand-crafted. All above mentioned basis function models were trained using two step complex training phases: learning of number of basis function neurons their RFs sizes and peak locations; learning of connection weights between the basis function neurons and the output neurons. For number of basis function neurons, RFs sizes and locations optimization: [23] used orthogonal least square algorithm, [18–20] used simplified node-decoupled extended Kalman filter algorithm, in [13, 14, 17, 24] basis function units with fixed RFs size and defined locations were used. To learn the network connection weights: [17] used least-mean-square (LMS) gradient descent learning technique, in [23] linear least square (LLS) algorithm was used, in [18–20] simplified node-decouple extended Kalman filter (SDEKF) algorithm was employed, [14] used delta rule gradient descent technique, [13] used recursive least square (RLS) algorithm and in [24] extended Kalman filter was used. Moreover, little work has been done on 3-D eye-head-arm coordination using basis function networks. Some methods have applied basis function networks to perform only the direct visuo-motor transformation [18, 23, 24]. Other work [13, 14, 17] has used basis function networks to implement bi-directional visuo-motor trans-

*Corresponding author. Email: syed.wasif@uog.edu.pk

formations, however to do so separate networks for direct and inverse transformations were required. In all these cases the head movement was restrained, resulting in head-centred and body-centred representations being the same. Furthermore, in [18] the visuo-motor transformations were performed in 2-D space. In [14] the bi-directional transformations were in 1-D space, however in [13, 17] they were in 3-D space.

In [25] we built a three stage Predictive Coding/Biased Competition-Divisive Input Modulation (PC/BC-DIM) basis function network to control eye movements which we extended by adding one more PC/BC-DIM stage for coordinated eyes-head control [26] . In this article, we extend the network model proposed in [26] using another PC/BC-DIM stage to control arm movements for 3-D coordinated eyes-head-arm movements and to demonstrate the working of the proposed PC/BC-DIM basis function model. The training of the PC/BC-DIM basis function network was simplified to one step online learning and optimization without involving any heuristics or complex learning algorithms. The proposed PC/BC-DIM basis function architecture is an omni-directional basis function network which has the ability to perform sensory-motor transformations in any direction. The proposed model controls coordinated eyes-head-arm movements in 3-D space for direct and inverse transformations. Specifically, to perform a direct visuo-motor transformation the proposed network transforms retinotopic visual information together with binocular eyes position information into a head-centred representation as in [25]. The head-centred representation is combined with information about the head position into a body-centred representation [26]. This body-centred representation can be used to control coordinated eyes and head movements to shift gaze and to bring a visual target onto the most sensitive part of the retina called the fovea. Moreover, we show that the body-centred representations can be used to control arm movements and hence allow the robot to reach a target identified visually. To perform an inverse visuo-motor transformation, information about the arm joint angles was used to determine the body-centred representation of the hand position. This body-centred representation was used to control coordinated eyes and head movements to shift gaze and to bring the hand onto the foveae of both eyes. We also show that the body-centred representations of multiple targets can be used to perform memory-based gaze shifts and arm movements in different directions to different targets of interest. To demonstrate this model, 3-D coordinated eyes-head-arm movements were performed using the iCub humanoid robot simulator having 7 DOFs for binocular eyes and head motor spaces along with 3 DOFs for non-redundant arm movements.

2. Methods

2.1. *The PC/BC-DIM Algorithm*

All experiments reported here were performed using the PC/BC-DIM hierarchical basis function neural network architecture proposed in [25–29]. Each level, or processing stage, in the hierarchy is implemented using the neural circuitry of three separate neural populations: the error, the prediction and the reconstruction neurons. The behaviour of the neurons in these three populations is determined by the following equations:

$$\mathbf{r} = \mathbf{V}\mathbf{y} \tag{1}$$

$$\mathbf{e} = \mathbf{x} \oslash (\epsilon_2 + \mathbf{r}) \tag{2}$$

$$\mathbf{y} \leftarrow (\epsilon_1 + \mathbf{y}) \otimes \mathbf{W}\mathbf{e} \tag{3}$$

Where $\mathbf{x}$ is a $(m \times 1)$ vector of input activations, $\mathbf{e}$ is a $(m \times 1)$ vector of error neuron activations; $\mathbf{r}$ is a $(m \times 1)$ vector of reconstruction neuron activations; $\mathbf{y}$ is a $(n \times 1)$ vector of prediction neuron activations; $\mathbf{W}$ is a $(n \times m)$ matrix of feedforward synaptic weight values; $\mathbf{V}$ is a $(m \times n)$ matrix of feedback synaptic weight values; ϵ_1 and ϵ_2 are parameters; and $\oslash$ and $\otimes$ indicate element-wise division

Algorithm 1 PC/BC-DIM Network Activations

procedure ACTIVATION($\mathbf{W}, \mathbf{V}, \mathbf{x}$)
  $\epsilon_1 = 1 \times 10^{-9}$
  $\epsilon_2 = 1 \times 10^{-9}$
  $\mathbf{y}$ =zeros$(n \times 1)$
  for $i = 1$ to iterations **do**
    $\mathbf{r} = \mathbf{V}\mathbf{y}$
    $\mathbf{e} = \mathbf{x} \oslash (\epsilon_2 + \mathbf{r})$
    $\mathbf{y} \leftarrow (\epsilon_1 + \mathbf{y}) \otimes \mathbf{W}\mathbf{e}$
  end for
  return $\mathbf{r}, \mathbf{e}, \mathbf{y}$
end procedure

and multiplication respectively. For all the experiments described in this paper ϵ_1 and ϵ_2 were both given the value 1×10^{-9}. The procedure used to determine the PC/BC-DIM network activations is provided in algorithm 1. The iterative process described in algorithm 1 was terminated after 150 iterations in all the experiments reported in this article.

The values of $\mathbf{y}$ represent predictions of the causes underlying the inputs to the network. The values of $\mathbf{r}$ represent the expected inputs given the predicted causes. The values of $\mathbf{e}$ represent the residual error between the reconstruction, $\mathbf{r}$, and the actual input, $\mathbf{x}$. The full range of possible causes that the network can represent are defined by the weights, $\mathbf{W}$ (and $\mathbf{V}$). Each row of $\mathbf{W}$ (which correspond to the weights targeting an individual prediction neuron) can be thought of as a "basis vector" or "elementary component" or "preferred stimulus", and $\mathbf{W}$ as a whole can be thought of as a "dictionary" or "codebook" of possible representations, or a model of the external environment. The activation dynamics described above result in the PC/BC-DIM algorithm selecting a (typically sparse) subset of active prediction neurons whose RFs (which correspond to basis functions) best explain the underlying causes of the sensory input.

The prediction neurons in a PC/BC-DIM network behave like basis function neurons [25, 26, 29]. Fig. 1 illustrates how a PC/BC-DIM basis function network can be used to perform a simple mapping from three input variables to an output variable. The algorithm used to perform transformation with the PC/BC-DIM basis function network for four variable case, shown in Fig. 1a (*i.e.*, provided inputs $\mathbf{x}_a$, $\mathbf{x}_b$ and $\mathbf{x}_c$ to determine $\mathbf{x}_d$), is described in algorithm 2. This algorithm will be used through this article for transformation between any number of input variables and for any inputs combination (*e.g.*, using $\mathbf{x}_a$, $\mathbf{x}_c$ and $\mathbf{x}_d$ transformation is performed to determine $\mathbf{x}_b$ as shown in Fig. 1b).

