Vision Research 51 (2011) 74–83

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

Recognizing people from dynamic and static faces and bodies: Dissecting identity with a fusion approach

Alice J. O'Toole [a,*], P. Jonathon Phillips [b], Samuel Weimer [a], Dana A. Roark [a], Julianne Ayyad [a], Robert Barwick [a], Joseph Dunlop [a]

[a] School of Behavioral and Brain Sciences, University of Texas at Dallas, 800 W. Campbell Rd, Richardson TX 75080, United States
[b] National Institute of Standards and Technology, United States

ARTICLE INFO

Article history:
Received 4 June 2010
Received in revised form 22 September 2010

Keywords:
Face
Dynamic
Gait

ABSTRACT

The goal of this study was to evaluate human accuracy at identifying people from static and dynamic presentations of faces and bodies. Participants matched identity in pairs of videos depicting people in motion (walking or conversing) and in "best" static images extracted from the videos. The type of information presented to observers was varied to include the face and body, the face-only, and the body-only. Identification performance was best when people viewed the face and body in motion. There was an advantage for dynamic over static stimuli, but only for conditions that included the body. Control experiments with multiple-static images indicated that some of the motion advantages we obtained were due to seeing multiple images of the person, rather than to the motion, *per se*. To computationally assess the contribution of different types of information for identification, we *fused* the identity judgments from observers in different conditions using a statistical learning algorithm trained to optimize identification accuracy. This fusion achieved perfect performance. The condition weights that resulted suggest that static displays encourage reliance on the face for recognition, whereas dynamic displays seem to direct attention more equitably across the body and face.

1. Introduction

In the real world, we interact with people *in motion*. These interactions typically begin at a distance and unfold over time, as a person approaches, and ultimately stands "face-to-face" with us. The recognition of a person in natural viewing conditions, therefore, begins with a glimpse at the overall shape of a person and builds toward more confident judgments as the particularities of the movements, body structure, and face are integrated and processed.

The human face has generally been regarded as the most easily accessible and accurate entry point into the task of determining a person's identity from visual cues. Despite evidence that humans excel at recognizing familiar faces (Burton, Bruce, & Hancock, 1999a), performance is less impressive for relatively unfamiliar faces (Hancock, Bruce, & Burton, 2000). In particular, there is evidence to indicate that recognition can be poor under viewing conditions that are non-optimal or are poorly matched to those in which a face is learned. The difficulties humans have with unfamiliar face recognition can be mitigated potentially by relying on a broader array of identity cues available in natural viewing conditions. These include the shape and structure of the body, as well as gait and other gesture-based movements of the body. Body motions and gestures that are idiosyncratic or "identity-diagnostic" have been referred to previously as *dynamic identity signatures* (O'Toole, Roark, & Abdi, 2002).

There is surprisingly little psychological work aimed at understanding the extent to which humans use visual information, *beyond the face*, to identify people. Most commonly, in past studies, identity perception from biological motion stimuli has been examined. For example, Kozlowski and Cutting (1977) found poor, but above chance performance, for identifying friends from point-light motion displays. Westhoff and Troje (2007) demonstrated that people could learn to discriminate a small number of individuals using their motions. Moreover, Loula, Prasad, Harber, and Shiffrar (2005) demonstrated that humans are most sensitive to point-light motions of themselves and friends, but are not able to discriminate the motions of strangers.

Using more natural dynamic viewing conditions, Burton, Wilson, Cowan, and Bruce (1999b) considered the relative contribution of the face versus body for recognition in dynamic viewing conditions. They looked at identification of people captured on CCTV as they walked through a door and found that observers performed quite poorly when they were unfamiliar with the person in the video, but were nearly perfect when the person was known to them. Davis and Valentine (2008) confirmed the finding that

* Corresponding author.
E-mail address: otoole@utdallas.edu (A.J. O'Toole).

0042-6989/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.visres.2010.09.035

matching unfamiliar identities in video is highly susceptible to error and found that this held across low-, moderate-, and high-quality video. Burton et al. (1999b) also found that identification performance declined substantially when the face in the video was obscured, but remained high when the body was obscured. This result suggests that even with more complete information about the face and body, recognition performance is supported more strongly by the face than by the other information in the video.

In static displays, Robbins and Coltheart (in press) likewise demonstrated the importance of the face in identifying relatively *unfamiliar* people. In that study, observers learned people from full body pictures and were tested with composite images made from the head of one person on the body of another person. People were more accurate at identifying people from their heads than from their bodies. Moreover, in integrating information from the combined face and body, Robbins and Coltheart (in press) found a greater degree of holistic processing across the right–left mid-line halves than across the top- and bottom- halves of the full body image. They conclude that the head is more important than the body for recognition, but that the body can also provide identity information, when the person is processed as an integrated whole.

In the context of viewing people in motion, Pilz, Bülthoff, and Thornton (2006) have also considered the question of how we integrate information across the face and body in making an identification decision. They placed three-dimensional head models from different people onto a single identical moving body, defined by an avatar. Observers responded more quickly to a target face when the body was approaching than when it was static. In a second experiment, they found that faces learned on an approaching avatar, were responded to more quickly than those learned on an avatar that was static. These findings suggest that natural approach motions may facilitate the processing of a face. However, the body information in the Pilz et al. (2006) study did not vary. Thus, it remains an open question if approach motion would likewise facilitate the processing of the body if it carried individuating information.

From a neural perspective, the visual processing of faces and bodies from dynamic and static displays is likely to involve a complex network of brain regions. Based on evidence from human neuropsychology and primate neurophysiology, Haxby, Hoffman, and Gobbini (2000) proposed a distributed neural network that divides the processing of the invariant and changeable aspects of *faces* into two streams. According to this model, the invariant features of faces, those useful for face identification, are processed in the ventral temporal areas of the cortex near the fusiform gyrus (cf. fusiform face area, FFA, Kanwisher, McDermott, & Chun, 1997). The changeable aspects of faces (e.g., expression, gaze), useful for social communication, are thought to be processed in the posterior Superior Temporal Sulcus (pSTS) along the dorsal stream of visual processing. (See Shultz & Pilz (2009) for a review of recent functional neuroimaging results for viewing natural face motions.)

As noted by Haxby et al. (2000), the invariant information in faces supports the function of identifying people, whereas the motion-based changeable information supports a social communication function. Given that the neural systems responsible for these functions are, to a first approximation, functionally and anatomically distinct, the question arises as to how facial motions contribute to *face recognition*. The task of recognizing someone is based presumably more on the invariant structure of a face. In theoretical terms, O'Toole et al. (2002) proposed two ways that motion could benefit face recognition. The *supplemental information hypothesis* posits that we represent dynamic identity signatures in addition to the invariant features of faces. The *representation enhancement hypothesis* posits that motion benefits face recognition by perceptual structure-from-motion processes that enable a better three dimensional representation of a face (O'Toole et al., 2002). To date, there is strong support for the supplemental information hypothesis, and hence the use of dynamic identity signatures for face recognition, but only limited support for the representation enhancement hypothesis (O'Toole & Roark, 2010).

