A Model Architecture for Mental Imagery

Alexander Kaiser, Wolfram Schenck, and Ralf Möller
Computer Engineering Group
Faculty of Technology
Bielefeld University

Abstract

Imagination or mental imagery is the capability of generating sensory experiences without actual sensory inflow. These sensory experiences might be the results of actions that are not carried out physically but only simulated within the cognitive system of an agent. We propose a model architecture for mental imagery that enables an artificial agent to generate views of parts of himself based on a set of given pose parameters and according to a specific viewing direction. The architecture is evaluated in a robotic study with an agent consisting of a robotic manipulator and a camera head. The agent adaptively learns to associate views of its gripper with its current pose. Furthermore, a visual forward model predicts visual changes according to the agent’s gaze direction.

Keywords: Mental imagery; visual prediction; cognitive modeling; robot study; neural networks

Introduction

Most simulation theories of cognition and perception are based on models involving the generation of internal sensory states as the effects of covert actions. These sensory states often relate to an internal simulation of interactions with the environment the effects of which are evaluated. Perception through anticipation (PtA) is an example of such a theory (Möller, 1999). In PtA, the agent perceives the world by an iterative simulation of actions and subsequent evaluation of their effects. The simulation starts with an initial sensory state (i.e. the actual current sensory state) and generates a set of new internal sensory states by covertly executing a set of certain actions. These internal sensory states are the basis for the next step of the simulation, and so on. The simulation is terminated at a certain point (e.g. after a specific number of iterations) and the final (internal) sensory states, along with the sequence of actions from which they result, are evaluated.

The results of an internal simulation can be used for perceptual tasks (like in the PtA approach) or action selection (Hesslow, 2002) (i.e. the action sequence leading to the most desirable final state is executed overtly). Thus, an artificial agent whose cognitive abilities are based on internal simulation must have a memory for holding covert motor and sensory states, a predictor for future sensory states based on the current state and a specific action (forward model), and an evaluator for action selection (Marques & Holland, 2009).

The internal sensory states which are generated during an internal simulation are termed mental images. Mental images, along with the process of generating them, is commonly termed mental imagery. Besides the obvious meaning of mental imagery, i.e. the generation of internal visual sensations, the term may also relate to other areas, e.g. motor imagery (Jeannerod, 1995). Neuroimaging studies suggest that cortical areas which are involved in the processing of perceived stimuli are also active during mental imagery (Kosslyn et al., 1993). This also holds for motor imagery (Jeannerod, 1995).

The subject of this article is a model architecture for mental imagery in the visuomotor context. The model enables an artificial agent to generate views of himself (or views of parts of himself) based on a set of pose parameters under a specific gaze direction. For a shorter presentation of this model from a mainly computational perspective, see Kaiser, Schenck, and Möller (2010).

The robotic setup that we use for our study consists of a robotic arm with an attached two-finger gripper and a stereo camera head. The mental imagery is restricted to views of the gripper based on the arm posture (encoded by the current angles of the joints) and the current gaze direction. The camera images are processed with a radial foveal mapping (Schenck, 2008) that mimics the distribution of photo receptors on the human retina, i.e. the resolution is higher towards the image center and lower in the periphery (retinal images).

In the technically oriented literature, the process of generating images based on a set of parameters is commonly termed view synthesis (Jägersand, 1997). There exist two main classes of methods to approach this problem: in model-based approaches, an explicit (often geometrical) model of the scene is used to render new views, whereas model-free approaches are based on interpolation / extrapolation of reference views. The model-based approach has a big drawback: it is not adaptive and requires a geometrical model of the scene. The model-free-approach on the other hand is based on collecting a set of reference views (i.e. the training set) which is used to generate new views (by using a specific method for interpolation, typically combined with a method for dimensionality reduction); the quality of generalization (i.e. the ability to generate views that were not present in the training set) relies heavily on the selection of the reference views and the used method.

For our architecture, we combine a sub-space-based model-free technique for view synthesis (Jägersand, 1997) with an image warping technique (Schenck & Möller, 2007). Thus, our approach does neither rely on geometrical models nor on any calibration of the sensors (i.e. the camera head); it is rather completely adaptive. The view synthesis combines methods for dimensionality reduction in order to represent images as low-dimensional appearance vectors. The association between an appearance vector and the corresponding set of pose parameters is established by non-linear regression (i.e. a 3-layer feed-forward network). The resulting images
are then warped into the view as it would appear from a desired gaze direction. This visual prediction step is performed by a visual forward model (Schenck & Möller, 2007). Decoupling the visual association from the visual prediction seems reasonable since it reduces the complexity of the visual association task (i.e. by reducing the number of degrees of freedom).