Algorithm 2 PC/BC-DIM Network Transformation

procedure TRANSFORMATION($\mathbf{W}, \mathbf{V}, \mathbf{x}_a, \mathbf{x}_b, \mathbf{x}_c$)
  $\mathbf{x}_d$ =zeros$(m_d \times 1)$
  $\mathbf{x} = [\mathbf{x}_a; \mathbf{x}_b; \mathbf{x}_c; \mathbf{x}_d]$
  $[\mathbf{r}, \mathbf{e}, \mathbf{y}]$ =ACTIVATION($\mathbf{W}, \mathbf{V}, \mathbf{x}$)
  $\mathbf{r}_d$=$\mathbf{r}$((length($\mathbf{x}_a$)+length($\mathbf{x}_b$)+length($\mathbf{x}_c$)) + 1 :end)
  return $\mathbf{r}_d$
end procedure

2.2. *The Proposed PC/BC-DIM Network for Eyes-Head-Arm Coordination*

The proposed eyes-head-arm coordination network utilizes the same eye-head coordination strategy as described in [26]. For clarity, the steps used to shift gaze and to move the arm to the target of interest are demonstrated with the help of the 1-D eye-head-arm control network shown in Fig. 2. The eye-head-arm coordination strategy is shown in Fig. 3 for the direct visuo-motor transformation and in Fig. 4 for the inverse transformation. The illustration of the 1-D eye-head-arm control network shown in Fig. 2

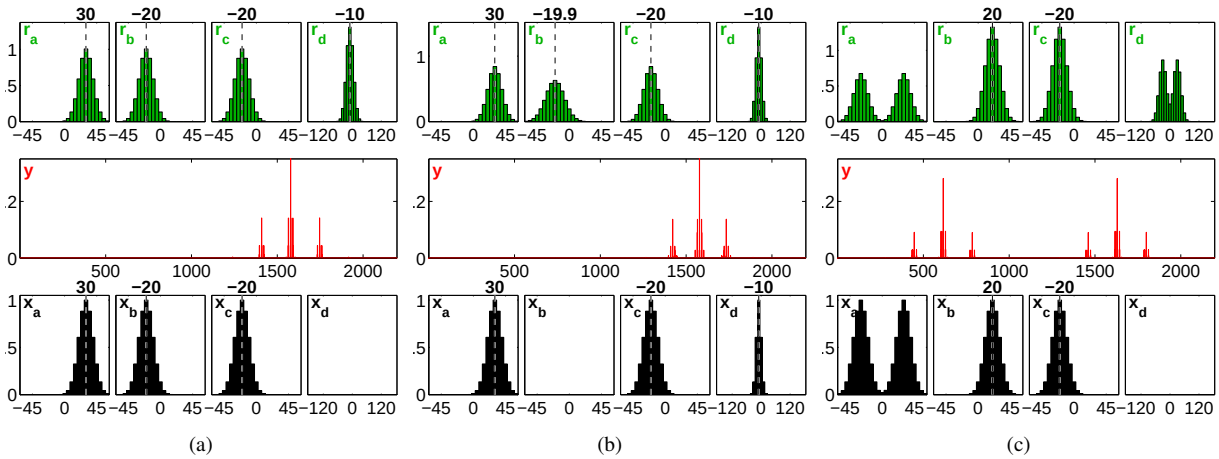


Figure 1.: Mapping between four variables using the PC/BC-DIM network illustrated. The PC/BC--DIM network has been wired-up to approximate the function $\mathbf{x}_d = \mathbf{x}_a + \mathbf{x}_b + \mathbf{x}_c$. In each sub-figure the lower histograms show the inputs, the middle histograms show the prediction neuron activations, and the upper histograms show the reconstruction neuron responses. The x-axis of each histogram is labelled with the variable value, except for the histogram representing the prediction neuron responses which is labelled by neuron number. The y-axes of each histogram are in arbitrary units representing firing rate. (a) When the three inputs representing $\mathbf{x}_a$, $\mathbf{x}_b$, and $\mathbf{x}_c$ are presented (lower histograms), the reconstruction neurons generate an output (upper histograms) that represents the correct value of $\mathbf{x}_d$ (as well as outputs representing the given values of $\mathbf{x}_a$, $\mathbf{x}_b$, and $\mathbf{x}_c$). (b) When the three inputs representing $\mathbf{x}_a$, $\mathbf{x}_c$ and $\mathbf{x}_d$ are presented (lower histograms), the reconstruction neurons generate an output (upper histograms) that estimates the correct value of $\mathbf{x}_b$ (as well as outputs representing the given values of $\mathbf{x}_a$, $\mathbf{x}_c$ and $\mathbf{x}_d$). (c) As (a) but with two values of $\mathbf{x}_a$ represented by a bi-modal input to the first partition. The network correctly calculates two values for $\mathbf{x}_d$ represented by the peaks of the bi-modal distribution produced by the reconstruction neurons in the last partition.

is simplified by superimposing the error and reconstruction neuron populations and by using double-headed arrow for the inputs/outputs to these populations. It is just the way of illustrating this model that has been simplified, however the mathematical model remains unchanged.

To implement the eye-head-arm coordination strategy for both the direct and inverse visuo-motor transformations, sensory-sensory and sensory-motor mappings were performed in five steps.

- For the direct transformation, in the first step (see Fig. 3a) the retina-centred information of the visual target coupled with the current eye position was provided as input to the first processing stage to perform a sensory-sensory transformation in order to produce a head-centred representation. This head-centred representation was provided as input to the second processing stage along with the current head position to perform another sensory-sensory transformation in order to produce a body-centred representation. Then this body-centred representation was provided as input to the third processing stage to perform the mapping between the body-centred representation and the arm joint angles. The arm joint angles required to reach the target of interest were determined as a result of this mapping.
- In the second step (see Fig. 3b), retinal foveal activity and the body-centred representation were used as input to perform a sensory-motor transformation to determine the value of eye position relative to the target in space. Using this eye position value, the eye performed a saccade to look at the visual target.
- In the third step (see Fig. 3c), the retinal foveal activity, the determined eye position relative to the target in space and the body-centred representation were used as inputs to perform another sensory-motor transformation to approximate the head position relative to the target in space. With this motor command the head was moved. The arm motor command determined in the first

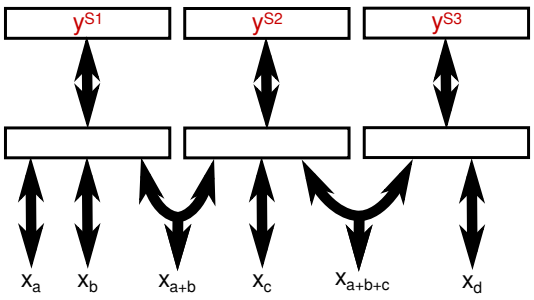


Figure 2.: A hierarchical architecture, consisting of three interconnected PC/BC-DIM networks, for mapping between four variables. The network is shown in a simplified format in which the error and reconstruction neurons are superimposed and double-headed arrows are used for inputs and outputs to and from these populations. This network can be used for 1-D coordinated gaze shifts and arm reaching movements. For the direct visuo-motor transformation, the network calculates $\mathbf{x}_d$ (*i.e.*, 1-D arm joint angles) and $\mathbf{x}_{a+b+c}$ (*i.e.*, body-centred representation) given $\mathbf{x}_a$ (*i.e.*, 1-D retina-centred representation), $\mathbf{x}_b$ (*i.e.*, 1-D eye position) and $\mathbf{x}_c$ (*i.e.*, 1-D head position). The first PC/BC-DIM network calculates an intermediate result ($\mathbf{x}_{a+b}$) in the third partition of it's reconstruction neurons: a head-centred representation. This intermediate result provides an input to the second PC/BC-DIM network. The second network's reconstruction of this intermediate representation is fed-back as input to the first PC/BC-DIM network. The second PC/BC-DIM network calculates an intermediate result ($\mathbf{x}_{a+b+c}$) in the third partition of it's reconstruction neurons: a body-centred representation. This intermediate result provides an input to the third PC/BC-DIM network. The third network's reconstruction of this intermediate representation is fed-back as input to the second PC/BC-DIM network. The third network calculates the arm position based on the correspondence between the body-centred representation and the arm position, since each arm position in body-centred space represents one body-centred location. For the inverse visuo-motor transformation, the network determines the $\mathbf{x}_{a+b+c}$ (*i.e.*, body-centred representation) and hence $\mathbf{x}_a$, $\mathbf{x}_b$ and $\mathbf{x}_c$ given $\mathbf{x}_d$ (*i.e.*, 1-D arm joint angles) as an input the third PC/BC-DIM network. The behaviour of the network is shown in Fig. 3 and 4.

step was executed to reach the target of interest.

