Self-recognition of avatar motion: how do I know it's me?

Richard Cook[1],*, Alan Johnston[1],[2] and Cecilia Heyes[3],[4]

[1]*Cognitive, Perceptual and Brain Sciences, Division of Psychology and Language Sciences, and* [2]*CoMPLEX, University College London, London, UK*
[3]*All Souls College, and* [4]*Department of Experimental Psychology, University of Oxford, Oxford, UK*

When motion is isolated from form cues and viewed from third-person perspectives, individuals are able to recognize their own whole body movements better than those of friends. Because we rarely see our own bodies in motion from third-person viewpoints, this self-recognition advantage may indicate a contribution to perception from the motor system. Our first experiment provides evidence that recognition of self-produced and friends' motion dissociate, with only the latter showing sensitivity to orientation. Through the use of selectively disrupted avatar motion, our second experiment shows that self-recognition of facial motion is mediated by knowledge of the local temporal characteristics of one's own actions. Specifically, inverted self-recognition was unaffected by disruption of feature configurations and trajectories, but eliminated by temporal distortion. While actors lack third-person visual experience of their actions, they have a lifetime of proprioceptive, somatosensory, vestibular and first-person-visual experience. These sources of contingent feedback may provide actors with knowledge about the temporal properties of their actions, potentially supporting recognition of characteristic rhythmic variation when viewing self-produced motion. In contrast, the ability to recognize the motion signatures of familiar others may be dependent on configural topographic cues.

Keywords: self-recognition; avatar; facial motion; inversion effect; mirror neurons

1. INTRODUCTION

People are better at recognizing their own walking gaits and whole body movements than those of friends, even when stimuli are viewed from third-person perspectives [1–4]. This self-recognition advantage is surprising because walking gaits and whole body movements are 'perceptually opaque' [5]; they cannot be viewed directly by the actor from a third-person perspective. While we sometimes view our movements in mirrors or video recordings, we see our friends' movements from a third-person perspective more often than our own. Therefore, if action perception depended solely on visual experience [6–9], one would expect the opposite result—superior recognition of friends' movements when viewed from third-person perspectives.

Superior self-recognition is important because it suggests that the motor system contributes to action perception [10–13]; that repeated performance of an action makes that action easier to recognize when viewed from the outside. However, while this implies that information about action execution can facilitate action recognition, it is unclear what kind of information plays this facilitating role, or how it is transferred from the motor to the perceptual system. The cues could be topographic—relating to the precise spatial configuration of limb positions and trajectories—or temporal—relating to the frequency and rhythm of key movement components. The transfer could depend on associative or inferential processes. An

associative transfer process would use connections between perceptual and motor representations established through correlated experience of executing and observing actions [14,15]. An inferential transfer process would convert motor programmes into view-independent visual representations of action without the need for experience of this kind [4,13,16]. If topographic cues are transferred from the motor to visual systems via an associative route, this raises the possibility that self-recognition is mediated by the same bidirectional mechanism responsible for imitation.

Here, we use markerless avatar technology to demonstrate that the self-recognition advantage extends to another set of perceptually opaque movements—facial motion. This is remarkable in that actors have virtually no opportunity to observe their own facial motion during natural interaction, but frequently attend closely to the facial motion of friends. Moreover, we show for the first time that while recognition of friends' motion may rely on configural topographic information, self-recognition depends primarily on local temporal cues.

