The Use of Dynamic Cues in Self and Familiar Face Recognition

Rachel Bennetts (rachel.bennetts@students.mq.edu.au)
Department of Biological Sciences, Macquarie University
Sydney, NSW 2109 Australia

Darren Burke (darren.burke@newcastle.edu.au)
School of Psychology, University of Newcastle
Ourimbah, NSW 2258 Australia

Kevin R. Brooks (kevin.brooks@psy.mq.edu.au)
Department of Psychology, Macquarie University
Sydney, NSW 2109 Australia

Rachel A. Robbins (dr.r.robbins@gmail.com)
Macquarie Centre for Cognitive Science, Macquarie University
Sydney, NSW 2109 Australia

Abstract

Familiarity plays an important role in face processing. The importance of familiarity is increased when facial form cues are degraded, so that a person must rely primarily on movement (dynamic) information to identify someone. It is, however, unclear which dynamic cues are used for face recognition of both familiar and unfamiliar faces. Furthermore, little work has been done on dynamic self-face recognition, and none has focused on the type of movement that facilitates this process. The current study used motion capture cameras to record and isolate facial movements in order to test recognition of self, familiar and unfamiliar faces. Participants completed a 2AFC same/different face matching task involving point-light displays of natural motion (i.e. both rigid and non-rigid motion), rigid motion only (e.g. nodding/shaking), non-rigid motion only (e.g. mouth/eyebrow motion) and still images to determine whether differences in familiarity resulted in the use of different movement cues. The manner (style) in which someone is speaking may also impact on whether they can be easily identified from dynamic cues. Consequently, speech style was either matched or mismatched between video clips. We found that matching performance was more accurate overall when speech style was matched than mismatched. Familiar face matching appears to use rigid, non-rigid and natural movement cues equally, but unfamiliar and self-face matching are more accurate for rigid than natural motion when speech style differs between clips. These results are discussed in relation to previous research on dynamic face recognition, and possible implications for current face processing models.

Keywords: face recognition; self recognition; familiarity; motion capture; biological motion; speech type; perception-action coupling

Face processing is an important cognitive and social task. Faces contain cues to the age, race, gender, emotional expression and identity of the person we are looking at. Furthermore, people tend to be remarkably good at decoding these cues accurately, particularly in the case of identity (Bruce, 1988). One factor that has been comprehensively investigated in the face recognition literature is the role of familiarity. While familiar face recognition is fairly robust to changes in lighting, viewpoint and expression (Bruce, 1988), unfamiliar face recognition appears to rely on lower level visual cues, and can be disrupted quite simply by changing basic image properties such as hue or lighting direction (Hancock, Bruce and Burton, 2000), viewpoint or facial expression (Bruce et al., 1999). There is also evidence to suggest that familiarity influences what facial information is used. Studies have found that unfamiliar face matching relies disproportionately on external features of faces, such as hair and head shape (Bruce et al., 1999) while familiar face recognition or matching is more accurate when internal, rather than external, facial features are presented (Ellis, Shepherd and Davies, 1979; Young, 1984). Furthermore, people are more sensitive to displacement changes of the eyes and nose in familiar faces than in unfamiliar faces (Brooks and Kemp, 2007).

The majority of studies have used static images such as photographs or sketches to study face recognition. However, in everyday life we generally encounter people in motion – for example, talking, laughing, nodding or shaking their heads. Research on perception of the motion of human bodies has shown that movement information can act as a reliable cue to identity (Cutting and Kozlowski, 1977; Loula, Prasad, Harber and Shiffrar, 2005; Prasad and Shiffrar, 2009). These studies isolated movement information from other identity cues by using point-light-walkers, in which reflective markers are placed on a person’s joints as they are filmed walking, and the subsequent videos are edited to show only the moving dots. This technique limits the amount of form-based information available, so that judgments are based primarily on motion cues. By extending this method to faces, and applying other techniques to degrade static images (e.g. blurring, negating, pixelating etc.), several studies have shown that facial movement information (dynamic cues) can also play a role in face recognition (see Roark, Barrett, Spence, Abdi and...
O’Toole, 2003 for a review). It has been suggested that dynamic cues aid recognition in two possible ways (O’Toole, Roark and Abdi, 2002). Firstly, people may be able to identify someone based purely on the characteristic ways they move. This would only be possible if the person being viewed was familiar enough that the viewer had learned their individual movement patterns. Secondly, movement cues may enhance the perception of someone’s 3D facial structure, which may in turn aid in recognition performance by building a more robust face representation. As this process relies on structure-from-motion, and not an inherent familiarity with a face, the second process could be applied to both familiar and unfamiliar face recognition.

