Recognizing People in Motion

Galit Yovel1,* and Alice J. O'Toole2

Natural movements of the face and body, as well as voice, provide converging cues to a person’s identity. To date, person recognition has been studied primarily with static images of faces. Face recognition, however, is part of a larger system, whose preeminent goal is to efficiently recognize dynamic familiar people in unconstrained environments. We present a comprehensive framework for understanding person recognition as it happens in the real world. In this framework, dynamic information plays the central role in binding multi-modal information from the face, body, and the voice to achieve robust and highly accurate recognition. The superior temporal sulcus (STS) integrates multisensory, dynamic information from the whole person for recognition, thereby complementing its role in social cognition.

Motion Plays a Central Role in Person Recognition

The utility of biological motion (see Glossary) for person recognition is seen most clearly when we consider the problem of recognizing familiar people in viewing conditions commonly experienced in the natural world. We argue that motion acts as the key element for binding together faces, bodies, and voices into a coherent representation of a person that supports recognition. In making this argument, we review recent evidence on the role of the body and biological motion in person recognition and consider studies showing that the integration of moving faces and voices facilitates recognition.

The idea that motion plays a central role in person recognition is not prominent in the cognitive science literature. This is due to three longstanding biases in this literature: the common practice of studying person recognition using static images of faces; the emphasis on studying the role of the body and body motion in action perception and social communication, rather than in recognition; and the strong emphasis on the study of human voices for speech perception, rather than for recognition. These biases are justified in decades past, when there was limited knowledge of the psychological, neural, and computational complexities involved in processing a moving and talking person in an unconstrained environment. Recent years have seen advances in all three disciplines that lead us to reconsider person recognition in this more complex in situ form. We will show that this complexity simplifies the problem of recognition by distributing the burden over multisensory identity cues and by using motion to bind these cues together. We will argue that the dynamic and multisensory nature of identity cues that apply in the real world point to the STS as the likely integration site for person recognition cues with acoustic and spatiotemporal extent.

Axiomatic to this argument is the observation that the visual and auditory person recognition systems in the brain evolved over hundreds of millions of years to work efficiently in the natural world. Recognizing people from a safe distance (at which the face is not sufficiently resolved to be helpful) is a distinct evolutionary advantage. Accordingly, we begin this discussion where perception begins: at a distance. From this perspective, it becomes clear that person recognition is a process that unfolds over space and time, not a decision made to a snapshot of a visual event (Figure 1A, Key Figure). At a distance, person recognition is strongly constrained by the viewing environment and the limits of human sensory systems. Identity cues from the body and

---

1School of Psychological Sciences, Sagol School of Neuroscience, Tel Aviv University, Tel Aviv, Israel
2School of Behavioral and Brain Science, University of Texas at Dallas, Dallas, TX, USA

*Correspondence: galyt@post.tau.ac.il (G. Yovel) and otoole@utdallas.edu (A.J. O’Toole).
**Key Figure**

The Roles of Body, Voice, and Biological Motion in Person Recognition

(A) Schematic shows the prominence and primacy of body and biological motion cues to identity at a distance, with the face becoming clearly resolved only at closer distances. This process unfolds over time, allowing for continuous accumulation of identity cues until a confident response is possible. The yoking of a dynamic face and the production of speech is shown below (reproduced, with permission, from [43]). The acoustic and spatiotemporal cues are perceptually bound together by motion. (B) Neural mechanisms of person recognition in the human brain: the posterior face and body areas, OFA and EBA, represent faces and bodies independently. The face and body areas in the fusiform gyrus (FFA, FBA) facilitate social communication by binding together cues from faces and bodies.

**Glossary**

**Biological motion**: motion generated by biological agents, such as humans and animals. It involves different neural and cognitive mechanisms than motion generated by nonbiological stimuli.

**Changeable facial aspects**: faces convey very rich information that varies from moment to moment. This includes facial expressions, mouth and eye movements. Current neural models posit that the STS mediates the processing of changeable facial aspects consistent with their important role in social communication [54].

