Research report

Fixation behavior in observation and imitation of human movement

Maja J. Matarić [a,*], Marc Pomplun [b]

[a] *Neuroscience Program and the Computer Science Department, University of Southern California, 941 West 37th Place, Los Angeles, CA 90089-0781, USA*

[b] *Department of Neuroinformatics, Faculty of Technology, Bielefeld University, 33501 Bielefeld, Germany*

Abstract

This paper describes experiments performed with 40 subjects wearing an eye-tracker and watching and imitating videos of finger, hand, and arm movements. For all types of stimuli, the subjects tended to fixate on the hand, regardless of whether they were imitating or just watching. The results lend insight into the connection between visual perception and motor control, suggesting that: (1) people analyze human arm movements largely by tracking the hand or the end-point, even if the movement is performed with the entire arm, and (2) when imitating, people use internal innate and learned models of movement, possibly in the form of motor primitives, to recreate the details of whole-arm posture and movement from end-point trajectories. © 1998 Elsevier Science B.V. All rights reserved.

Keywords: Perceptual–motor interaction; Eye-tracking; Movement imitation

1. Introduction

Imitation is one of the most ubiquitous forms of human learning. What appears to be a simple process of 'monkey see, monkey do' involves intricate interaction between several cognitive systems. The work described in this paper focuses on studying the relationship between perception and motor control in imitation.

Recent work in neuroscience, cognitive science, and developmental psychology has provided preliminary but growing evidence for a link between the perceptual and motor systems. A shared neural substrate between imagined and executed movements was shown [17,35], as was increased activity of motor cortex motor evoked potential during movement observation, both in humans [24,34] and monkeys [20], the latter involving 'mirror neurons' in the pre-motor cortex [66]. Psychophysical data have led to similar conclusions [57,75], as has evidence from developmental studies [9], lesion studies [27], and performance theories about skills such as squash and catching [1,22,67,68].

This convergence of evidence is largely generated in experiments not focused on imitation. In order to address the perceptual end of sensori–motor integration within the context of imitation, we asked the following question:

Do subjects observe and attend to movement stimuli differently depending on whether they are just watching or watching with the intention to subsequently imitate, i.e., generate the observed movement?

Our experiment would present subjects with stimuli and monitor their fixations in order to address two key issues:
1. Is there a difference between watching to imitate vs. just watching?
2. When watching to imitate, what features are fixated on, in order to obtain sufficient information to repeat the observed movement?

If we assume strong sensori–motor integration, we would expect to see no difference in fixation patterns between the imitation and no-imitation conditions, since the unconscious process of early stages of movement preparation would be uniformly active in both. A result providing a negative answer to the first question would add overt fixation behavior to other data supporting the sensori–motor connection. The second question addresses issues of underlying mechanisms for motor control by considering sparse information provided by fixation data and its use in subsequent movement generation which humans perform with excellent imitative proficiency [15,40,72].

To address these questions, we designed an eye-tracking experiment in which the subjects were shown videos of different types of natural but unfamiliar finger, hand, and

arm movements, were told in advance they would either imitate or just watch, and had their fixations recorded. Although fixations are only one part of complex perception, the results lend insight into the underlying sensori–motor integration question.

2. Materials and methods

2.1. Subjects

The subject pool consisted of 34 university students, and six faculty, research, and support staff, totaling 40, 12 females and 28 males. One subject was left handed, the rest were right handed. Each participated in a one-time session lasting approximately 20 min, varying only in the time required to fit and calibrate the eye-tracker. All subjects received a standard payment.

2.2. Equipment

The experiments were performed with subjects wearing a head-mounted eye-tracking apparatus while watching short videos presented on a computer screen (Fig. 1). The 'Omnitrack1' eye-tracker [70] was used, consisting of a light metal head-mounted frame that points an infra-red camera at the subject's right pupil and another camera mounted on the front of the frame and directed at the stimulus screen. The stimuli were short videos (moving displays), presented on a 17″ computer screen equipped with four infra-red emitters used for eye-tracker head-set alignment. The distance between the screen and the subject's eyes was adjusted to 60 cm, to provide a constant visual angle of 30.2° with respect to the horizontal screen size, as this setting provides the most accurate Ominitrack1 results. The right eye gaze position was measured within the screen area; head movement of up to about 15° results in no loss of accuracy. Ideal calibration leads to a mean error of 0.5° of visual angle which corresponds to 5 mm on the screen. Head-set shifts introduce deviations resulting in mean accuracy between 0.7° and 1.0°. The cameras are mounted at the front of the head-set, and counterweighted at the back. Due to variability in subject head shape, head-set balance is not perfect in most cases, resulting in small vertical shifts of the head-set, and thus leading to errors in the vertical (y) measurement. Repeated movements (such as our deferred imitation sessions interspersed between the stimuli), can also induce those head-set errors.

The eye-tracker's computational equipment stores the (x, y) (horizontal, vertical) coordinates, the pupil size, and the duration of the fixation at the screen resolution of 640 by 480 pixels at 60 Hz. Fixations are detected using a 30°/s velocity threshold. The gaze velocity is computed over a 33 ms window and must remain below the threshold for at least three successive video frames in order to be
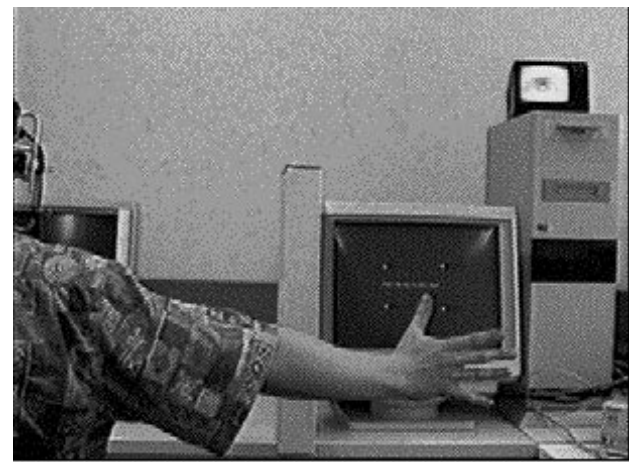

Fig. 1. The experimental eye-tracking apparatus; top: the head-mounted frame with two cameras; middle: the complete three-monitor set-up and the position of the subject; bottom: the imitation condition.

registered as a fixation. Thus, the minimum duration of detected fixations is 83 ms.

