IEEE International Workshop on Analysis and Modeling of Faces and Gestures
October 17 - 17, 2003
Nice, France
pp 36-43.

Human Recognition of Familiar and Unfamiliar People in Naturalistic Video

D. A. Roark, A. J. O'Toole & H. Abdi
School of Behavioral and Brain Sciences, GR4.1
University of Texas at Dallas
Richardson, Texas 75083-0688

Abstract

Understanding the human performance factors that mediate successful person identification can be helpful in the development of automatic face recognition algorithms. Face familiarity and facial motion are two factors that seem especially useful when subjects make recognition decisions from challenging viewing formats. We tested the effects of these two factors on person recognition from naturalistic, surveillance-like video. Subjects learned faces from either static photographs or facial speech videos and were asked to recognize people from whole body gait videos. We found that the more experience participants had with a face during learning (i.e., 1-view, 2-view, and 4-view conditions), the better their recognition performance for people in the whole body video gait clips. Thus, familiarizing subjects with high-resolution images or videos of faces was sufficient to improve recognition from low-resolution, whole-body images. Moreover, participants who learned faces from dynamic video clips were more accurate than participants who learned the faces from static images, but only when they were familiar with the faces. Facial motion and face familiarity may therefore play a role in understanding recognition when there are photometric inconsistencies between learning and test stimuli.

1. Introduction

It is well known that human memory for *familiar* faces and people is impressive. This is true even when recognition decisions must take place under less-than-perfect viewing conditions. These conditions include dim lighting, changes in viewpoint, viewing from a distance, and seeing the person in motion. By contrast, psychophysical studies of human memory for *unfamiliar* faces have indicated that even relatively minor photometric inconsistencies make recognition performance surprisingly error-prone. Changes in viewpoint, lighting, and image resolution are especially problematic for human subjects trying to recognize unfamiliar faces (see [8, 24] for reviews).

Automatic face recognition algorithms have been deployed in a variety of applied settings over the past decade (cf. [16, 17]). The performance of these algorithms is not unlike human subjects on the task of unfamiliar face recognition. Namely, computer-based face recognition systems are remarkably successful when the photometric parameters of the test stimuli match those of the database stimuli (see [16]). However, as is the case for human subjects recognizing unfamiliar faces, when the photometric parameters of target and database stimuli do not match, the performance of most algorithms decreases substantially (see [17]).

The most useful current and future applications of face and person recognition research lie in naturalistic contexts. For these applications, algorithms are required to operate on data from video cameras that capture the natural movements of people, often in public places [7, 23]. In most of these settings, the illumination comes from a natural source such as sunlight, for which the direction and strength of the light varies by time of day, season of the year, and weather conditions. This problem of recognition in natural illumination has been identified as an important area of future research [17].

Adapting current algorithms to address the problem of recognizing moving people in naturalistic viewing conditions is currently an important focus of research on face and person recognition systems. Human performance with familiar faces stands as an example of a face recognition system that can operate robustly under naturalistic viewing conditions and with people in motion. It is therefore potentially important to understand the human performance factors that operate when humans recognize *familiar* faces and people. These factors somehow make it possible for humans to overcome the photometric transformations that typify most "real world" recognition decisions.

The purpose of the present study was to assess the effects of face familiarity and facial motion on recognition performance in naturalistic video. We were interested in knowing how these two factors mediate subject performance using viewing conditions similar to those we encounter when recognizing people in everyday life. Therefore, we used surveillance-like, whole-body images of the targets. In doing so, we hoped to learn more about the relative contribution of face familiarity and facial motion to recognition decisions made in natural environments.

1.1. Previous Literature

Previous research supports two general findings that form the logical basis for our study. First, facial motion benefits face recognition more for familiar than unfamiliar faces. These results are most easily assessed in poor viewing conditions. Second, information about the face alone is useful in recognition decisions made from whole-body videos.

There is compelling evidence that facial motion is an effective cue for familiar face recognition under a variety of non-optimal viewing conditions. Specifically, when subjects are asked to view either video clips or static images of well-known faces that have been degraded in some way (e.g., negated, inverted, or blurred), the moving images are recognized more accurately than the static images (e.g., [11, 12, 13, 3]).

