Perceiving the unusual: Temporal properties of hierarchical motor representations for action perception

Yiannis Demiris *, Gavin Simmons

Biologically Inspired Autonomous Robots Team (BioART), Intelligent Systems and Networks Group, Department of Electrical and Electronic Engineering, Imperial College London, South Kensington Campus, London SW7 2BT, UK

Abstract

Recent computational approaches to action imitation have advocated the use of hierarchical representations in the perception and imitation of demonstrated actions. Hierarchical representations present several advantages, with the main one being their ability to process information at multiple levels of detail. However, the nature of the hierarchies in these approaches has remained relatively unsophisticated, and their relation with biological evidence has not been investigated in detail, in particular with respect to the timing of movements. Following recent neuroscience work on the modulation of the premotor mirror neuron activity during the observation of unpredictable grasping movements, we present here an implementation of our HAMMER architecture using the minimum variance model for implementing reaching and grasping movements that have biologically plausible trajectories. Subsequently, we evaluate the performance of our model in matching the temporal dynamics of the modulation of cortical excitability during the passive observation of normal and unpredictable movements of human demonstrators.

© 2006 Elsevier Ltd. All rights reserved.

Keywords: Imitation; Minimum variance; Hierarchical structures; Action recognition; Corticospinal excitability

1. Introduction

An increased interest in computational mechanisms that will allow robots to observe, imitate and learn from human actions has resulted in a number of computational architectures that allow the matching of demonstrated actions to the observer robot’s equivalent motor representations (Alissandrakis, Nehaniv, & Dautenhahn, 2002; Billard, 2000; Demiris & Hayes, 2002; Schaal, Ijspeert, & Billard, 2003). These architectures, whilst sharing common computational components such as modules for processing and classifying visual information and retrieving motor representations, differ in the way that the perceptual information is coded and classified, the organisation of the motor system, and the stage at which the motor representations are used. The final aspect, at what stage the motor representations are used, differentiate architectures that follow the general ‘observe, classify, imitate’ decomposition (Kuniyoshi, Inaba, & Inoue, 1994), from those that advocate a stronger involvement of the motor systems in the perception process, through a ‘rehearse, predict, observe, reinforce’ decomposition (Demiris & Hayes, 2002; Demiris & Johnson, 2003; Schaal et al., 2003). In the latter, the observer robot invokes its motor systems to rehearse potential actions, predicting and confirming incoming observed states during the demonstration. This approach has gained biological credibility with the discovery of the mirror system in monkeys and humans (Grezes, Armony, Rowe, & Passingham, 2003; Rizzolatti, Fadiga, Gallese, & Fogassi, 1996). Not all theoretical models advocate the actual rehearsal of candidate actions as our previous work has done (Demiris & Hayes, 2002), opting instead for a weaker version of this motor theory of perception, usually termed ‘motor resonance’, in which the motor representations are retrieved through a resonance mechanism rather than a generative mechanism.

For imitation approaches that advocate the use of motor systems during the perception stage it becomes crucial to have a clear and flexible motor system organisation. Hierarchical representations, involving primitive motor structures at the lowest level, while increasing their complexity in higher levels, have been proposed (Demiris & Johnson, 2003; Wolpert, Doya, & Kawato, 2003), and tested in robotic systems (Demiris & Johnson, 2003), which successfully learned and used sequences of actions by observation. However, little has been done with respect to the temporal dimension of these representations, including how they can be coordinated, as well as their relation to biological data.
In this paper, we will examine in detail the issue of hierarchical representations, and in particular examine how higher level models can be composed from (and coordinate) lower levels primitives. Our approach will use representations based on the biologically plausible minimum variance model of movement control (Harris & Wolpert, 1998; Simmons & Demiris, 2005), which leads to principled and biologically plausible coordination of the underlying components. We subsequently compare a partial instantiation of our hierarchical attentive multiple models for execution and recognition (HAMMER) architecture (Demiris & Khadhouri, in press) for reaching and grasping actions, with transcranial magnetic stimulation (TMS) data from humans during the passive observation of grasping movements by a demonstrator (Gangitano et al., 2004).