- However at the end of head movement the eye position relative to target could be incorrect. The fourth and fifth steps were used to correct the eye gaze using similar approach as reported in [26]. In the fourth step (see Fig. 3d) for the correct eye position approximation, a sensory-sensory transformation was performed to update the head-centred representation using updated retinal activity after gaze shift and the current eye position.
- In the fifth step (see Fig. 3e), a sensory-motor transformation was performed with retinal foveal activity, updated head-centred representation and the body-centred representation as input to determine the correct eye position in head.

The third processing stage performs the mapping between the body-centred representation and the arm joint angles and has no role in executing coordinated eyes and head movements. Therefore, it is logical to disconnect the third processing stage from the second stage during the second and the third steps. As a result the third processing stage did not take part further in the transformations described in all other steps except the step one. To perform the inverse visuo-motor transformation, the efferent copy of the arm joint angles was used as an input to the third processing stage to perform a sensory-sensory transformation in order to determine the body-centred representation of the hand (see Fig. 4a). Then steps two to five, as described for the direct visuo-motor transformation with the disconnected third processing stage, were performed to shift the gaze and to view the hand in visual space (see Fig. 4b and Fig. 4c, however the gaze correction steps will be similar as shown in Fig. 3 hence are not shown here).

The 3-D eyes-head-arm coordination network shown in Fig. 5 uses a five processing stage PC/BC-DIM neural hierarchy to learn body-centred representations of visual space and the correspondence between body-centred locations and arm joint angles. The proposed network is shown in the simplified format in which the error and the reconstruction neuron populations are shown as a single population

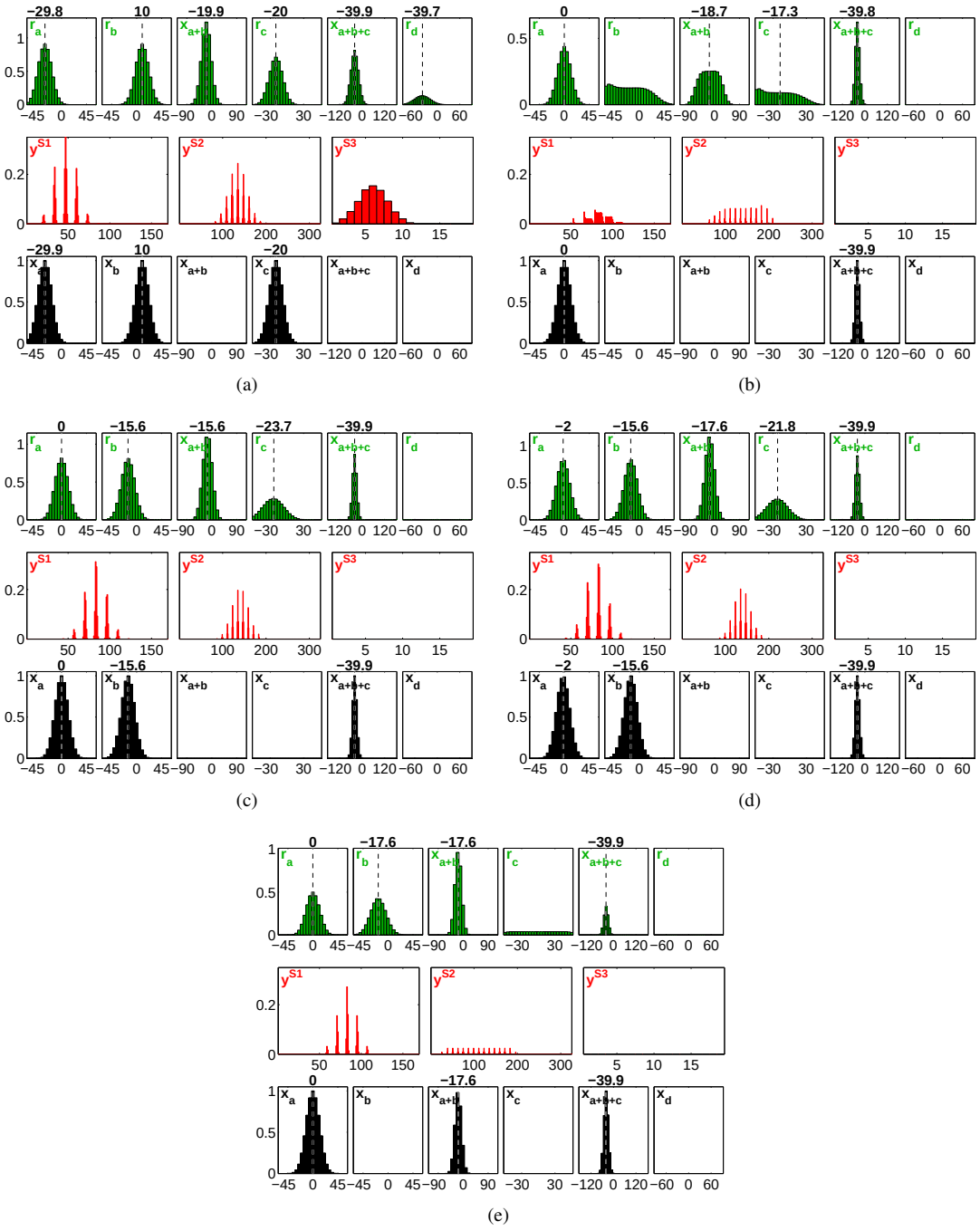


Figure 3.: The 1-D hierarchical PC/BC-DIM network shown in Fig. 2 performs the eye-head-arm coordination strategy for the direct visuo-motor transformation. The black histograms in each sub-plot show the input provided to the network whereas the red histograms show the prediction neuron activations and the green histograms show the response of the reconstruction neurons. (a) The population coded input was provided at $\mathbf{x}_a$ (*i.e.*, 1-D retina-centred input), $\mathbf{x}_b$ (*i.e.*, 1-D eye position) and $\mathbf{x}_c$ (*i.e.*, 1-D head position) to approximate $\mathbf{x}_{a+b}$ (*i.e.*, 1-D head-centred representation) in the first stage, $\mathbf{x}_{a+b+c}$ (*i.e.*, 1-D body-centred representation) in the second stage and $\mathbf{x}_d$ (*i.e.*, 1-D arm position) in the third stage as shown in the upper histograms. The arm joint angle (*i.e.*, $\mathbf{x}_d$) required to reach the target was determined in this step. (b) Using retinal foveal activity $\mathbf{x}_a$ (*i.e.*, a peak centered at zero) and known body-centred representation $\mathbf{x}_{a+b+c}$, the eye position $\mathbf{x}_b$ was computed. (c) The retina foveal activity $\mathbf{x}_a$, eye position $\mathbf{x}_b$ computed in the previous step and the body-centred representation $\mathbf{x}_{a+b+c}$ were provided as input to compute the head position $\mathbf{x}_c$. Using the eye position $\mathbf{x}_b$ and head position $\mathbf{x}_c$ gaze was shifted. The arm motor command (*i.e.*, $\mathbf{x}_d$) determined in the first step was executed to reach the target. (d) Using the current updated retinal activity $\mathbf{x}_a$ and the current eye position $\mathbf{x}_b$, a new head-centred representation $\mathbf{x}_{a+b}$ was computed as shown in mapping. (e) Then using this head-centred representation $\mathbf{x}_{a+b}$ and retina foveal activity $\mathbf{x}_a$ as input, the correct eye position in head $\mathbf{x}_b$ was produced by the network.