Although the Haxby et al. (2000) and O'Toole et al. (2002) models were proposed to account for face processing, some essential elements of these perspectives may apply analogously to the recognition of people from natural viewing of full bodies. It has been known for sometime that the pSTS plays an important role also in processing body motion as well as the motion of individual body parts (e.g., hands) (cf. Allison, Puce, & McCarthy, 2000; Pinsk, DeSimone, Moore, Gross, & Kastner, 2005). As noted, for the face, and possibly body, the role of pSTS may be primarily for processing social communication movements (Haxby et al., 2000). By extension, the pSTS may also have a role in recognition via dynamic identity signature processing (O'Toole et al., 2002).

The extra-striate body area (EBA) may likewise contribute to the recognition of people from static images of bodies and body parts (Downing, Jiang, Shuman, & Kanwisher, 2001). This region, located in the lateral occipital cortex, responds to still images of bodies and body parts more strongly than it responds to a variety of control images, including faces. Downing et al. (2001) have suggested a role for the EBA in representing the visual appearance of bodies. In particular, they suggest a role for EBA in identification when viewing conditions are poor and the face is not easily accessible due to poor lighting, occlusion, or viewing direction. Some studies have also proposed a role for EBA in processing body motions with the goal of understanding actions and intent (Astafiev, Stanely, Shuman, & Corbetta, 2004), but this finding remains controversial (Downing, Peelen, Wiggett, & Tew, 2006; Peelen & Downing, 2005).

Combined, the data from functional neuroimaging studies indicate a widely distributed network of neural regions involved in processing faces and bodies, both from static and dynamic stimuli. These studies also suggest that neural regions may differ in the extent to which they subserve different tasks, including the processing of social signals (pSTS), the recognition of intent (pSTS, EBA), and person recognition (FFA, EBA, and pSTS). The complexity of the neural processing belies a simpler question about how humans use the information in faces and bodies for identifying someone under natural viewing conditions, when a face is attached to a body and is experienced intermittently in motion and at rest. A better understanding of how humans identify people from static and dynamic information in the face and body can constrain the interpretation of the neural data.

The goal of the present study was to systematically assess the contribution of the face and body for making an identity judgment in static versus dynamic presentation conditions. We also tested the extent to which identification advantages in video could be accounted for by the presentation of "more information about a person" from the multiple-static images that comprise the video sequence. We carried out a series of experiments in which participants matched "person identity" (same or different?) in pairs of static images/videos. We used this identity matching task to assess the quality of information available perceptually, without requiring longer-span memory resources. For all experiments, the task was to determine whether two images/videos were of the same person or of different people. The experiments differed only in the type of stimulus used for the identity match. In Experiments 1a, 2a, and 3a, participants viewed pairs of videos. In Experiments 1b, 2b, and 3b, identifications were made on the "best" image extracted from the videos. The stimuli used in Experiments 1a and 1b included both the face and body. For Experiments 2a and 2b, only the face was visible and for Experiments 3a and 3b, only the body was visible. As we shall see, the face and body and body-only

Table 1
In this table, we give a summary of the experiments, with their presentation and information-type conditions. *N* is the number of participants in each experiment, divided between the CC, CG, and GG conditions. The main effect of match type (CC, CG, and CG) is reported in the last column and is significant in all but two cases (see text for details).

Experiment	Information	Presentation	*N*	Main effect
1a	Face and body	Video	48	$F(2,45) = 9.21$, $p < .001$
1b	Face and body	Static	30	$F(2,27) = 1.25$, $p < ns.$
1c	Face and body	Multi-static	30	$F(2,27) = 3.37$, $p < .05$
2a	Face-only	Video	30	$F(2,27) = 4.54$, $p < .001$
2b	Face-only	Static	36	$F(2,33) = 12.12$, $p < .001$
3a	Body-only	Video	30	$F(2,27) = 10.03$, $p < .001$
3b	Body-only	Static	31	$F(2,28) = 9.39$, $p < .001$
3c	Body-only	Multi-static	30	$F(2,27) = .36$, $p < ns.$

experiments yielded a video advantage. Therefore, we carried out multi-static control experiments (Experiments 1c and 3c) to test the extent to which the video advantage could be accounted for by the extra image-based information in the video. Table 1 gives a summary of stimulus conditions in each experiment.

Within each experiment, we also varied the types of videos presented for identity matching. In one condition, participants saw pairs of "gait" videos, picturing a person walking toward a camera. In a second condition, they saw pairs of "conversation" videos, picturing the subject conversing with another person. In a third condition, participants had to match the identity of the two people between a conversation and gait video. We expected performance to be best for the gait stimuli, because the quality and resolution of the final frames of these videos were better than any of the images in the conversation videos. The primary reason we used different types of match conditions was to diversify the stimulus types, allowing for a more general test of the main questions of the study. These general questions focused on video versus static presentations and recognition from the face versus body.

Next, we applied a *fusion strategy* to the task of quantitatively and qualitatively assessing how to optimally combine human identity judgments based on different information (face and/or body, viewed in static or dynamic displays) to improve identification. Fusion has been used widely in computer vision applications to improve biometric identifications by combining information from multiple sources (e.g., face and fingerprint, or face and iris) (Ross, Nandakumar, & Jain, 2004). In general, the idea is that when partially independent information about a person's identity is available from multiple sources, the information can be combined to improve accuracy over that of the best performing source. Fusion algorithms vary in complexity from simple averaging of the judgments from different sources to pattern classification algorithms that learn a statistical mapping from the source judgments to the identification status (e.g., same or different person). Here we used a pattern classifier based on partial least squares (PLS) regression to implement the fusion. PLS combines elements of principal components analysis (PCA) and multiple regression and provides a set of weights for the optimal combination of information across sources. These weights can be used to assess the role of different information sources in creating an optimal identity judgment. As such, they can provide insight into the extent to which the information used by humans across these presentation modes is complementary, redundant, or independent.

2. Experimental methods

The methods were similar for all experiments, and so for brevity, we describe them once and include a brief section that details the stimulus manipulations undertaken in each experiment. We conducted these experiments independently using different observers so that we could use the same set of identity pairings in each experiment. This allows for the fusion across experiments to be based on independent participant judgments for single viewings of each identity pair.

2.1. Participants

Volunteers for the experiments were recruited from the undergraduate student population enrolled at The University of Texas at Dallas (UTD). Students received research credit as part of a course requirement for psychology majors. A minimum of 30 volunteers participated in each experiment. Exact numbers of participants for each experiment are indicated in Table 1. None of the participants had any previous familiarity with the people filmed in the images/videos.