2. Background

2.1. Hierarchies

Hierarchies are computationally interesting since they advocate a logical representational decomposition: motor primitives at the lower levels take care of the executional details while progressively higher levels shift their emphasis towards exerting temporal, contextual and cognitive control. From a robotics point of view, this allows for easier task planning and execution. In action understanding and gesture recognition, hierarchical representations have been regularly used since they allow the processing of information in multiple levels of detail (Hu, Tan, Wang, & Maybank, 2004). There is increasing evidence that the human brain might be organised in such hierarchical fashion (Essen & Maunsell, 1983; Fuster, 2004; Koechlin, Ody, & Kouneiher, 2003) and evidence for such hierarchical arrangements have been found, for example, in rapid movement sequences such as keypresses with the fingers (Rosenbaum, Kenny, & Derr, 1983).

2.2. From primitives to composite inverse models

When it comes to the representation of human and robot actions, there are two aspects to hierarchical organisation: a somatotopic one and a functional one. A node at a certain level in the hierarchy has to assemble a motor plan deciding which body components are to be used, as well as how they are going to move, individually and in coordination to each other. Options include having the underlying components arranged sequentially (Demiris & Hayes, 2002), or additionally allow parallel execution (Demiris & Johnson, 2003). In the second case, an arbitration mechanism is needed that checks whether the underlying components running in parallel have any overlapping controlled degrees of freedom, using the somatotopic representation. There are no cross-inverse model constraints, such as speed of execution, for example.

Hierarchical organisation of motor structures (for selecting effectors and checking consistency) should be distinguished from their use during execution. Although structures can be arranged in the different ways described above, there are distinct possibilities regarding the use of the hierarchy during execution:

- Hierarchical structures are used during the planning phase, but once the sequence of commands have been planned (for example, by calling the lower levels with common parameterizations), and the higher levels have been populated, there is no further communication or online adjustment during execution between the different levels.
- There is an active recruitment and adjustment during the execution phase, where execution is passed to the lower levels; higher levels are coordinating them, for example, by determining their start and end times.

Although interesting data regarding the organisation of the mirror system have been reported (Rizzolatti et al., 1996; Umilta et al., 2001) there is little information regarding the temporal properties of the mirror system. We will review some recent evidence for this temporal dimension next.

2.3. Temporal coupling between action observation and execution

While fMRI and PET studies have demonstrated the existence of a mirror system in humans (Grezes et al., 2003), the temporal resolution limitations of brain scanning technology means that we are less knowledgeable regarding the temporal aspects of the mirror system. Recent experiments with transcranial magnetic stimulation (TMS) have, however, shed some light into the temporal coupling between action observation and execution (Gangitano, Mottaghy, & Pascual-Leone, 2001; Gangitano et al., 2004). Previous computational models of the mirror system have shown that such temporal coupling is crucial (Demiris, 1999; Demiris & Hayes, 2002; Oztop & Arbib, 2002). For example Demiris (1999), derived a set of testable predictions, most important of which was that monkey mirror neurons would not fire (or fire less) when the demonstrated movement was performed at speeds unattainable by the observer monkey. In Gangitano et al. (2001), it was shown that the amplitude of the motor evoked potentials (MEP) induced by TMS in humans observing a reaching-grasping action was modulated by the amount and timing of the observed finger aperture. A strict temporal coupling between corticospinal excitability and the dynamics of the reaching and grasping movement when passively observed was clearly demonstrated (Gangitano et al., 2001). A followup study (Gangitano et al., 2004) shed further light into the temporal characteristics of this coupling. The modulation in corticospinal excitability profiles during the observation of reach and grasp actions was studied under three experimental visual stimuli:

- Observation of natural reaching and grasping (RnG) actions.
- Observation of a RnG action where the appearance of the maximal finger aperture was significantly delayed.
- Observation of a RnG action where an unexpected finger closing and opening action was inserted before the final grasp portion of the demonstration.
The first condition replicated the results of Gangitano et al. (2001) in that the observer’s cortical excitability profile was in concordance with the kinematic profile of the demonstrated finger movements. The second condition did not show any modulation in the corticospinal excitability profile, while for the last condition the initial profile remained the same (for as long as the two stimuli remained the same), but upon sight of the unexpected finger closing and opening action, there was a slow decay in the initial activation. We will use our model to reproduce these results computationally, using the same types of stimuli.