2.3. Task and stimuli

Three types of stimuli/moving displays were used: films of finger, hand, and arm movements against a black

background (Fig. 2). The different stimuli were chosen in order to avoid stimulus-specific behavior and to control both for the effects of stimulus size in the image and for speed of movement. The images in each stimulus film were sized to fill the screen; in finger-movement films the hand filled the screen, while in hand and arm movement films a much smaller percent of the screen was devoted to the hand/fingers/end-point. Each stimulus film lasted between 4 and 6 s; the stimulus pool consisted of 19 different films. The movements in the films were designed to be stylized and arbitrary, thus unfamiliar and different

from gestures used in spoken and sign language and common signaling systems. Finger movement films featured closeups of a motionless hand, either upright (wrist below hand) or on its side (wrist on the right, see top of Fig. 2), with fingers extending out and curling in, in discrete steps. Hand movement films featured a hand and lower forearm (the elbow was never shown, see middle of Fig. 2) following a smooth trajectory or moving through a sequence of discrete positions. Arm movement films featured the entire bare arm, from the tips of the fingers to the shoulder (see bottom of Fig. 2), in which the hand and the arm changed position at the wrist and elbow, using discrete flexion and extension movements at each of the joints.
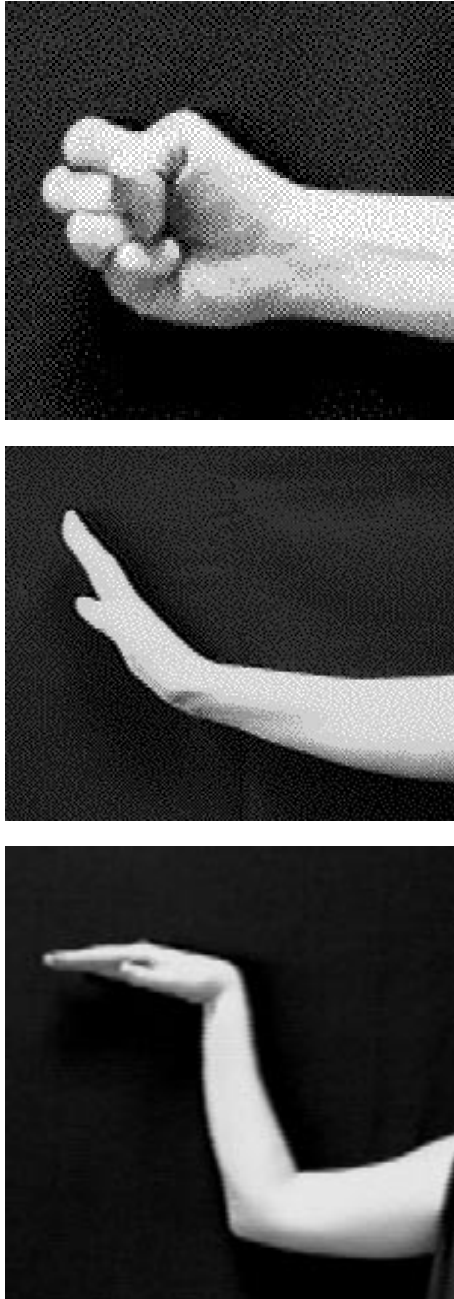
2.4. Design and procedure

Each subject was instructed to avoid head movements and imitate as well as possible. The subject was shown a sequence of 25 films, and instructed before each film whether to just watch or to watch and then imitate as soon as the film finished. The computer screen was black between the stimuli and during the imitation phase. In the imitation condition, the subject extended the right arm and imitated without visual feedback.

The 25-film experimental block consisted of four parts, covering the different conditions: (1) just watching five films; (2) watching and imitating five films; (3) just watching three films; (4) watching and imitating six films. The specific number of films shown was based on the total number available, and on the target experiment duration: the entire experiment took no more than 30 min, in order to prevent subject fatigue. Stimuli within each of the four parts were drawn randomly from the stimulus pool; the order of the parts was randomized across subjects.

2.5. Data analysis

For cross-validation purposes, we used two methods of data analysis, a qualitative and a quantitative one. In the qualitative method, a single naive observer watched the eye-tracker output on a video screen that superimposed the fixation point over the stimulus films. The observer collected qualitative data about the position of fixation relative to the hand. The data were then combined to compute the total percent of fixations on and off the hand; the hand included the fingers, palm, and wrist. As this was an extremely time-intensive procedure (it took a week of full-time film viewing on the part of the observer), we devised a more efficient as well as a more precise, automated method. In this automated, quantitative data analysis method, we digitized all of the original film stimuli (without fixation points) and marked in each frame the following salient points, i.e., features: the tips of the fingers, the centers of palm, wrist, elbow, and shoulder. Not all of the

Fig. 2. Examples of finger, hand, and arm movement stimuli, shown as single frames taken from the videos.

salient points were visible in all stimuli; all visible ones were marked. This information was used to automatically compute the distance between the fixation point on a given stimulus film segment and the visible features in the film.

In the analysis, we accounted for the distance on the *x*-axis, because the subjects' imitation movements caused vertical shifts of the head-set leading to deviations in the gaze-position measurement. Due to the geometry of the head-set, these errors mainly affected the *y*-values, so the restriction to *x*-values yielded considerably higher precision. Note that the *y*-value was not completely ignored; rather, it was only disregarded with respect to the measurement of distance between gaze and feature position. It was still used to decide the distance of the fixation relative to the features (i.e., the fingers, hand, the elbow) in situations of corresponding *x*-positions. Due to large distances between features, the *y*-coordinate was a reliable basis for discrimination. Since arm, hand, and fingers were presented from different angles, there was no essential information loss, i.e., the *y*-values were not as meaningful as the *x*-values.

3. Results

In order to address the original questions about sensori–motor interaction, we performed a series of analyses of the eye-tracking data. Specifically, we asked the following questions.

(1) Where do subjects fixate when watching movements? (Section 3.1)

(2) How long do subjects fixate on specific features while watching movements? (Section 3.2)

(3) Are there patterns of transitions between fixations, i.e., are specific feature sequences more frequent? (Section 3.3)

(4) Does fixation behavior change over the presentation of a movement video, i.e., do subjects fixate on different features early on compared to later in the stimulus presentation? (Section 3.4)

(5) Does fixation behavior differ between the no-imitation and imitation conditions, i.e., do subjects fixate different when they are just watching compared to when they plan to subsequently imitate? (Section 3.5)

The results of the data analysis addressing each of the above questions are presented next.

Table 1
The results of analysis by observation; rows indicate the percent of fixations on and off the hand for each stimulus type

Feature fixated on	Type of stimulus film			Mean %
	Finger	Hand	Arm	
On the hand	74.6	86.0	83.4	81.34
Off the hand	25.4	14.0	16.6	18.66

Table 2
The results of automated analysis; rows indicate the average percent of total viewing time, with standard error, spent fixating on each feature

Feature fixated on	Type of stimulus film		
	Finger	Hand	Arm
Hand	90.7 ± 0.7	67.4 ± 0.9	61.3 ± 1.8
Elbow	–	–	21.9 ± 1.3
Shoulder	–	–	7.0 ± 1.4

3.1. Location of fixations

In Table 1, we show the qualitative data; the rows give the position of fixation for each type of stimulus and demonstrate that the largest number of fixations is on the hand. The relatively high instance of fixations off the hand in finger stimuli is due to qualitative scoring; as shown in Table 2, more accurate automated analysis, which incorporated an error margin due to head movements, eliminated this effect.

Table 2 shows the details for each stimulus feature demonstrating the percent of total time the subjects spent looking at each; the rows list the features for each of the stimulus types. We defined 'target areas' around the features (fingers, palm, wrist, elbow, shoulder) that contained the features themselves and a tolerance area derived from the eye-tracker motion error margin. Importantly, the target areas for the features were properly scaled to match the associated targets: the fingers, palm, and wrist target areas were the same size. In arm movement films, the additional elbow and shoulder target areas were larger (diameter of 60 pixels each) than the fingers, palm, and wrist areas (diameter of 30 pixels each). This arrangement was chosen because elbow and shoulder are rather large and coarse fixation 'attractors', while the other three areas are located close to each other and represent small and detailed targets. The results are consistent with the qualitative analysis. The automated analysis was based on an error margin, as stated earlier, while the qualitative analysis was based on the observer's subjective decision as to whether a fixation is on or off the hand, respectively. Not surprisingly, a quantitative difference exists between the exact values in the two analyses, but the qualitative result is consistent.