**Dynamic identity signature**: identity cues that reside in idiosyncratic elements of face and body movements. These are nuanced, person-specific expressive gestures that punctuate social and pragmatic face and body movements. For these dynamic identity signatures to facilitate person recognition, the visual system must represent and remember, not only the categorical content of biological motions (e.g., smiling, walking, getting out of a car), but also person-specific variations of these; that is, the way he smiles, walks, or gets out of a car.

**Extrastriate body area (EBA)**: a body-selective brain area in the lateral occipital cortex [61]. The EBA is defined functionally within each individual as the cluster of voxels in the extrastriate cortex that shows significantly higher response to body than to object stimuli.

**Form-from-motion**: information from faces and bodies can benefit person recognition by enhancing the quality of shape representations of a face or person. These cues are rooted in basic perceptual abilities to see structure from motion, where shape is revealed more clearly through motion.

**Fusiform body area (FBA)**: a body-selective brain area in the middle of the fusiform gyrus. The FBA is defined functionally within each individual as the cluster of voxels in the fusiform gyrus that shows significantly higher response to body than to object stimuli [62].

**Fusiform face area (FFA)**: a face-selective brain area in the middle of the fusiform gyrus [63]. The FFA is defined functionally within each individual as the cluster of voxels in the fusiform gyrus that shows...
biological motion are visually robust and informative across a wide range of distances [1], even when we see people from the back, whereas a closer view of the face is needed for identification. Efficient person recognition depends, therefore, on the ability to optimize a trade-off between the diagnostic value of the identity information from different sources (faces, bodies, voices, and biological motion) and the corresponding perceptual quality of these cues at various distances and in variable/noisy environments.

We begin by reviewing recent evidence on the role of the body and body motion in person recognition. (See Box 1 for a discussion of biological motion.) Next we show that dynamic information also contributes to person recognition based on the face and voice. We then propose that the STS, which has been primarily studied in the context of social cognition, may play an important role in real-life recognition of the dynamic, talking whole person. At the outset, we note that the relevant literature for this review is sparse. We would argue, however, that it is sufficiently informative to sketch out a theoretical framework for multisensory, dynamic person recognition. This framework will help to identify gaps in our knowledge and to provide structure for existing and forthcoming data on the topic.

The Contribution of the Static Body to Person Recognition

Beginning with static images, when viewing conditions are good and when the face is informative for identity, people rely primarily on the face for recognition, even when the body supports accurate identification [2]. When viewing conditions are suboptimal (e.g., due to poor illumination), however, there is striking evidence of the role of the body in person identification. In these cases, people shift reliance from the face to the body, even when close-up views of the face are available [3]. In a recent study, participants judged whether two images, taken under very different viewing conditions, were of the same person or different people. In this identity-matching study, image pairs were chosen by face recognition algorithms to be ‘challenging’ and showed a person from the waist up. The algorithms estimated challenge level by computing the similarity of the faces, ignoring the bodies. For these challenging ‘face pairs’, the face and body contributed equally to human identification accuracy [3]. (See Box 2 for a discussion of computational approaches to visually based person recognition.)

An even more challenging task demonstrated that, in some cases, the body can account entirely for identification accuracy [4], with people remarkably unaware that they are relying on the body, rather than the face. Again, with images that pictured people from the waist up, face recognition algorithms were used to select extremely challenging face stimuli. These consisted of image pairs that machines always confused (100% incorrect) (i.e., same-identity face pairs always labeled as different people; different-identity pairs labeled as the same person). Human performance was well above chance. Two follow-up conditions, in which people matched identity with the face or body obscured, provided a clear account of human superiority over the machines. Human performance with the face alone was barely above chance; human performance with the body equaled performance with the entire image. Notwithstanding, people systematically reported using internal face features, rather body features, for identification, indicating their limited awareness of relying on the body. Eye movements, however, clearly showed a shift to the body, and away from the face, when the body was a better cue to identity information than the face. This illustrates the adaptive nature of the visual system in person identification.