2.4. From the biological data to a computational model

We will use a biologically inspired model of motor control (Simmons & Demiris, 2005) based on the minimum variance model (Harris & Wolpert, 1998) and describe how more complex actions can be formulated from lower ones in a principled manner, and used for action recognition under the HAMMER architecture. We will concentrate on grasping movements, for which there are good biological (Rizzolatti et al., 1988) and modelling (Fagg & Arbib, 1998) data, and show how they can be formed from combinations of reaching and opening and closing finger movements. We will subsequently compare the performance of the resulting model, HAMMER–MV, for the explanation of the patterns of corticospinal excitability recording during the observation of human normal and unpredictable reaching actions.

3. HAMMER

3.1. Building blocks

The HAMMER family of architectures uses inverse and forward models (Karniel, 2002; Narendra & Balakrishnan, 1997; Wolpert & Kawato, 1998) as the basic building blocks. An inverse model is a module that takes as inputs the current state of the system and the target goal(s) and outputs the control commands that are needed to achieve or maintain those goal(s). The functional reverse to this concept is that of a forward model of a controlled system: a forward model is a module that takes as inputs the current state of the system and a control command to be applied on it and outputs the predicted next state of the controlled system.

Pairing an inverse model with a forward model in the way shown in Fig. 1 results in a structure that can be used both for executing an action as well as perceiving it.

When HAMMER is asked to rehearse or execute a certain action, the inverse model module receives information about the current state (and, optionally, about the target goal(s)), and it outputs the motor commands that it judges as necessary to achieve or maintain these implicit or explicit target goal(s). The forward model provides an estimate of the upcoming states should these motor commands get executed. If, instead of feeding the current state of the imitator to the inverse model, the imitator feeds into it the current state of the demonstrator, the inverse mode will generate the motor commands that it would output if it was in that state and wanted to execute this particular action. By inhibiting the motor commands from being sent to the motor system, the forward model can output an estimated next state, which is a prediction of what the demonstrator’s next state will be.

This predicted state is compared with the demonstrator’s actual state at the next time step. This comparison results in an error signal that can be used to increase or decrease the behaviour’s confidence value, which is an indicator of how closely the demonstrated action matches a particular imitator’s action.

3.2. Distribution

HAMMER consists of multiple pairs of inverse and forward models that operate in parallel (Demiris & Hayes, 2002). Fig. 2 shows the basic structure.

When the demonstrator agent executes a particular action the perceived states are fed into all of the imitator’s available inverse models. Following the algorithm described above, this generates multiple motor commands (representing multiple hypotheses as to what action is being demonstrated) that are sent to the forward models. The forward models generate
predictions about the demonstrator’s next state: these are compared with the actual demonstrator’s state at the next time step, and the error signal resulting from this comparison affects the confidence values of the inverse models. At the end of the demonstration (or earlier if required) the inverse model with the highest confidence value, i.e. the one that is the closest match to the demonstrator’s action is selected.

Although in theory all of the imitator’s available inverse models can be activated and processed in parallel, in practice only a subset is activated, based on context, the attention focus of the observer (Demiris & Khadhouri, in press) and the perceived state and viewpoint of the demonstrator (Johnson & Demiris, 2005).

In this paper, we will concentrate on one inverse model (for cases involving multiple ones see Demiris and Hayes (2002), Demiris and Johnson (2003), and Demiris and Khadhouri (in press)), that for grasping objects, focusing in detail on the hierarchical aspect of the architecture.