It is less clear that facial motion facilitates recognition of unfamiliar faces. Specifically, although some studies report a recognition advantage for moving faces ([21, 11]), other studies find no effect ([22, 2]), and one study even reports a recognition advantage for static faces over moving faces ([6]). Thus, the literature indicates that a certain amount of experience with a face may be necessary to use facial motion effectively as a recognition cue (for reviews see [15, 19]).

It is worth noting that the above studies defined "familiar faces" as "famous faces." Bruce and colleagues explored face familiarity and facial motion more systematically by manipulating the amount of exposure subjects were given to previously unknown faces ([3]). Bruce et al. briefly familiarized subjects with videotaped faces presented either once or twice during learning and found no difference in subjects' subsequent ability to match target faces from static photographs. They concluded that relatively small amounts of prior familiarization with a face are insufficient to support a familiarity advantage for face matching. Thus, although there is evidence that motion benefits recognition when faces are familiar, it is less certain how much and what kind of familiarity is needed to support this advantage.

A recent study by Burton and colleagues ([4]) addressed this question by investigating the extent to which face versus body information can support recognition performance from whole-body video clips. In this study, familiarity was defined as "personal familiarity," with targets who were professors (un)known to subjects prior to the experiment. Burton et al. carried out two experiments. In the first one, participants viewed surveillance video clips of the professors entering a building. These black and white videos were typical of many low-cost security systems with poor illumination and low resolution. At test, participants were shown high-quality color photographs and were asked to pick out the faces they remembered seeing in the video clips. Burton et al. found that participants who were familiar with the professors performed almost perfectly, while participants unfamiliar with the professors performed at chance level. In the second experiment, the familiar participants were tested with videos that had been edited to obscure either the face or the body of the targets with a black square. Performance with "face-obscured" videos was severely impaired by comparison to performance with the "gait-obscured" videos. The intriguing aspect of these results is that even in poor-quality video, in which both the bodies and faces of the targets were difficult to see, the face proved important in the recognition decision.

1.2. Rationale

Based on the findings of both Bruce et al. ([3]) and Burton et al. ([4]), we investigated the effects of different levels of familiarity with a face (i.e., 1-, 2-, or 4-views during the learning trials) on recognition from naturalistic, surveillance-like video. Face familiarity and facial motion are both factors that seem to help subjects overcome photometric inconsistencies between learning and test stimuli. These are potentially relevant for adapting computational models to robust recognition tasks.

Specifically, we addressed two questions about the role of familiarity and motion in person recognition. First, can familiarity with the face alone transfer to an increased ability to recognize the person from whole body videos? This could be informative for computational models of person recognition from video in that it might suggest that valuable identity information from the face can be mapped onto the whole-body gait video. Second, do the beneficial effects of motion on recognition increase as familiarity with the face increases? This could have interesting implications for computational models in deciding when recognition is likely to be superior from moving video versus static frames extracted from the video.

2. Experiment

Subjects and Design. Eighty-four undergraduate students participated in the experiment as part of a class requirement. We manipulated two independent variables in our design: 1.) presentation format (moving vs. static) and 2.) level of familiarity (one, two or four exposures to each face). The former varied "between subjects," so that any given subject learned either moving or static faces. The latter varied "within subjects" so that all subjects saw some faces at each level of familiarity. All subjects were tested with the same whole-body gait videos. Recognition accuracy was assessed using the signal detection measure of d', defined as the Z-score for the hit rate minus the Z-score for the false alarm rate.

Stimuli. The stimuli for the static learning trials consisted of digitized color photographs with the subject looking directly toward the camera (see Figure 1, top image).