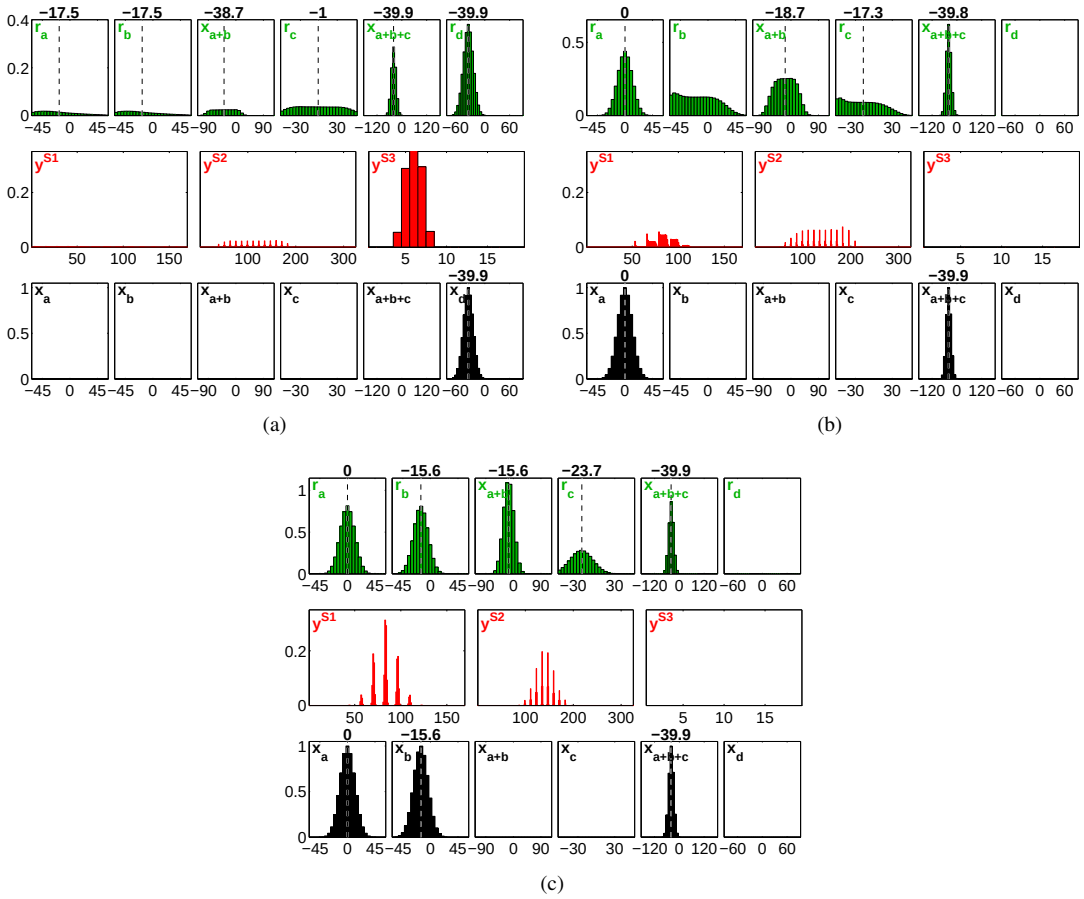


Figure 4.: The 1-D hierarchical PC/BC-DIM network shown in Fig. 2 performs the eye-head-arm coordination strategy for the inverse visuo-motor transformation. The black histograms in each sub-plot show the input provided to the network whereas the red histograms show the prediction neuron activations and the green histograms show the response of the reconstruction neurons. (a) The population coded input was provided at $\mathbf{x}_d$ (*i.e.*, 1-D current arm position) in the third stage to approximate the $\mathbf{x}_{a+b+c}$ (*i.e.*, 1-D body-centred representation). (b) Using retina foveal activity $\mathbf{x}_a$ (*i.e.*, a peak centered at zero) and known body-centred representation $\mathbf{x}_{a+b+c}$, the eye position $\mathbf{x}_b$ was computed. (c) The retina foveal activity $\mathbf{x}_a$, eye position $\mathbf{x}_b$ computed in previous step and body-centred representation $\mathbf{x}_{a+b+c}$ were provided as input to compute head position $\mathbf{x}_c$. Using the eye position $\mathbf{x}_b$ and head position $\mathbf{x}_c$ gaze was shifted to view the hand. Similar eye-head gaze correction steps were performed as shown in Fig 3.

and the inputs and outputs to these populations are also combined together (similar to the network shown in Fig. 2). The mathematical model remains unchanged. The proposed eyes-head-arm coordination network contains a PC/BC-DIM processing stage (shown on the left of Fig. 5) that performs mappings between the position of a visual target on the left retina, the position of the left eye in the skull (the left eye pan and tilt), and the head-centred location of the visual target relative to the left-eye. An identical PC/BC-DIM processing stage, shown in the second position of Fig. 5, performs the same transformations for the right eye. A third PC/BC-DIM processing stage translates between the individual head-centred representation centred on the left and the right eyes, and a global head-centred representation of visual space, that can be driven by targets viewed by either or both eyes. The fourth processing stage in Fig. 5 uses the global head-centred representation and an efferent copy of head position (*i.e.*, the head pan, tilt and swing) as input to produce a body-centred representation of visual targets. The fifth and last processing stage in Fig. 5 performs the mapping between the body-centred representation of the visual target and the arm joint angles to reach the target at that body-centred location. The same

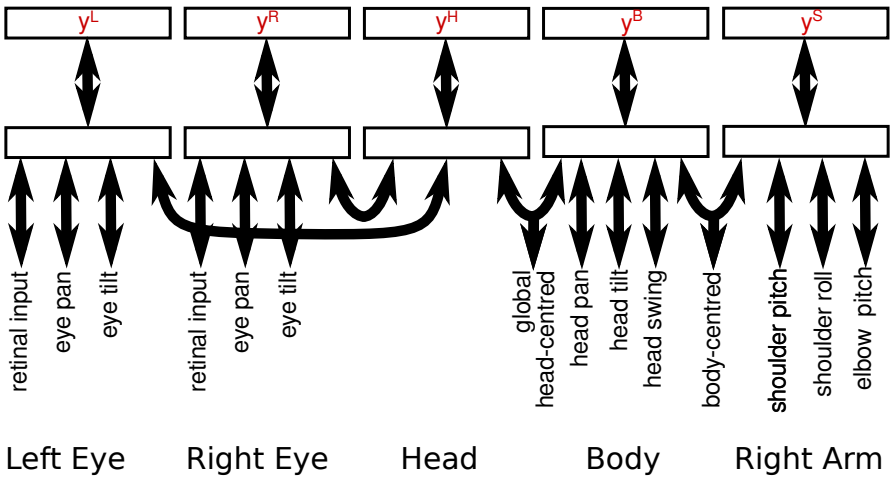


Figure 5.: The hierarchical PC/BC-DIM network for 3-D eyes-head-arm coordination drawn using the simplified format.

eye-head-arm coordination strategy was used in the 3-D PC/BC-DIM eyes-head-arm coordination network as described for the 1-D case in Fig. 3 for the direct and in Fig. 4 for the inverse visuo-motor transformations, however now sensory-sensory and sensory-motor mappings were performed with 2-D retinal activities and the efferent copy of pan and tilt from both eyes, pan, tilt and swing for the head and the right arm joint angles. If a corrective saccade was required then the correction was made using steps four and five of the eyes-head coordination strategy reported in [26].

The retinal input (*i.e.*, $\mathbf{x}_a$) to both the first and the second processing stages was encoded using a 2-dimensional uniform array of neurons with Gaussian RFs as used in [25, 26]. For a given visual target, the responses of each retinal neuron was proportional to the overlap of the visual target with its receptive field. These responses were concatenated into a vector to provide the input to the PC/BC-DIM network.