The two processes underlying the recognition of moving faces may be driven by different types of movement. For example, structure-from-motion cues are more likely to be facilitated by rigid movements such as nodding, shaking and tilting of the head, because these allow more “views” of the face which can be extrapolated into 3D representations. By contrast, characteristic motion patterns of a face could be made up of both rigid head movements and non-rigid face movements (e.g. speech and expression). Only a handful of studies have attempted to investigate the role of rigid and non-rigid facial motion in recognition. Bruce and Valentine (1988, experiment 2) used the point-light technique to compare recognition performance for six familiar faces that were depicted either pulling facial expressions (non-rigid) or moving the head rigidly. They found that recognition performance was above chance, and equivalent for both types of motion. Lander and Chuang (2005) used contrast and blurring manipulations to test for recognition of familiar faces that were moving rigidly (looking up and down) or non-rigidly (expressing or talking). In contrast to Bruce and Valentine’s (1988) results, they found no significant benefit for rigid motion compared to still frames, but a significant benefit for both forms of non-rigid motion. In contrast to these studies, Hill and Johnston (2001) created animated heads from motion capture data of unfamiliar faces. This technique enabled them to isolate the rigid and non-rigid motion of the same motion sequence (a joke being told). They found that performance in an identity categorization task was significantly better for rigid motion than for non-rigid motion. Given that each study used different methodologies, different levels of familiarity, and different ways of presenting rigid and non-rigid motion, it is difficult to directly compare the results, or draw any firm conclusions about the relationship between motion cues and familiarity. Consequently, one of the aims of the current study is to directly compare performance on a face-matching task for familiar and unfamiliar faces when the rigid and non-rigid movements of the face and head are isolated.

Self-recognition can be viewed as a special case of familiarity. There is a large amount of neurological evidence to suggest that self-recognition activates brain networks distinct from those involved in other face recognition, particularly the right prefrontal cortex (Keenan, Wheeler, Gallup & Pascual-Leone, 2000). Studies involving whole-body point-light displays have found that recognition of oneself is better than recognition of familiar others, which is in turn better than recognition of strangers for walking and other whole body movements (Loula et al., 2005; Prasad and Shiffrar, 2009). Furthermore, self-recognition of point-light walkers is less viewpoint-dependent than recognition of familiar individuals in point-light walker displays (Jokisch, Daum and Troje, 2006; Prasad and Shiffrar, 2009). This has led several authors to suggest that motor processes influence the visual analysis of human movement, resulting in an increased sensitivity to one’s own motion patterns (Jokisch et al., 2006; Knoblich and Flach, 2003; Loula et al., 2005; Prasad and Shiffrar, 2009). Currently, very few studies have tested these predictions with facial movement patterns, and none have attempted to isolate the role of rigid and non-rigid motion in self-recognition. This study aims to address this gap in the literature by testing recognition of own-face point-light displays, and comparing these to recognition of familiar and unfamiliar face point-light displays. Participants completed a face-matching task involving point-light displays of natural motion (i.e. both rigid and non-rigid motion), rigid motion only, non-rigid motion only and still images. In these displays, participants were shown talking normally, telling a joke or reading some dramatic dialogue. Previous research using cross-modal stimuli has found that people are able to match unfamiliar voices and faces, even when speaking different syllables, as long as the style of speech (e.g. casual speech, clear speech, conversational speech) is the same (Kamachi, Hill, Lander and Vatikiotis-Bateson, 2003; Lander, Hill, Kamachi and Vatikiotis-Bateson, 2007). As the previous research on this topic used cross-modal stimuli, this study aimed to investigate whether these effects also apply to purely visual motion displays of unfamiliar faces, and whether changing the speech style has the same effect when recognising unfamiliar faces, self-faces (where participants would have extensive motor-based experience in all speech types) or familiar faces (where participants may have visual experience with a range of speech types). Therefore, the two point-light displays in each trial were either of the same speech type (e.g. two jokes, two dialogue excerpts, two candid speech excerpts) or differing speech types (joke-dialogue, candid-joke, dialogue-candid). The final design consisted of three variables: familiarity (self; familiar; unfamiliar), motion type (natural; rigid; non-rigid; static) and speech style (matched or mismatched).

