Article

Consistency and Variability Among Latent Print Examiners as Revealed by Eye Tracking Methodologies

Tom Busey [1]
Chen Yu [1]
Dean Wyatte [2]
John Vanderkolk [3]
Francisco Parada [1]
Ruj Akavipat [1]

Abstract: We recorded the eye positions of 18 expert latent print examiners and 18 novice participants across two separate experiments that were designed to represent abbreviated latent print examinations. In the first experiment, participants completed self-paced latent and inked comparisons presented on a computer monitor while their eyes were tracked with a commercial eye tracker. The similarity of eye fixation patterns was computed for each group of subjects. We found greater variability under some conditions among the experts than the novices in terms of the locations visited. However, experts spent approximately 50% longer than novices inspecting the images, which may have led to differences in strategies adopted by the two groups. A second experiment used trials that always lasted 20 seconds and found that under these time-controlled circumstances, experts were more consistent as a group than novices. Experts also had higher accuracy, spent a greater proportion of time inspecting the latent prints, and had shorter saccades than novices. However, the two groups spent an equal time looking at regions that contained minutiae. The results are generally consistent with experts relying on a common set of features that they choose to move their gaze to under time-limited conditions.

[1] Department of Psychological and Brain Sciences, Indiana University, Bloomington
[2] University of Colorado, Boulder, CO
[3] Indiana State Police Laboratory, Fort Wayne, IN

Received June 28, 2010; accepted November 10, 2010

Journal of Forensic Identification
60 / 61 (1), 2011

Introduction

Latent print examinations involve a process by which a latent print, often recovered from a crime scene, is compared against a known standard or sets of standard prints. Despite advances in computer matching technology, latent prints are still compared by human experts. In the United States and in many other countries, there is no fixed number of matching points or details that is mandated by the courts or forensic science community. This implicitly gives the examiners some latitude in terms of the details they choose to use in order to determine whether the two prints come from the same source. For example, instead of just relying on matching minutiae, the examiner is free to use what details he or she feels are relevant, including what is known as first-level detail of general direction of ridges, second-level specific ridge paths, and third-level detail or the texture and shape of individual ridge elements.

Although this practice takes full advantage of the vast capabilities of the human perceptual system, it does leave open the question of what details experts actually rely on when conducting examinations. In addition, experts may choose to rely on different types of detail or information depending on the circumstances and their training, which may raise issues with respect to the nature of the evidence presented in court. A recent National Academy of Sciences report [1] was somewhat critical of the language used by examiners when testifying about their results and called for more training and research on the nature of the latent print examinations. The report revealed weaknesses in our current knowledge about what information experts rely on when performing identifications and exclusions. Part of the difficulty resides in the fact that much of the processes of perception are unconscious and can be difficult to translate into language [2, 3] and examinations may be subject to extra-examination biases [4–6].

Forensic comparative examinations are visually demanding and typically require both magnification as well as eye movements to bring the high acuity region known as the fovea within the eye onto regions deemed relevant or diagnostic in the image. The eye tends to move three to four times a second, with intervals known as fixations separated by rapid movements called saccades. The eye can only acquire visual information during fixations, because the processing of visual information tends to be suppressed during saccades. The distribution of fixations over an image gives some idea of what information the participant deemed relevant when conducting an examination.

The goal of the present work is to describe the degree to which experts rely on similar or different sources of information when performing examinations and whether they are more or less consistent as a group when selecting particular details as novices are. We collected eye tracking data from latent print examiners and novices and analyzed their moment-by-moment eye movements during a fingerprint examination as one means to address these questions. Although this study was the first to use eye tracking to address the expertise of fingerprint examiners, eye tracking techniques have been successfully applied in several other scientific fields to assess implicit knowledge from human experts. The field of mammography research has adopted similar eye tracking methodologies. Krupinski and colleagues [8–10] have used eye tracking to investigate not only what features radiologists rely on when inspecting mammograms, but also to suggest cognitive mechanisms such as holistic processing when experts are viewing mammograms [11]. Similar work with chest x-rays demonstrated that dwell times were longer on missed tumors than at other locations [12], suggesting that errors in radiology are due to identification problems rather than detection problems [10]. The field of questioned documents has also benefited from an eye tracking approach [13], which has helped to delimit the visual features that experts rely on when comparing signatures.