**Model Architecture**

Figure 1 shows a schema of the overall architecture. The different modules are depicted as boxes; arrows indicate information flow either in the image or in the motor domain (i.e. joint angles or viewing direction), respectively. Images are represented by sketches of the robot gripper (lateral view): $I$ corresponds to the output of the visual association; the gripper appears as it would be fixated by the cameras. The kinesthetic values corresponding to this gaze direction, termed gripper viewing direction, are the output of the kinesthetic association.

The desired viewing direction can be regarded as another input of the model. Together with the gripper viewing direction, generated by the visual association, the desired viewing direction is used to calculate a saccade. In the following, a saccade is regarded as a mere change in viewing direction, symbolized by the arithmetic operation in figure 1.

The saccade is used to drive a visual prediction step which takes as input the initially associated view ($I$ in figure 1) and generates an image $\hat{I}$. The visual prediction is conducted by an internal model (forward model) of the agent’s oculomotor apparatus. The final output $\hat{I}$ is an image of the gripper corresponding to a given arm posture and seen from a given viewing direction.

The decomposition of the overall problem into these three sub-problems seems reasonable under computational aspects: the visual and kinesthetic association can be regarded as functional mappings and can thus be implemented as feedforward networks. Visual prediction is a more challenging task since image data is typically high-dimensional. For this reason, we employ a warping technique which is based on a mapping between pixel positions rather than on directly predicting the image’s pixel values (Schenck & Möller, 2007).

**Robotic Agent**

The robotic agent used throughout this study resembles roughly the upper torso of a human: it consists of a camera head with two cameras, each mounted on a pan-tilt unit (PTU), and a 6-degrees-of-freedom serial manipulator with an attached two-finger gripper. The cameras are mounted above the arm. The whole setup faces a table which serves in our experiments only as a background.

**Camera Head**

The camera head consists of two analog cameras which provide RGB images of $320 \times 240$ pixels. Each of the cameras is mounted on a separate pan-tilt unit (PTU) that allows for control over the camera’s gaze direction. Note that for our present study, only the image from the left camera is used. The overall gaze direction of the camera head (both PTUs) is encoded by a 4-dimensional vector, denoted by $v$. Instead of using the pan and tilt angles of both PTUs directly, we employ a sophisticated vergence model for gaze control (see Schenck (2008) for details).

**Manipulator**

The agent is furthermore equipped with a robot manipulator with six rotatory degrees of freedom. The joint angles are denoted by $\theta_1, \ldots, \theta_6$. The manipulator can be decoupled into arm (joints 1–3) and wrist (joints 4–6). Thus, its inverse kinematics can be calculated in closed form. Note that this computation is just a shortcut which is needed in order to collect a large sample of training patterns.

Attached to the manipulator is a gripper with two fingers. The longitudinal axis of the gripper is always kept parallel to the ground while it can take three different orientations about the vertical axis, i.e. $\{0^\circ, 15^\circ, 30^\circ\}$.\(^1\)

**Kinesthetic Association**

The kinesthetic associative model directs the agent’s gaze towards the center of its gripper by associating an arm posture (i.e. a set of joint angles) with a gaze direction of the camera head. The purpose of this model is two-fold: during the training of the visual associative model it is used to collect training images of the gripper for different arm postures. In the application phase, the corresponding arm posture is used to generate saccades which are then used to drive the visual prediction (visual forward model).

The associative model is implemented as a 3-layer feedforward neural network (Rumelhart, Hinton, & Williams, 1986) with 6 inputs (corresponding to the 6 joint angles), a hidden layer with 40 units and 4 outputs (corresponding to those values that encode the gaze direction). Linear activation functions were used for the output layer and sigmoid functions (tanh) for the hidden layer. The training data was collected by approaching points within a regular grid of end effector coordinates. The cameras were controlled by a saccade controller (Schenck, 2008) such that the gripper was fixated for every grid point.

We chose an adaptive solution to this problem, although we are aware that there exist simple engineering solutions which rely on computing the inverse kinematics of the camera head. However, we think that using an adaptive approach here is more appropriate for a biologically oriented model like ours.

**Saccade Control**

A saccade controller (Schenck, 2008) is a controller which directs the gaze towards a salient object. In terms of control theory, the reference corresponds to the image center, the system input are the pan / tilt angles, and the measured output is the centroid of the salient object in image coordinates. Thus,\(^1\)At an orientation of $0^\circ$, the gripper is pointing away from the cameras.
the measured error is the deviation between the image center and the object’s centroid.