---

generate an integrated representation of the static face and body [57,58]. The STS is sensitive to dynamic information from the face, body, and voice [55]. The posterior STS represents dynamic information from the whole person (e.g., [49]). A similar model has been proposed recently for the functional specialization of macaque face patches to the dynamic whole person [65,66]. Dynamic face and voice areas reside along the STS [51,55]. Abbreviations: EBA, extrastriate body area; FBA, fusiform body area; FFA, fusiform face area; OFA, occipital face area; pSTS, posterior superior temporal sulcus; STS, superior temporal sulcus.
Box 1. Biological Motion of Faces and Bodies

Faces move in both rigid and non-rigid ways. Rigid motions include nods and head turns; non-rigid motions include facial expressions and gaze. Both types of motion cue social interactions. Head turns can direct attention to novel events in the environment; nods can be used to communicate assent. Smiles, frowns, and other non-rigid facial gestures can indicate the emotional state and intent of others. The motion of faces can also support recognition in limited cases (see main text and [14,15]).

Body motion, by contrast, is always non-rigid. People walk, dance, gesture, manipulate objects, sit down, and strike full-body poses that capture or evade attention. Some body motions communicate social and emotional information [67–69]. A crest-fallen stance or arms raised in victory tell us more about the emotional state of another person than a facial expression [70]. However, most body motion is motivated by the simple need to solve problems (e.g., getting from point A to B, or pouring a cup of coffee).

Face and body motions can contribute to person recognition if they offer information that specifies or constrains hypotheses about a person’s identity. Two factors may underlie the contribution of motion to person recognition: form-from-motion processes, whereby motion augments the representation of body shape, and dynamic identity signature processes, whereby recognition is enhanced by representing idiosyncratic body and facial movements [15]. Figure 1 shows how the relative contribution of these two types of process to person recognition can be isolated. If person recognition from the static image of a person is better following the presentation of dynamic rather than static images, then form-from-motion processes contribute to person recognition (Figure 1A). If recognition of the dynamic whole person is better for dynamic than static images, than dynamic information is used for recognition (Figure 1B). These two processes are not mutually exclusive, although their relative contribution may vary in different situations. For example, dynamic identity signatures are more likely to facilitate recognition of familiar rather than unfamiliar people, where we have prior experience of their idiosyncratic movements and gestures. Additionally, people with distinctive movements and gestures are more likely to be recognized based on dynamic information. Finally, based on studies with PLDs, the contribution of motion to person identification is greater for certain types of action, such as dancing and boxing, than for other types of movement, such as walking [7].

![Figure 1. How to Determine the Contribution of Form-from-Motion and Dynamic Identity Signature in Person Recognition](image_url)

Experimental designs that can be used to determine the role of form-from-motion and dynamic identity signature in person recognition (see [12]). In both cases, different videos/images of the same person that were taken on different days are presented at study and test. (A) If learning a person from video produces better recognition of a static image of a person than learning the person from static images, then form-from-motion processes contributed to person recognition. (B) When a person is learned in motion, if dynamic identity signatures contributed to person recognition, person recognition from a dynamic video will be better than from static images.
Box 2. Computational Models of Person Recognition

Computer-based face recognition systems operate on static images of faces and are mainstream technology in social media, law enforcement, and identity management [71–73]. State-of-the-art algorithms are based on deep convolutional neural networks (DCNN) (cf. [74]) and currently recognize faces across variable viewing conditions. Face recognition systems have improved markedly in recent years due to the availability of vast amounts of human-labeled training data (e.g., “This is Brad Pitt”) from the web (cf. [82]). These data sets include many variable images of each identity, with a recent system employing more than 4 million training images of 4000 identities [75]. This training resembles one small part of human visual experience with faces. Consequently, machines are beginning to show performance that generalizes over photometric variation. Although machine performance compares favorably to humans in limited cases [4,76–78], for the most challenging cases, human performance is still far superior to that of machines [4,79].