3.3. Hierarchy

More recently, we have designed and implemented a hierarchical extension (Demiris & Johnson, 2003) to the arrangement described above: primitive inverse models are combined to form higher, more complex sequences (Fig. 3), with the eventual goal of achieving increasingly more abstract inverse models (Johnson & Demiris, 2004). This helps in dealing with what is known as the correspondence problem (Nehaniv & Dautenhahn, 2002), the matching of actions between differently embodied agents.

However, our previous work has not placed special emphasis on the temporal characteristics of the actions. In this paper, we will describe our particular implementation of HAMMER, termed HAMMER–MV, using the minimum variance model for movement (Harris & Wolpert, 1998), and demonstrate how reaching and grasping primitives can be combined into grasping actions. Subsequently we will compare its performance in perceiving normal and unpredictable grasping actions against TMS data on the perception of such stimuli by humans (Gangitano et al., 2004).

4. The HAMMER–MV implementation

HAMMER–MV follows the general architecture of HAMMER, but uses minimum variance controllers as lower level inverse models, and coordinated combinations of these at the higher ones. We will start by giving an overview of our implementation of the minimum variance model, and show results on how it can be used to generate biologically plausible reaching trajectories. Subsequently, we will describe how we implement a particular instance of a hierarchical representation for a grasp using the minimum variance model, and show results on how it can be used to grasp an object with biologically plausible trajectories. Finally, we describe how the same hierarchical representation can be used to recognise the movement when done by a human demonstrator, in Section 4.1.

4.1. Minimum variance

4.1.1. Introduction

A common approach in developing computational models of human movement is to consider the motor system as planning and executing movements by optimizing some quantity or characteristic of the movement (Flash & Hogan,
saccadic eye movements have been successfully predicted by trajectories and velocity profiles of both arm movements and since fast-movements also require large control signals. The accuracy trade-off also emerges naturally from this approach, resulting in a less accurate movement overall. The speed–smoothing criteria, as non-smooth movements require large quadratic regulator optimal control scheme, which we important, as it has to be of a form suitable for fitting into control scheme suitable for controlling a robot arm (Simmons 1985; Nakano et al., 1999; Uno, Kawato, & Suzuki, 1989). Optimization criteria are usually chosen such that the trajectories predicted by the model match the empirically observed features of human movement, such as smoothness (Morasso, 1981) and speed–accuracy trade-off (Fitts, 1954). However, few models account for the small variations between repeated movements. Harris and Wolpert (1998) noted that motor commands sent to the muscles, like all neural signals, were subject to signal-dependent noise (where the variance of the noise is proportional to the amplitude of the signal). Noise on the motor command leads to variation in the intended movement. Over the course of a movement these variations build up, leading to inaccuracy of the final position and possibly resulting in task failure. They proposed that the goal of the motor system is to minimise inaccuracy at the end-point of movement. Over the course of a movement these variations are characterised by spatial coordinates and a temporal position within the course of the movement: $V_i = [x_i, y_i, n_i],\quad 0 < n_i < T$ where $i$ is $i$th via-point, $x_i$ and $y_i$ are the spatial coordinates and $n_i$ is the time step at which the trajectory must pass through the via-point. At this time step, the state cost $Q_n$ takes a form similar to that of $Q_T$, ensuring the via-point is reached. Setting the costs for the target, post-movement period and any via-points results in a set of state cost matrices $Q$: $Q = \{Q_0, \ldots, Q_T, \ldots, Q_{T+N}\}$ (4) the cost function matrices $Q$ are used to calculate a ‘cost-to-go’ for each time step. This is an estimate of the total remaining cost from the next state. It is calculated recursively, ‘offline’ before any movement takes place, starting from the fact that the remaining ‘cost-to-go’ at $n = T + N$ is simply the final state cost $Q_{T+N}$. This results into another set of matrices $P$: $P = \{P_1, \ldots, P_{T+N}\}$ (5) The optimal control scheme uses state feedback to produce a motor command; that is, the current state is multiplied by a gain term to produce the control signals. The set of state feedback gains $K$ are calculated from the ‘cost-to-go’ matrices, giving a set of matrices $K$: $K = \{K_0, \ldots, K_{T+N}\}$ The motor command is then calculated as: $u_n = K_n s_n,\quad n = 0, \ldots, T + N$ (6) Following the minimum variance model, this control signal is subject to white noise with zero mean and variance proportional (with coefficient $k$, set to 0.01 in these experiments) to the amplitude of the signal: $\hat{u}_n = u_n + w_n,\quad w_n \sim N(0, k u_n^2)$ (7) Finally, the noisy motor command is used to update the state vector according to the state dynamics matrices $A$ and $B$: $s_{n+1} = A \cdot s_n + B \cdot \hat{u}_n$ (8) 4.1.3. Reaching results We used a two-link planar arm (as the one shown in Fig. 6 but without the fingers) to demonstrate our implementation,
successfully capturing the smooth movements and speed-accuracy trade-off characteristics of point-to-point reaching (Simmons & Demiris, 2005), as shown in Fig. 4. A straight movement is shown, moving the end point of the arm from coordinates (0.2,0.2) to coordinates (0.4,0.5).