Previous studies comparing recognition of self-produced and friends' actions have focused on whole body movements, employing point-light methodology [8] to isolate motion cues [1–4,17]. This technique is poorly suited to the study of self-recognition because point-light stimuli contain residual form cues indicating the actor's build and, owing to the unusual apparatus employed during filming, necessarily depict unnatural, idiosyncratic movements. In contrast, we used an avatar technique that completely eliminates form cues by animating a common facial form with the motion derived from different actors [18,19]. Because this technique does not require individuals to
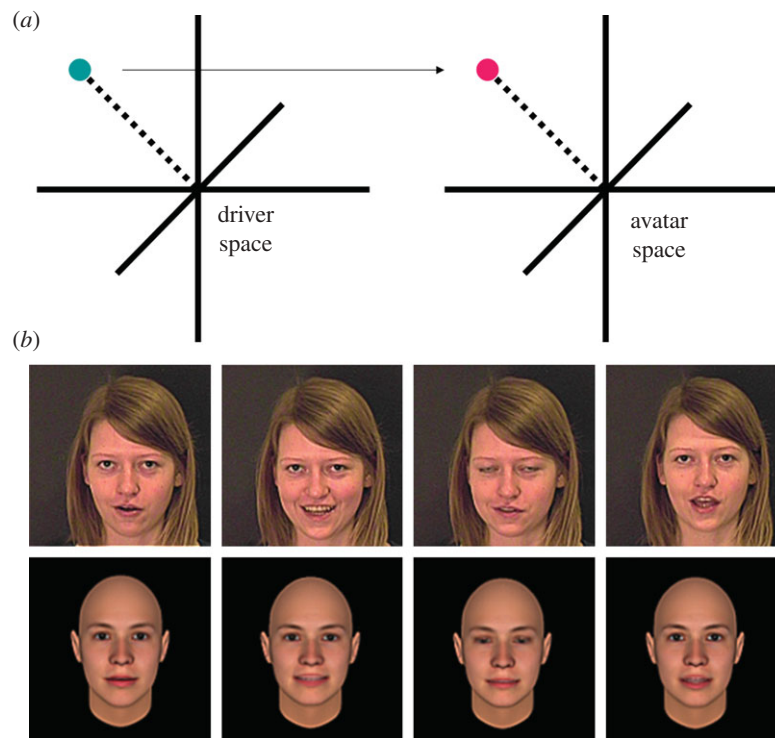
Figure 1. (*a*) Schematic of the animation process employed in the Cowe Photorealistic Avatar procedure. Principle components analysis (PCA) is used to extract an expression space from the structural variation present within a given sequence of images. This allows a given frame within that sequence to be represented as a mean-relative vector within a multi-dimensional space. If a frame vector from one sequence is projected into the space derived from another sequence, a 'driver' expression from one individual may be projected on to the face of another individual. If this is done for an entire sequence of frames, it is possible to animate an avatar with the motion derived from another actor. This technique was used to project the motion extracted from each actor's sequences onto an average androgynous head. (*b*) Examples of driver frames (top) and the resulting avatar frames (bottom) when the driver vector is projected into the avatar space. Example stimuli and a dynamic representation of the avatar space are available online as part of the electronic supplementary material accompanying this article.

wear markers or point-light apparatus during filming, it is also better able to capture naturalistic motion than the methods used previously.

2. EXPERIMENT 1

Experiment 1 sought to determine whether there is a self-recognition advantage for facial motion, and whether this advantage varies with the orientation of the facial stimuli. Visual processing of faces is impaired by inversion [20,21], and this effect is thought to be due to the disruption of configural cues [22–24]. If the recognition of self-produced facial motion is mediated by configural topographic information—cues afforded by the precise appearance of the changing face shape—the self-recognition advantage should be greater for upright than for inverted faces.

(a) *Methods*

Participants were 12 students (four male, mean age = 23.2 years) from the University of London comprising six same-sex friend pairs. Friends were defined as individuals of the same sex, who had spent a minimum of 10 h a week together during the 12 months immediately prior to the experiment [3]. Participants were of approximately the same ages and physical proportions.

Each member of the friendship pairs was filmed individually while recalling and reciting question and answer jokes [19]. The demands of this task—to recite the jokes from memory, while aiming to sound as natural as possible—drew the participants' attention away from their visual appearance. These naturalistic 'driver sequences' were filmed using a digital Sony video camera at 25 frames per second (FPS). Suitable segments for stimulus generation were defined as sections of 92 frames (3.7 s) containing reasonable degrees of facial motion, and in which the participant's gaze was predominantly fixated on the viewer. The majority of clips contained both rigid and non-rigid facial motion. Facial speech was also present in most, but exceptions were made when other salient non-rigid motion was evident.