2.2. Stimuli

A database of video clips and static images of faces and people (O'Toole, Harms, Snow, Hurst, Pappas, Ayyad, and Abdi, 2005) served as the source of stimuli for these experiments. There were multiple *gait* and *conversation* videos available for each person in the database. A *gait video* showed a person walking parallel to the line of sight of a stationary camera, starting at a distance of 10 m. The person is filmed as they walk toward the camera and veer off to the left to pass the camera (see Fig. 1 for a multi-frame example of these videos). The gait videos varied across individuals from 8 s to 11 s, depending on how quickly the individual walked. The average duration of the videos was 9.6 s. We decided not to edit these videos to a common duration in order to preserve natural differences in walking speed and style for individuals. A *conversation video* showed a person conversing with a laboratory staff member. The lab member stands with his/her back to the camera and the subject faces the lab member. The distance between the camera and the center point of the subject's trajectory was 10.4 m. The videos were filmed from the top of a short flight of stairs at a height of 3.5 m, looking down on the subject and the lab member. To encourage gesturing in the videos, the subject was asked to give directions to a building on campus. For the experiments, these 10 s videos were edited to be 9.6 s in length to match the average of the gait videos. Both types of videos were filmed in a building foyer with high ceilings, enclosed entirely on one side with glass windows. This environment approximates outdoor lighting and makes for variable lighting conditions across the set of videos because the position and intensity of the light (mostly the sun) varies on a stimulus-by-stimulus basis. There were two sets of images and videos for each person: an original set and a second, duplicate set of images and videos collected between one week and six months subsequent to the original set. Thus, across the two filming sessions, there are natural variations in the person's appearance including hairstyle, clothing, etc. This ensured that participants in the identity matching experiments could not base their decisions on transient cues such as clothing, or other artifacts.

To create stimuli for the *body-only* experiments, we obscured the face by blurring a circular region around and including the face in each frame of the video. To create stimuli for the *face-only* experiments, we applied a black-out mask to the entire image in each frame, exclusive of a circular bubble around the face. For the static presentations, we extracted the "best" still image from each video as follows. For the gait videos, this was the image taken closest to the camera that showed the face from the frontal view. For the conversation video, we chose a good image that showed the face from as close to a frontal view as possible. See Fig. 2 for examples of the stimuli.

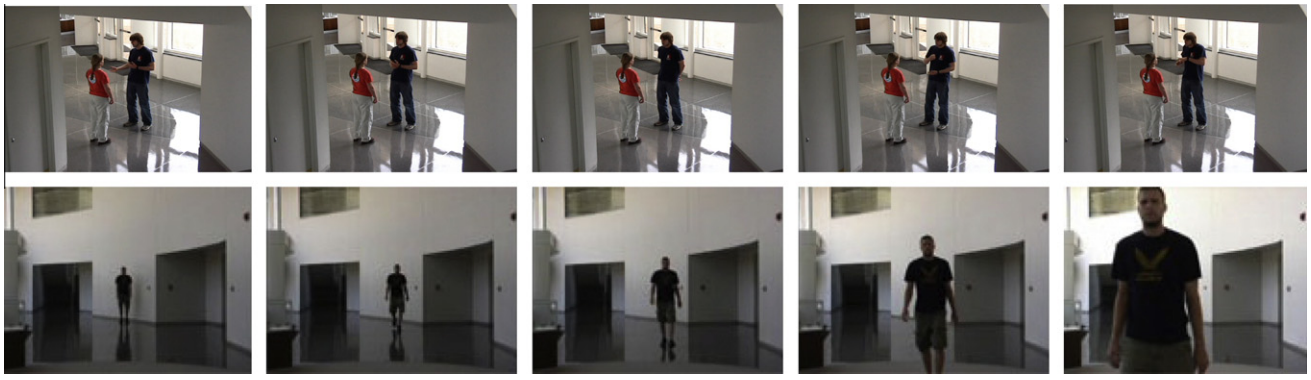

Fig. 1. Five frames extracted from an example of the conversation videos (top) and the gait videos (bottom).

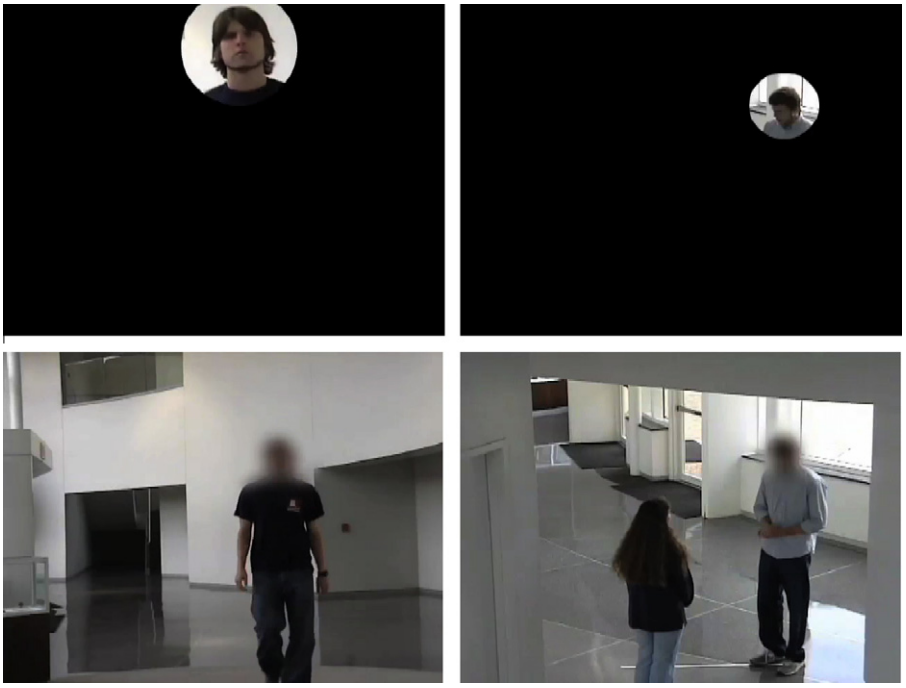

Fig. 2. Five frames extracted from an example of the conversation videos (top) and the gait videos (bottom).

To make stimuli for the multi-static control experiments, we extracted one image per second for each of the videos. Specifically, we took the first frame of each second of the video and presented these frames in sequence at a rate of one image per second. As noted previously, the original videos varied in length from 8 to 11 s. We did not shorten the videos to preserve the walking speed of the individuals. Thus, for comparability, we likewise allowed the number of multi-static frames to vary (between 8 and 11), keeping the sampling rate constant. The images were presented in the sequence in which they actually occurred in the video (i.e., not in random order), the image sampling was far enough apart in the video to eliminate apparent motion.

For all stimuli, in all experiments, the images subtended a visual angle of approximately 22.18° horizontally and 15.03° vertically. These figures are approximate, because participants were free to move their heads or the chair while they viewed the computer.

In all, there were 60 unique identities represented in the videos. All were young adult males between 19 and 30 years of age. Twenty identities were used to create identity-match pairs (i.e., two videos of the same person-presented in *match trials*). The remaining 40 identities were used to create no-match pairs (i.e., two videos of different people).

2.3. Procedure

Participants in each experiment were assigned randomly to one of three conditions. In the gait–gait (GG) condition, they matched identity in a pair of stimuli created from the gait videos (with the exact stimulus type determined by the experiment). In the conversation–conversation (CC) condition, participants matched identity in a pair of stimuli created from the conversation videos. In the conversation–gait (CG) condition, participants matched identity between stimuli created from a conversation video and a gait video.