Table 3
Fixation patterns within the hand, as demonstrated in the finger and hand films; rows indicate the average percent of total viewing time, with standard error, spent fixating on each feature

Feature fixated on	Type of stimulus film		
	Finger	Hand	Arm
Fingers	63.9 ± 2.4	31.6 ± 1.1	38.4 ± 1.3
Palm	20.0 ± 1.7	19.6 ± 1.0	9.4 ± 0.7
Wrist	6.8 ± 1.3	16.3 ± 1.0	13.5 ± 1.0

Table 3 uses a format consistent with Table 2 to show the details of the fixation behavior within the hand, breaking them down by fingers, palm, and wrist. The empty entries in the tables indicate features that were not present in some stimulus videos; specifically, finger and hand videos did not include the upper arm, so no fixations went to the elbow and shoulder.

Most of the fixations were spent on the hand, and within the hand, most fixations were spent on the fingers. The percent values for all of the features were entered into a one-factorial ANOVA with repeated measures for each type of movement film (i.e., finger, hand, and arm). Note that two-factorial ANOVAs could not be calculated due to the different numbers of levels for each type of stimulus; namely, the finger and hand stimuli necessarily contained no data for the shoulder and elbow. Finger movement films demonstrated a significant main effect of the fixated feature on the percent values ($F = 175.51$; $p < 0.0001$) and meaningful differences between all three levels (fingers, palm, and wrist). In hand movement films, a significant main effect was found as well ($F = 43.39$; $p < 0.0001$), again with three significant between-level contrasts. Here, the order was hand, wrist, then palm. The main effect in arm movement films was significant as well ($F = 93.25$; $p < 0.0001$), with the fingers being attended longer than the other features. The only non-significant contrast was found between the two least fixated features, namely palm and shoulder. These data are consistent with the qualitative analysis; both show a significant pattern of fixation on the end-point of the movement, i.e., the hand. The significance of the different parts of the hand varies considerably between stimulus types, with the hand being the target of most of the fixations.

As a null hypothesis, we automatically generated random fixations for all of the stimulus films, and analyzed them in exactly the same way as we did with the human fixation data. The number and duration of these random fixations was identical with those produced by all 40 subjects taken together. Table 4 shows the random fixation data for all five features. As before, finger and hand films did not include the elbow and shoulder features, so no fixations went there. As predicted, the distribution of random fixations is homogeneous, showing obvious differences from that generated by human subjects and thus

Table 4

The results of the automatically generated random fixations, for each feature in the stimulus

Feature fixated on	Type of stimulus film		
	Finger	Hand	Arm
Fingers	13.7 ± 1.1	10.5 ± 0.6	7.1 ± 0.9
Palm	14.1 ± 0.9	9.9 ± 0.7	7.1 ± 0.8
Wrist	13.5 ± 0.9	10.1 ± 0.6	6.8 ± 0.5
Elbow	–	–	13.2 ± 1.4
Shoulder	–	–	12.6 ± 1.2

Table 5

The results of the automated analysis showing the average fixation duration with standard error (in ms) for each feature in the stimulus

Feature fixated on	Type of stimulus film		
	Finger	Hand	Arm
Fingers	629.2 ± 26.0	314.0 ± 17.3	389.4 ± 11.5
Palm	568.2 ± 27.7	366.3 ± 23.9	453.6 ± 26.5
Wrist	398.3 ± 25.9	322.8 ± 16.6	370.5 ± 15.3
Elbow	–	–	388.5 ± 14.6
Shoulder	–	–	338.2 ± 17.2

demonstrating the significance of our results. This analysis verifies that our eye-tracking results are not biased by our analysis methods, and that there is nothing inherent in the video sequences that produces clustering of fixations on the particular features over time. Note that the columns here need not add up to a 100% because we introduced the described target areas and some fixations could not be attributed to any of the salient features within those areas.

3.2. Duration of fixations

Automated analysis measured the duration of fixations on fingers, palm, wrist, elbow, and shoulder. Table 5 presents the average duration in ms with the rows indicating the salient points for stimulus type. Fig. 3 gives a visual illustration of the data: it shows a single frame from an arm movement video, with the locations of the dots indicating fixations and their radii indicating fixation duration.

The durations were entered into a one-factorial ANOVA with repeated measures. In finger movement films, distribution of gaze position varied significantly between the fingers, palm, wrist ($F = 42.95$; $p < 0.0001$) with signifi-
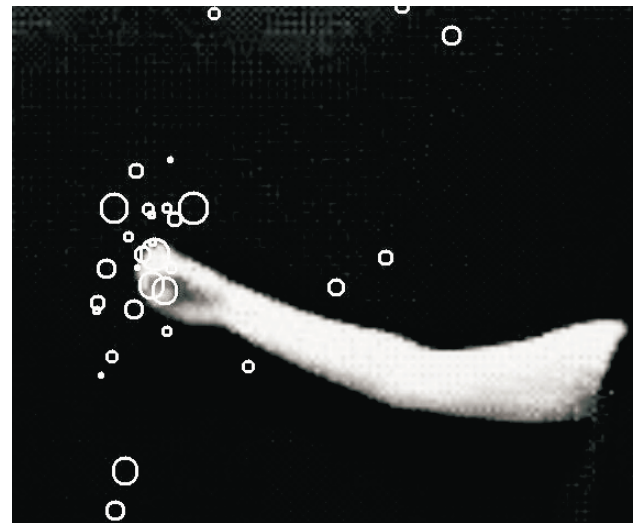

Fig. 3. An example data frame from an arm movement video; the dots mark the positions of the subjects' fixations, the dot radii indicate fixation duration.

cant differences in duration between all of the above three levels. Hand movement films revealed a significant main effect as well ($F = 6.71$; $p < 0.002$), but only the data contrasting fingers versus palm and palm versus wrist were significant. Arm movement films also demonstrated significant differences ($F = 7.68$; $p < 0.0001$). Fingers, wrist, and elbow did not induce different fixation durations but the palm and shoulder differences were significant: palm fixations were longer, shoulder fixations shorter. Finally, no delay was found between the hand position in the arm movement films and the subjects' corresponding gaze position. This is not surprising, since the movement speed in the stimuli was slow enough to allow for smooth pursuit [77].

3.3. Fixation transitions between features

We measured the 'probabilities of transition' between all pairs of features for the three different types of stimulus films in order to assess whether some transitions are more likely than others. As shown in the figures below, the data indicate no significant preferred direction of saccades. Furthermore, the data confirm a strong tendency to fixate toward the end-point, i.e., the hand, in case of the arm videos, and the fingers the rest of the videos. This further supports the fixation location and duration data; subjects tend to fixate at the end-point and once there, to remain there for some time.