Methods

Participants

Fifteen participants (9 female, mean age 32 years), including three of the authors, had their facial movements recorded for this study. Participants were acquaintances of the authors who took part in the study for financial compensation. Of the fifteen participants, seven (including two of the authors) took part in the perceptual testing. The remaining eight participants served as familiar and
unfamiliar stimuli for the other participants. Familiar pairs consisted of people of the same gender who had known each other for at least one year (minimum 14 months, maximum 28 years). All participants had normal or corrected-to-normal vision, and provided informed consent before commencing the experiment.

Apparatus
Movement sequences were captured using an 8-camera Vicon motion capture system (Oxford Metrics), calibrated to less than 1mm error. The motion capture system recorded data at 200Hz from 28 markers placed on the face (figure 1 illustrates the configuration used). Any errors or gaps in data collection were corrected using a spline algorithm. Point-light display videos were created using Vicon BodyBuilder (Oxford Metrics) and Matlab (Mathworks). After editing, each video measured 560x420 pixels. The experiment was run on a MacBook Pro 15 inch laptop, using SuperLab 4.07 ( Cedrus).

Stimuli
Each participant was recorded telling a joke, reading some dramatic dialogue (excerpt from Hitchhikers Guide to the Galaxy, the Original Radio Plays) and engaging in casual conversation. Each speech style was repeated twice, to ensure that videos presented in the final experiment were never the same clip. For each participant, two clips of between 4 and 10 seconds were selected to represent each speech style.

For each clip, four stimuli were created, displaying natural (rigid and non-rigid) motion, rigid motion only, non-rigid motion only, and a single static image.

To create the natural motion videos, the motion capture data were processed in Matlab to present the movement as black dots on a white background, with a 500ms black screen before and after the clip.

To create the rigid and non-rigid motion videos, the movement of the forehead, temples and nose bridge was isolated using Vicon BodyBuilder (Oxford Metrics), and used to approximate rigid head movement. For the rigid videos each of the 28 points was restricted to only display rigid head movements, whereas for the non-rigid videos, the rigid head movement was removed from all the marker data. As with the natural motion videos, the edited motion capture data were then processed in Matlab.

To create the static images, the 32nd frame of each combined motion video was extracted using Matlab.

Procedure
Participants were tested between 4 and 6 weeks after the initial capture session. They were informed that they would be shown point-light videos of themselves, their familiar partner (who they knew the identity of), and an unfamiliar person. Each trial consisted of two point-light videos or stills presented simultaneously on the screen (see figure 1 for an illustration). The participant was asked to indicate via key press whether the images were of the same person or different people. The trial repeated itself every 10 seconds until a response was recorded, and participants could respond at any time throughout the trial.

The experiment had three variables: familiarity (self; familiar; unfamiliar), motion type (natural; rigid; non-rigid; static) and speech style (matched or mismatched). In matched speech style trials, two clips of the same type were played (e.g. two jokes), whereas for mismatched trials, the videos depicted two different speech styles (e.g. joke and dialogue). Matched and mismatched trials were presented for all three speech types. In total, each participant viewed 144 trials, half of which were “same” trials.

Order of trial presentation was randomized for all variables but motion type, which was blocked. Natural, rigid and non-rigid motion blocks were counterbalanced between A

B

Figure 1: A: Motion capture data were collected from 28 points on the face. B: The data were then shown to participants as moving black dots on a white background. Each trial consisted of 2 point-light displays on the screen at once. Participants could respond at any time throughout the trial.
participants. The static images were always presented last, to prevent participants learning and using static cues rather than motion cues to match the faces. Participants viewed 6 practice trials (natural motion, depicting faces that did not appear elsewhere in the experiment) before beginning the first block.

Results

Accuracy was calculated for the same-identity trials only, as errors in the different identity trials could have resulted from the misperception of either of the identities. Accuracy data for matched and mismatched speech type are displayed in Figure 2. A 4 (motion) x 3 (familiarity) x 2 (matching speech type) repeated measures ANOVA revealed a main effect of matching speech type, $F(1,6) = 8.48, p = .027$, eta$^2 = .586$. Participants were significantly more accurate matching identity when the two videos depicted the same speech type (M=0.885, SD=0.059) than when they depicted differing speech types (M=0.75, SD=0.148). Participants were significantly above chance levels of responding (50%) both for trials where the speech type was matched, $t(6) = 13.72$, $p < .0005$; and where type was mismatched, $t(6) = 4.80$, $p = .003$. No other main effects were significant.