The participants viewed pairs of videos (images) and were asked to determine if the people pictured were the "same person" or "different people". On each trial, they viewed the first video in the pair on the left side of the screen, followed by the second video presented on the right side of the screen. The screen went blank at the end of each video. For the best-static image experiments, the

first image appeared on the left side of the screen for 9.6 s (the average duration of the videos) and the second image appeared for 9.6 s on the rights side of the screen. Again, the screen went blank at the end of each image presentation. Next a prompt appeared with the following response choices: "(1) sure they are the same person; (2) think they are the same person; (3) do not know; (4) think they are not the same person; (5) sure they are not the same person." The prompt remained visible until the participant pressed a response key.

There were 40 trials in all: 20 matched identity trials and 20 non-matched identity trials. The order of stimulus presentation was randomized for each participant.

2.4. Results

The confidence ratings for the identity match task enabled the construction of ROC curves. These appear in Fig. 3 and offer an overview of performance across stimulus conditions (tests of statistical significance using the *d*'s computed from these data follow). Identification performance appears best when both the face and the body were presented in motion. In more detail, these curves suggest three kinds of results. First, comparing the right–left ROC curve pairs suggests an advantage for the video over static presentations for the face and body and body-only conditions, but not for the face-only condition. Second, performance for the face and body appears to be better than performance for either the face or body alone. Moreover, identification with the face-only is far better than identification with the body-only. Third, the relative placement of the ROC curves within each experiment indicates better performance for the GG condition over the CC and CG conditions, in all but the face-only video match condition. The CC and CG conditions were roughly equivalent in all but the static face-only condition.

Although the ROC curves provide a complete account of the data, these curves are difficult to test for statistical significance. Thus, to test for statistical significance, we computed a d' for discriminating matched and mismatched identity pairs each individual in each condition of the experiments. As indicated in Fig. 3, the bow-shape of the curves suggest that d' is an appropriate summary measure. To calculate a d', the responses must be divided into correct matches (hits) and incorrect matches (false alarms), which requires placing a somewhat arbitrary break in the confidence rating scale to define match and non-match responses.[1] The *d*'s were calculated by dividing the rating scale into "match judgments" (ratings of 1 or 2, in which the participant said "sure or think" same person) and non-match judgments (ratings of 3, 4 or 5, do not know and sure or think different people). We tabulated the proportion of *hits* and *false alarms* as follows. *Hits* were defined as match pairs that received ratings of 1 or 2 (i.e., sure or think that they are the same person). False alarms were defined as non-match pairs that received ratings of 1 or 2.

2.4.1. Overview experimental results

To examine the effects of video versus static presentation, as well as the kind of information presented (face and body, body, or face), we conducted a two-factor (video/static and information type) meta-anova, combining data across the six experiments.[2] An overview of the means for these conditions appears in Fig. 4. Consistent with the figure, there was a main effect of video versus static presentation, $F(1,199) = 17.17$, $p < .0001$, with video better than static. There was also a main effect of the information presented, $F(2,199) = 54.88$, $p < .0001$, with face and body best, followed by face-only and then body-only. Both main effects were qualified by the presence of a significant interaction between video/static presentation and information type, $F(2,199) = 4.81$, $p < .009$. The source of this interaction can be found in two results involving the face-only conditions. First, static and dynamic presentations were equivalent when only the face was presented. This was supported by simple main effects tests of the effect of presentation mode (video/static) in each information-type condition (face and body, face-only, body-only). These showed a significant effect of presentation mode in the face and body condition ($F(1,72) = 19.76$, $p < .0001$) and in the body-only condition ($F(1,55) = 9.71$, $p < .01$), but not in the face-only condition ($F(1,60) < 1$, *ns*). Thus, we conclude that observers did not benefit from seeing multiple images of the face from the video, or from the motion of the face in the videos. The lack of a motion effect for the face condition is not surprising as the videos show only rigid rotational and translation movements of the head.

The second component of the interaction is more interesting. This is the equivalence of the static face-only condition ($M = 1.75$, $SE = .11$) and the static face and body condition ($M = 1.78$, $SE = .10$) (Tukey HSD test, *ns*). It is worth noting that the face images from which the judgments were made in the static condition were included (identically) in the face and body static images. By identically, we mean that the size of the face image in the static face-only presentation was identical to the size of the embedded face in the face and body image. This finding suggests that when observers looked at the full person in a static image, they use only the face for the identity decision. By contrast, in the video presentation conditions, performance was better with the face and body than with the face alone. The fusion data we present shortly offers insight into this interaction.

Combined, the two components of the interaction result in three conditions with roughly equal levels of performance: (1) static presentation of the face and body; (2) static presentation of the face only; and (3) dynamic presentation of the face-only. These conditions stand in contrast to a substantial performance advantage for video presentations of the face and body together. Substantially lower performance is seen for the conditions that eliminate the face. These body-only conditions also show a video advantage.

2.4.1.1. Multi-static controls. Given the video advantage found for the face and body and body-only conditions we conducted a multiple-static image version of each of these two conditions. The equivalent performance for the face in the best-static image and video conditions suggests that more information about the face (i.e., more images/frames) would not improve performance.[3]

Across the video, static, and multi-static experiments, three patterns of performance are possible. A "pure motion advantage" should yield equivalent performance for the best-static and multi-static conditions. If both the motion and the additional static images contribute to the video advantage, performance in the multi-static control should fall between the video and best-static condition performance. If the video performance can be accounted for by the multiple images in the video, then the multi-static control condition will be at the same level as the video condition. We found examples of all three patterns in our findings.

The results of the two multi-static control experiments are plotted in Fig. 5 along with the video and best-static image results. Performance in the multi-static condition, relative to the video and

[1] We divided the scale to assign rating of 1 or 2 to "match" and responses of 3, 4 or 5 to non-match judgments, but we verified that the results were the same with the second obvious break point between 3 and 4.

[2] Because the effects of match type (GG, CC, and CG) were relatively consistent across experiments, for simplicity we omitted match type from the meta-anova.

[3] Although it is logically possible to find better performance in the multi-static condition without a dynamic advantage, if this were to occur, it would likely reflect a preference for presentation style (e.g., short exposures to multiple images might be helpful in attending to the images). Here, we set aside that possibility to focus on understanding the source of the motion advantages we obtained.