Figs. 4–6 demonstrate the data in pictorial form. For each of the three types of stimulus films (finger, hand, and arm), we demonstrate the state transition diagram that shows how probable/likely fixation transitions are between each of the features. The probabilities are expressed in percent values. The highlighted areas indicate the feature being fixated on: fingers, palm, or wrist in finger and hand films, and fingers, palm, wrist, elbow, or shoulder in
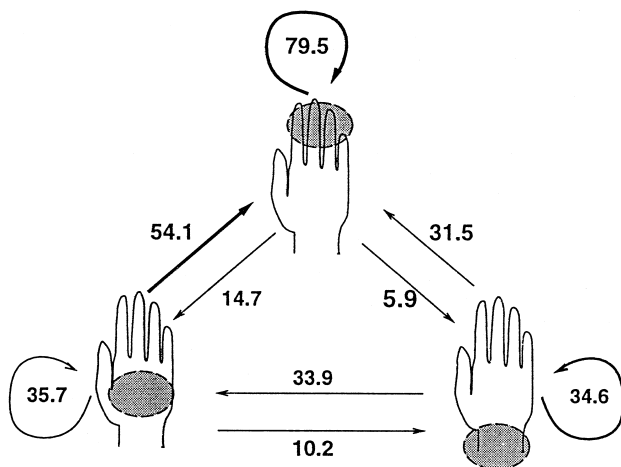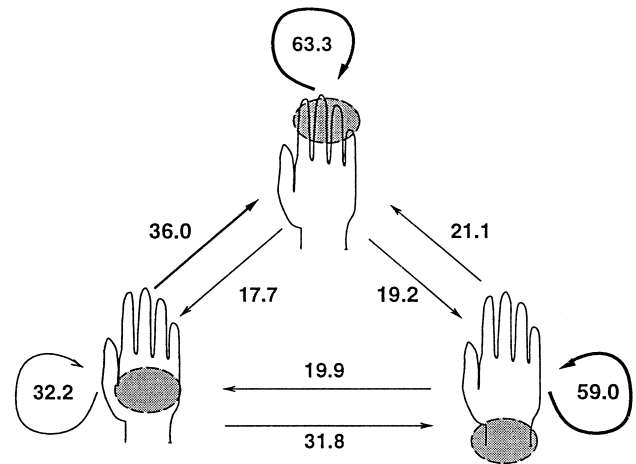

Fig. 5. Transition probabilities in hand films. The highlighted areas indicate the feature being fixated on: fingers, palm, or wrist. The numbers on the arrows indicate the percent probability of saccading in that direction. Darker arrows indicate most likely transitions.

arm films. The numbers on the arrows indicate the transition probability, i.e., the likelihood of saccading in that direction. The most likely transitions for each type of stimulus film are also marked with darker arrows. The data demonstrate the subjects' tendency to fixate on the hand region. The arm video transitions also show that subjects have a tendency to, when not fixating on the hand, move toward the hand or remain in the region of the fixated feature, rather than fixate between non-hand features.

3.4. Temporal distribution of fixations

Another interesting question to ask is whether the observed fixation pattern changes over the course of stimulus presentation. We addressed this question by dividing the video sequences into two segments and comparing the results of one-factorial ANOVA with repeated measures for each type of movement film (finger, hand, and arm). We found that the data show differences in the subjects' distribution of fixations between the two halves of the videos. Since the informational content of the videos is designed to be relatively uniform, i.e., the presented movements do not change in content, nature, or richness throughout the duration of each video, the observed difference must stem from other factors.

To examine these differences, we performed analyses of variance for all the combinations of features and stimulus types. The analyses were based on the mean values for each of the 40 subjects. Table 6 demonstrates the composite, consisting of 11 analyses, indicating the average percent of total viewing time spent on each feature during the 1st and 2nd half of the videos. The associated statistics are included in the table to indicate the significance of the observed differences.

In finger movement stimuli, the fingertips attracted fewer fixations during the 1st half than during the 2nd half

Fig. 4. Transition probabilities in finger films. The highlighted areas indicate the feature being fixated on: fingers, palm, or wrist. The numbers on the arrows indicate the percent probability of saccading in that direction. Darker arrows indicate most likely transitions.
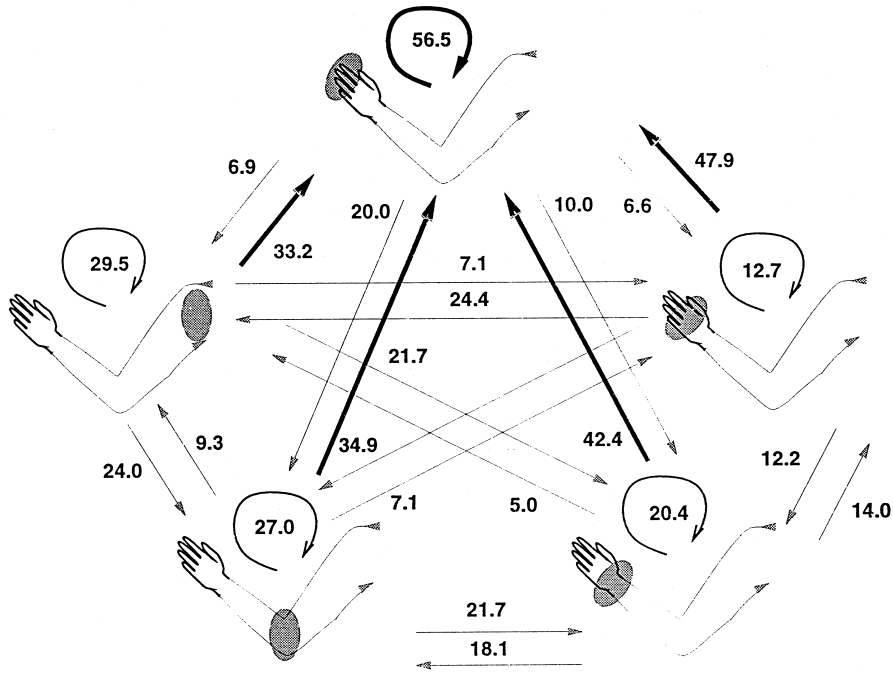
Fig. 6. Transition probabilities in arm films. The highlighted areas indicate the feature being fixated on: fingers, palm, wrist, elbow, or shoulder. The numbers on the arrows indicate the percent probability of saccading in that direction. Darker arrows indicate most likely transitions.

of the videos. Conceivably, shortly after the start of a video sequence subjects 'scanned' the whole scene and subsequently directed their attention at the fingertips as the most salient feature. In hand movement stimuli, most fixations in the 1st half are aimed at the fingertips, while during the 2nd half the majority shifts to the wrist. Similarly, arm movement films show a weak shift of attention from the fingertips in the first part, to the palm and the shoulder in the 2nd half.