The ANOVA also revealed a significant motion x familiarity interaction, $F(6,36) = 2.65$, $p = .031$, eta$^2 = .306$. Planned pairwise comparisons (Bonferroni-corrected) between motion and familiarity revealed that, averaged across matching speech type, static images were recognized better than natural motion for unfamiliar faces ($F(1,36) = 5.39, p = .026$). This may reflect a ceiling effect for static images, suggesting that participants learnt shape-based differences throughout the experiment, and were able to apply them in the static condition. For mismatched speech type, unfamiliar faces were recognized more accurately than self-faces ($F(1,36) = 7.86, p = .008$). For both self and unfamiliar faces, rigid motion was more accurately matched than natural motion in the mismatched speech condition (self $F(1,18) = 5.71, p = .028$; unfamiliar $F(1,18) = 5.97, p = .025$). No other pairwise comparisons were significant.

Discussion

In the current study, participants completed a face matching task featuring moving point-light displays of themselves, someone familiar to them, and an unfamiliar person. The main aim of the experiment was to investigate how rigid and non-rigid motion contribute to face matching, and whether these contributions are modulated by familiarity.

Our results support the hypothesis that dynamic facial cues can be used as a source of information about a person’s identity (O’Toole et al, 2002). Participants’ responses were above chance when shown video clips depicting either the same or different speech styles.

The motion x familiarity interaction supports the hypothesis that familiarity may change the dynamic cues that people use when trying to match faces. Pairwise comparisons did not reveal any significant effects that could explain where the differences in cue use arise, however, this may be due to a ceiling effect for displays that depicted the same speech style, or a lack of power in the analysis. Only the analysis of mismatched speech style trials reveals a pattern of responding that resembles the results of previous studies. Bruce and Valentine (1988) found that identification of familiar faces from point-light-displays was...
equally accurate when either rigid or non-rigid motion was shown. Likewise, in the current study familiar faces were matched equally well in the rigid and non-rigid motion conditions. Furthermore, participants were just as accurate matching faces with natural (both rigid and non-rigid) motion as they were matching faces where one motion type was isolated. These results contrast with Lander and Chuang’s (2005) findings that familiar face recognition was more accurate for videos showing expressions or speech than for videos of rigid head movements. However, the rigid head movements in Lander and Chuang’s study were confined to a 1.5 second clip of the actor looking up and down once. It is possible that the movement information provided was not sufficient for participants to recognize an individual’s characteristic movement patterns or extract 3D structural information. Bruce and Valentine (1988) and the present study both included longer clips (minimum four seconds) with a wider range of rigid head motion, which may have facilitated these processes, and allowed participants to use either source of motion information (rigid or non-rigid) to complete the task.

In contrast to familiar faces, unfamiliar faces were matched more accurately when rigid motion was isolated, compared to when the whole range of natural movement was present. Once again, this effect was only apparent for mismatched speech style trials. Previous research involving unfamiliar faces has found an advantage for rigid head movement over non-rigid facial movement, but not over natural movements of the face and head (Hill and Johnston, 2001). There are several reasons why the current results may differ from those of the previous study. Firstly, Hill and Johnston (2001) used full-face animations, rather than point-light displays as used here. Hill, Jinno and Johnston (2003) compared the use of animations and point-light-displays in a gender-matching task, and found no difference. However, it is possible that the type of display does matter for identity judgments. Identity-matching tasks require fine-grained comparisons of movement, and it is possible that these comparisons are impeded by the presence of complex rigid and non-rigid cues together. In other words, point-light-displays that depict both head and face motion may be harder to match than rigid-only displays due to the sheer amount of dynamic information present. This effect may be reduced in full-face animations, which appear more like the faces of people we see every day, and may consequently be processed more efficiently. A second point of difference between the studies is that Hill and Johnston only used clips depicting joke-telling, whereas the current experiment showed clips of joke-telling, dramatic dialogue and casual speech. It is possible that the pattern of results found by Hill and Johnston is specific to joke-telling, and the use of movement cues for identification may change depending on what speech style is depicted, or when people are required to generalize between speech styles as in the current study.

Matching of one’s own face appears to use similar dynamic cues to matching unfamiliar faces. In mismatched speech style trials, rigid motion alone resulted in more accurate matching performance than natural motion, but rigid and non-rigid motion were equally beneficial for self-face matching. However, unfamiliar faces were matched significantly better than self faces in these trials, and overall there was no difference in accuracy between self and familiar faces. This is surprising, given that many studies have found a recognition advantage for one’s own body movements over those of familiar and unfamiliar others (e.g. Loula et al., 2005; Jokisch et al.; 2006; Prasad and Shiffrar, 2009). It is possible that point-light faces are harder to recognize overall than point-light walkers, or that the displays used in the current study were more complex than those used in the point-light walker studies mentioned above. Alternatively, facial motor processes may be harder to access and recognize as one’s own than whole-body processes. If this were the case, self-face recognition would rely primarily on visual experience. In either case, further study is necessary to clarify the extent, if any, that perception-action coupling may play in motion-based self-face recognition.