For the purpose of the simulations reported in section 3 the retinotopic input to the model, the input encoded by the retinal neurons described above, are images captured from the iCub cameras. However, the environment in which the iCub is placed is very impoverished consisting of one or two highly salient objects in front of a blank background. In more realistic environments, it would be necessary to process the raw images to derive a retinotopically organised representation to act as the input to the model. This retinotopic input would encode the locations of targets for possible saccades. It is assumed that this could be achieved by processing the images to form a saliency map [30].

The eye position signals (*i.e.*, the pan and tilt of both eyes), the head position signals (*i.e.*, head pan, tilt and torsion/swing) and the right arm position signals (*i.e.*, shoulder pitch, shoulder roll and elbow pitch) were each encoded using a 1-dimensional array of neurons with Gaussian RFs that were uniformly distributed between the maximum and minimum values. Decoding these values was performed using standard population vector decoding to find the mean of the distribution of responses [31].

2.3. *Training the Eyes-Head-Arm Coordination Control PC/BC-DIM Network*

The 1-D eye-head-arm coordination network used above to illustrate how the proposed network can perform simple mappings (*i.e.*, the results shown in Fig. 3 and Fig. 4) was hard-wired to perform the proposed eye-head-arm coordination strategy. The 3-D PC/BC-DIM eyes-head-arm coordination network involves more complex mappings, and hence requires some method of learning the appropriate connectivity. This can be achieved using an unsupervised approach for training the weights as used in our previous work [25, 26].

The first three processing stages in Fig. 5 were trained to learn the head-centred representation of visual targets as described in [25]. The fourth processing stage was trained to learn the body-centred representation of visual space as described in [26]. The fifth processing stage was used to learn the correspondence between the arm joint angles and the visual location of the hand in body-centred coor-

dinates. With the iCub body stationary, the iCub arm joint angles were randomly selected to position the hand at random positions in body-centred visual space. For each hand position, motor babbling was performed with both eyes pan, tilt and head pan, tilt and swing to search for the hand. The hand palm was made salient by giving it a distinct colour during training. Hence, once the hand palm came into view, retinal inputs were generated. Using these retinal activities and an efferent copy of the eye positions, the global head-centred representation was obtained. The global head-centred information combined with an efferent copy of head position was used to produce the body-centred representation of the robot hand position. Each unique correspondence between the body-centred representation of the hand position and the arm joint angles was represented by a different prediction/basis function neuron in the fifth processing stage. Therefore for one set of arm joint angles the hand will be at one unique body-centred position and one unique prediction/basis function neuron will show activity. Repeating this process for large number of random trails with different hand positions enabled the fifth processing stage of the eyes-head-arm coordination network to learn the correspondence between body-centred representations and arm joint angles for all reachable locations.

One issue with the above method of training is to decide how finely the arm joint angles are required to change in order to position the hand at unique body-centred location. Certainly, the hand should appear at all locations in body-centred space that robot needs to learn for correspondence. But theoretically there are infinite values of arm joint angles between full allowable range of each joint which will result in infinite hand positions in body-centred space and infinite number of basis function neurons. To address this issue the following procedure was used. For each set of arm joint angles the network initially does not learn the correspondence between the hand body-centred location and the arm joint angles but in fact the inverse visuo-motor transformation was performed as described in section 2.2 using the current joint angles as input. If with these eyes-head motor commands the hand was in view of both eyes producing binocular retinal activities then no learning was performed. If unsuccessful, then the hand would now be at new body-centred location and the network learnt the correspondence between the body-centred location and the hand position as mentioned above. Using this online learning and optimization procedure the network set basis function RFs sizes, peak locations and connection weights.

For each new correspondence a new prediction neuron was added to the network. This prediction neuron was given weights corresponding to the inputs received by the fifth processing stage. The learning rule followed to update the network connection weights $\mathbf{W}$ and $\mathbf{V}$ is:

$$\mathbf{W}_i \leftarrow \mathbf{W}_i + \tilde{\mathbf{x}} \tag{4}$$

$$\mathbf{V}_j \leftarrow \mathbf{V}_j + \hat{\mathbf{x}} \tag{5}$$

Where i represents the index of row vector or basis vector in weights $\mathbf{W}$ whereas j represents the index of column vector in weights $\mathbf{V}$. The $\tilde{\mathbf{x}}$ is copy of input vector $\mathbf{x}$ normalized to have sum value equal to one, whereas $\hat{\mathbf{x}}$ is copy of $\mathbf{x}$ normalized with maximum value in $\mathbf{x}$. In particular, the feedforward weight matrix $\mathbf{W}$ was normalized with sum of input vector $\mathbf{x}$. The feedback synaptic weight matrix $\mathbf{V}$ is transposed copy of the feedforward weight matrix $\mathbf{W}$ and then normalised such that each column has a maximum value of one. During all training process the arm joint angles were given 100,000 random values but correspondence between the hand position and body-centred locations was learnt for only 19,614 locations.

3. Results

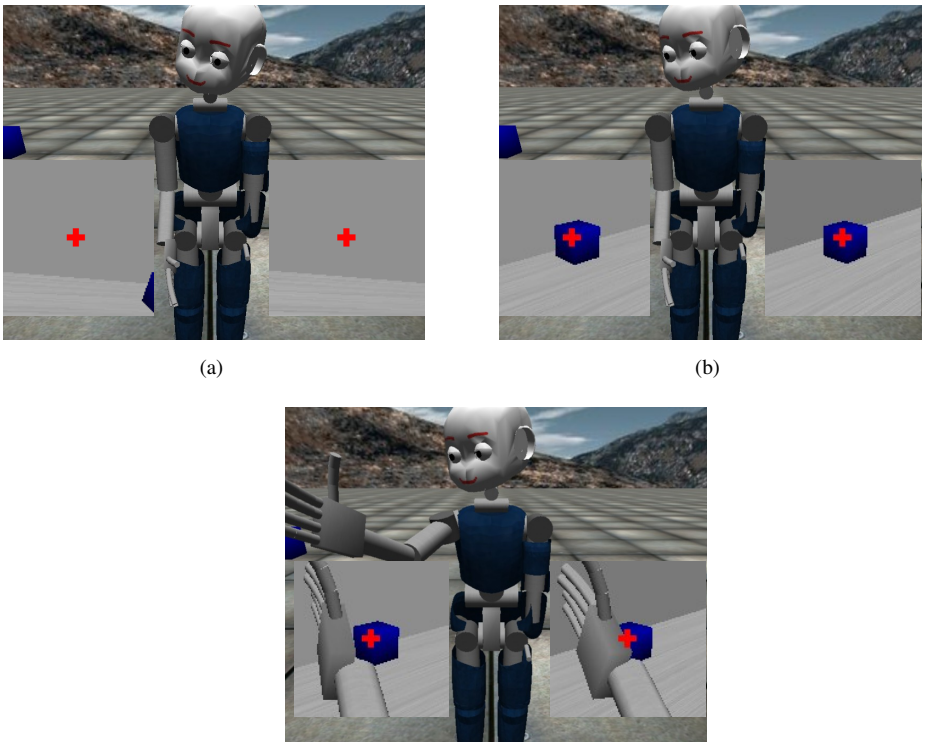
The performance of the proposed 3-D eyes-head-arm coordination network was examined using the iCub humanoid robot simulation platform [32, 33] with stationary body and free head and right arm. Visual targets of box shape were created without gravity effect and with a width, height and length of 0.038 in iCub Simulation World Units (SWUs). Targets could be placed at depth between 0.1 and 0.3 SWUs. The retina of both eyes were populated with uniform RFs distribution and the standard deviation

of each RF was $\sigma = 7$ pixels, the peak spacing between RF centres was 14 pixels, and in total 81 RFs were used to uniformly tile the input image as used in [25, 26]. The size of each iCub retinal image was 128x128 pixels, which corresponds to 25.6x26.4 degrees of visual angle. The right arm of the iCub was used, employing only three degrees of freedom *i.e.*, shoulder pitch, shoulder roll and elbow pitch. The range of arm shoulder pitch was -90° to -30°, shoulder roll was +15° to +90° and elbow pitch ranged from +20° to +100° and were varied in steps of 1° during training. The head pan signal ranged from -40° to +40°, tilt from -30° to +30° and head swing had range of -20° to +20° as in [26]. Eye pan had a range of -20° to +20° and tilt ranged from -12° to +12°. The eyes, head and arm position signals were encoded with 1-dimensional Gaussian RFs evenly spaced every 4° and with $\sigma = 2°$ as in our previous work [25, 26].