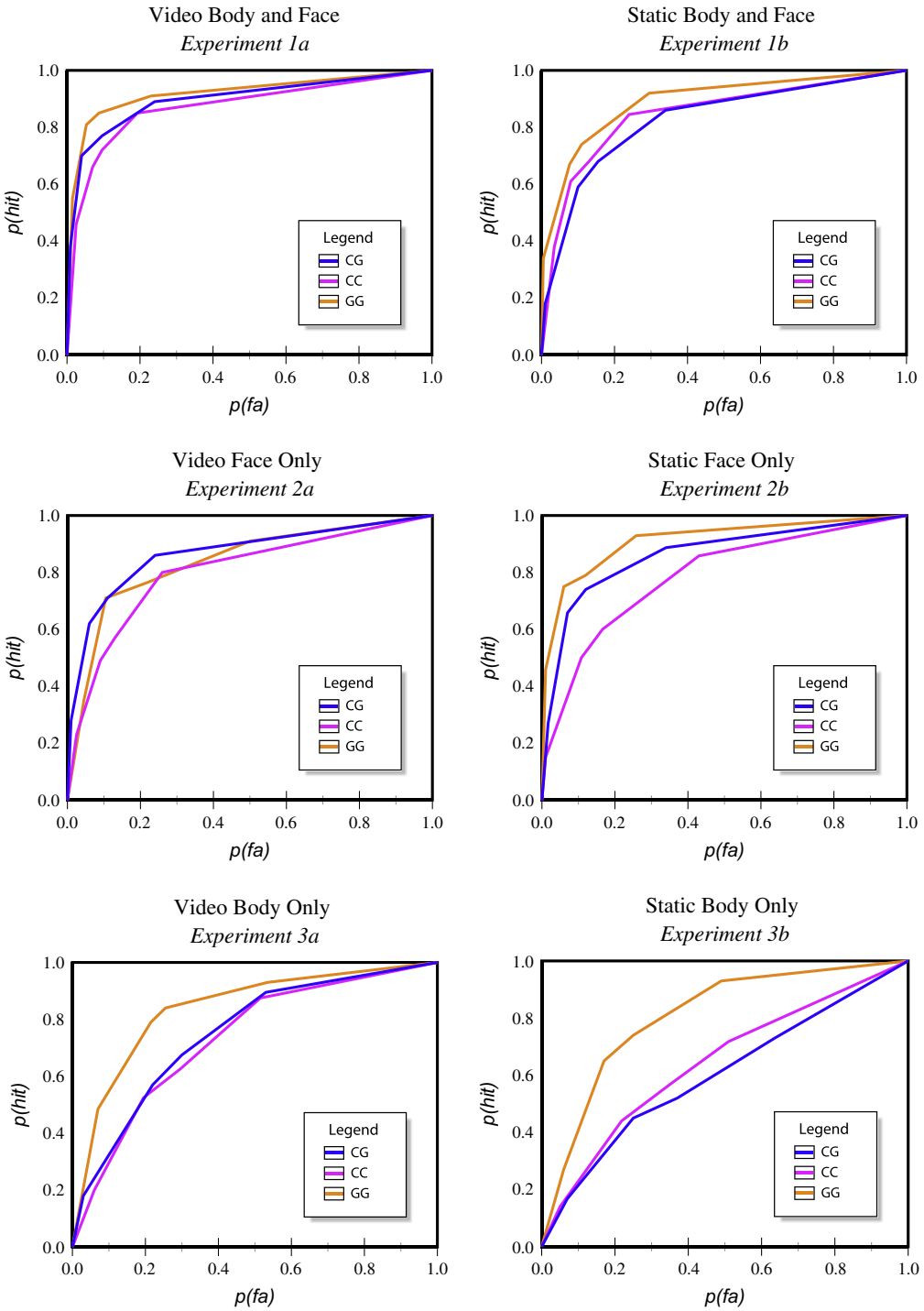

Fig. 3. ROC curves for the experiments show a video advantage for the face and body conditions and the body-only condition, but not the face-only condition. They also show a small advantage for the face and body conditions over the face and a stronger advantage for the face-only over body-only condition. There is a reasonably consistent GG advantage over the CG and CC conditions.

static presentations, yielded no "general" result. Starting with the face and body and body-only presentations, the GG comparison showed a pure video advantage, with the multi-static performance well below the video at the level of the single static presentation. This indicates that the video advantage in the GG conditions comes from using inherently dynamic information for identification. The fact that the pure video advantage appears only in the GG condition, where the motion in both videos (e.g., walking style) is similar enough to be useful for identification, is a further indication of the use of dynamic identity signatures.

At the opposite extreme, in the body-only CG and CC conditions, presentation of multiple-static images completely accounted for the video advantage. In the CG face and body condition, both the motion and the extra information in multiple-static images contribute to the video advantage. Again, the fusion simulations offer insight into these findings.

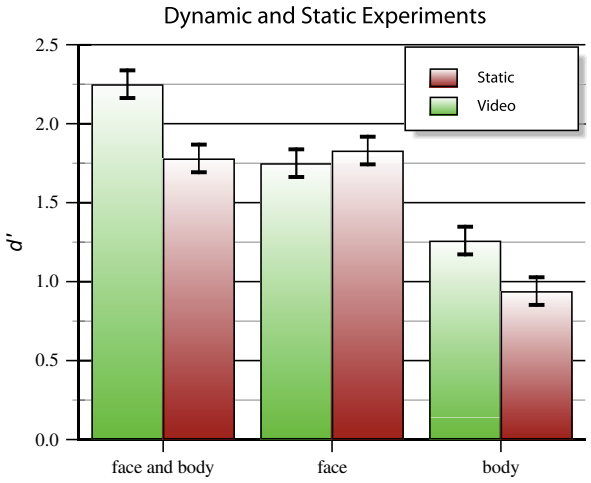

Fig. 4. An overview of the means for the first six experiments shows a video advantage for the face and body and body-only conditions. There is also an advantage for the face-only over the body-only conditions. Of note, the interaction between factors highlights the equivalent performance for a static presentation of the whole person and a static presentation of the face-only. Error bars indicate standard error of the mean.

Up to this point, the results show that human identification is at its best when the whole person was seen in motion. This indicates that people can benefit from complementary information about the face and body and that seeing the whole person in motion can, in some cases, add to the accuracy of the identification judgment. There was also evidence that performance with the face-only was far better than with the body-only. An interaction between body part and presentation mode suggests that the face "carries" identification in static presentations that include both face and body. Next, we consider the effects of match mode within the experiments.

2.4.1.2. Within-experiment match mode comparisons. As noted initially, the primary reason we used different types of match conditions was to diversify the stimulus types, allowing for a more general test of motion versus static presentations and the use of face versus body information. We assumed that differences in this variable would be due to the specifics of the information each provides. To determine the effects of the matching condition (CC, GG, and CG), in each experiment the data were submitted to a one-factor analysis of variance (ANOVA) with pair type as a between-subjects factor and d' as the dependent variable. A summary of results appears in Table 1 and shows statistically different performance across the CC, CG, and GG conditions in all but Experiment 1b, the static face and body match, and in Experiment 3c, the multi-static body match.

As expected, performance with the GG stimuli was generally best. This is likely due to the fact that the person in the gait video was quite close to the camera in the final frames of the video, offering a better view than any of the frames in the conversation videos. Notably, the performance using face-only in the CC and CG conditions was above chance, even given the small size and and low quality of the images. This performance may be based more on participants' accurate rejection of non-matched pairs than on confident judgments of matched pairs. The GG advantage was found in all but the video face-only experiment. We are uncertain why the video face-only experiment differed from the others for the GG advantage. Across the experiments the ordering of the CG and CC conditions varied, but was largely undifferentiated. Of note, for the static presentation of the face there was a relatively strong advantage for the CG multi-modal face comparison over the CC comparison. This seemingly odd result, where matching between images of higher and lower quality is better than matching between two lower quality images, is consistent with previous work (Lui, Seetzen, Burton, & Chaudhuri, 2003). Combined, these results suggests that the higher quality image can bootstrap face processing from the lower quality image.