Computational models of person recognition from biological motion are almost exclusively focused on gait and employ several approaches for extracting motion and form features from walkers in PLDs and in natural video [80]. The state-of-the-art in gait recognition is roughly a decade behind that seen for faces. Although there has been steady progress in quantity of walker motion for identification, coding motion features that generalize across viewpoint, footwear variation (e.g., heels versus flats), and weight load (e.g., carrying a backpack or shopping bag) remains challenging. DCNNs have the potential to overcome some of these problems, but only if they have access to the quantity and quality of gait-training examples now available for static images of faces. It is highly likely that these kinds of gait data set will emerge in the next decade.

Beyond face and gait, dynamic identity signatures superimposed on actions (idiosyncratic ways of smiling, gestulating with hands while talking, nodding and head turning while listening to music) are largely unexplored. Computational progress on these problems may await a solution to the prerequisite problem of categorizing actions (e.g., typing, dancing) [81].

The Contribution of the Body in Motion to Person Recognition

When the body is seen in motion, longstanding evidence points to the importance of the face in recognizing familiar people. A closer look at the evidence, however, tells a more complicated story. In a now classic study [5], people saw closed-circuit TV (CCTV) videos of people walking through a doorway. Identification accuracy decreased dramatically when the face was obscured, but only slightly when the body was masked. However, people viewed the entire video in this study, allowing them to wait for a closer view of the face before responding. When a recognition response is required from a distant vantage point, the contribution of the body resembles that seen for closer views of static people in “suboptimal viewing conditions” [3,4], highlighting the fact that a distant view of a face is suboptimal. In a recent study [1], participants were familiarized with people using close-up and distant videos. Recognition was tested with videos of familiar and unfamiliar people approaching from a distance. The test conditions varied so that participants viewed the whole person, the face only, or the body only. Responses made at a distance showed an independent contribution to recognition from both the face and body. At closer vantage points, recognition accuracy for the entire person did not exceed accuracy for the face alone [1], consistent with [5].

Although these results suggest a gradual shifting of reliance from the body to the face as viewing distance decreases, the study also indicated that, when forced to rely on the body, person recognition was reliable [1]. Specifically, recognition from the body alone remained stable and above chance at all distances, whereas recognition from the face alone increased with proximity. People showed remarkably accurate meta-knowledge of the quality of information in the face versus body over distance. When participants knew that a closer view of the face would be available later in the video, they responded more conservatively than when they knew that only a distant view would be seen.

The use of natural moving walkers to test recognition leaves open the question of whether performance in the body-only condition benefited from dynamic identity signatures in gait [1] or came from body structure. Next, we consider studies that assessed more directly, the role of biological motion in person recognition.
Box 3. Person Recognition from Point Light Displays

PLD of people in motion have demonstrated that body motion provides information about human actions, as well as the emotion, gender, and identity of a person [7,83,84]. Here, we focus on studies of person identification from PLDs. Overall, these studies rely on a small number of identities and report relatively weak recognition levels. One of the first studies on recognition showed that people can recognize their friends as point light walkers [8]. Recognition was above chance level (16.6%), but relatively low (28%), suggesting that identity information cannot be extracted easily from point light walkers. Another study revealed moderately accurate performance (75%) in a naming task for participants who were highly familiar with two point light walkers (over 20 hours a week of interaction). However, recognition rate did not exceed chance level (50%) for participants who were less familiar with the walkers (5 hours/week of interaction or less) [8]. Other studies have started with previously unfamiliar people and indicate that recognition from PLDs requires extensive pretraining and familiarization with the individuals to be recognized [8,9]. Interestingly, caricature effects may also enhance motion-based person recognition. Following familiarization of six individuals with unexaggerated arm movements, recognition was better when arm movements were exaggerated [10].