All experiments reported below were performed by following the eyes-head-arm coordination strategy sequentially described in section 2.2. The proposed eyes-head-arm coordination network is not only capable of shifting gaze to the target of interest but also performs convergent eyes movements to focus on the target as we have shown for saccade and vergence control and eyes-head coordination in previous work [25, 26]. Moreover, the proposed network also has the ability to perform memory-based gaze shift and arm reach to different visual targets positioned at different body-centred locations.

3.1. *Direct Visuo-motor Transformation*

The performance of the network was measured in terms of gaze shift and arm reach accuracy after each gaze shift and arm movement to the target of interest. To quantitatively measure the gaze shift and the arm accuracy with the iCub simulator, the robot's eyes and head were placed at a random pose whereas the right arm was placed at its home location (*i.e.*, should pitch, roll and elbow pitch at $0°$), and then a visual target was generated at a random location and depth but so that it was visible to at least one eye as shown in Fig. 6a. The arm joint angles could also be set to random values but this created a chance of the hand position starting at the visual target position, for this reason the arm started at the home position. The eyes-head-arm coordination strategy for the direct visuo-motor transformation was adopted as mentioned in section 2.2. The visual input corresponding to the target, together with the efferent copy of eyes pan/tilt and head pan/tilt/swing positions were used to determine the body-centred representation and corresponding arm joint angles to reach the target. This body-centred representation and the binocular retina foveal activities were used to compute eyes positions required to foveate the target. Using the retina foveal activities, the calculated eyes position and the computed body-centred representation, the desired head position was also computed. These determined eyes and head position motor commands were used to shift the gaze. After the shift of gaze the robot arm motor command was executed to reach the target of interest. Fig. 6 shows an example simulation of the 3-D eyes-head-arm coordination network for coordinated gaze shift and arm reaching movement using the direct visuo-motor transformation with the iCub robot. The post-gaze distance was measured between the foveal locations and the position of target in the binocular retinal images. The accuracy of arm reach was measured in the iCub simulation environment by calculating the distance between the target center coordinates and the hand positions in world coordinates. The mean and standard deviation of post-gaze error and arm reach error was calculated for 100 trials and the results are shown in Fig. 7. The mean value of post-gaze distance was 1.92° and SD was 0.87° which compares to an accuracy for large gaze shifts in primates of $< 3°$ [34]. The mean value of arm reach error was 0.12 and SD was 0.05 SWUs. There were two reasons behind the arm reach error. First is that the centroids of the target and the hand can not coincide due to the physical extent of the target object and the hand. Secondly the trajectory of the arm reaching movement (which depends on the target location and the initial arm position) can bring the fingers or the thumb into contact with the target before the hand palm, after which the arm reaching movement was stopped with the hand touching the target boundary. Therefore there was always a difference between the target coordinates and the hand position as the hand palm was 0.022 thick, 0.069 long and 0.065 wide in SWUs.



(a) (b)

(c)

Figure 6.: Example simulation of gaze shift and reaching to a target of interest with the right arm using the direct visuo-motor transformation. The two windows to the left and right of the iCub show the views of both eyes. The box within these windows is the visual target and the cross hairs mark the location of the fovea in middle of each retina (the cross hairs were not visible to the robot). (a) The initial eyes, head and right arm position before gaze shift. (b) After gaze shift to the target. (c) After the right arm moved to reach the target.

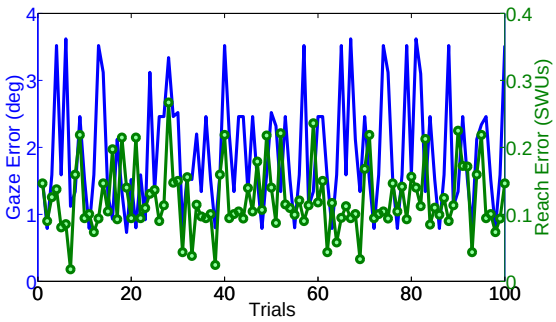
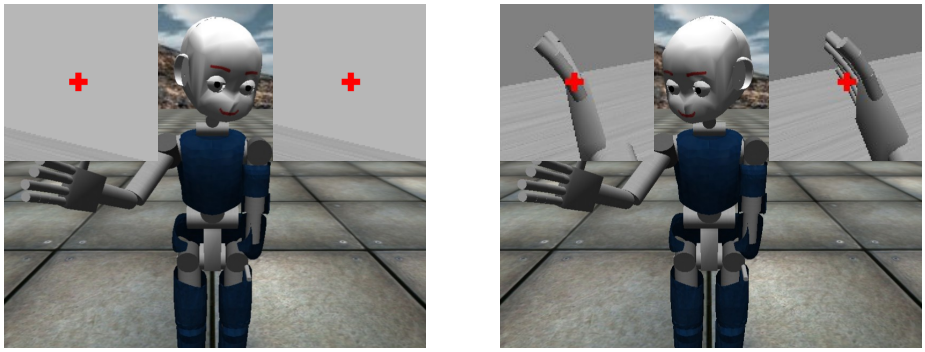


Figure 7.: Gaze accuracy in terms of post-gaze shift error and arm reach accuracy for the trained 3-D PC/BC-DIM eyes-head-arm coordination network. The arm reach error was measured in terms of iCub simulator world coordinate units (SWUs) by calculating the distance between the palm position and the visual target location.

3.2. *Inverse Visuo-motor Transformation*

To test the 3-D eyes-head-arm coordination network for the inverse visuo-motor transformation, the robot right arm, eyes and head were positioned at a random pose. The eyes-head-arm coordination strategy for the inverse visuo-motor transformation was adopted as mentioned in section 2.2. The efferent copy of the right arm joint angles was used to determine the body-centred representation of the right



(a) (b)

Figure 8.: Example simulation of the inverse visuo-motor transformation. (a) The initial eyes, head and the right arm position before gaze shift. (b) Gaze shift to view right hand.

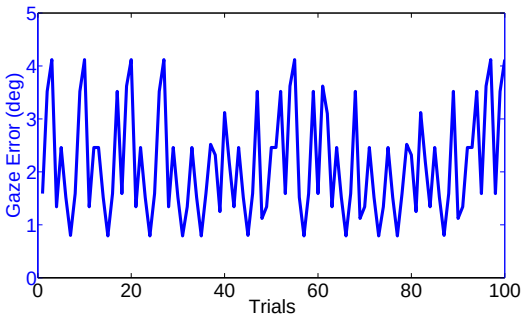


Figure 9.: Gaze accuracy in terms of post-gaze shift error for the trained 3-D PC/BC-DIM eyes-head-arm coordination network. The post-gaze error was measured after performing inverse visuo-motor transformation to view the hand and the error shows the position of hand relative to the foveae.

hand in body-centred space. This body-centred representation along with the retina foveal activities were used to compute the required eyes and head movements. The determined eyes and head motor commands were used to shift the gaze to view the right hand. These simulation results are shown in Fig. 8. The accuracy of the gaze shift to view the hand was determined in a similar way as described for the direct visuo-motor transformation for 100 trails and gave similar results with mean post-gaze shift distance of 2.14° and SD of 1.02° as shown in Fig. 9.