4.1.4. The coupled inverse-forward model formulation

As stated above, our implementation of the minimum variance model is also suitable for producing an instance of the HAMMER architecture (Simmons & Demiris, 2004). To do this, we require a movement to be encoded as a pair of forward and inverse models. The forward model aspect is the state update equation, since it takes the current state and a motor command and produces a prediction for the next state, as required. We store the dynamics matrices $A$ and $B$ as the forward model.

The inverse model is required to produce a motor command to move the current state towards the goal. This function is fulfilled in the optimal control scheme by the motor command equation (Eq. (7)). Here, we store the set of cost-to-go matrices $P$ as the inverse model. Both these and the dynamics matrices can be calculated and stored ‘offline’ without having to execute a movement. They are calculated using the parameters of the movement, such as target and number of time steps (Simmons & Demiris, 2004).

With these structures in place, our optimal control implementation of the minimum variance model can be fitted directly into the HAMMER architecture. Pairs of forward and inverse models exist for individual movements of individual effectors and can be combined into hierarchical structures, both somatotopically and functionally. We now describe one such hierarchy used for reach-to-grasp movements.

4.2. Combining primitives to form higher representations

The ‘classical’ view of grasping considers reach-to-grasp movements to be made up of two independent visuomotor processes—transporting the hand to the object, and performing the grip (Jeannerod, 1981). This hypothesis has been used to explain a wealth of experimental data on human grasping, and in particular to describe several key features. Among these are the observation that maximum grip size (between thumb and finger) increases with the size of the object, with a slope of about 0.8 (Jeannerod, 1981), and that this maximum grip aperture occurs at around 60–80% of the movement time (Jeannerod, 1984). More recently, Smeets and Brenner (1999) proposed that reaching does not require two separate processes, but can be explained as pointing with the digits to target positions on an object. They used the minimum jerk model of human movement to produce digit trajectories for a grasping movement, avoiding any consideration of the mechanics of the joints and limbs. Their model was able to predict the key features of grasping without the need for separate visuomotor channels for the transport and grip components. Following the view of Smeets and Brenner (1999), we extended our implementation of the minimum variance model to grasping. However, in our model we take the system actuators explicitly into account since these are the parts of the system subject to signal-dependent noise. We regard grasping as two separate components (transport, grip) that are nonetheless planned using the same principle (pointing to a target using the minimum variance model).

The grip itself is also comprised to two separate movements, that of the thumb and that of the finger. To perform a reach to grasp movement using HAMMER–MV we assemble a hierarchy of inverse models, as shown in Fig. 5. The individual components of the hierarchy are parameterized from the highest level using information about the object and the required movement time. In line with the design philosophy of the HAMMER architecture, this hierarchical representation can be used to execute a grasping movement or perform online action recognition of a demonstrator performing a grasp, as we will demonstrate later in the paper.