To date, the most extensive study on the role of motion in person recognition used point light walkers, as well as PLDs of actions such as jumping, dancing, boxing, or ping-pong playing [7]. Consistent with previous studies, performance for discriminating friends from strangers based on gait was modest. Recognition was much higher, however, for PLDs of dancing, boxing, ping-pong playing, and jumping people. These findings suggest that dynamic identity signatures can be extracted more easily from actions that convey spatiotemporal variations that are richer than those available from gait. A remarkable finding in this study was that recognition of self (69%) was far better than recognition of friends (47%), suggesting the important contribution of the motor system to person recognition in motion and highlighting the need to study the roles of both perceptual and motor processes in dynamic person recognition (see Outstanding Questions).

Biological Motion from Bodies Contributes to Person Recognition at a Distance

Multiple lines of evidence converge to indicate that biological motion facilitates person recognition via the processing of dynamic identity signatures (Box 1). Early studies with point light displays (PLD) demonstrated the basic human ability to recognize oneself and familiar people from gait [6,7] and to learn to discriminate new identities based on gait [8,9]. Biological motion is even subject to caricature effects that improve recognition when arm movements were exaggerated over time [10] (for more detailed discussion on PLD, see Box 3).

Using more natural stimuli, a direct role for motion in facilitating identification was seen in a study in which people viewed videos or static frames of a person approaching from a distance or engaged in a conversation with active gesturing [11]. Performance when viewing the entire person was compared with conditions where the body or face was obscure. Identification was always best when the entire person was seen in motion. Recognition from the whole person exceeded recognition of face alone only when the person was seen in motion (Figure 2). Thus, when a stationary person is seen, we attend to the face, at the cost of processing relevant information in the body. In motion, attention is more equally distributed to the entire person and facilitates the use of a wider array of identity cues available from the body and biological motion.

The mediating role that motion plays in binding together faces and bodies was also demonstrated in a recent study where participants viewed either a whole person in motion or multiple still images, followed by a test image of a face or whole body [12]. (See Figure IA in Box 1 for the whole body condition.) When the video was seen first, the body contributed significantly to person recognition, above and beyond the face. After exposure to the still images, only the face contributed to recognition. Thus, the general advantage for identifying people in dynamic over static presentations was mediated by information from the body that went unnoticed when only static images were available for familiarization. Consistent with this idea of motion in binding the face and body together into a person, face recognition is faster when the face is attached to a body in motion than when seen alone, even when the body provides no diagnostic identity information [13].

In summary, when considering the dynamic whole person, the body and motion contribute significantly to person recognition beyond the static image of a face. The contribution of the body
is particularly evident when seeing someone from a distance or when viewing conditions are poor. Importantly, person recognition is best when we see the whole person in motion (Figure 2).

**Biological Motion from Faces Contributes to Person Recognition ‘Close-Up’**

The role of facial motion in person recognition from close-up views has been reviewed elsewhere [14–16], including in two recent papers [17,18], the latter of which examines the role of motion in the development of face recognition skills. Thus, we provide only a short synopsis of the findings here. The available data point to the conclusion that we rely on dynamic identity signatures from faces for recognition, when we know a person well and when image quality is poor [19–21]. It is important to remember that, in evolutionary terms, the sudden appearance of a high resolution, close-up view of a face is a relatively uncommon event. In more natural viewing conditions, a familiar person approaches from a distance. In logical terms, therefore, dynamic identity signatures from facial motion are needed for recognition only if biological motion from the body and body shape cues fail from a distance and the face fails closer up. Consequently, facial motion is rarely needed for familiar person recognition, despite evidence that it can be reliable when all other cues fail [15,19,20,22].

In summary, visually based person recognition begins at a distance and evolves over time and across viewing proximity. The certainty of a recognition decision can increase as additional identity information becomes perceptually available. Ultimately, the distance at which we recognize someone depends on their familiarity and the quality of the viewing environment. Despite limited direct evidence, familiarity is likely to prove an important mediating variable in the integration of multisensory identity cues. For now, the evidence indicates that motion can act both as a direct cue to identity and as a mechanism for integrating the face and body into a person [23].