Several authors have used eye tracking methodologies to address visual expertise in domains other than medical imaging. Chess experts have been shown to fixate relevant pieces a greater proportion of time and make saccades to more important pieces for a given position, suggesting a configural approach [14]. Busey and Vanderkolk [15] provided both behavioral and electrophysiological evidence for configural processing in latent print examinations that is consistent with chess experts and the holistic processing seen in radiologists [9]. In a different forensic discipline, Bond [16] has use eye tracking to address the nature of expertise in deception detection and has argued that much of the task relies on nonverbal observations. The two experts that were studied differed in their strategies; one looked primarily at facial features, whereas the other looked at the lower limbs and torso. Despite this variability among the two experts, both were shown to be highly accurate in a screening study. Chi [17] summarized the strengths exhibited by visual experts, including superior feature detection and recognition, better cognitive monitoring, and strategy selection. However, weaknesses also exist, including over-confidence, an inability to generalize outside their domain of expertise, and vulnerability to cognitive biases.

Not all of visual expertise may be bound up in superior search strategies. Abernethy and Russell [18] argued that expertise in badminton is characterized primarily by better use of available information rather than the choice of eye gaze location. This may be a function of this particular domain, where the choice of eye gaze is dictated primarily by one object. However, it is likely that elements of visual expertise may reside in the interpretation of particular visual information rather than the choice of what to acquire. For example, several authors have argued that visual expertise creates novel feature detectors by changing the nature of the perceptual representations that experts employ [19, 20].

Mello-Thomas et al. [21] examined agreement among radiologists' eye tracking behavior. Digital mammograms were divided up into foreground and background regions. Although they were unable to examine the agreement among scanpaths, they could examine the ratio of foreground to background sampling and found high agreement among radiologists, which supports their proposal that expert radiology examination involves integrating information of features with background information into a "bigger picture" for diagnosis. In our experiments, the contextual information provided by the overall ridge flow is also the same detail that defines local sources of information, making it difficult to segment fingerprints into foreground and background regions. Thus we propose a novel approach of comparing spatial distributions of fixations using the Earth Mover metric, which provides more detail about the similarity of the regions visited, not just the identity of the regions as classified as foreground or background.

The strength of the eye tracking approach is that it is completely noninvasive with the exception that participants know that they are being tested. The limitation of the eye tracking approach is that we only know where the eyes move, not necessarily what information they are actually gathering [22]. With the exception of blinks, the eyes always point somewhere even if the participant is not actively acquiring visual information. To address this, we used relatively brief, time-limited displays to encourage the participants to make every eye fixation count and move their eyes as quickly as possible to the most diagnostic regions of the fingerprints.

In this paper, we describe the results of two experiments that are designed to compare experts and novices in order to determine which group is more consistent in acquiring relevant data with respect to the locations of eye fixations. We focus primarily

on the consistency question, in part because it bears on the issue of whether there is a commonly accepted sufficiency standard used in fingerprint identification. The fixation locations are distributed over the images and different participants may have different numbers of fixations depending on how quickly they move their eyes. We will compare any two participants with each other by using a comparison procedure known as the Earth Mover metric, which we will describe in a subsequent section. This metric essentially provides a measure of similarity between two sets of eye fixations coming from two participants. Thus, our measure focuses on the similarity of their overall looking patterns when conducting an examination but not particular fixed regions in a particular moment.