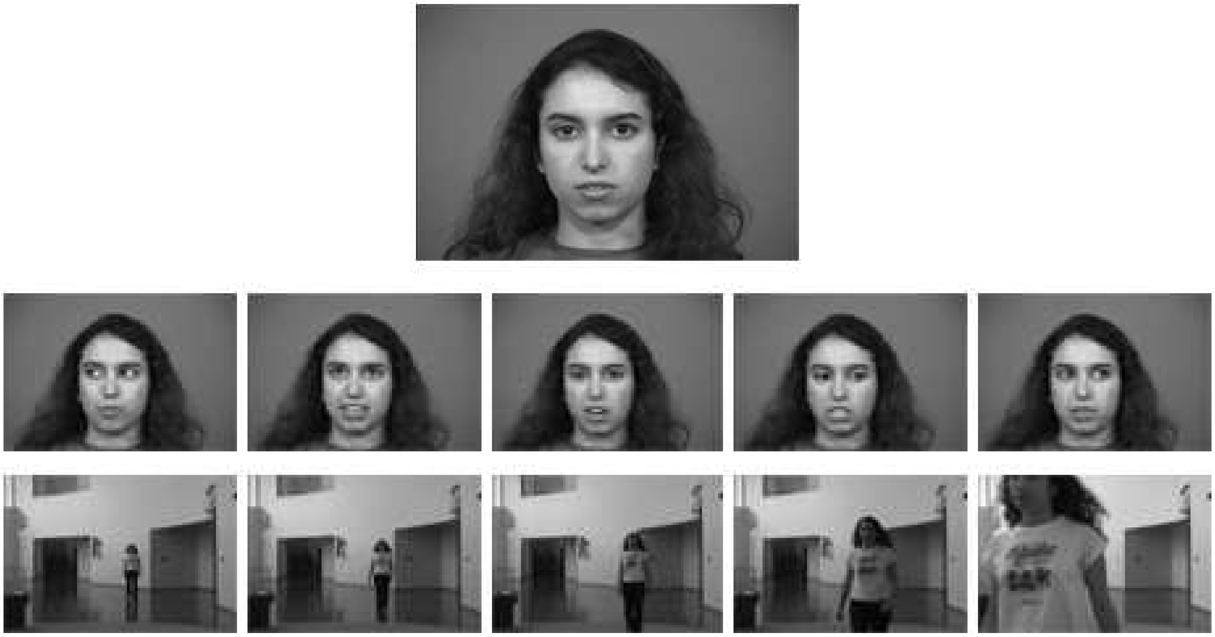

Figure 1: Sample face used in the static learning trials (top). Images extracted from a sample video stream used in the moving learning trials (middle). Images extracted from a sample gait video used in the test trials (bottom).

The stimuli for the moving learning trials consisted of 5-second digital video clips with the subject talking directly toward the camera (with no audio) (see Figure 1, second row). All of these learning trial stimuli were high resolution, close range images or videos captured under controlled illumination.

The "gait videos" used for the test trials were filmed under uncontrolled illumination conditions, similar to outdoor lighting. Specifically, these videos were taken in a building foyer with high ceilings, enclosed entirely on one side with glass windows. This environment approximates outdoor lighting conditions, while protecting the subjects and the cameras from the elements. It also makes for variable lighting conditions across the videos because the position and intensity of the light source (mostly the sun) varies on a stimulus by stimulus basis. The 9-second videos depict a subject walking parallel to the line of sight of the camera starting at a distance of 10 meters away. The subject is shown walking toward the camera and veering off to the left in the final few paces. The face of the subject is clearly visible only in the final 2-3 seconds of the video clip. In general, due to the lighting variability and relatively short temporal exposure to the face, recognition from these videos is challenging (see Figure 1, third row).

Procedure. The participants were given instructions that explained that they would first view a series of either moving or static faces and would be asked subsequently to remember the people seen in this learning session. Each participant was assigned randomly to either the "static" or "moving" learning condition. For the "static" condition $(N = 42)$, participants viewed 27 still frontal images of faces presented for 5-seconds. For the "moving" condition $(N = 42)$, participants viewed 27 5-second video clips of speaking faces.

In addition, we varied the familiarity of the faces so that one-third of the pictures/videos were viewed only once, one-third were viewed twice, and one-third were viewed four times. Faces in these different familiarity levels were interspersed randomly during the learning phase. In summary, each participant saw a total of 63 images (of 27 faces), consisting of 9 different faces in each of the 1-view, 2-view, and 4-view familiarity conditions (i.e., 9 faces viewed once; 9 faces viewed two times; 9 faces viewed four times).

At test, all subjects viewed 54 gait videos (27 targets and 27 distracters). After each video was presented, participants responded at the computer keyboard using appropriately labeled keys. They responded "old," to indicate a person whose faces was seen previously or "new," to indicate a person whose faces was not seen previously.