**The Contribution of the Dynamic Face to Voice Recognition**

Our ability to identify people does not depend only on visual information from the face and body, but also on our ability to recognize people based on their voices (for reviews, see [24–27]). At
what stage of processing is identity information from the face and voice integrated? According to one model, faces and voices are processed in parallel, each within its own modality-specific perceptual module, and are integrated only at a postperceptual, a-modal hub, providing access to semantic information known as the person identity node (PIN) [28] (Figure 3A). According to an alternative model, face-voice integration is enabled by direct connections and/or correlated activations of unimodal face and voice-selective mechanisms and may already take place at perceptual stages [29,30] (Figure 3B) (for recent reviews, see [27,31,32]). These models, however, do not consider the important role dynamic information plays in face-voice integration. Voices are typically generated by dynamic rather than by static faces and, therefore, face-voice integration is likely to be mediated by mechanisms that are specialized for the processing of the temporal variations from the dynamic face and voice, prior to the generation of an a-modal semantic representation (Figure 3C).

The critical role that motion plays in determining the identity of a person based on its face and voice has been intriguingly demonstrated in studies that revealed that participants can match the identity of a video of a mute speaking face and the speaking voice of unfamiliar people even when the voice and mute face say different sentences [33–35]. Such matching between voices and their corresponding static face images is below chance level, indicating the crucial role of motion in matching the identity of a face to its voice. The various features that may be used for such dynamic face-voice identity matching were systematically assessed in a later study [36]. Matching performance was reduced when the intonation of the voice and unheard speech of the mute speaking face were different (e.g., a question versus a sentence or conversational versus clear speech), but not when they differed by the content of the sentence or were speaking at different speeds. This ability to match audiovisual dynamic identity signatures may also account for findings showing that speech recognition is more accurate following the presentation of PLD of face movements of the same individual [37,38]. Thus, similar dynamic changes can be extracted from faces and voices and matched within individuals. The correspondence
between the dynamic face and voice is further demonstrated by studies that assessed the synchronicity of face and voice onset times. Voice recognition was best near synchrony and up to a 100-ms auditory lag [39]. These findings are consistent with speech perception studies [40,41] and, therefore, may reflect a general temporal integration mechanism of the voice with the dynamic face.

The fact that dynamic identity signatures of speaking mute faces and voices can be matched for unfamiliar faces predicts that such correspondence will be even greater for familiar faces. Indeed, in a study that assessed voice recognition with dynamic or static faces, subjects were asked to discriminate between personally familiar and unfamiliar voices [42]. Voices were presented either alone, with a congruent face (same identity), or with an incongruent face (different identity). Performance on the voice recognition task was lower when it was presented with an incongruent dynamic face, whereas an incongruent static face did not interfere with voice recognition. These effects were larger for familiar than unfamiliar voices. In a follow-up electrophysiological study, dynamic face-voice stimuli of familiar people generated much stronger neural responses than face or voice alone as early as 50–80 ms after stimulus onset [43]. Differences between identity congruent and incongruent face-voice stimuli emerged at about 250–600 ms after stimulus onset. This difference was long lasting, possibly reflecting the accumulating information that can be extracted from the dynamic face-voice stimuli. Thus, face-voice integration takes place at early perceptual stages (see also [30]), followed by mechanisms that evaluate the correspondence between the identity of the face and the voice.

In summary, in natural settings, voices are emitted by dynamic faces. The studies discussed above suggest that voices and their corresponding dynamic faces generate unique temporal patterns for each individual that can be used for linking the dynamic face to its voice in the process of person recognition. We posit, therefore, that identity information from the dynamic face and voice is integrated by multisensory, perceptual mechanisms that extract dynamic identity signatures from the face and voice for both familiar and unfamiliar people. For familiar people, this dynamic multi-modal perceptual representation is further integrated with conceptual information in postperceptual semantic hubs. This facilitates our ability to match a face to its voice (Figure 3C). Next, we will discuss the neural mechanisms that may underlie this process.