If we postulate that the end-point of the arm is the source of the majority of information about the task being demonstrated, and thus the most relevant for imitation, then we can explain the data above as follows. In finger movement stimuli films, the screen is filled with the image of a hand with moving fingers. After an initial scan, subjects quickly learn that the fingers are the only moving feature and continue to fixate on them. In contrast, in the hand and arm movement films, multiple features can move simultaneously, i.e., the movement of the fingers is not the only source of information. Consequently, subjects fixate on the fingers when they are introducing novel information, but may become familiarized with the general form of

Table 6
The entries in each table cell indicate the percentages of 1st and 2nd half of the stimuli (shown as 1st/2nd) and the associated statistics

Feature fixated on	Type and portion of stimulus film		
	Finger (1st/2nd)	Hand (1st/2nd)	Arm (1st/2nd)
Fingers	62.07/74.1	61.90/35.16	46.53/41.19
	$F = 49.91$; $p < 0.0001$	$F = 177.53$; $p < 0.0001$	$F = 7.29$; $p = 0.0102$
	MSe = 0.00671	MSe = 0.00806	MSe = 0.00779
Palm	23.81/16.47	14.81/21.07	7.34/9.29
	$F = 42.38$; $p < 0.0001$	$F = 19.81$; $p = 0.0001$	$F = 6.12$; $p = 0.0178$
	MSe = 0.00254	MSe = 0.00396	MSe = 0.00125
Wrist	15.12/9.52	23.29/43.77	15.09/15.57
	$F = 15.15$; $p = 0.0004$	$F = 161.66$; $p < 0.0001$	not significant
	MSe = 0.00415	MSe = 0.00519	
Elbow	–	–	24.49/22.66
	–	–	not significant
	–	–	
Shoulder	–	–	6.56/11.29
	–	–	$F = 16.59$; $p = 0.0002$
	–	–	MSe = 0.00270

Table 7
Relative pupil dilation with standard error, shown in the number of pixels in the camera's image, for three types of stimuli and two imitation conditions

Type of stimulus film	Experimental condition	
	Imitation	No imitation
Finger	1874.0 ± 109.6	1680.7 ± 102.3
Hand	1934.4 ± 134.1	1832.4 ± 107.0
Arm	2188.4 ± 132.4	1858.7 ± 114.1

the stimulus and thus capable of briefly fixating elsewhere in the moving image without significant loss of information.

3.5. Imitation vs. no-imitation condition

Finally, we investigated whether the prospect of subsequent imitation influenced fixation behavior. Interestingly, neither fixation duration nor the percent of total time spent on each feature proved to be significant. However, another parameter, pupil dilation, was found to depend on the condition of imitation. Table 7 presents the average pupil dilation for the two conditions and the three different film stimuli.

A two-factorial ANOVA on pupil dilation revealed significant main effects for both condition of imitation ($F = 90.06$; $p < 0.0001$) and type of stimulus ($F = 11.4$; $p < 0.0001$). The differences between the three levels of the second factor were all significant. No interaction between the two factors was found ($F = 1.83$; $p = 0.167$). Since our stimuli were of constant brightness, the variation of pupil dilation can be assumed to be positively correlated with the subjects' cognitive activity [37]. In this context, the data reveal that higher stimulus complexity increases the subjects' attention, i.e., pupil dilation increases from finger to hand to arm movement films. Furthermore, the dilation is higher in the imitation case, implying that the task of subsequent imitation leads to higher cognitive activity during stimulus presentation.

In summary, our data show that subjects tend to fixate at the end of the moving manipulator, i.e., the hand, the finger(s), or the thumb, that pupil size increases with higher movement complexity and the imitation condition, and that the temporal pattern of fixation changes over the duration of stimulus presentation. In all cases, both fixation frequency and duration are maximized at the end-point; the results are significant across different image scales and levels of stimulus complexity. [1]

---

[1] We have conducted follow-up experiments at the ATR Human Information Processing Research Laboratory involving stimuli films featuring movements with a pointer. The data are currently being analyzed; preliminary results suggest that subjects tend to fixate and track the end-point of the held object.

4. Discussion

4.1. Relation to other eye-tracking work

Eye-tracking studies, performed on static images, have been applied to a variety of motor tasks, including demonstrations of behaviors such as boundary tracing [41] and learning skills from picture-text combinations [14]. Video is a better mode of motor task presentation, and eye-tracking work with video has lately been applied largely to the study of memory representations [5,29,71]. In particular, Ballard et al. [5] used a head-mounted eye-tracker to record saccades during a block copying task and, in subsequent work [6], an assembly task. While the latter involved an imitation component, the study focused on working memory limitations as manifested through information-gathering fixation behavior. The experiments we report on are the first to use eye-tracking in video to study imitation of human movement. In general, eye-tracking experiments to date have been aimed at addressing perceptual behavior, either in itself or as a means of studying memory, while we have focused on the connection between the perceptual and motor systems.

4.2. Relation to selected theories of imitation

Much work has been done in classifying the types and stages of imitation [15,62,72], as well as in the study of its role in social behavior [7,8,49] and more recently, in evolution of communication [21,34,66]. In all cases, imitation is recognized as a complex form of learning and its evolution and relation to other forms of learning are being explored [53,54]. The issue of sensori–motor integration in imitation, addressed here, is also a key focus of Meltzoff and Moore's work [47,48], which proposes the idea that imitation is a fundamental human capability, found in newborns, and based on an innate link between the perceptual and motor systems [49]. This link enables young children to imitate facial expressions and hand movements without visual feedback. The authors hypothesize supramodal representations of human movements and postures as a part of the innate imitation mechanism [50,51]. The data we have described is complementary with this work, as well as with a growing body of neuroscience evidence for generative perception of movement, whether it be specific for imitation, or as a general aspect of visuo-motor processing.

4.3. Theories of motor control

The data we report allow for postulating that people derive the specification of a movement task largely by tracking the trajectory of the end-point (hand or pointer). This result is intuitive: the end-point of a manipulator typically carries the most information about the task. Within this view, the precise definition of 'end-point' depends on

the task itself, and can vary from the tip of a pointer or a thumb, to the entire hand including the wrist, as our data show. The observed results are consistent with the research supporting end-point motion planning [55,25], in that the perceptual behavior can be interpreted to be acquiring the information for forming an end-point trajectory plan. The data are also related to results by Epelboim et al. [23], who found that eye-movements of subjects increased when tested on looking at targets while tapping vs. just looking; the difference is presumably due to the heavier memory load in the former case. Our lack of difference between the imitation and no-imitation conditions supports the postulate of a general movement observation strategy that involves a generative process.

Consistent fixation on the stimulus' end-point, even if the stimulus is defined over the entire arm, has several important implications for motor control and perception. First, it appears that people are capable of filling-in details about posture and control of the arm from very little information, indicating that they may map internal movement primitives onto the observed behavior. To roughly assess the accuracy of the subjects' perception, we showed the subjects four pairs of arm movement videos and asked them to decide in each case whether the videos were the same or different. Eleven subjects were each given four instances of same-different decisions, resulting in a total of 44 tasks and only nine errors (six incorrect judgments of 'same', three incorrect judgments of 'different'). Thus the data show that in the majority of cases (79.55%), the subjects answered the same-different question correctly, thus demonstrating the ability to discriminate movements that were not fixated on while watching the arm-movement videos. Filling-in movement details would require internal models involving the kinematics of the arm in order to transform the observed task from external space into joint space. This result is consistent with the work of Johansson [36] demonstrating human capability to recognize biological motion from a small number of structured visual cues.

Williams [78] describes a related imitation experiment in which subjects were shown videos of simulated throwing motions using a point-light indicating the kinematics of the model arm. His analysis showed that all subjects produced the correct sequences (consistent with Johansson [36]) and spatial parameters, but did not correctly model the timing of the present movement. Instead, they tended to impose their own timing. This result could be consistent with the existence of motor primitives that encode movement timing, and thus bias movement production in imitation. The notion of motor primitives has been suggested in lower animals, based on convergent force fields found in frogs and more recently in rats [10,56].