3.3. *Memory-based gaze shift and arm reach*

The proposed network can perform memory-based gaze shifts and arm reaching movements and these movements can be performed with different targets. To test this, two targets were presented in visual space. The robot eyes and head were positioned in a random pose and the right arm at its home location (due to the reason described in section 3.1), two visual targets were generated at random locations but such that both were visible. The sensory-sensory transformation was performed using the visual input along with an efferent copy of the eyes and head position to determine the body-centred representation of the visual targets. The body-centred representation of the visual targets is shown in Fig. 10b in the form of a body-centred map with the neural activities of the prediction neurons in the fourth processing stage. Using one body-centred representation and steps two and three of the eyes-head-arm coordination strategy (as described in 2.2) the gaze was shifted to the first target of interest while the second body-centred representation was stored in memory. Using the stored body-centred representation the mapping between the body-centred representation and the arm joint angles was performed (similar as described in the first step of the eyes-head-arm coordination strategy for the direct visuo-motor transformation

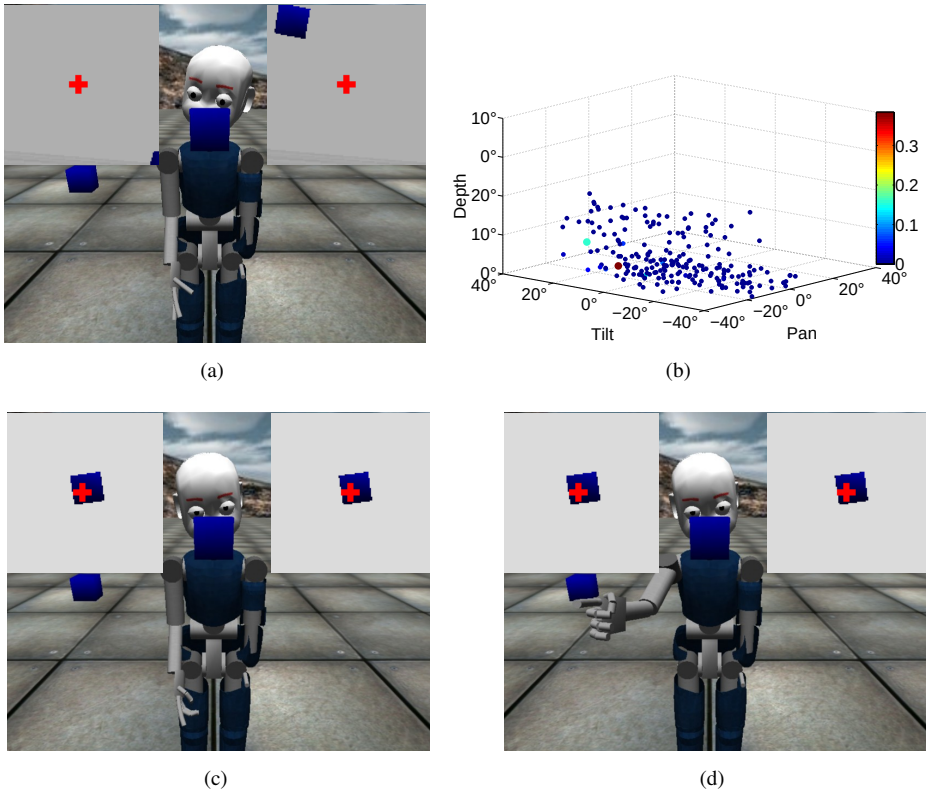


Figure 10.: Example simulation of a gaze shift to one visual target and a memory-based reach to the second visual target. (a) The initial eyes, head and right arm position before gaze shift. (b) The body-centred representation of visual targets in body-centred map showing neural activities of reconstruction neurons with given colour map scale. (c) Gaze shift to the visual target to bring the target onto binocular foveae. (d) The right arm moved to reach the second target of interest using memorized body-centred representation of the target.

in 2.2) and the arm joint angles generated at the fifth processing stage were read out. These arm joint angles were used to reach the second target of interest with the right arm. An example simulation is shown in Fig. 10. The post-gaze shift distance and the arm reach error were measured for 100 trails after the memory-based gaze shifts and the arm reaching movements to two different targets. The mean post-gaze distance being 2.13° and the SD being 0.40°, whereas the mean arm reach error being 0.12 and the SD was 0.05 SWUs similar to those results reported in section 3.1.

4. Discussion

In this article we used an omni-directional basis function type neural network for coordinated eyes-head-arm movements. The proposed basis function network is based on the PC/BC-DIM computational substrate employing multiple instances of the PC/BC-DIM processing stage. The proposed network is hierarchically structured with independent eyes, head and arm control circuits capable of performing the direct and the inverse visuo-motor transformations. We showed that the visual sensory information imaged in retinotopic space together with the eyes position can be transformed to head-centred representation. The determined head-centred representation can be combined with the head position signal for transformation to a body-centred representation. Then we showed that the determined body-centred representation can be used to perform the mapping between the body-centred representation and the arm joints angles. The arm joint angles determined as a result of this transformation were executed to perform the arm reaching task. The same body-centred representation was also used to determine the eyes and head movements for a coordinated gaze shift to the targets. This shows that the sensory-motor trans-

formation can be performed from the visual sensory information to arm motor space in various stages as shown in literature for biological systems [1–3]. We described the eyes-head-arm coordination strategy and also showed in (section 2.2) how bi-directional sensory-motor transformations can be achieved. Using visual information as the driving signal the network performed the direct visuo-motor transformation, whereas the same network performed the inverse visuo-motor transformation when driven by the efferent copy of arm joint angles. The proposed eyes-head coordination network performed accurate large gaze shifts to targets of interest and convergent eyes movements to fixate on the targets with biological comparable accuracy similar as in [25–27]. The arm always accurately reached the target of interest. We also showed that for the arm reaching movement the hand visibility is not required. The arm reach task was also achieved without involving the proprioceptive signals of current arm joint angles. The proposed eyes-head-arm coordination network has the following novel and distinct features compared to the previously reported work.

4.1. *Network Architecture*

The PC/BC-DIM basis function network has key architectural differences compared to all previously proposed basis function networks. Specifically, in previously models the basis function layer neurons used radial, typically Gaussian, activation functions [4, 9, 13, 14, 17, 18, 20, 23, 24], and the parameters of these Gaussian activation functions (the centre and spread of each basis function neuron RF) were defined through some heuristic or optimization procedure. However, the PC/BC-DIM basis function network does not set the response profile of the basis function neurons through a pre-defined Gaussian activation function. Instead, the RF of each prediction neuron is defined by the weights it receives and the non-linear interactions with other prediction neurons.

4.2. *Learning and Optimization*

The learning process of the PC/BC-DIM basis function network was made fast and simple due to network connection weights being defined as rescaled copies of the inputs. Furthermore, to optimize the number of basis function neurons involved in the approximation of any non-linear transformation an online optimization step was performed. This optimization step was not an extra algorithm, in fact it was just a sensory-motor transformation to determine motor commands using the network *i.e.*, shifted eyes-head gaze to view the hand. If this action was unsuccessful then a basis function neuron was added. Therefore, with the PC/BC-DIM basis function network both learning and optimization processes were performed in one step instead of two separate learning phases as reported in previously published basis function models [4, 9, 13, 14, 17, 18, 20, 23, 24].

4.3. *Omni-directional Transformation*

The same proposed PC/BC-DIM basis function network was used to perform bi-directional transformation without adding new connections for the inverse transformation as in [4, 9] or without using separate basis function networks for each direction in the bi-directional transformation as in [13, 14, 17].

4.4. *PC/BC-DIM Spatial Auto-encoder*

The proposed eyes-head-arm coordination network is a deep neural network architecture, employing PC/BC-DIM processing stages as the computational substrate. Each PC/BC-DIM stage functions like a spatial auto-encoder which acted as an encoder for the transformation in one direction whereas the same PC/BC-DIM stage acted as a decoder for the transformation in the opposite direction. The PC/BC-DIM auto-encoder thus has a profound difference compared to all previously reported work for similar eyes-hand coordination tasks where the auto-encoder employed separate encoder and decoder neural circuitry [36–39].