Gathering data from latent print examiners is somewhat challenging because they work in laboratories throughout the country. We will present two sets of data, one of which was collected using a commercial eye tracker and one of which was collected using a custom built eye tracking application that was developed for off-site data collection. The general experimental paradigm was the same in those two experiments. Human observers (experts or novices) were asked to sit in front of a computer screen in which pairs of fingerprints were displayed one by one. Participants were asked to visually examine those fingerprints and decide whether the two fingerprints displayed simultaneously matched each other or not. There was no particular instruction about where they should look during the matching task so that they could freely move their eyes on a fingerprint image. Although there were modest differences in the eye tracking devices and stimuli across the two experiments, the primary difference between the two experiments was the amount of time that participants were allowed to spend before moving to the next fingerprint. Across the two experiments, the combined data reliably illustrates the conditions under which experts show more consistency or variability than novices.

Experiment 1

Experiment 1 tested six experts and six novices using a commercial eye tracking system with traditional latent and inked print comparisons. Figure 1 illustrates an example stimulus pair, along with the eye fixations and eye trace for one observer. The typical latent print examination can take hours or even days to complete for difficult prints. Examiners will usually start with an inspection of the latent print, which may be augmented

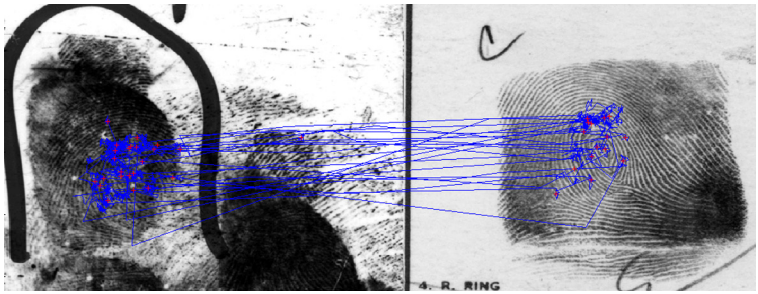


Figure 1

Example stimuli with fixation and eye trace from the Tobii eye tracker used in Experiment 1. The red crosses represent eye fixations.

by photographs, notes, or drawings. They then move on to the inked print. In order to obtain a complete data set from each participant, we limited the amount of time that each participant could spend on each fingerprint. In our first experiment, participants could spend no more than a minute on each image pair, although if they finished before then, we asked them to move to the next image in order to avoid corrupting our database with uninformative eye movements as they waited for the next print. There is a tradition of using relatively short presentations in eye tracking studies; for example, Charness et al. [14] analyzed only 1 to 2 seconds of gaze data in their study of chess experts in order to focus on perceptual expertise rather than the decision-making process. These relatively brief viewing times are likely to tap into whatever visual search strategies experts have developed, but may not reflect a true latent print examination. There is an important difference between chess and the present study, in that chess has positions that become familiar to experts over time, while there are an unlimited number of potential configurations of ridge flow in fingerprints. Nodine et al. [23] studied radiologists interpreting mammograms, which have similar characteristics to fingerprints in that the number of visual configurations is essentially infinite. This may still allow experts to tap into their body of expertise during the initial viewing of an image although they will never find an exact match to items stored in memory and therefore must compute some measure of similarity. Rapid decisions based on brief, 200 msec presentations were also explored by Kundel and Nodine [24]. It is important to note that lesion detection is a categorization task in which subjects attempt to classify regions as belonging to a

particular class. This task favors the acquisition of information that is common to the class of lesions. Latent print individualization, on the other hand, works best when the rarest information in the print is identified, because this will be most diagnostic when excluding other nonmatching prints.

We are most interested in where they direct their eyes during these abbreviated examinations rather than whether they achieve a certain level of accuracy. Note that the ability to terminate a trial before the full 60 seconds can affect the strategy adopted by experts and novices, especially if novices adopt an earlier stopping criterion. This may create interpretational difficulties with some of the statistics as discussed later. In Experiment 2, we fixed the viewing times for all subjects at 20 seconds in part to alleviate these concerns.