Finally, counterbalancing was implemented to assure that across all participants each face appeared equally often in the 1-view, 2-view, and 4-view conditions. The entire experiment took about 30 minutes.

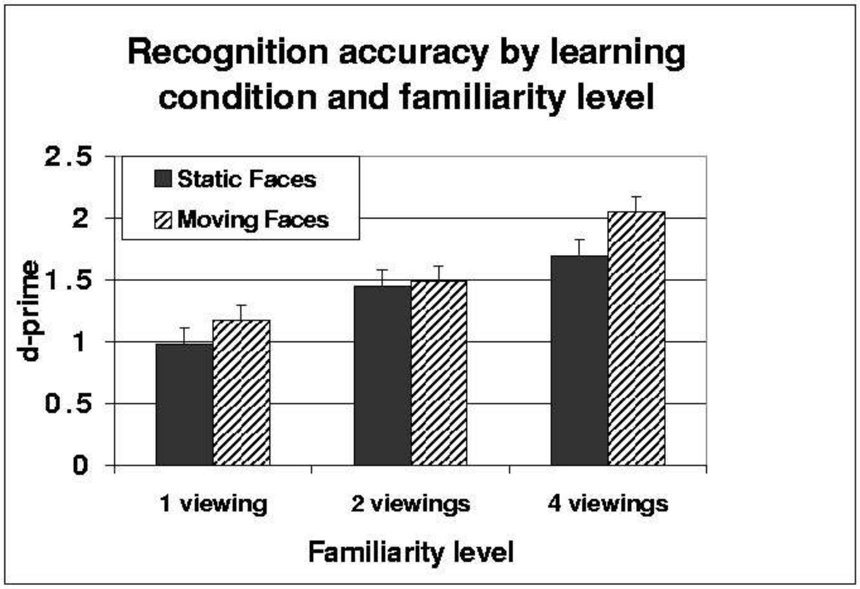

Figure 2: Recognition performance for moving and static learning conditions as a function of familiarity.

2.1. Results

For each subject in each condition, recognition accuracy was calculated as d'. The d's were analyzed with a two-factor ANOVA with presentation format (moving vs. static) and familiarity (1, 2, or 4 views at learning) as independent factors.

The average d's across the different conditions appear in Figure 2. As can be seen in this figure, recognition performance increased with familiarity ($F(2,144) = 50.18$, $MSE = 10.05$, $p < .0001$). This answers our first question in the affirmative: Familiarity with the face alone can translate into increased accuracy when recognizing people in full body gait videos.

Also, we found a statistical advantage for the motion presentation condition over the static presentation condition, [$F(1,72) = 3.7$, $MSE = 2.69$, $p < .05$]. This statistical result should be interpreted with caution, however, given the interaction trend between familiarity and presentation [$F(2,144) = 2.35$, $MSE = .47$, $p < .09$]. Specifically, as illustrated in Figure 2, the motion advantage is more salient in the 4-view familiarity condition than in either the 1- or 2-view conditions. This interaction pattern suggests that motion becomes more beneficial for recognition as familiarity with a faces increases. We analyzed this interaction more precisely with a contrast analysis to test our initial hypothesis that motion should be beneficial only with the most familiar faces. This contrast confirmed that recognition performance differed only for the 4-view condition, [$F(1,246) = 73.79$, $MSE = .43$, $p < .001$].

3. Discussion

The results from our experiment provide clear answers to the questions we posed originally. Our first question was whether familiarity with the face alone would translate into an increased ability to recognize the person from whole-body gait videos. The strong effect of familiarity we obtained in this experiment confirms that experience, defined as multiple exposures to the face alone, can improve person recognition at a distance.

Why does face information predominate in this type of recognition task? We can offer a couple of speculations. First, the performance advantage that familiarity with the face provides speaks more generally to the *quality* of the identity information in human faces. Faces seem to contain a robust type of identity information that "maps" well, and for human subjects, easily, onto whole moving bodies that are unfamiliar. Indeed, we were a bit surprised that subjects in the 4-view condition performed as well as they did, given the photometric inconsistencies between high quality face images/videos, and low resolulation whole body gait videos. And yet, the ability our subjects showed in making the leap from one image format (face-only) to another (whole-body), is consistent with the robust person recognition we exhibit for people we know well. Faces may therefore provide good working identity templates of individuals that may be mapped onto a variety of more complex scenes. The more experience we have with these templates, the more flexible and helpful they become in a variety of viewing situations.