4.5. *Multiple Functions*

One additional advantage of the proposed model is that the network connections and weights were set for bi-direction visuo-motor transformations, but the same network was also capable to perform memory-based gaze shift and arm reaching movements. This provided added ability in the usage of the PC/BC-DIM basis function network for sensory-motor transformations.

4.6. *Scalability*

A very common problem with basis function type networks is that the network size increases exponentially with the number of input variables. In existing work on visuo-motor transformations this problem has been avoided by reducing the number of inputs involved in the sensory-motor transformations. In the work of [13, 14, 17, 18, 23] the head movement variables were discarded by restraining the head movements. Furthermore in [18] the visuo-motor transformation was performed in a 2-D workspace whereas in [14] the transformation was achieved in 1-D space which also reduced the number input variables. However, the proposed network solved the scalability problem by dividing the whole problem into five subtasks involving subsets of all input variables. We showed that the visuo-motor task can be achieved with such a decomposition and we showed in [25, 27] that the network size scales linearly after such decomposition.

4.7. *Multiple Stimuli*

Previous work did not consider at all multiple visual targets, which is a common situation in everyday life. The proposed network showed successful memory-based visuo-motor transformations for gaze shift and arm reaching movement to different targets of interest. The proposed model is a comprehensive model of eyes, head and arm movement control for gaze shift and arm reaching.

Acknowledgements

This work was partially funded by Higher Education Commission Pakistan under grant No. PM(HRDI-UESTPs)/UK/HEC/2012.

References

[1] Buneo C, Jarvis M, Batista A, Andersen R. Direct visuomotor transformations for reaching. Nature. 2002;416(6881):632-636.
[2] Carrozzo M, McIntyre J, Zago M, Lacquaniti F. Viewer-centered and body-centered frames of reference in direct visuomotor transformations. Experimental Brain Research. 1999;129(2):201-210.
[3] Crawford J. Spatial transformations for eye-hand Coordination. Journal of Neurophysiology. 2004;92(1):10-19.
[4] Pouget A, Deneve S, Duhamel J. A computational perspective on the neural basis of multisensory spatial representations. Nature Reviews Neuroscience. 2002;3(9):741-747.
[5] Deneve, Sophie and Latham, Peter E and Pouget, Alexandre. Efficient computation and cue integration with noisy population codes. Nature Neuroscience. 2001;4(8):826-831.
[6] Deneve S, Pouget A. Basis functions for object-centered representations. Neuron. 2003;37(2):347-359.
[7] Pouget A, Sejnowski T. Spatial transformations in the parietal cortex using basis functions. Journal of Cognitive Neuroscience. 1997;9(2):222-237.
[8] Pouget A, Sejnowski T. A neural model of the cortical representation of egocentric distance. Cerebral Cortex. 1994;4(3):314-329.
[9] Pouget A, Snyder L. Computational approaches to sensorimotor transformations. Nature Neuroscience. 2000; 3 :1192-1198.
[10] Salinas E Abbott L. Transfer of coded information from sensory to motor networks. The Journal of Neuroscience. 1995; 15(10) :6461-6474.

[11] Salinas E, Sejnowski T. Gain modulation in the central nervous system: where behavior. Neurophysiology, and Computation Meet. The Neuroscientist. 2001;7(5):430-440.
[12] van Rossum M, Renart A. Computation with populations codes in layered networks of integrate-and-fire neurons. Neurocomputing. 2004;58-60:265-270.
[13] Antonelli M, Grzyb B, Castelló V et al. Plastic representation of the reachable space for a humanoid robot. From Animals to Animats 12. 2012;167-176.
[14] Chinellato E, Antonelli M, Grzyb B, del Pobil A. Implicit sensorimotor mapping of the peripersonal space by gazing and reaching. IEEE Transactions on Autonomous Mental Development. 2011;3(1):43-53.
[15] Molina-Vilaplana J, Pedreño-Molina J, López-Coronado J. Hyper RBF model for accurate reaching in redundant robotic systems. Neurocomputing. 2004;61:495-501.
[16] Kim D, Huh S, Seo S et al. Self-organizing radial basis function network modeling for robot manipulator. International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems. 2005;579–587.
[17] Marjanovic M, Scassellati B, Williamson M. Self-taught visually-guided pointing for a humanoid robot. From Animals to Animats: Proceedings of. 1996;35-44.
[18] Meng Q, Lee M. Automated cross-modal mapping in robotic eye/hand systems using plastic radial basis function networks. Connection Science. 2007;19(1):25-52.
[19] Meng Q, Lee M. Error-driven active learning in growing radial basis function networks for early robot learning. Neurocomputing. 2008;71(7-9):1449-1461.
[20] Chao F, Wang Z, Shang C et al. A developmental approach to robotic pointing via human–robot interaction. Information sciences. 2014;283:288–303.
[21] Zhang P, Lü T, Song L. RBF networks-based inverse kinematics of 6R manipulator. The International Journal of Advanced Manufacturing Technology. 2004;26(1-2):144-147.
[22] Cornelius Webera, Mark Elshawb, Jochen Triescha, Stefan Wermterb. Neural control of actions involving different coordinate systems. Humanoid Robots: Human-like Machines. 2007;.
[23] Sun G, Scassellati B. A fast and efficient model for learning to reach. International Journal of Humanoid Robotics. 2005;02(04):391-413.
[24] Chao F, Zhang X, Lin H, et al. Learning robotic hand-eye coordination through a developmental constraint driven approach. International Journal of Automation and Computing. 2013;10(5):414-424.
[25] Muhammad W, Spratling M. A neural model of binocular saccade planning and vergence control. Adaptive Behavior. 2015;23(5):265-282.
[26] Muhammad W, Spratling M. A neural model of coordinated head and eye movement control. Journal of Intelligent & Robotic Systems. 2016;1-20.
[27] Muhammad W. Omni-directional basis function network for sensory-sensory and sensory-motor transformations. Ph.D. dissertation. 2016.
[28] Spratling M. Predictive coding as a model of cognition. Cogn Process. 2016;17(3):279-305.
[29] Spratling M. A neural implementation of bayesian inference based on predictive coding. Connection Science. 2016;28(4):346–383.
[30] Niebur E. Saliency map. Scholarpedia. 2007;2(8):2675.
[31] Georgopoulos AP, Schwartz AB, Kettner RE. Neuronal population coding of movement direction. Science. 1986;233(4771):1416-1419.
[32] Tikhanoff V, Cangelosi A, Fitzpatrick P, et al. An open-source simulator for cognitive robotics research: the prototype of the iCub humanoid robot simulator. In Proceedings of the 8th workshop on performance metrics for intelligent systems. 2008;19:57-61.
[33] Metta G, Sandini G, Vernon D, et al. The iCub humanoid robot: an open platform for research in embodied cognition. In Proceedings of the 8th workshop on performance metrics for intelligent systems. 2008;19:50-56.
[34] Tomlinson RD. Combined eye-head gaze shifts in the primate. III. Contributions to the accuracy of gaze saccades. Journal of Neurophysiology. 1990;64(6):1873-1891.
[35] Fanello S R, Pattacini U, Gori I, et al. 3d stereo estimation and fully automated learning of eye-hand coordination in humanoid robots. Humanoid Robots (Humanoids), 2014 14th IEEE-RAS International Conference on. 2014;1028–1035.
[36] Finn C, Tan X Y, Duan Y, et al. Deep spatial autoencoders for visuomotor learning. Robotics and Automation (ICRA), 2016 IEEE International Conference on. 2016;512–519.
[37] Pinto L, Gandhi D, Han Y, et al. The curious robot: learning visual representations via physical interactions. European Conference on Computer Vision. 2016;3–18.
[38] Ghadirzadeh A, Maki A, Kragic D, et al. Deep predictive policy training using reinforcement learning. arXiv preprint arXiv:1703.00727. 2017.

[39] Finn C, Levine S, Abbeel P. Guided cost learning: deep inverse optimal control via policy optimization. Proceedings of the 33rd International Conference on Machine Learning. 2016;48.