We employed a P-type (linear) controller for this task. The controller equation is given by $\Delta v = Ge$, where $\Delta v$ denotes the change in gaze direction (i.e. the saccade), $G$ denotes a gain matrix (see Schenck (2008) for details), and $e$ is the error vector (i.e. the deviation between the centroid and the image center). In order to avoid oscillations, the controller tolerates errors of 1 pixel in each direction.

**Training**

For the training we defined a rectangular workspace of size $150\text{mm} \times 120\text{mm} \times 300\text{mm}$ that was sampled by a $7 \times 6 \times 15$ regular grid. Thus, the distance between adjacent grid positions is approximately $20\text{mm}$ in each direction. The whole grid was sampled for each gripper orientation $\alpha \in \{0^\circ, 15^\circ, 30^\circ\}$ separately. The six joint angles were calculated from the Cartesian coordinates using inverse kinematics (IK).\(^2\) From the 8 theoretically possible IK solutions of the given arm only those belonging to a specific family (i.e. elbow down) were selected. Furthermore, a collision detector was used in order to avoid collisions of the arm with itself or its environment. Thus, the actual number of approached grid points deviates from the theoretical number (630) for the different orientations: $0^\circ (563)$, $15^\circ (483)$, $30^\circ (355)$.

During the collection of the training patterns, the gripper held a salient target object (i.e. a stick with a reddish ball of small diameter attached) that was fixated using the saccade controller; this was repeated for every grid position. A training pattern is a pair $(\theta, v)$, where $\theta \in \mathbb{R}^6$ denotes the vector of joint angles and $v \in \mathbb{R}^4$ denotes the gaze direction. There are 1381 such examples in total. The three-layer network was trained off-line using resilient propagation (RProp) (Riedmiller & Braun, 1993). Furthermore, the total number of patterns was divided into a training set (70%), and disjoint test and validation sets (both 15%). If the error on the validation set did not decrease for 200 epochs, the training was terminated (early stopping). Early stopping of the training is a heuristic to impose a regularization onto the network weights to circumvent overfitting.

\(^2\)Note that using the regular grid and inverse kinematics for the collection of the training examples is just a technical shortcut.

**Visual Association**

The visual associative model takes the joint angles $(\theta)$ as input and returns the corresponding (fixated) view of the gripper, i.e. an image. Image data is usually high-dimensional (depending on the resolution of the images) which would require a large associative network. For this reason, we propose a model that has a two-stage architecture (see figure 2): the images are represented as low-dimensional appearance vectors which are then associated with the corresponding arm postures. Seeking a low dimensional representation for the gripper images seems reasonable, because it is most likely that they occupy a sub-space of relatively low dimensionality within the full image space.

The association between the joint angles and the appearance vectors is performed by a three-layer feed-forward network with linear output units and sigmoid hidden units. During the training phase, the images are transformed into appearance vectors which serve as target values for the network training. During the application phase, the images are reconstructed from the network output by applying the inverse transformation. Irrelevant information is discarded during this transformation, resulting in a small reconstruction error.

The two-fold structure of the visual associative model can be regarded as the composition of two functions $g \circ f$, where $f : \mathbb{R}^n \to \mathbb{R}^m$ and $g : \mathbb{R}^m \to \mathbb{R}^M$. Function $f(\cdot)$ maps the agent’s $n$ pose parameters onto an appearance vector of dimensionality $m$. Note, that there is no closed-form solution for this function; thus it must be modeled by a non-linear regression technique, e.g. a feed-forward neural network. The estimated appearance vector is transformed into an image of dimensionality $M$ by function $g(\cdot)$.
Where an image be denoted by \( i \in \mathbb{R}^M \), where \( M \) is the number of pixels, then its appearance vector is given by
\[
a = g^{-1}(i) = W^T(i - \bar{i}).
\]
Here \( W \in \mathbb{R}^{M \times m} \) is a projection matrix with orthonormal columns and with \( m \ll M \), and \( \bar{i} \) is the mean image. Note that if \( W \) is chosen such that the elements of \( a \) are mutually uncorrelated and the variance \( \text{Var}[a_i] \) is maximized, then the elements of \( a \) are the principal components of \( i \).