Second, the familiarity manipulation we implemented is suggestive of the relatively limited requirements needed to evoke a familiarity advantage in human subjects. The results of this study suggest that even a relatively small amount of increased exposure time to newly-learned faces is enough to elicit a familiarity benefit in a recognition task that requires bridging across photometric changes. The kind of familiarity manipulation we implemented involved simply repeating presentations of the same face in the same format (e.g., frontal image or facial speech video). The fact that we found a familiarity advantage despite the lack of *variety* of experience subjects had with the faces, suggests that the parameters of face experience needed to improve recognition performance may be linked to increasing the memory strength of face representations that we encounter frequently. Future experiments that address how different kinds of familiarity with a face might alter performance are obviously of interest. Does viewing a variety of poses provide a different kind of benefit than simply viewing the same pose repeatedly?

The second question we addressed concerned the beneficial effects of motion with increased familiarity. In the present experiment, learning the moving face was more effective for person recognition than learning the static face, *only* in the high familiarity condition. This suggests that motion provides recognition benefits when we have seen a face multiple times, but not when the face is relatively new to us.

Why might the benefits of motion be limited to faces/people with whom we are familiar? One possibility is that it is important to become familiar with the structure of a face, before motion information can be dealt with effectively.

A second possibility is that with multiple exposures to faces we can form a representation that generalizes more efficiently across changes in photometric inconsistencies. If this is the case, then we would expect this kind of familiarity to aid recognition under nearly any kind of photometric parameter change. We have recently tested this hypothesis by asking subjects to recognize familiar and unfamiliar faces that have changed in pose or viewpoint from the originals. In this case, we found no benefit for motion as familiarity increased [20].

A third possibility is that motion processing in the visual system is set up in a way that allows for more efficient access among representations formed from moving stimuli than from static stimuli. Indeed, there is good evidence that the visual system processing can be divided into parallel static and motion streams [14]. There is also evidence to suggest that the changeable (i.e. facial motions) and invariant components (i.e., static facial features) of faces are processed separately in the human brain [9]. This separation of processing has implications for recognizing faces and people in motion [9, 15]. It has been argued that the function of processing face and body motion differs from the function of processing static facial and body features. Processing motions may serve a social function (e.g., processing expressions, and body language), whereas processing static features may support identity judgments via their invariant nature. If there are separate systems, as hypothesized, there may be more efficient cross-access within motion-based representations. Future work is needed to consider this hypothesis in more detail.

4. Summary and Conclusions

How good must a particular face recognition algorithm be in order to compete with a human observer? The answer to this question depends on whether or not we refer to human observers who are familiar or unfamiliar with the people they are trying to detect and recognize. Current algorithms compete well with unfamiliar human face recognition performance but badly with familiar human face recognition performance. We know of no current algorithms that could perform the task we gave to the human subjects in this experiment.

We examined the role of familiarity and motion in this experiment and found that familiarity mediates successful recognition in difficult viewing situations. Progressively more familiarization with a faces can promote recognition accuracy even when the presentation modes between learning and test are quite different. Second, we found that the effectiveness of facial motion as a recognition cue is tightly bound to the amount of prior experience that a person has with a face. Facial motion thus seems to facilitate recognition across photometric inconsistencies when subjects are sufficiently familiar with the people to be recognized.

Acknowledgments

This work was supported in part by grants from ONR and DOD/DARPA awarded to A.O'T. and H.A.

References

[1] V. Bruce & T. Valentine, "When a nod's as good as a wink: The role of dynamic information in facial recognition," *Practical Aspects of Memory: Current. Research and Ideas*, 1, pp. 169-174, 1988.

[2] V. Bruce, Z. Henderson, K. Greenwood, P. J. B. Hancock, A. M. Burton & P. Miller, "Verification of face identities from images captured on video," *Journal of Experimental Psychology: Applied*, 5, pp. 339-360, 1999.