Avatar stimuli were produced from this footage using the Cowe Photorealistic Avatar technique [25,26] (figure 1). The avatar space was constructed from 721 still images derived from Singular Inversions' FACEGEN MODELLER 3.0 by placing an approximately average, androgynous head in a variety of poses. These poses sampled the natural range of rigid and non-rigid facial motion, but were not explicitly matched to real images. The resulting image set included mouth variation associated with speech, variations of eye gaze, eye aperture, eyebrow position and blinking, variation of horizontal and vertical head position, head orientation and apparent distance from camera. Fourteen 3.7 s avatar stimuli were produced for each actor by projecting each of the 92 frames of the driver sequence into the avatar space, and
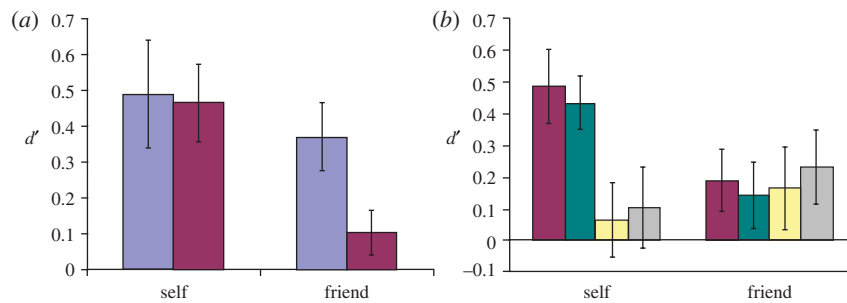
Figure 2. (*a*) Results from experiment 1. Whereas discrimination of friends' motion showed a marked inversion effect, participants' ability to discriminate self-produced motion was insensitive to inversion. (*b*) Results from experiment 2. When presented with inverted avatar stimuli, participants could correctly discriminate their own veridical motion (i.e. without any disruption) and sequences of anti-frames. However, when the temporal or rhythmic properties were disrupted either through uniform slowing, or random acceleration/deceleration, self-discrimination did not exceed chance levels. Error bars denote standard error of the mean in both figures. (*a*) Purple bars, upright; maroon bars, inverted. (*b*) Maroon bars, inverted veridical; green bars, anti-sequence; yellow bars, rhythm disrupted; grey bars, slowed.

converting the resulting vector into movie frames (figure 1). This process is described in full by Berisha *et al.* [25]. The resulting avatar stimuli were saved and presented in uncompressed audio-video-interleaved (AVI) format.

Friend pairs were required to complete a three alternative forced choice (3-AFC) recognition test. In each trial, participants were shown a single avatar stimulus, in an upright or inverted orientation, and were required to indicate whether the motion used to animate the head had been taken from themselves, their friend or a stranger. The stimuli derived from each actor appeared once as 'self', once as 'friend' and once as 'other'. The experiment was completed over two sessions: In session 1, participants completed a block of upright trials followed by an inverted block; in session 2, block order was reversed. Different strangers were allocated across the first and second sessions to ensure that effects were not artefacts of the particular stranger allocations.

Experimental trials began with a fixation dot presented for 750 ms, followed by an avatar stimulus looped to play twice. Following stimulus offset 'self, friend or other?' appeared at the display centre. Participants were required to press S, F or O keys to record their judgement. No feedback was provided during the experiment. Participants were informed that trial order was randomized, but a third of trials would present their own motion, a third the motion of their friend and a third the motion of a stranger. Each stimulus was presented twice, making a total of 84 trials per block. Participants were seated at a viewing distance of approximately 60 cm. Avatar stimuli subtended $6 \times 4°$ of visual angle.

Testing for experiment 1 commenced five to six months after filming. The delay was longer than that which is typically imposed in studies of self-recognition [1–4] to minimize any risk that test performance would be influenced by episodic recall of idiosyncratic movements made during filming. As a further precaution, participants were informed only a few minutes prior to testing that they would be required to discriminate their own motion. These steps, together with the measures taken to prevent encoding of idiosyncrasies during filming, ensured that the effects observed were due to recognition of actors' motion signatures and not attributable to episodic recall of the filming session.

For each condition, d-prime (d') statistics were calculated to measure participants' ability to discriminate

self-produced and friends' motion from the motion of strangers [27]. Hits were therefore correct identifications (self-response to self-stimulus/friend response to a friend stimulus), whereas false alarms were incorrect judgements of the stranger stimuli (self-response to stranger stimulus/friend response to stranger stimulus). The analyses reported were conducted on the resulting distributions of d-prime values.