Let \( X = [i_1, \ldots, i_N] \) denote the data matrix of all \( N \) images, where \( i_k = i_k - \bar{i} \) denotes the \( k \)th mean-centered image. The projection can be calculated by performing an eigen-decomposition of the implicit covariance matrix \( XX^T = U \Lambda U^T \). Note that the implicit covariance matrix is of size \( N \times N \), i.e. its size scales quadratically with the number of images. Thus, for the usual case when \( N \ll M \), this matrix requires significantly less memory than the explicit covariance matrix. Computing \( \tilde{W} = X^T U \) and normalizing the columns of \( \tilde{W} \) yields the desired \( W \) (Turk & Pentland, 1991).

The inverse transformation which is used to reconstruct an image from its appearance vector is given by
\[
\hat{i} = g(a) = Wa + \bar{i} = \sum_{i=1}^{m} a_i w_i + \bar{i},
\]
where \( w_i \) denotes the \( i \)th column of \( W \) (i.e. the \( i \)th eigen-image). Thus, the reconstructed image is a linear combination of \( m \) eigen-images \( w_i \) with coefficients \( a_i \). Note that the reconstruction is only perfect for \( m = M \); in the case considered here, i.e. \( m \ll M \), the reconstruction is only an approximation. A measure for the deviation between an original image and its reconstruction is the reconstruction error \( E_{\text{reco}} = ||i - \hat{i}||^2 \).

### Image Processing and Training

The objective of the training process is to “learn” a mapping from posture space into appearance space. The training was conducted in analogous fashion to the kinesthetic association. This time, the gripper fingers were in an open position as it would be before grasping. For each grid position, the inverse kinematics is used to calculate the corresponding joint angles and the arm is moved into this position. The previously trained kinesthetic associative model then generates a gaze direction such that both cameras fixate the gripper. The PTUs are moved and the camera images are stored.³

The original \( 320 \times 240 \) images were cropped to a small \( 45 \times 35 \) region around the gripper (which is roughly located in the image center). The gripper was then separated from the background based on HSV color information: the S (satura-)
ation channel was thresholded and the result used as a mask. The RGB images were converted to grayscale. Afterwards these gray scale images were warped by a retinal mapping (Schenck & Möller, 2007; Schenck, 2008), expanding the central region. Thus, the resolution is higher in the central part of the image and lower in the periphery (fovea effect).

The retinal images were once again cropped to \( 85 \times 69 \) pixels around the center. These cropped images were used for computing the appearance vectors (using the eigen-image approach). In contrast to the \( M = 5865 \) image dimensions, only a number of \( m = 10 \) eigen-images (see figure 3) were used for reconstruction. The average reconstruction error, i.e. the deviation between the original images and their reconstructions, amounted to \( E_{\text{reco}} = 3.47 \).

The visual associative network has a similar structure as the kinesthetic associative model (i.e. a 3-layer topology):
\[
\hat{a}_i = f_i(\Theta) = \sum_{j=1}^{n_h} w_{ij} \varphi(\sum_{k=1}^{6} w_{jk} b_k + b_j^h) + b_i^h,
\]
³For the results presented in this paper, only the left camera is used.
where $\varphi(\cdot) = \tanh(\cdot)$ denotes the activation function, $w_{ji}^h, b_i^h$ denote the hidden weights/biases, and $w_{oi}^n, b_i^n$ denote the output weights/biases. We chose a relatively large hidden layer, consisting of $n_h = 100$ sigmoid units. A training pattern is a pair $(\theta, a)$, consisting of an arm posture $\theta \in \mathbb{R}^6$ and the appearance vector of the corresponding gripper view $a \in \mathbb{R}^{10}$; the total number of such patterns was $N = 1381$. For the training we used the RProp algorithm with early stopping to minimize the quadratic error $E = \frac{1}{2} \sum_{j=1}^{N} \|a_j - f(\theta_j)\|^2$.

Figure 4 shows the original views of the gripper (upper row) for 3 different grid position together with their PCA reconstructions and the reconstructions from the visual association (bottom row). The PCA reconstruction and the visual association do not differ noticeably. However, both reconstructions contain artifacts such as the white shadow underneath the right gripper finger (see rightmost column in figure 4).

Figure 4: Example images from three different grid positions. Original images (top row), PCA reconstruction (middle row), and output of the associative model (bottom row). Adapted from Kaiser et al. (2010).

### Visual Forward Model

The visual forward model (Schenck & Möller, 2007) is a predictor for the visual consequences of a saccade, i.e. a forward model of the agent’s oculomotor apparatus. It receives as input the retinal images computed from the current view of the cameras and a saccade (motor command). From these inputs, the visual forward model predicts an image according to a given change gaze direction, $\Delta \theta$ and the current view.