The Superior Temporal Sulcus: The Neural Hub for Dynamic Person Recognition

In the previous sections, we considered the important role motion plays in the recognition of the whole dynamic, talking person. We argued that this is achieved by the extraction of dynamic identity signatures from the face, body, and voice and by binding this diverse information into a unified, multisensory representation of the whole person. Which neural mechanisms underlie our ability to recognize people in dynamic real-life settings? The processing of identity has been traditionally attributed to the fusiform face area (FFA) [44,45]. This is likely due to the fact that motion has not been considered important or useful for person recognition. Given recent findings of the limited selectivity of FFA for dynamic facial information ([46]; see also Box 4), it seems unlikely that it would play a major role in recognizing dynamic people. A more likely candidate for recognition of the whole dynamic talking person is the STS, which activates strongly in response to dynamic faces [46,47], biological motion [48–50], and human voices [51,52] (Figure 1B).

Why has STS not been considered previously in the context of person recognition? There are two reasons. The first is the longstanding tradition of studying person recognition with static images of faces: a stimulus that activates the STS far less effectively than dynamic faces or bodies [46]. The second reason is the emphasis on studying the role of the STS in social cognition [53–55]. Social cognition refers to the social and emotional information that can be inferred from people’s behavior. This topic has included understanding others’ actions,
Box 4. Neural Models of Face Processing

Neural models of face processing (Figure 1) provide a functional framework for a so-called core system of face-selective brain areas in the lateral occipital cortex (OFA), the fusiform gyrus (FFA), and the posterior STS [pSTS face area (pSTS-FA)] [44,56]. In this classic distributed face model, the core face network is composed of two pathways that extract different types of information from faces. The ventral pathway processes invariant information from faces and includes the OFA and FFA. Invariant facial information is proposed to be useful for identifying and visual categorization (e.g., by sex, age, and race). The dorsal core system resides in the pSTS-FA and processes changeable information from faces (e.g., expression and gaze) [56]. This division of labor for processing invariant and changeable facial aspects has been challenged recently by multiple findings indicating that the FFA is also involved in the processing of facial expressions (cf. [60] for a review).

The distributed face model [56] was amended soon after it was proposed to account for our ability to remember dynamic identity signatures (Box 1) [15]. Specifically, the putative function of the pSTS-FA was expanded to include a representation of the idiosyncratic gestures and movements of faces, which can be used for identification. This proposal is consistent with recent studies that find much stronger response to dynamic than static faces in the pSTS and additional areas in the anterior STS [46,86], but similar responses in the OFA and FFA [46,86].

The combination of the findings suggests the need for caution in developing functional neural models of recognition without data that specifically address differences in neural responses to static and dynamic faces [86,87]. In particular, the division of labor between the STS and FFA should be re-evaluated to determine whether ‘motion-versus-form’ or ‘changeable-versus-invariant’ better describes the functional distinction between dorsal and ventral face pathways, respectively (see also [60]).

![Figure 1](image_url)

**Figure 1. Two Neural Pathways of Face Processing.** (A) According to the classic distributed model, the face system is composed of two pathways, a ventral pathway that extracts invariant facial aspects and a dorsal pathway that extracts changeable facial aspects [44,56]. (B) According to an alternative model, the division of labor between the dorsal and ventral face streams is to motion and form, respectively [15,60,86,87]. Abbreviations: FFA, fusiform face area; OFA, occipital face area; STS-FA, superior temporal sulcus face area; MT, see Glossary.

intentions, emotional, and mental states, but has not included person recognition. The exclusion of identity processing from other types of social communication skill is evident also in current neural models of face recognition [44,56]. These models dissociate the processing of invariant facial aspects, such as identity and gender (attributed to the FFA) and changeable aspects, such as expression, eye gaze and mouth movements (attributed to the STS) (Box 4). The supporting data for these models come largely from responses to static images of faces. In the broader context of a dynamic talking person, it seems clear that the STS has the necessary machinery for processing person identity, including the extraction of unique dynamic signatures from the face, body, and voice; the extraction of invariant shape cues from form-from-motion processing; and the multisensory response characteristics needed to combine these cues.