4.3. Arm construction

The arm model used in these experiments was similar to that used in Simmons and Demiris (2005). The model was a two-link planar arm with two rotational degrees of freedom corresponding to the shoulder and elbow. The lengths of the two links were 30 cm, roughly corresponding to the lengths of the upper arm and forearm of the demonstrator. The hand was constructed from two similar components, both two-link models with two rotational degrees of freedom. The link

Fig. 4. (a) Repeated point-to-point reaching movements, showing variation between trajectories, (b) end-point variances for movements performed with different movement times.
lengths were set to 10 cm, accounting for the hand and the digit. The arrangement of the arm and hand are shown in Fig. 6.

4.3.1. Grasping results

Fig. 7 shows ten trials of our model generating finger trajectories using the minimum variance model. These trajectories follow the characteristic patterns of human grasping (Jeannerod, 1984), including an overshoot in the grip aperture at approximately 70% of the movement time. Note that although the initial grip aperture is known to affect the kinematics of the grasping, in here we keep the initial grip aperture constant following the experimental data of Gangitano et al. (2004)

5. Experiments

In our final set of experiments, human demonstrations of reaching actions were recorded and given as input to the 2D six degree of freedom simulated arm, controlled using HAMMER–MV. In the following sections, we will describe...
the visual stimuli we recorded, and the equations governing the matching of the model arm’s performance against the human data.

### 5.1. Visual stimuli

Three different types of reaching and grasping movements (Fig. 8, 9 and 10) were recorded using a human demonstrator, closely following the experimental approach and stimuli types of Gangitano et al. (2004).

The movements of the human demonstrator were restricted into a 2D plane parallel to the table surface, and a 100 mm foam ball was used as the grasping target. Data were captured with a Unibrain firewire camera at the rate of 30 frames per second. Colour markers were placed at the thumb and index fingers of the demonstrator and a coloured arm band at his wrist. These three points (end of thumb and pointer fingers, and center of mass of the wrist) were tracked using the CamShift algorithm, and the coordinates of these points in successive frames along with time stamps were saved to a file, to be used as input to HAMMER-MV. Periods of inactivity at the start and end of the demonstration were removed from the file.

### 5.2. Error calculation and confidence update

The confidence was updated at each time step by comparison between the imitator’s prediction and the demonstrator’s trajectory. We specifically compared the predicted change in both the distance to the target, \( r \), and the grip aperture, \( g \), as shown in Fig. 6, making the confidence for the top-level of the hierarchy a function of the confidences of its component parts. We first found the coordinates of the mid-point of the grip, as shown in Eq. (9)

\[
m_n = \frac{f_n - t_n}{2} + t_n
\]

where \( m_n \) is the position of the grip mid-point, and \( f_n \) and \( t_n \) are the positions of the tips of the finger and thumb, respectively. We defined the distance to the target as the absolute difference between the coordinates of the centre of the object and the mid-point.
point of the grip, as shown in Fig. 6:

\[ r_n = \sqrt{(\text{object} - m_n)^2} \]

we also calculated the absolute value of the grip aperture at each time step, also shown in Fig. 6:

\[ g_n = \sqrt{(f_n - t_n)^2} \]

using Eqs. (9)–(11), values of \( m_n \), \( r_n \), and \( g_n \) were calculated both for the demonstration trajectory and the predicted trajectory. Given these values, the confidence value at each time step was given as

\[ C_n = C_{n-1} + \Delta C_r + \Delta C_g \]

where

\[ \Delta C_r = \text{sgn}(r_n - r_{n-1}) \times \text{sgn}(r_{n-1}^{\text{pred}} - r_n^{\text{pred}}) \times (r_{n-1}^{\text{pred}} - r_n^{\text{pred}}) \times w_r \]