[3] V. Bruce, Z. Henderson, C. Newman & A. M. Burton,"Matching identities of familiar and unfamiliar faces caught on CCTV images," *Journal of Experimental Psychology: Applied*, 7, pp. 207-218, 2001.

[4] A. M. Burton, S. Wilson, M. Cowan & V. Bruce, "Face recognition in poor-quality video," *Psychological Science*, 10, pp. 243-248, 1999.

[5] A. M. Burton, V. Bruce, & P. J. B. Hancock, "From pixels to people: A model of familiar face recognition," *Cognitive Science*, 23, pp. 1-31, 1999.

[6] F. Christie & V. Bruce, "The role of dynamic information in the recognition of unfamiliar faces," *Memory and Cognition*, 26, pp. 780-790, 1998.

[7] R. T. Collins, A. J. Lipton, and T. Kanade, "Introduction to the special section on video surveillance," *IEEE Trans. Pattern Anal. and Machine Intell.*, vol. 22, no. 8, pp. 745 -746, 2000.

[8] P. J. B. Hancock, V. Bruce, & A. M. Burton,"Recognition of unfamiliar faces," *Trends in Cognitive Sciences*, 4, pp. 330-337, 2000.

[9] J. V. Haxby, E.A. Hoffman, & M.I. Gobbini. "The distributed human neural system for face perception," *Trends in Cognitive Sciences*, 4, pp. 223-233, 2000.

[10] Z. Henderson, V. Bruce, & A. M. Burton, "Matching the faces of robbers captured on video," *Applied Cognitive Psychology*, 15, pp. 445-464, 2001.

[11] B. Knight & A. Johnston, "The role of movement in face recognition," *Visual Cognition*, 4, pp. 265-273, 1997.

[12] K. Lander, F. Christie, & V. Bruce, "The role of movement in the recognition of famous faces," *Memory & Cognition*, 27, pp. 974-985, 1999.

[13] K. Lander, V. Bruce, & H. Hill, "Evaluating the effectiveness of pixelation and blurring on masking the identity of familiar faces," *Applied Cognitive Psychology*, 15, 101-116, 2001.

[14] W. H. Merigan, "P and M pathway specialization in the macaque," *From Pigments to Perception*, Eds., A. Valberg and B. Lee, pp. 117-125, 1991.

[15] A. O'Toole, D. Roark, & H. Abdi, "Recognizing moving faces: A psychological and neural synthesis," *Trends in Cognitive Sciences*, 6, pp. 261–266, 2002.

[16] P. J. Phillips P. J., Moon H., Rizvi S., Rauss, P. "The feret evaluation methodology for face recognition algorithms," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1090–1103, 2000.

[17] P. J. Phillips, P. Grother, R. Micheals, D. M. Blackburn, E. Tabassi, and J. M. Bone, "Face Recognition Vendor Test 2002: Evaluation Report," *NISTIR 6965*, www.frvt.org, 2003.

[18] G. Pike, R. Kemp, N. Towell, & K. Phillips, "Recognizing moving faces: The relative contribution of motion and perspective view information," *Visual Cognition*, 4, pp. 409–437, 1997.

[19] D. Roark, S. Barrett, M. Spence, A. O'Toole & H. Abdi, "Psychological and neural perspectives on the role of motion in face recognition," *Behavioral and Cognitive Neuroscience Reviews*, 2, pp. 15-46, 2003.

[20] D. Roark, A. O'Toole, & H. Abdi, "Person recognition from naturalistic video: Effects of familiarity and facial motion," *Journal of Vision*, 3, 2003.

[21] I. M. Thornton & Z. Kourtzi, "A matching advantage for dynamic human faces," *Perception*, 31, pp. 113-132, 2001.

[22] S. Snow, G. Lannen, A. O'Toole, & H. Abdi, "Memory for moving faces: Effects of rigid and non-rigid motion," *Journal of Vision*, 2, 600a, 2002.

[23] L.A. Wang, W.M. Hu, and T.N. Tan, "Recent developments in human motion analysis," *Pattern Recognition*, vol. 36, pp. 585-601, 2003.

[24] W. Zhao, R. Chellappa, A. Rosenfeld, and J. Phillips, "Race recognition: A literature survey," Technical Report CS-TR-4167-R, University of Maryland, 2002.