(b) *Results and discussion*
The mean *d*-primes from experiment 1 are shown in figure 2*a*. Participants were able to successfully discriminate their own motion both in upright ($M = 0.49$, $t_{11} = 3.25$, $p = 0.008$) and inverted ($M = 0.47$, $t_{11} = 4.34$, $p = 0.001$) orientations, as well as their friends' motion when presented upright ($M = 0.37$, $t_{11} = 3.95$, $p = 0.002$). However, recognition of friends' motion failed to exceed chance levels when stimuli were inverted. Whereas friend-recognition was substantially impaired by inversion ($t_{11} = 2.84$, $p = 0.016$), self-recognition was not ($t_{11} = 0.24$, $p > 0.80$). Consequently, evidence of superior self-recognition was seen only when stimuli were inverted ($t_{11} = 2.84$, $p = 0.016$). When stimuli were presented upright, discrimination of self-produced and friends' motion was comparable ($t_{11} = 0.62$, $p > 0.50$).

These results show that people are able to recognize their own facial motion under remarkably cryptic conditions—when it is mapped onto an inverted average synthetic head. They also indicate that, under these conditions, self-recognition is superior to friend recognition. Stimulus inversion impaired friend recognition but not self-recognition, suggesting that people are not only better at self-recognition, but that they use different cues to identify self-produced and friends' motion. More specifically, this pattern of results raises the possibility that recognition of self-produced and friends' motion may depend on different cues: while configural topographic cues, known to be disrupted by inversion [22–24], may be necessary for the recognition of familiar others, such cues may play a less significant role, if any, in self-recognition.

3. EXPERIMENT 2
Experiment 2 investigated directly the role of topographic and temporal cues in self-recognition of inverted facial

motion. The same participants completed the 3-AFC test while viewing inverted stimuli under three additional conditions: anti-sequence, rhythm disrupted and slowed. In the anti-sequence condition, stimuli were transformed in a way that selectively disrupted their topographic properties, whereas the rhythm disrupted and slowed manipulations disrupted the temporal characteristics of the avatar stimuli.

(a) Methods

Experiment 2 was completed in a single session, conducted 10–11 months after filming. The stranger allocations were identical to those employed during the second session of experiment 1. Data from the inverted condition in this session, where the stimuli were 'veridical' rather than temporally or spatially distorted, were used for comparison with the results of experiment 2.

Sequences of anti-frames were created which depicted the 'mirror' trajectory through avatar space (figure 3). For a given frame, the corresponding anti-frame is the equivalent vector projected into the opposite side of the avatar space. Thus, each anti-frame was derived by multiplying each veridical frame vector by −1. Because frames and anti-frames are equidistant from the mean avatar posture, sequences of anti-frames preserve the relative magnitude and velocity of the changes in expression space over time, but reverse the direction of the rigid and non-rigid changes, radically distorting their appearance (see the electronic supplementary material). It was anticipated that participants, who were naive to the nature of the manipulation, would be unable to recover from the anti-sequences, the topographic features characteristic of particular individual's facial motion.

Rhythm disrupted stimuli were created by inserting 46 pairs of interpolated frames between 50 per cent of the original frame transitions. The resulting 184 frames were converted into uncompressed AVI movie files using MATLAB and played at 50 FPS (twice the original rate). Runs of interpolated transitions were encouraged by biasing the decision to interpolate (chance ±25%) contingent on whether frames had or had not been inserted on the previous transition. Inserting pairs of interpolated frames at half the transitions and playing the rhythm disrupted stimuli and twice the original frame rate ensured that they were of the same duration as the veridical stimuli. Moreover, biasing the insertions so that they clustered together ensured salient rhythmic disruption: segments containing frequent interpolations appeared slower than the veridical; segments with few insertions appeared faster than the veridical. In the slowed condition, stimulus duration was increased by a constant parameter chosen at random from one of seven levels ranging from 120 per cent of veridical to 180 per cent in 10 per cent intervals.