and

\[ \Delta C_g = \text{sgn}(g_n - g_{n-1}) \times \text{sgn}(g_{n-1}^{\text{pred}} - g_n^{\text{pred}}) \times (g_{n-1}^{\text{pred}} - g_n^{\text{pred}}) \times w_g \]

these equations mean that if both the demonstration and the prediction are moving in the same direction, the confidence that the inverse model has correctly recognised the movement will increase. If the two trajectories move in different directions, the confidence will decrease. The change in confidence is modulated by both the step-by-step change in the prediction and by a weighting term, which sets the relative importance of each component of the movement to the overall confidence that the inverse model is predicting correctly.

5.3. Results

Fig. 11 shows the confidence plot over time for an example trajectory of the three different conditions, with all the trajectories shown in Fig. 12, along with the averages for all instances of the three different conditions shown in Fig. 13.

The results closely follow the characteristics of the data reported by Gangitano et al. (2004). For example, in Fig. 11, the confidence progression during the observation of a normal movement over time follows the increasing trend reported by Gangitano et al. (2004), reaching its maximum at the point of maximum aperture, and subsequently reducing slightly as the movement closes to an end. The confidence progression during the observation of the second stimuli type (a grasp movement with delayed aperture opening), does not give any significant increase in confidence values during the demonstration, including during the delayed aperture opening (similarly to what is reported in Gangitano et al. (2004)). Finally, the confidence progression during the open–close–open grasp follows the general trend of the normal movement, until the point of the sudden finger closure, with a general descending trend from then on, again matching the general trend as reported by Gangitano et al. (2004). This pattern is largely reproduced amongst all captured trajectories, as shown by the average confidences of all demonstrations in Fig. 13. The amplitudes of the variances are also largely expected, with smaller variations in both normal and delayed cases, and up to closing in the open–close–open stimuli case. After the unpredictable behaviour of closing the fingers mid-way, the confidence variance is higher. The variance in the delayed aperture opening is also expectedly low given that the forthcoming states are mostly predictable, increasing towards the end.

6. Discussion

It has been advocated earlier in this paper that hierarchical representations can be a useful engineering tool when
structuring motor systems. The HAMMER–MV implementation of this concept demonstrates why: hiding the details of the lower level details into higher level structures allows for easier task planning than that achieved with flat, non-hierarchical representations—only the details of the goal and desired task parameters need to be supplied and the higher inverse model will recruit and coordinate the appropriate lower level primitives. Additionally, the higher level models also modulate the contribution of each of the underlying primitives when predicting future states: apart from judging success individually, predictions from lower levels are rated, and modulated according to how much they contribute to the higher goal (Eqs. (12)–(14)). The minimum variance implementation we employed was particularly suited for this, since it allows the calculation of the confidence of the lower level level primitives (e.g. finger movements) individually as well as with respect to each other (e.g. finger aperture), and propagates their weighted values upwards, for example, to allow higher level nodes to calculate the progress of the overall model towards the final goal.

Using the biologically inspired minimum variance model (Harris & Wolpert, 1998; Simmons & Demiris, 2005), the grasp model propagates to its components the time in which to
complete the task, and the object to be grasped, and receives back the cost-to-go matrices from them for those particular requirements. Planning of the movement is done off-line, and there are no parameters to be adjusted during execution in the minimum variance model, therefore, there is no way to temporally ‘morph’ the plan to what is being observed. Although that is technically possible (as it was demonstrated in previous experiments we have performed with adaptive PID controllers (Demiris & Hayes, 2002) which allowed online adaptation to different speeds of the demonstration, within limits), the data of Gangitano et al. (2004) suggest that humans do not morph the ongoing motor plan temporally, and the minimum variance implementation of our model captures that aspect more closely than (Demiris & Hayes, 2002). It would also be interesting to see whether other criteria (for example, smoothness of movement) that result in human-like trajectories could also replicate the effects observed in our experiments. Although the motor plan generated off-line by HAMMER–MV can be interrupted, and a new one re-generated to possibly better capture the demonstration, it is unlikely that for short demonstrations such as the ones we examined here (and used in the experiments of Gangitano et al. (2004)) there is sufficient time to do so. It would be interesting, both for the neurophysiology and the computer modelling sides to repeat such experiments but for longer actions, including possibly sequences of them, to determine whether such a ‘resetting’ mechanism (as termed in Demiris (2002)) is indeed present in humans and study its characteristics.