4.4. Generative perception and motor primitives

A mapping from observed movement to internal movement primitives implies that there exists a recognition system for movement which makes use of related principles as the system for movement generation. This postulate directly connects to several lines of research suggesting that neural information processing may be based on the interaction between recognition and generative models in the brain.

In neuroscience, combined data from mental chronometry, autonomic responses, and cerebral blood flow experiments demonstrate that imagined and executed actions share the same neural substrate [13,17,35]. These results point toward a sensori–motor connection between movement imagination and the early stages of movement preparation. Analogously, a number of imaging experiments (e.g., Refs. [24,34]) provide data in support of the connection between observed and executed movement. Measuring motor evoked potential (MEP) levels in the motor cortex of normal subjects under different movement observation and static object observation conditions shows significant MEP increase during movement observation, suggesting the existence of a system for matching action observation and execution [24]. Earlier work found analogous results in monkeys observing a human experimenter [20]. So-called 'mirror neurons', located in the pre-motor cortex, are involved, and fire in response to observed movement, based on data from monkeys watching other monkeys, and consistent with previous data on monkeys watching humans [66].

Psychophysical experiments have provided data suggesting similar conclusions about the connection between observed and executed movement. A study of closed-eye copying sine/cosine wave functions on paper showed no improvement after immediate rehearsal and no effect from distractions. This may imply the absence of an intermediate level between perception and motor control and the presence of a generative perception module [75]. Earlier experiments in hand-pointing adjustments without visual feedback provide consistent conclusions [57]. Furthermore, experiments in form perception of geometric trajectories [74] demonstrate some perceptual biases toward what may be commonly perceived modes of movement generation.

Carey [12] briefly summarizes the various work demonstrating 'monkey see, monkey do' premotor cells which fire in response to perceiving specific hand movements. Recently, Altschuler et al. [2] describe 'person see, person do' EEG-wave activity suggesting that the mu wave responding to visual input of movement may be 'the human electrophysiologic analog to a population of neurons in area F5 of monkey premotor cortex'. Arbib and Rizzolatti [3] even postulate the role of such an 'observation/execution matching' system as a gestural basis for the evolution of language.

A common internal representation used to both interpret and generate movement would explain the lack of difference between fixation behavior in the imitation and no-imitation conditions. It should also reflect itself in the correctness of imitation, i.e., the faithfulness with which

movements are recreated. We expect to find results showing consistent errors, complementary for example to those shown in experiments that studied the subjects' ability to mimic the rotary movement of hand-held three-dimensional objects by changing their orientation [73]. We are currently analyzing subjects' motor data in order to better address the connection between perception and motor control.

4.5. Theories of movement perception

Consistent with the notion of primitives is data by Perrett et al. who performed a series of studies of movement perception, with the overall suggestion of the existence of specialized neural detectors (and predictors) for specific postures and movements as well as goal-oriented behavior. They describe data from the superior temporal sulcus cells of the macaque showing selectivity for movement type and stimulus form, i.e., for specific body and body part movements [61], demonstrate similar view-independent results, implying the potential of viewer-centered and goal-centered descriptions of the movement [58], describe population cells involved in the recognition of specific actions of others [59], and show that expectation of movement indirectly correlates with the amount of resulting neural firing [60].

Computational theories and implementations of movement perception also deal with classifying movement, largely by employing time-series analysis tools, specifically so-called Hidden Markov Models [65], originally used to automatically segment and recognize speech [64], but more recently generalized to human and robot movement analysis. For example, Pook and Ballard [63] use it to analyze a video of a robot flipping a fried egg, and Brand et al. [11] describe a coupled Hidden Markov Model (HMM) approach to recognizing processes such as object assembly. Lee and Xu [43] use a similar HMM-based approach to recognize several hand-shaped letters while Yang et al. [79] use the HMM representation to model human skill learning by observation and skill transfer to articulated robot arms. They demonstrate recognition and copying of written digits as well as part replacement by teleoperation.

4.6. Computational implementations of imitation

The existence of movement primitives could significantly simplify the process of automated movement perception as well as generation, and is thus being explored by Matarić [45] and Schaal [69]. More generally, a great deal of work in the field of neural computation and connectionism has addressed the problem of modeling either perception or motor control, but work combining the two is rare [4], and work addressing imitation per se even more so. Exceptions include work by Kawato [39] and related implementations [26,52] of the bi-directional theory

of motor control applied to the task of learning dynamic tasks by observation. Wada et al. [76] describe the computational theory in which movement generation and recognition are 'two aspects of a single function', and discuss it in the context of handwriting and speech. The proposed model of the integrated perception and generation process consists of three stages: the lowest level is a neural network, the middle level is a 'via-point' estimation algorithm, and the top level is a symbolic pattern recognition scheme. The idea of via-points has also been successfully applied to artificial demonstrations of learning by imitation, in the domain of Kendama, a Japanese game of impaling a ball on a stick [52].

The first robotics work to address imitation was focused on assembly from observation, where series of visual images of human movement were segmented, interpreted, and then repeated by a robotic arm [33,32,42,31,38]. While this work was aimed at repeating behavior, more recent efforts, including our own, have been oriented toward analyzing the underlying mechanisms of imitation and modeling those on artificial systems [44,69]. For example, Demiris and Hayes [18] have implemented demonstrations of artificial imitation in two kinds of robots; in one set of experiments, a robotic head observes and imitates the head movements of a human demonstrator [19] by incorporating the idea of visual feature detectors akin to specialized neurons found by Perrett et al. [58,61] and an innate visuo-motor map suggested by Meltzoff [46] and Meltzoff and Moore [49]. In another experiment, the imitation mechanism is used as the basis for maze learning on a mobile platform [28].

More abstract computational work, which combines detection and generation, by Hinton et al. [30] and by Dayan et al. [16] demonstrates a connectionist computational model combining a top-down generative and a bottom-up recognition systems in a so-called 'wake–sleep' algorithm. In the 'wake' phase, the neural network is driven by the recognition process, while the generative component is adapted to better reconstruct the incoming data. In the 'sleep' phase, the network is driven by the generative process, while the recognition component is analogously adapted. The goal of the system is to construct economical 'minimal description length' representations of the input.

Imitation continues to be addressed by different research communities. As suggested in the introduction, the goal of this work has been to provide additional insight into the connection between perceptual and motor systems in human imitation, in hopes of bridging some of those relevant fields in order to aid that study.

References

[1] B. Abernethy, Expertise, visual search and information pick-up in squash, Perception 19 (1990) 63–78.

[2] E.L. Altschuler, A. Vankov, V. Wang, V.S. Ramachandran, J.A. Pineda, Person see, person do: human cortical electrophysiological correlates of monkey see monkey do cells, in: Proceedings of the Neuroscience Society Meeting, New Orleans, 1997.

[3] M. Arbib, G. Rizzolatti, Neural expectations: a possible evolutionary path from manual skills to language, Communication and Cognition 29 (2–4) (1996) 393–424.

[4] M.A. Arbib, Visuomotor coordination: neural models and perceptual robotics, in: Visuomotor Coordination Amphibians, Comparisons, Models and Robots, Plenum, 1989, pp. 121–171.

[5] D.H. Ballard, M.M. Hayhoe, F. Li, S.D. Whitehead, Hand–eye co-ordination during sequential tasks, in: Proceedings of the Royal Society of London, 1992.