(b) Results and discussion

The mean *d*-primes from experiment 2 are shown in figure 2b. If self-recognition is mediated by topographic cues, one would expect participants to be unable to recognize themselves in the anti-sequence condition. However, despite the profound changes to the rigid and non-rigid topographic cues, a marginally significant self-advantage was again observed ($t_{11} = 2.17$; $p = 0.053$), replicating that seen in experiment 1. Participants showed better
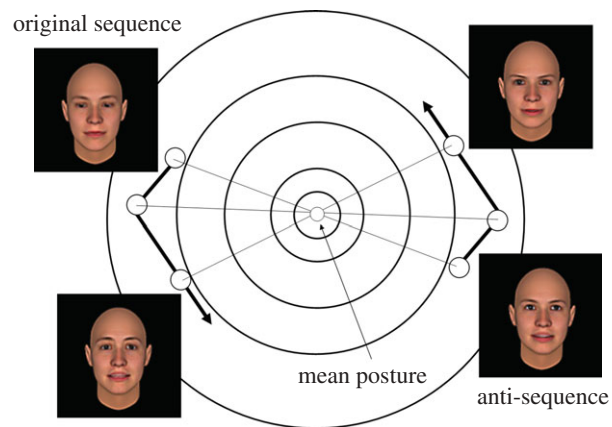

Figure 3. Schematic of three frames and their corresponding anti-frames within avatar space. Anti-frames are derived by projecting a veridical frame vector into the diametrically opposite side of the avatar space, across the mean posture. For example, a frame in which an actor is raising their eye-brows, pronouncing the phoneme /ooh/ and tilting their head to the front-right, becomes an anti-frame where the actor is frowning, pronouncing the phoneme/ee/and tilting their head backwards towards the left. As a result the topographic cues contained within a sequence are grossly distorted, while leaving the temporal and rhythmic structure intact.

than chance discrimination of their own motion ($M = 0.43$, $t_{11} = 5.04$, $p < 0.001$), comparable with the inverted veridical condition ($M = 0.48$, $t_{11} = 0.50$, $p > 0.60$), whereas friend recognition failed to exceed chance levels ($M = 0.14$, $t_{11} = 1.33$; $p > 0.20$). Thus, participants continued to recognize their own motion when the feature trajectories and configurations were grossly distorted, suggesting that self-recognition does not rely on the identification of familiar topographic cues.

In contrast, changes to the temporal properties of the stimuli eliminated the self-recognition advantage, and reduced recognition of self-produced motion to chance levels. Participants could no longer discriminate their own motion in either the rhythm-disrupted ($M = 0.06$, $t_{11} = 0.53$; $p > 0.60$) or the slowed ($M = 0.10$, $t_{11} = 0.76$; $p > 0.40$) conditions. Self-recognition under both rhythm-disrupted ($t_{11} = 3.15$, $p = 0.009$) and slowed ($t_{11} = 2.48$, $p = 0.031$) conditions was poorer than under veridical conditions. These findings indicate that self-recognition is not mediated by cues such as frequency of eye-blinks or gross head movements, which are unaffected by rhythmic disruption and slowing. Taken together, the results of experiment 2 suggest that self-recognition depends on the temporal characteristics of local motion. Friend recognition again failed to exceed chance levels in either rhythm disrupted ($M = 0.16$, $t_{11} = 1.22$; $p > 0.20$) or slowed manipulations ($M = 0.23$, $t_{11} = 1.97$; $p > 0.07$). This is not surprising given that participants could not discriminate friends' inverted veridical motion.

4. GENERAL DISCUSSION

Inversion of faces is thought to impair perception by disrupting configural representation [22–24]. That discrimination of friends' motion was impaired by inversion, therefore, suggests that configural 'motion signatures' [9],

integrated from multiple features across the face, mediate friend recognition. In contrast, participants' ability to recognize their own motion was found to be insensitive to inversion. Strikingly, participants were able to discriminate their own anti-sequences as well as they could their own inverted veridical motion. Discrimination of inverted self-produced motion was impaired only by stimulus manipulations, which altered the temporal properties of the stimuli. Together, these findings suggest that recognition of self-produced motion is mediated by temporal information, extracted from local features. Such cues might include the rhythmic structure afforded by the onsets and offsets of motion segments and characteristic variations in feature velocities.