3. Fusion

The purpose of the fusion simulations was to assess more quantitatively how the information presented to participants across the different experiments can be combined to support more accurate identity judgments. As noted, fusion methods are commonly used in computer vision and biometrics applications when there are multiple, but imperfect, sources of information that are useful for identification. Fusion can improve performance when the contributing information is at least partially independent and when an optimal formula for combining the information generalizes across exemplars. In other words, fusion will improve performance when

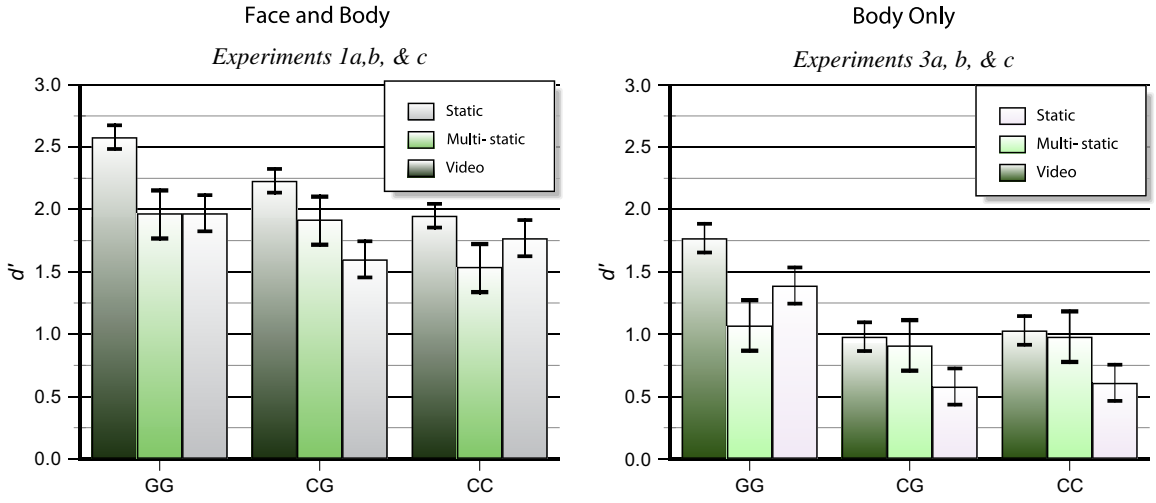

Fig. 5. The multi-static control experiments show a range of results from a clear demonstration of the movement in improving performance (GG) to a clear demonstration of multi-static images accounting for the video advantage (CG body-only, CC body-only), to a contribution from both movement and multiple images (CG face and body). Error bars indicate standard error of the mean.

the information or strategies humans employ in different conditions are complementary. We used fusion here as a tool for assessing how information across these sources is used by humans and to see how the presentation modes (dynamic or static) affect this pattern of use.

It is perhaps worth stressing that even if information is duplicated in conditions (e.g., static face and static face-body), it is nonetheless still possible to improve human performance with fusion. This could occur, for example, if viewing a particular type of stimulus affects the way humans allocate attention to different parts of the stimulus. We will see evidence of this type of effect in the fused combination of conditions.

4. Methods

Fusion of the experiments was accomplished with PLS regression, a technique that combines elements of principal component analysis and multiple regression (Abdi, 2003; Naes, Isaksson, Fearn, & Davis, 2004). The technique is used to predict a set of dependent variables from a set of independent variables. In other words, PLS is a standard pattern classification algorithm. In the present application, we used PLS to predict the true match status of the 40 pairs of faces (i.e., same person or different people) from the estimates made by humans (i.e., their ratings) under different conditions (e.g., dynamic face-only, dynamic body-only, etc.). Specifically, the classifier was trained to learn a statistical mapping from the human estimates to the ground-truth identity match status of the face pairs. The goal is to find a way to combine the human estimates in different conditions to improve performance.

The choice of PLS is in part arbitrary and we would expect other pattern classification algorithms to give similar results. We used PLS because it gives a set of easily interpretable weights for individual predictors. PLS yields a set of orthogonal factors, called latent vectors $t_1,\ldots,t_l$ from the covariance matrix of the predictors and dependent variables. The latent vectors (factors) are used to predict the dependent variable(s) by appropriately weighting the predictors. The set of weights is referred to as B_{pls} in the PLS-regression literature. Like other types of multivariate pattern analyses, PLS solutions are specified in terms of the number of factors (latent vectors) used (e.g., 2-, 3-, 4-factor solutions) for the prediction. We report the solution that improves performance most.

Also, as is the case for other pattern classifiers, the PLS should be tested for generalization using a cross-validation procedure. To cross validate classifier results, the test of classification is made on a stimulus (or stimulus set) not used in training the classifier. We implemented cross-validation by training the classifier with $n-1$ face pairs (i.e., 39) and testing it with the left-out pair. This procedure was implemented 40 times, iterating the left-out face pair through the set of available pairs. Thus, the performance we report is based on the proportion of times the correct match status of the left-out pair was predicted by the classifier.

The fusions we report are as follows. First, we carried out a fusion that combined identity judgments across all conditions of the six video and static experiments. Based on the results of this first fusion, three additional subset fusions were undertaken, combining data from within the stimulus type conditions (GG, CG and CC) across the body information conditions (face and body, face-only, and body-only).

4.1. Six-experiment fusion

The predictors used in this fusion were the estimates of the match status of the 40 pairs of identities (20 matched identities and 20 mismatched identities) from each of the three conditions (CC, CG, and GG) of the video and static experiments (Experiments 1a and 1b, Experiments 2a and 2b, Experiments 3a and 3b). For each pair of images/videos in each condition of each experiment, we averaged the response ratings (i.e., 1: sure the same person to 5: sure different people) across participants for the individual identity pairs. These averages were used as real-valued predictors that retain information both about the human participants' estimates of identity and their certainty. To equate the stability of the averages across the different experiments which varied somewhat in number of participants, we averaged the first 10 participants in each condition of each experiment. Recall that there were between 30 and 48 participants in each experiment, divided roughly equally between the CC, CG, and GG conditions. Thus, the minimum number of participants in each condition was 10, and so we used data from the first 10 participants in each component of the fusion. Thus, the predictor for each pair was the average of the participants' ratings of the likelihood that the people were the same. We had 18 such estimates (six experiments, three estimates per experiment) for each pair, that varied based on the type of information (face and body, body, or face) and presentation type (video, static) used in the different experiments. The dependent variable was the actual match status of the pair (same person/different people), quantified as 1 or 0.

A robust estimate of the fusion performance was determined in a cross-validation test in which the PLS regression was computed n times with $n-1$ identity pairs and tested with the nth "left-out" pair. The fusion performance we report is based on the proportion of correct match status classifications of the 40 face pairs. We tested a range of retained PLS factors to find the best performance.