It is important to note that the comparisons between the results of Gangitano et al. (2004) and the computational results reported here rely on the assumption that the confidence of an inverse model can be mapped to changes in MEPs of the controlled body part in humans. This relies on the intuition that the neural substrate of a more confident inverse model (i.e. a model that explains the observed movement better) will be more active than that of a less confident model. Elsewhere (Demiris & Khadhouri, in press) we have linked the confidence of inverse models to the attentional resources that are allocated to them. Since attention to action increases activity in prefrontal, premotor and parietal cortices (Rowe, Friston, Frackowiak, & Passingham, 2002), we can envisage a link between inverse models and MEPs, although the exact nature of this link will require more detailed neural modelling than is within the scope of this paper, and is the subject of future work.

Interestingly, sight of the static setup with all the components (arm and ball) in place, gives rise to corticospinal excitability (Gangitano et al., 2004), even before movement begins. Similar data have also been reported in Kilner, Vargas, Duval, Blakemore, and Sirigu (2004). In our model instantiation here we have not addressed that, but elsewhere we have argued that such initial activation can be due to a bottom-up initialization of the available inverse models according to the scene observed (Khadhouri & Demiris, 2005). Applicable inverse models will receive a ‘pre-boosting’ while non-applicable ones will be eliminated from the pool of available ones, and not considered subsequently.

It is also important to note that although in the experiments reported here we match the observed action with internally generated actions at the trajectory level (the kinematics of the movement), we do not advocate that the mirror system works exclusively at this level. Indeed in our work reported elsewhere (Johnson & Demiris, 2005) we have also used more abstract qualitative states (for example, ‘hand closer to target’) to match the predicted and perceived states of an action. We have also encoded inverse models higher in the hierarchy in an increasingly abstract way, for example, by considering whether the goals of inverse models lower in the hierarchy have been achieved, even if the kinematics of the demonstrated and internally generated action do not match. Such a goal-oriented account would be more suitable for explaining neural recording data on the recently discovered tool-use mirror neurons (Ferrari, Rozzi, & Fogassi, 2005), as well as the mirror neuron data on partially-occluded action demonstrations (Umilta et al., 2001). We envisage that a multi-level system that incorporates both goal oriented and movement oriented matching of perceived and internally generated actions (preliminary results in Johnson and Demiris (2005)) will better explain the operation of the mirror neural substrate at the system level in...
humans and monkeys. A particular challenge we are trying to address is the developmental course of this system, incorporating psychological data that infants use their own experience in the perception of others’ actions (Sommerville, Woodward, & Needham, 2005) including reasoning about the rationality of those actions (Gergely, 2003).

7. Conclusions

The neurophysiological data mentioned in this paper lend support to the notion that the human brain does not passively observe actions but actively forms hypotheses and predicts forthcoming states. In Gangitano et al. (2004), it was shown that there is no temporal fragmentation of the action plan in the motor representation of the observer. The computational implementation of the HAMMER architecture described in this paper reproduced these results, using a hierarchical controller based on the minimum variance principle, and shed some light into possible ways for implementing such strategies in robotic systems, towards implementing biologically plausible mechanisms for human action understanding and imitation.

Acknowledgements

The first author acknowledges the support of UK Engineering and Physical Sciences Research Council (EPSRC Grant GR/S11305/01) and the Royal Society. The second author is supported by a EPSRC/DTA doctoral scholarship. Thanks to all the BioART members for their valuable feedback, and especially Anthony Dearden for his help in capturing and analysing the visual stimuli.

References