[6] D.H. Ballard, M.M. Hayhoe, P.K. Pook, R.P.N. Rao, Deictic codes for embodiment of cognition, Behavior and Brain Science, to appear, 1997.

[7] A. Bandura, Social Learning Theory, Prentice-Hall, Englewood Cliffs, NJ, 1977.

[8] A. Bandura, R.H. Walters, Social Learning and Personality Development, Holt, Rinehart and Winston, New York, 1963.

[9] B.I. Bertenthal, Origins and early development of perception, action and representation, Annual Review of Psychology 47 (1996) 431–459.

[10] E. Bizzi, F.A. Mussa-Ivaldi, S. Giszter, Computations underlying the execution of movement: a biological perspective, Science 253 (1991) 287–291.

[11] M. Brand, N. Oliver, A. Pentland, Coupled hidden markov models for complex action recognition, In: Proceedings, CVPR, IEEE Press, 1997, pp. 994–999.

[12] D.P. Carey, 'Monkey see, monkey do' cells, Current Biology 6 (9) (1996) 1087–1088.

[13] D.J. Crammond, Motor imagery: never in your wildest dream, Trends in Neuroscience 20 (2) (1997) 54–57.

[14] R. Daugs, K. Blische, N. Olivier, Scanning habits and visuo-motor learning, in: J.K. O'Regan, A. Levy-Schoen (Eds.), Eye Movements, From Physiology to Cognition: Selected/Edited Proceedings of the Third European Conference on Eye Movements, 1978, pp. 323–332.

[15] J.M. Davis, Imitation: a review and critique, in: Bateson and Klopfer (Eds.), Perspectives in Ethology, Vol. 1, Plenum, 1973, pp. 43–72.

[16] P. Dayan, G.E. Hinton, R.M. Neal, R.S. Zemel, The helmholtz machine, Neural Computation 7 (5) (1995) 889–904.

[17] J. Decety, Do imagined and executed actions share the same neural substrate?, Cognitive Brain Research 3 (1996) 87–93.

[18] J. Demiris, G. Hayes, Do robots ape? in: Proceedings of the AAAI Fall Symposium on Socially Intelligent Agents, Technical Report FS-97-02, AAAI Press, Cambridge, MA, 1997, pp. 28–30.

[19] J. Demiris, S. Rougeaux, G.M. Hayes, L. Berthouze, Y. Kuniyoshi, Deferred imitation of human head movements by an active stereo vision head, in: Proceedings of the 6th IEEE International Workshop on Robot Human Communication (RoMan97), Sendai, Japan, IEEE Press, 1997.

[20] G. di Pellegrino, L. Fadiga, L. Fogassi, G. Rizzolatti, Understanding motor events: a neurophysiological study, Experimental Brain Research 91 (1992) 176–180.

[21] M. Donald, Precis of origins of the modern mind: three stages in the evolution of culture and cognition, Journal of Behavioral and Brain Sciences 16 (1993) 737–791.

[22] J. Elkins, The Object Stares Back: On the Nature of Seeing, Simon & Schuster, New York, 1996.

[23] J. Epelboim, R.M. Steinman, E. Kowler, M. Edwards, Z. Pizlo, C.J. Erkelens, H. Collewijn, The function of visual search and memory in sequential looking tasks, Vision Research 35 (23–24) (1995) 3401–3422.

[24] L. Fadiga, L. Fogassi, G. Pavesi, G. Rizzolatti, Motor facilitation during action observation: a magnetic stimulation study, J. Neurophysiol. 73 (6) (1995) 2608–2611.

[25] T. Flash, N. Hogan, The coordination of the arm movements: an experimentally confirmed mathematical model, Journal of Neuroscience 7 (1985) 1688–1703.

[26] H. Gomi, M. Kawato, Equilibrium-point control hypothesis examined by measured arm stiffness during multijoint movement, Science 272 (1996) 117–120.

[27] M.A. Goodale, L.S. Jakobson, A.D. Milner, D.I. Perrett, P.J. Benson, J.K. Hietanen, The nature and limits of orientation and pattern processing supporting visuomotor control in a visual form agnosic, Journal of Cognitive Neuroscience 6 (1) (1994) 46–56.

[28] G. Hayes, J. Demiris, A robot controller using learning by imitation, in: A. Borkowski, J.L. Crowley (Eds.), Proceedings of the International Symposium on Intelligent Robotic Systems, Grenoble, France, LIFIA-IMAG, 1994, pp. 198–204.

[29] M.M. Hayhoe, D.H. Ballard, S.D. Whitehead, Memory use during hand-eye coordination, in: Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society, Boulder, Colorado, 1993, pp. 534–538.

[30] G.E. Hinton, P. Dayan, B.J. Frey, R.M. Neal, The wake–sleep algorithm for unsupervised neural networks, Science 268 (1995) 1158–1161.

[31] G.E. Hovland, P. Sikka, B.J. McCarragher, Skill acquisition from human demonstration using a hidden markov model, in: Proceedings, IEEE International Conference on Robotics and Automation, Minneapolis, MN, 1996, pp. 2706–2711.

[32] K. Ideuchi, M. Kawade, T. Suehiro, Assembly task recognition with planar, curved, and mechanical contacts, in: Proceedings of IEEE International Conference on Robotics and Automation, Atlanta, GA, 1993.

[33] K. Ikeuchi, T. Suehiro, P. Tanguy, M. Wheeler, Assembly plan from observation, Technical report, Carnegie Mellon University Robotics Institute Annual Research Review, 1990.

[34] M. Jeannerod, M.A. Arbib, G. Rizzolatti, H. Sakata, Grasping objects: the cortical mechanisms of visuomotor transformation, Trends in Neuroscience 18 (1995) 314–320.

[35] M. Jeannerod, J. Decety, Mental motor imagery: a window into the representational stages of action, Current Opinion in Neurobiology 5 (1995) 727–732.

[36] G. Johansson, Studies on visual perception of locomotion, Perception 6 (1977) 365–376.

[37] D. Kahneman, Attention and Effort, Prentice-Hall, Englewood-Cliffs, NJ, 1973.

[38] M. Kaiser, Transfer of elementary skills via human–robot interaction, Adaptive Behavior 5 (3–4) (1997) .

[39] M. Kawato, Bi-directional theory approach to integration, in: T. Inui, J. McClelland (Eds.), Attention and Performance XVI, MIT Press, Cambridge, MA, 1996, pp. 335–367.

[40] B. Kolb, B. Milner, Performance of complex arm and facial movements after focal brain lesions, Neurophychologia 19 (1981) 491–504.

[41] H. Kudo, K. Uomori, N. Yamada, N. Oinishi, N. Sugie, Binocular fixation-point shifts induced by a limb occlusion, in: Proceedings of the IEEE/EMBS, 1992, pp. 1670–1671.

[42] Y. Kuniyoshi, M. Inaba, H. Inoue, Learning by watching: extracting reusable task knowledge from visual observation of human performance, IEEE Transactions on Robotics and Automation 10 (6) (1994) 799–822.

[43] C. Lee, Y. Xu, Online, interactive learning of gestures for human/robot interfaces, in: Proceedings, IEEE International Conference on Robotics and Automation, Minneapolis, MN, Vol. 4, 1996, pp. 2982–2987.