On the question of face–body integration, again the neural substrates of person perception have been studied primarily with static images. This has revealed an important role for the FFA, but not the occipital face area (OFA), in the representation of a static image of the whole person [57,58] (Figure 1B). Studies using dynamic whole bodies as a measure of biological motion, however, often present PLDs of moving people and compare them to scrambled dots [50,55], or focus on the processing of action rather than recognition (for review see [59]).
The STS is also likely to play a central role in recognizing people based on their voices. Current neural models of face-voice integration suggest that this process may be mediated by direct connections between the FFA and voice areas in the STS (Figure 3B) [27,29]. However, given the important role that facial motion plays in voice recognition, and the evidence that the FFA is not sensitive to dynamic information from faces, we suggest that person recognition of the dynamic speaking person is more likely to be mediated by the face and voice areas in the STS (Figure 1B). As noted, the temporal variations of speaking faces and voices can be matched within individuals, even for unfamiliar faces. This requires a mechanism that is highly sensitive to subtle dynamic variations in the face and voice across time. For familiar people, the multisensory dynamic identity signature is integrated with semantic information in the anterior temporal lobe [27,31]. Thus, the multi-modal representation of person identity takes place at the perceptual rather than post-perceptual (i.e., conceptual) stages with the STS playing a central role in this process.

To summarize, in contrast to current neural models that highlight the important role of the FFA in person recognition, considering person recognition in real-life settings reveals that motion plays a critical role in recognition. The STS is well equipped to process dynamic identity signatures from the face, body, and voice. This new view implicates the STS in the processing of multisensory dynamic information for both person recognition and social/emotional cognition. It further emphasizes the possibility that the division of labor between the STS and FFA may be more driven by the distinction between motion and form than by the type of tasks served by changeable versus invariant information about people (Box 4; for a review, see [50]).

Concluding Remarks

In this review, we considered how person recognition is accomplished in the dynamic unconstrained environments that characterize real-life settings. We suggest that, in such settings, person recognition depends on all available cues, including face, body, voice, and body motion, and that motion is critical in gluing together this diverse and complex multisensory information. Because the majority of research on person recognition still focuses primarily on the perception of static images of faces, our understanding of the recognition of the whole dynamic, talking person remains limited. However, as more findings converge to suggest the importance of diverse multisensory cues to recognition, we believe that the field of face recognition will slowly turn toward the study of the whole person, taking into account all of the available information for successfully recognizing people (see Outstanding Questions).

References


Outstanding Questions

We have posited a neural system in the STS responsible for extracting identity information from a dynamic, speaking person. How does this system link with brain areas in the anterior temporal lobe that have been implicated in the processing of semantic information about the identity of familiar people?

How do mechanisms of dynamic person recognition in STS interact with the action recognition system that activates mirror neurons?

What is the extent of overlap, or mechanisms of interface, between person recognition via motion-based identity signatures and action recognition?

What is the relative contribution of dynamic identity signature and form-motion processes in familiar whole-person recognition?

What is the functional organization of dynamic face and voice areas in the STS? What is the extent of overlap and what role do they play in extracting identity information from the dynamic face and voice?

How can the time course of person recognition be characterized? We have argued that the perceptual utility of multisensory person recognition cues is variable based on the visual and acoustic environment. Does evidence from different cues accumulate in a centralized process or are recognition systems guided by the most reliable of the available cues at any given time?

What new neural and psychological methods can be brought to bear in understanding the nature of cue integration in person recognition? Computational vision and biometrics researchers have relied strongly on statistical fusion methods for combining multiple (uncertain) cues to identity. Can fusion methods serve as a model for understanding neural and psychological cue integration?

Can neural case studies in prosopagnosia offer insight into the connection between static and dynamic recognition systems? Hypothetically, a motion-based recognition system in the STS could operate when static face-processing areas in the ventral stream face areas are impaired. More
41. Schwartz, J.L. and Savariaux, C. (2014) No, there is no 150 ms lead of visual speech on auditory speech, but a range of audio-visual asynchronies varying from small audio lead to large audio lag. PLoS Comput. Biol. 10, e1003743