The visual forward model has a two-fold structure: the image warping is conducted by a mapping model (i.e. a predictor of pixel locations) and a so-called validator model which indicates if a pixel can be faithfully predicted. Each of these models receives as inputs the saccadic motor command along with each pixel position of the output image frame (i.e. the image frame corresponding to the image after the saccade). The mapping model predicts for each pixel position in the output frame a corresponding pixel position in the input image (inverse mapping); the pixel is set to this value if indicated as valid by the validator model.

Both models, i.e. mapping and validator model, are adaptively learned (Schenck & Möller, 2007). Thus, the visual forward model does neither require a priori information nor a calibrated camera system. Therefore it integrates nicely into the overall architecture.

### Results

We use the associated images of three arm postures (see figure 4, bottom row) to demonstrate the overall model in the following. The arm postures were selected such that each of the three angles ($0^\circ, 15^\circ, 30^\circ$) is represented once. Note that the angle space is sampled too coarse to allow for generalization, e.g. the model would not be able to generate an image corresponding to an angle of $22.5^\circ$.

The three different images were fed into the visual forward model (see A, B, and C in figure 5). For each image, we used the visual forward model to predict the result of 9 different saccades (i.e. roughly $10^\circ$ in each direction); the images in each center of figures 5 A, B, and C correspond to the output of the visual association before a saccade.

It can be seen from figure 5 that the resulting images are smooth and consistent, i.e. the images are not noticeably distorted. The borders in most of the outer images is marked as not predictable by the validator model. This is due to the fact that the input image does not contain any pixel information for these regions.

### Conclusion and Outlook

We approached the problem of generating internal visual sensory states by decomposing it into the visual association of images of a gripper based on a set of pose parameters and the prediction of views according to a specific gaze direction. The visual association is performed by a model-free approach to view synthesis. In this approach, an image is reconstructed from its appearance vector (i.e. a dimensionality-reduced representation of the image) which is associated with a set of corresponding pose parameters by a neural network. The views generated by the visual association always appear as if the gripper is fixated. Therefore, we employ a visual forward model to warp these images according to an arbitrary gaze direction.

The architecture is fully adaptive since it is based on artificial neural networks and subspace methods from pattern recognition. We presented some examples which suggest that the association capabilities result in consistent and valid images. Nevertheless, this has still to be analyzed quantitatively. Furthermore, we expect that, in its present form, the model will not be able to generalize over the angle space. For the present study we used a very coarse grid of 3 different angles. Using more angles leads to a higher variability in the training images and thus more eigen-images would be required, and consequently a larger association network. One possible solution would be to replace the (linear) PCA by a non-linear technique for extracting the appearance vectors.

Several technical shortcuts were used for the experiments in this study, e.g. the number of exploration trials was re-
duced by defining a regular grid of spatial positions, and the neural networks were trained off-line. These shortcuts are problematic from a modeling perspective; in a more realistic setting, the agent would learn its sensory-motor associations on-line while performing exploration movements. Furthermore, we assumed a priori knowledge about the characteristics of the gripper and performed a color-based segmentation of the camera images (i.e. sensory input) in order to extract the gripper images.

In contrast, Philipona, O’Regan, and Nadal (2003) propose in the context of sensorimotor contingencies a scenario in which simple agents learn internal representations of “physical space” with as little a priori knowledge as possible. This is done by sending random motor commands to the agent’s actuators and measuring the correlation between (the unlabeled) sensory and proprioceptive inputs. Following a similar route, we could improve our approach regarding gripper identification by performing small gripper movements around a certain (possibly random) position which is fixated by the camera head. If we assume that the environment is static, all changes in the sensory input would be due to gripper movements. Thus, portions that belong to the agent (i.e. the gripper) could be clearly separated from irrelevant content (i.e. background) without a priori knowledge.

The presented model is intended to be used as a building-block of more complex architectures which rely on the processing of sensorimotor (e.g. visuomotor) data. Consider a scenario in which the agent interacts with different objects. The objective is now to predict a possible displacement of the object (in the visual domain) based on the motor commands carried out by the agent. This can be achieved by detecting changes in the visual input. The movement of the agent’s manipulators during the interaction also leads to changes in the visual scene. These changes can be predicted by the proposed architecture and subsequently canceled out. In the context of simulation theory, e.g. the PtA approach, the proposed model could be used to generate a multi-step prediction in which the gripper and the objects can be distinguished from each other.

References