[44] M.J. Mataric, A computation model of motor imitation, in: AAAI-94 Fall Symposium on the Control of the Physical World by Intelligent Agents, 1994.

[45] M.J. Mataric, Behavior-based robotics as a tool for synthesis of artificial behavior and analysis of natural behavior, Trends in Cognitive Science 2 (3) (1998) 82–87.

[46] A.N. Meltzoff, Molyneux's babies: cross-modal perception, imitation and the mind of the preverbal infant, in: N. Eilan, R. McCarthy, B. Brewer (Eds.), Spatial Representation: Problems in Philosophy and Psychology, Blackwell, 1993, pp. 219–235.

[47] A.N. Meltzoff, M.K. Moore, Newborn infants imitate adult facial gestures, Child Development 54 (1983) 702–709.

[48] A.N. Meltzoff, M.K. Moore, Imitation of facial and manual gestures by human neonates, Science 198 (1988) 75–78.

[49] A.N. Meltzoff, M.K. Moore, Imitation, memory, and the representation of persons, Infant Behavior and Development 17 (1994) 83–99.

[50] A.N. Meltzoff, M.K. Moore, Infants' understanding of people and things: From body imitation to folk psychology, in: J.L. Bermudez, A. Marcel, N. Eilan (Eds.), The Body and the Self, MIT Press/Bradford Books, 1995, pp. 44–69.

[51] A.N. Meltzoff, M.K. Moore, Explaining facial imitation: a theoretical model, Early Development and Parenting 6 (2) (1997) 157.1–157.14.

[52] H. Miyamoto, S. Schaal, F. Gandolfo, H. Gomi, Y. Koike, R. Osu, E. Nakano, Y. Wada, M. Kawato, A kendama learning robot based on bi-directional theory, Neural Networks 9 (6) (1996) .

[53] B.R. Moore, Avian movement imitation and a new form of mimicry: tracing the evolution of a complex form of learning, Behavior 122 (1992) 614–623.

[54] B.R. Moore, The evolution of imitative learning, in: C.M. Heyes, B.G. Galef (Eds.), Social Learning in Animals: The Roots of Culture, Academic Press, New York, 1996, pp. 245–265.

[55] P. Morasso, Spatial control of arm movements, Experimental Brain Research 42 (1981) 223–227.

[56] F.A. Mussa-Ivaldi, S.F. Giszter, E. Bizzi, Linear combinations of primitives in vertebrate motor control, Proceedings of the National Academy of Sciences 91 (1994) 7534–7538.

[57] D. Pelisson, M.A. Goodale, C. Prablanc, Adjustments of hand pointings to visual targets do not need visual reafference from the moving limb, in: J.K. O'Regan, A. Levy-Schoen (Eds.), Eye Movements, From Physiology to Cognition: Selected/Edited Proceedings of the Third European Conference on Eye Movements, 1978, pp. 115–121.

[58] D.I. Perrett, M.H. Harries, R. Bevan, S. Thomas, P.J. Benson, A.J. Mistlin, A.J. Chitty, J.K. Hietanen, J.E. Ortega, Frameworks of analysis for the neural representation of animate objects and actions, Journal of Experimental Biology 146 (1989) 87–113.

[59] D.I. Perrett, M.H. Harries, A.J. Mistlin, A.J. Chitty, Recognition of objects and actions: frameworks for neuronal computation and perceptual experience, in: O. Guthrie (Ed.), Higher Order Sensory Processing, Manchester Univ. Press, 1989.

[60] D.I. Perrett, M.H. Harries, A.J. Mistlin, J.K. Hietanen, P.J. Benson, R. Bevan, S. Thomas, M.W. Oram, J. Ortega, K. Brierley, Social signals analyzed at the single cell level: someone is looking at me, something touched me, something moved!, International Journal of Comparative Psychology 4 (1) (1990) 25–55.

[61] D.I. Perrett, P.A.J. Smith, A.J. Mistlin, A.J. Chitty, A.S. Head, D.D. Potter, R. Broennimann, A.D. Milner, M.A. Jeeves, Visual analysis of body movements by neurones in the temporal cortex of the macaque monkey: a preliminary report, Behavioral Brain Research 16 (1985) 153–170.

[62] J. Piaget, Play, Dreams and Imitation in Children, W.W. Norton, New York, 1962.

[63] P.K. Pook, D.H. Ballard, Recognizing teleoperated manipulations, in: Proceedings of IEEE International Conference on Robotics and Automation, Atlanta, GA, Vol. 2, May 1993, pp. 578–585.

[64] L.R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, IEEE 2 (1989) 257–286.

[65] L.R. Rabiner, B.H. Juang, Introduction to hidden markov models, IEEE ASSP Magazine, 1986, pp. 4–16.

[66] G. Rizzolatti, L. Gadiga, V. Gallese, L. Fogassi, Premotor cortex and the recognition of motor actions, Cognitive Brain Research 3 (1996) 131–141.

[67] G.J. Savelsbergh, H.T.A. Whiting, J.R. Pijpers, A.A.M. van Santvoord, The visual guidance of catching, Experimental Brain Research 93 (1993) 148–156.

[68] G.J.P. Savelsvertgh, R.J. Bootsma, Perception–action coupling in hitting and catching, International Journal of Sport Psychology 25 (1994) 331–343.

[69] S. Schaal, Learning from demonstration, in: M.C. Mozer, M. Jordan, T. Petsche (Eds.), Advances in Neural Information Processing Systems 9, The MIT Press, 1997, pp. 1040–1046.

[70] D.M. Stampe, Heuristic filtering and reliable calibration methods for video-based pupil-tracking systems, Behavioral Research Methods, Instruments, and Computers 25 (1993) 137–142.

[71] W.H. Teichner, D. LeMaster, P.A. Kinney, Eye movements during inspection and recall, in: J.W. Senders, D.F. Fisher, R.A. Monty (Eds.), Eye Movements and the Higher Psychological Functions, 1978, pp. 259–277.

[72] M. Tomasello, A.C. Kruger, H.H. Rather, Cultural learning, The Journal of Behavioral and Brain Sciences 16 (3) (1993) 495–552.

[73] H.A.H.C. Van Veen, Visually Acquired Information about Rotating Objects, PhD thesis, Universiteit Utrech, 1996.

[74] P. Viviani, N. Stucchi, The effect of movement velocity on form perception: geometric illusions in dynamic displays, Perception and Psychophysics 46 (3) (1989) 266–274.

[75] S. Vogt, Imagery and perception–action mediation in imitative actions, Cognitive Brain Research 3 (1995) 79–86.

[76] Y. Wada, Y. Koie, E. Vatikiotis-Bateson, M. Kawato, A computational theory for movement pattern recognition based on optimal movement pattern generation, Biological Cybernetics 73 (1995) 15–25.

[77] J.D. Weir, J.F. Stein, R.C. Miall, Cues and control strategies in visually guided tracking, Journal of Motor Behavior 21 (3) (1989) 185–204.

[78] J.G. Williams, Visual demonstration and movement production: effects of timing variations in a model's action, Perceptual and Motor Skills 68 (1989) 891–896.

[79] J. Yang, Y. Xu, C.S. Chen, Human action learning via hidden markov model, IEEE Transactions on Systems Man, and Cybernetics—Part A: Systems and Humans 27 (1) (1997) 34–44.