A second aim of this study was to investigate whether familiarity with a face (including self-faces) would facilitate matching across different speech styles. Regardless of whether a face was familiar or not, presenting two video clips depicting the same style of speech resulted in significantly better matching performance than clips depicting different speech styles. Despite this detriment, matching performance remained above chance even when speech style was mismatched. These findings support Lander et al.’s (2007) conclusions that cues to identity are somewhat linked to the style of speech, and extend their findings to a unimodal matching task involving familiar, self and unfamiliar faces. The results suggest that some variation in movement patterns of a face can be accounted for by the manner in which the person is speaking, and some variation is a result of the individual’s characteristic movement patterns. People are sensitive to both forms of variation, regardless of familiarity, as evidenced by the fact that matching identity was still possible when speech style was mismatched, but was at its best when the speech styles matched (even though the movement was not identical as different clips were used). In fact, performance was almost at ceiling when speech styles were the same.

As noted above, the effect of matching speech type was the same when a face was unfamiliar, familiar, or even one’s own. This suggests that visual (familiar faces) and/or motor (self faces) experience with a range of speaking styles does not provide any extra benefit when trying to generalize across speech styles. It is difficult to account for the fact that unfamiliar faces were matched as well as faces with which people had extensive personal experience, even in the mismatched speech type condition. It is possible that characteristic facial movement patterns are extracted very quickly, leading to a rapid familiarization effect. Alternatively, people may be comparatively poor at encoding differences arising from variation in speech style, and simply rely on extracting the same identity-specific
movement patterns on each encounter. The use of a matching, rather than explicit identification task, may have artificially increased performance in the unfamiliar condition, since people only had to extract and compare identity cues, rather than having to extract them and match them to the memory of a specific person. Investigation of the familiarization process, and the use of more memory-based identification tasks, may help clarify this point in the future.

One last point of interest in these results is the high level of performance in the static condition. In order to prevent participants learning to identify their target faces from form (rather than dynamic) cues, the static pictures were always presented as the final experimental condition. However, this resulted in extremely high accuracy for the static images, with an average of 79% correct. In comparison, other studies that have shown static point-light images have found chance levels of responding (Bruce and Valentine, 1988; Loula et al., 2005). It appears that participants used the moving stimuli to encode form-based information, which they were then able to apply in the static matching condition. While this makes it hard to quantify the influence of form cues in the moving conditions, it does support the hypothesis that moving faces provide enough 3D information to build a more robust static representation of the face and head.

Current models of face recognition have emphasized that dynamic information may help the identification process in several ways: by building up more robust 3D representations via structure-from-motion processes (for familiar and unfamiliar faces), and/or by providing an alternative route to familiar face recognition via characteristic facial movements (for familiar faces only) (O’Toole et al., 2002). The results of the current study have several possible implications for this model. Firstly, it appears that both familiar and unfamiliar faces benefit from rigid and non-rigid motion equally. Since structure-from-motion processes rely primarily on rigid transformations of an object, it is unclear how people were able to match unfamiliar faces using non-rigid motion alone. It is possible that the face motion in this experiment became familiar more rapidly than usual, as it was the only available cue to identification. Alternatively, people may be able to extract and match the characteristic motion patterns of unfamiliar, as well as familiar, faces. If this is the case, the results imply that characteristic motion patterns of the face consist of a combination of rigid and non-rigid cues, rather than being biased towards one form of motion. Secondly, the relationship between self-face recognition and face recognition in general is not currently addressed in O’Toole et al.’s (2002) model. Based on the present results, it appears that self-face recognition may follow a similar path to unfamiliar face recognition. However, further behavioural and neural studies will be needed to examine the relationship between self and other motion-based face processing, the different types of motion involved, and the role of perception-action coupling in this context. Finally, keeping the style of speech constant significantly improved matching performance for all faces tested. Currently, the model of movement-based face recognition does not address how non-identity based dynamic cues impact on recognition, but future work on speech, expression and eye movements may focus on the interactions between these face processing systems and identity, rather than the functional separation between them.

Acknowledgments

The authors wish to thank Bill Thompson for access to the motion capture facilities, and Alex Chilvers for providing technical assistance.

References


**Citation details for this article:**
DOI: 10.5096/ASCS20094