The self-recognition advantage is puzzling because people have relatively few opportunities to observe their own perceptually opaque movements and thereby to acquire knowledge about the topographic features of their own actions. We have suggested that the results of the present study solve this puzzle by showing that, in both upright and inverted conditions, people use temporal rather than topographic cues for self-recognition. However, it could be argued that our results are consistent with an alternative interpretation—that participants typically use configural topographic cues to recognize themselves in the upright orientation, but then resort to a temporal strategy when forced to do so by stimulus inversion. This is a coherent interpretation, but it lacks theoretical and empirical motivation. At the theoretical level, it remains unclear how participants could acquire the topographic knowledge assumed by this hypothesis, or why the visual system would use hard-to-derive topographic knowledge, when readily available temporal cues permit self-recognition in both orientations. At the empirical level, we are not aware of any evidence that topographic rather than temporal cues mediate self-recognition in either orientation.

That self-recognition depends on temporal cues is consistent with previous reports of a self-recognition advantage for highly rhythmic actions such as walking [1,2,4]. It is also consistent with the observation that participants cannot accurately discriminate self-produced and friends' motion when the stimuli depict walking or running on a treadmill [3]. The artificial tempo imposed by a treadmill reduces natural variation in the temporal properties that define an individual's gait. Similarly, the importance of temporal cues is suggested by studies showing that participants can recognize their own clapping both from degraded visual stimuli depicting just two point-lights [28] and from simple auditory tones matched with the temporal structure of actions [29].

If self-recognition was found to be dependent on configural topographic cues, it would suggest that the motor system contributes to action perception via an inferential route. We rarely see our own actions from a third-person perspective. Therefore, we have little opportunity to learn what our bodies look like from the outside as we act. Since such sensorimotor correspondences could not be learned through correlated experience of observing and executing the same action, they would have to be inferred; a complex but unspecified process would be needed to generate view-independent visual representations of actions from motor programmes [4,13,16]. That self-recognition depends on temporal rather than topographic

cues indicates that such an inferential process is unnecessary; the information required for self-recognition can be acquired during correlated sensorimotor experience. We have the opportunity to learn the temporal signatures of our actions via first-person visual, proprioceptive, somatosensory and vestibular experience. Once acquired, this temporal knowledge may subsequently support self-recognition from third-person perspectives.

REFERENCES

1 Beardsworth, T. & Buckner, T. 1981 The ability to recognize oneself from a video recording of ones movements without seeing ones body. *Bull. Psychon. Soc.* **18**, 19–22.

2 Jokisch, D., Daum, I. & Troje, N. F. 2006 Self recognition versus recognition of others by biological motion: viewpoint-dependent effects. *Perception* **35**, 911–920. (doi:10.1068/p5540)

3 Loula, F., Prasad, S., Harber, K. & Shiffrar, M. 2005 Recognizing people from their movement. *J. Exp. Psychol. Hum. Percept. Perform.* **31**, 210–220. (doi:10.1037/0096-1523.31.1.210)

4 Prasad, S. & Shiffrar, M. 2009 Viewpoint and the recognition of people from their movements. *J. Exp. Psychol. Hum. Percept. Perform.* **35**, 39–49. (doi:10.1037/a0012728)

5 Heyes, C. & Ray, E. D. 2000 What is the significance of imitation in animals? *Adv. Study Behav.* **29**, 215–245. (doi:10.1016/S0065-3454(08)60106-0)

6 Bulthoff, I., Bulthoff, H. & Sinha, P. 1998 Top-down influences on stereoscopic depth-perception. *Nat. Neurosci.* **1**, 254–257. (doi:10.1038/699)

7 Giese, M. A. & Poggio, T. 2003 Neural mechanisms for the recognition of biological movements. *Nat. Rev. Neurosci.* **4**, 179–192. (doi:10.1038/nrn1057)

8 Johansson, G. 1973 Visual-perception of biological motion and a model for its analysis. *Percept. Psychophys.* **14**, 201–211. (doi:10.3758/BF03212378)