Methods

Stimuli

The stimuli for Experiment 1 came from the National Institutes of Standards and Technology Special Database 27, which had previously determined identifications of latent and inked prints that are typical of what is traditionally found during casework. We created three lists of images, each of which has 30 pairs of images. List 1 had five nonmatches (exclusions), List 2 had three nonmatches, and List 3 had eight nonmatches. The matching versus nonmatching dichotomy was used to determine false alarm rates and compute estimates of response criterion shifts.

Participants

Our experts were recruited from forensic science laboratories in Indiana, Illinois, and Nevada that were associated with state or large metropolitan agencies. They had an average of 15.3 years working as latent print examiners and were an average of 45 years old. There were four men and two women. Four of the six had trained other examiners. Two wore glasses. The novices were recruited from the Bloomington (Indiana) community, had no prior experience with latent prints, and tended to be younger, with a mean age of 23 years. There were three men and three women. None wore glasses or contacts.

All participants were tested according to the procedures of the Human Subjects Protection committee of Indiana University.

Procedures

Participants were seated approximately 36 inches away from a 17″ LCD monitor set at a resolution of 1024 x 768 pixels. The images were scaled so that the latent and inked prints together filled the horizontal dimension, with 138 pixel horizontal borders on the top and bottom of the prints. These viewing conditions imply that one pixel subtends 0.0200 degrees of visual angle, and to convert pixels into units of degrees of visual angle, simply multiple the number of pixels by 0.0200 for Experiment 1. The monitor was part of a model 1750 Tobii eye tracking system (Tobii Technology, Falls Church, VA), which uses infrared cameras positioned on the monitor to track the position of the eye gaze by monitoring both eyes. The Tobii eye tracker asks the observer to move his or her eyes to a known location on the monitor as indicated by a bull's eye. The position of both eyes is measured and this procedure is repeated for a total of nine locations. This establishes the relation between the observer's eye position and positions on the screen and establishes a relation between the eye position and a position on the monitor.

After the calibration procedure, the participant was shown pairs of prints and asked to determine whether they came from the same source. They were given up to one minute to make this determination, and if they came to a conclusion sooner, they stated this conclusion and proceeded to the next image. An experimenter manually recorded their response as either identification or exclusion. Observers were allowed to use as little or as much of the allotted minute to respond as they wished. Because of a disk crash and a corrupted backup disk, the behavioral responses for three novices were lost, although we will show that the extant data is still sufficient to demonstrate large accuracy differences between experts and novices. The eye tracking data was not lost for these subjects.

Results

We first report the number of basic statistics, such as the average duration of eye fixations, the number of fixations, and the length of the saccades.

Fixation Statistics

The raw gaze data was split into fixations and saccades using the Tobii fixation finding algorithm that is based on a fixation radius and a minimum duration. If the eye has a set of consecutive raw gaze locations within a circle of fixed size (30 pixels or

.6 degrees of visual angle for our setup) for a minimum duration (we specified 100 msec), the software labels that portion of a raw eye gaze data as one fixation. These fixations are then assumed to be separated by saccades.

We calculated the average duration of each fixation for each subject, as well as the proportion of time that each group spent on the latent print. However, neither of these statistics yielded significant group differences (all $p>0.05$).

Some approaches to latent print examinations are characterized by integrating visual information from nearby areas. This may result in many fixations near each other, separated by relatively short saccades. We might expect, therefore, that the saccade length for experts would be smaller than those for novices. We computed the average length of saccades within the latent prints and within the inked prints for both groups. Consistent with this expectation, we found that experts had much smaller saccades than novices on both types of prints (latent prints: 38.8 vs 54.8 pixels; $F(1,34)=11.0$; $p<0.01$, cohen's $d=1.14$; inked prints: 36.4 vs 54.1 pixels; $F(1,34) = 11.9$; $p<0.01$, cohen's $d=1.18$). Experts, therefore, seem to make smaller jumps between locations on the inked and latent prints when looking within each type of print.