4.2. Fusions for GG, CG, and CC conditions

Three additional fusions within the stimulus type conditions (GG, CG, and CC) were also conducted. For each of these, we extracted the appropriate stimulus type across Experiments 1a, 1b, 2a, 2b, and 3a, 3b. Each of these fusions used six predictors (video and static presentations of face and body, face-only and body-only conditions).

5. Results

The cross-validation six-experiment fusion classified the match status of the face pairs with 100% accuracy for both the 3-factor and 4-factor solutions. The weight patterns for these solutions were similar and showed that high-valued weights (i.e., those contributing most strongly) were concentrated in the GG conditions. This is likely due to the general performance advantage for the GG conditions across the experiments. For this reason, we divided the fusions into the GG, GC, and CC subset fusions.

The cross-validation fusion for the GG conditions, by itself, yielded perfect match classification accuracy, again for the 3- and 4-factor solutions. Thus, perfect performance was achievable from the information presented in the GG conditions. Again, the pattern of weights for the two solutions were similar, and so we averaged them. These averaged weights appear in Fig. 6 and show an intriguing result. The strongly weighted components for the static presentations are from the conditions that include the face (face and body, face-only). For dynamic presentations, the conditions that include the body (face and body, body-only) are strongly weighted. The result suggests that in the static presentation, the face dominates, and the body seems to add little useful information for identification. In the dynamic presentation, however, the body dominates with little independent or complementary contribution from the face. The result also suggests that the combination of the information humans assess most readily from the static presentations (the face) and information assessed most readily from the dy-

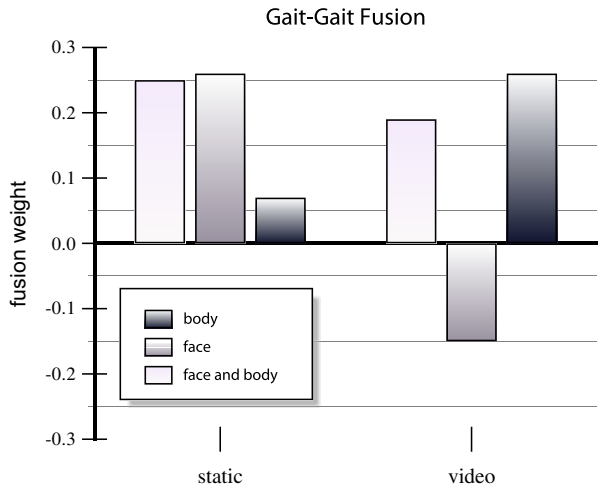

Fig. 6. The weights on the fused conditions indicate that in static presentations, the face dominates with the body adding little to the identification. In the video presentation, the body dominates with little independent or complementary information from the face.

namic presentations (the body) produced perfect identification. Note that the fusion does not indicate how humans combined information across static and dynamic presentations, but rather, how they *might combine* independent judgments made from the two presentation modes to optimize identification accuracy.

The cross-validation fusion for the CG condition did not achieve perfect match classification, although it did improve classification over the next best condition. The weights in this case, however, were roughly equivalent across all six sub-conditions used in the fusion, suggesting that observers rely on complementary information in each of the six conditions. The CC condition fusion did not improve match status classification accuracy, but rather, in all cross-validation solutions, proved worse than the best input condition. This suggests that there was no formula for combining the identity information across these conditions in a way that generalized across the face pairs. More likely, different combinations of condition-based estimates might be better suited to different subsets of the identity pairs.

6. Discussion

When we recognize a person in the real world, we see the *whole person*, in motion and at rest. In this study, we examined the effects of dynamic and static presentations of the face and body for recognizing people in relatively natural viewing conditions. The primary finding of this study is that human identification is at its best when the whole person was seen in motion. This indicates that identification can benefit from both the face and body, and that seeing the whole person in motion can add to the accuracy of the identification judgment. In other words, recognition in the present study was most accurate when the conditions approximated natural viewing conditions, that include a person approaching.

In dissecting this natural viewing condition advantage, a striking finding was the equivalence of the static face and body and static face-alone conditions. Consistent with previous studies (Burton et al., 1999b; Robbins & Coltheart, in press), the present data confirm human reliance on the face for identification in static viewing conditions, even when the body is available. From this result, it is tempting to conclude that the static body does not, or cannot, provide useful information for human identification. This conclusion is at odds, however, with the solid performance (i.e., $d' \approx 1.0$) we found in the static body-alone condition, indicating that humans *can* use the body for identification. Rather, a better interpretation of the combined findings is that body-based identity information (i.e., structure) is more likely to be used when the face is unavailable, or in real world terms, when viewing conditions for the face are poor.

In neural terms, areas in the inferior temporal cortex, including FFA and OFA, are the likely neural sub-strates for face processing from static images. Concomitantly, the use of static body information in the present study accords well with the function proposed for the EBA by Downing et al. (2001). Based on the particular responsiveness of EBA to static bodies, Downing et al. (2001) proposed a role for EBA in representing the visual appearance of bodies when viewing conditions are poor or when the face is not easily accessible due to poor lighting, occlusion, or viewing direction.

The second component of our empirical findings concerns the effects of motion on identification. Motion improved identification accuracy when the body was visible. This suggests that the body motions we see in natural viewing conditions can contribute to the visual representation of identity. Of note, these body-based video advantages came from different sources, which we probed by comparing performance in the best-static and dynamic conditions to a multi-static image control condition. The pure motion benefit we found in the gait-to-gait comparisons indicates the use of dynamic identity signatures for identification and fits with the supplemental information hypothesis (O'Toole et al., 2002), and thus a role for the pSTS in in person recognition. A prerequisite for using this information is that in the gait-to-gait comparisons, there was a match between the types of motion signatures available. Thus, stereotyped walking motions may have provided the supplemental motion-based identity information. Sarkar, Phillips, Lui, Grother, and Bowyer (2005) provide an overview of the computational issues involved in using gait for identification.

Other body-based video advantages could be accounted for entirely by seeing multiple images of the person. This was clearest for the body-only conversation–conversation and conversation–gait comparisons. This latter is a cross-modal comparison requiring observers to match across rather different image formats. In these cases, we found roughly equal performance for the video and multi-static conditions, at a level that exceeded performance for the best-static image condition. Of note, the video/multi-static advantage for the cross-modal case could not have been due to direct image matching processes between the comparison pair. In fact, the images embedded in the conversation and gait videos differed markedly in viewpoint, illumination, distance, and resolution. Rather the match task required observers to compare video/images between a higher quality (gait) and lower quality (conversation) stimulus. Consequently, the video and multi-static advantage had to have been based on active internal processing, whereby multiple images in the sequence are used to create a more robust representation than would be possible with the single image. Previous studies (Lui et al., 2003; Roark, Barrett, O'Toole, & Abdi, 2005) have likewise shown a kind of bootstrapping from lower to higher quality face recognition. All three findings suggest a process that actively constructs a more robust representation from low-quality stimuli, using internal resources from long term experience with faces and bodies. A computational illustration of combining images to improve the quality of a face representation for recognition can be found in a recent paper (Jenkins & Burton, 2008).