9 O'Toole, A. J., Roark, D. A. & Abdi, H. 2002 Recognizing moving faces: a psychological and neural synthesis. *Trends Cogn. Sci.* **6**, 261–266. (doi:10.1016/S1364-6613(02)01908-3)

10 Blake, R. & Shiffrar, M. 2007 Perception of human motion. *Annu. Rev. Psychol.* **58**, 47–73. (doi:10.1146/annurev.psych.57.102904.190152)

11 Blakemore, S. J. & Frith, C. 2005 The role of motor contagion in the prediction of action. *Neuropsychologia* **43**, 260–267. (doi:10.1016/j.neuropsychologia.2004.11.012)

12 Schutz-Bosbach, S. & Prinz, W. 2007 Perceptual resonance: action-induced modulation of perception. *Trends Cogn. Sci.* **11**, 349–355. (doi:10.1016/j.tics.2007.06.005)

13 Wilson, M. & Knoblich, G. 2005 The case for motor involvement in perceiving conspecifics. *Psychol. Bull.* **131**, 460–473. (doi:10.1037/0033-2909.131.3.460)

14 Heyes, C. 2001 Causes and consequences of imitation. *Trends Cogn. Sci.* **5**, 253–261. (doi:10.1016/S1364-6613(00)01661-2)

15 Heyes, C. 2010 Where do mirror neurons come from? *Neurosci. Biobehav. Rev.* **34**, 575–583. (doi:10.1016/j.neubiorev.2009.11.007)

16 Meltzoff, A. N. & Moore, M. K. 1997 Explaining facial imitation: a theoretical model. *Early Dev. Parenting* **6**, 179–192. (doi:10.1002/(SICI)1099-0917(199709/12)6: 3/4<179::AID-EDP157>3.0.CO;2-R)

17 Cutting, J. E. & Kozlowski, L. T. 1977 Recognizing friends by their walk: gait perception without familiarity cues. *Bull. Psychon. Soc.* **9**, 353–356.

18 Daprati, E., Wriessnegger, S. & Lacquaniti, F. 2007 Kinematic cues and recognition of self-generated actions. *Exp. Brain Res.* **177**, 31–44. (doi:10.1007/s00221-006-0646-9)

19 Hill, H. & Johnston, A. 2001 Categorizing sex and identity from the biological motion of faces. *Curr. Biol.* **11**, 880–885. (doi:10.1016/S0960-9822(01)00 243-3)

20 Knight, B. & Johnston, A. 1997 The role of movement in face recognition. *Vis. Cogn.* **4**, 265–273. (doi:10.1080/713756764)

21 Yin, R. K. 1969 Looking at upside-down faces. *J. Exp. Psychol.* **81**, 141–145. (doi:10.1037/h0027474)

22 Maurer, D., Le Grand, R. & Mondloch, C. J. 2002 The many faces of configural processing. *Trends Cogn. Sci.* **6**, 255–260. (doi:10.1016/S1364-6613(02)01903-4)

23 Tanaka, J. W. & Farah, M. J. 1993 Parts and wholes in face recognition. *Q. J. Exp. Psychol.* **46**, 225–245.

24 Young, A. W., Hellawell, D. & Hay, D. C. 1987 Configurational information in face perception. *Perception* **16**, 747–759. (doi:10.1068/p160747)

25 Berisha, F., Johnston, A. & McOwan, P. W. 2010 Identifying regions that carry the best information about global facial configurations. *J. Vis.* **10**, 1–8. (doi:10.1167/10.11.27)

26 Cowe, G. 2003 *Example-based computer generated facial mimicry.* London, UK: University of London.

27 Macmillan, N. A. & Creelman, C. D. 1991 *Detection theory: a user's guide.* New York, NY: Cambridge University Press.

28 Sevdalis, V. & Keller, P. E. 2010 Cues for self-recognition in point-light displays of actions performed in synchrony with music. *Conscious. Cogn.* **19**, 617–626. (doi:10.1016/j.concog.2010.03.017)

29 Flach, R., Knoblich, G. & Prinz, W. 2004 Recognizing one's own clapping: the role of temporal cues. *Psychol. Res. Psychol. Forsch.* **69**, 147–156. (doi:10.1007/s00426-003-0165-2)