It is worth noting that the lack of motion benefit with *faces* should not be over-interpreted to suggest that we have no representation of facial motion in the identity code. Indeed, previous studies have demonstrated that *non-rigid facial motions* can be used for identifying someone (Hill & Johnston, 2001; Knappmeyer, Thornton, & Bülthoff, 2003). Under normal conditions, these non-rigid motions are visible only when we view a face from a short

distance. At this close distance, movement is generally not needed for identification, because of the high quality of the pictorial codes. Rather, consistent with distributed model, the primary function of non-rigid facial motions is likely to be social.

In the introduction, we proposed that a better understanding of how humans identify people from static and dynamic information in the face and body could constrain the interpretation of the complex neural network of brain areas that respond to faces and bodies. The fusion data offer a functionally based mechanism for applying these constraints to a complex data set. It is worth stressing that the fusion applied here does not tell us specifically how humans used the information in the various conditions, but rather how human identification judgments made in different stimulus and viewing conditions *could be* combined to optimize accuracy. The fusion results suggest that humans access non-redundant identity information from the face versus body to differing degrees from moving versus static stimuli. Specifically, it indicates that optimal performance can be achieved by combining human observer judgments from static viewing conditions that include the face and dynamic viewing conditions that include the body. One reason for the differential access of face versus body information from moving and static stimuli, may be based on the complex structure of neural areas processing face and body information for different reasons.

In summary, human judgments of identity are likely to be based on the collaborative computations of multiple representations of face and body, and their associated motions in the high-level visual processing network. A knowledge of how humans identify people in natural viewing environments can ground theories of how this identity information interacts in these neural networks.

Acknowledgments

This work was supported by funding from the Technical Support Working Group, US Dept. of Defense, to A.J. O'Toole. P. Jonathon Phillips was supported, in part, by the US Federal Bureau of Investigation.

References

Abdi, H. (2003). Partial least squares regression (pls-regression). In M. L. Beck, A. Bryman, & T. Futing (Eds.), *Encyclopedia for research methods for the social sciences* (pp. 792–795). CA, Sage: Thousand Oaks.

Allison, T., Puce, A., & McCarthy, G. (2000). Social perception from visual cues: The role of the STS region. *Trends in Cognitive Science, 4*(7), 267–278.

Astafiev, S., Stanely, C., Shuman, G., & Corbetta, M. (2004). Iextrastriate body area in human occipital cortex responds to the performance of motor actions. *Nature Neuroscience, 7*.

Burton, A. M., Bruce, V., & Hancock, P. J. B. (1999a). From pixels to people: A model of familiar face recognition. *Trends in Cognitive Sciences, 23*, 1–31.

Burton, A. M., Wilson, S., Cowan, M., & Bruce, V. (1999b). Face recognition in poor-quality video. *Psychological Science, 10*, 243–248.

Davis, J. P., & Valentine, T. (2008). Cctv on trial: Matching video images with the defendant in the dock. *Applied Cognitive Psychology, 23*(4), 482–505.

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science, 293*(5539), 2470–2473.

Downing, P. E., Peelen, M., Wiggett, A., & Tew, B. D. (2006). Is the extrastriate body area involved in motor actions? *Nature Neuroscience, 1*(1), 52–62.

Hancock, P. J. B., Bruce, V., & Burton, A. M. (2000). Recognition of unfamiliar faces. *Trends in Cognitive Sciences, 4*, 330–337.

Haxby, J., Hoffman, E., & Gobbini, M. (2000). The distributed human neural system for face perception. *Trends in Cognitive Science, 20*(6), 223–233.

Hill, H., & Johnston, A. (2001). Categorizing sex and identity from the biological motion of faces. *Current Biology, 11*, 880–885.

Jenkins, R., & Burton, A. M. (2008). 100% accuracy in automatic face recognition. *Science, 319*, 435.

Kanwisher, N., McDermott, J., & Chun, M. (1997). The fusiform face area: A module in human extrastriate cortex specialized for face perception. *Journal of Neuroscience, 17*, 4302–4311.

Knappmeyer, B., Thornton, I., & Bülthoff, H. H. (2003). The use of facial motion and facial form during the processing of identity. *Vision Research, 43*(18), 1921–1936.

Kozlowski, L. T., & Cutting, J. E. (1977). Recognizing the sex of a walker from a dynamic point-light display. *Perception & Psychophysics, 21*, 575–580.

Loula, F., Prasad, S., Harber, K., & Shiffrar, M. (2005). Recognizing people from their movements. *Journal of Experimental Psychology: Human Perception and Performance, 31*, 210–220.

Lui, C. H., Seetzen, H., Burton, A., & Chaudhuri, A. (2003). Face recognition is robust with incongruent image resolution. *Journal of Experimental Psychology: Applied, 9*, 33–44.

Naes, T., Isaksson, T., Fearn, T., & Davis, T. (2004). *Multivariate calibration and classification*. New York: NIR Publications.

O'Toole, A. J., & Roark, D. A. (2010). Memory for moving faces: The interplay of two recognition systems. In C. Curio, H. H. Bülthoff, & M. Giese (Eds.), *Dynamic faces: Insights from experiments and computation*. Cambridge, MA: MIT Press.

O'Toole, A., Roark, D., & Abdi, H. (2002). Recognition of moving faces: A psychological and neural perspective. *Trends in Cognitive Science, 6*, 261–266.

Peelen, M., & Downing, P. (2005). Is the extrastriate body area involved in motor actions? *Nature Neuroscience, 8*(125), 6996–7001.

Pilz, K. S., Bülthoff, H. H., & Thornton, I. M. (2006). Looming facilitates short-term face processing. *Journal of Vision, 8*, 1–13.

Pinsk, M., DeSimone, K., Moore, T., Gross, C., & Kastner, S. (2005). Representations of faces and body parts in macaque temporal cortex: A functional MRI study. *Proceedings of the National Academy of Sciences, 102*(19), 6996–7001.

Roark, D. A., Barrett, S. E., O'Toole, A. J., & Abdi, H. (2005). Learning the moves: The effect of facial familiarity and facial motion on person recognition across large changes in viewing format. *Perception, 35*, 761–773.

Robbins, R., & Coltheart, M. (in press). Heads, bodies and holistic processing in person recognition. *Journal of Vision*.

Ross, A., Nandakumar, K., & Jain, A. (2004). *Handbook of multibiometrics*. New York: Springer-Verlag.

Sarkar, S., Phillips, P. J., Lui, Z., Grother, P., & Bowyer, K. (2005). The humanid gait challenge problem: Data sets, performance, and analysis. *IEEE: Transactions on Pattern Analysis and Machine Intelligence, 27*, 162–177.

Shultz, J., & Pilz, K. S. (2009). Natural facial motion enhances cortical responses to faces. *Experimental Brain Research, 194*(3), 465–475.

Westhoff, C., & Troje, N. (2007). Kinematic cues for person identification from biological motion. *Perception and Psychophysics, 69*(2), 241–253.