Overall Viewing Times

Recall that participants had up to one minute to view each latent and inked print pair, but could terminate the trial earlier if they felt they had enough information to make a decision. We allowed this early termination because eye fixation data is less meaningful if participants no longer have a task they are working on. However, the two groups may have viewed the prints for different amounts of time, which may have consequences for our statistics we report below. To assess whether there are trial duration differences, we calculated the duration of each trial across all three lists and found that the experts tended to view the prints for approximately 50% longer than novices on average. Experts spent 34.2 seconds on average and novices spent 21.3 seconds on average viewing the prints. This difference was statistically significant ($t(17)= 3.12$, $p= 0.006$, cohen's $d = 0.732$). The design of Experiment 2 addresses this difference, and below we discuss the consequences of this difference with respect to different analyses.

Behavioral Accuracy

Would these longer viewing times and shorter saccades translate into higher accuracy for experts? Despite losing the behavioral data for three of the novices, we found a large and significant difference between the experts and novices in terms of behavioral accuracy. We computed a measure of discriminability called d', which is based on signal detection theory [25]. This analysis is complicated by the fact that there were relatively few nonmatching print pairs (five, three, and eight in Lists 1, 2, and 3, respectively) and only one expert erroneously called a nonmatching print a match. This means that the false alarm rate for most experts and one of the novices was zero. To correct for this, we set the false alarm rate to be 0.03125, which is 1/(16*2) or half the size of the smallest possible false alarm rate that is nonzero. This is a conservative measure of d' and tends to reduce the difference between experts and novices because it was applied to more experts than novices. Despite this, and despite the loss of data for three novices, we found a large and statistically significant difference in terms of accuracy for the two groups. Experts had a mean d' of 2.47, whereas novices had a mean of 1.3 ($t(7) = 2.50$; $p<0.05$, cohen's $d = 2.36$). Thus, our experts outperformed our novices on these abbreviated latent print examinations, by an almost 2 to 1 ratio. We also computed beta, a measure of the response shift in the observers. The mean beta for experts was 4.17 and the mean beta for novices was 2.49. This difference was not statistically significant ($t(7) = 1.41$; $p=.2$, cohen's $d = 1.13$).

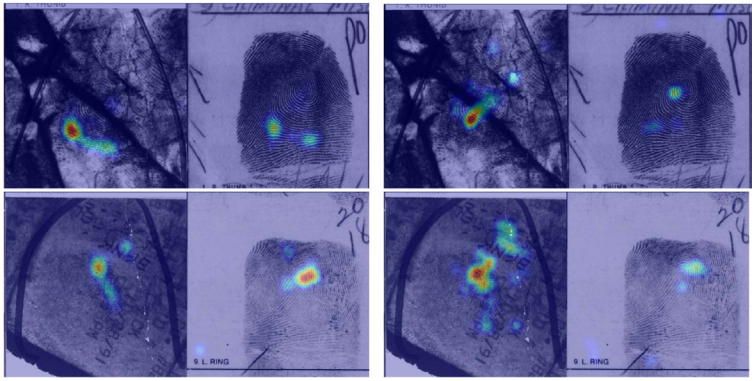
Earth Mover Analysis

In our next analysis we looked to see whether the regions visited by experts were similar across our six experts, using an analysis called the Earth Mover metric [26]. The Earth Mover metric is a comparison technique that computes the similarity between two sets of points and has successfully been used in image retrieval. The basic idea behind the Earth Mover analysis is that each fixation has a particular location that we can visualize as a small dot on a map of the fingerprint. The difficulty in matching arbitrarily numbered sets of points is that one must compute some form of correspondence between the two sets of points. Computationally, it is too expensive to try all possible combinations of points looking for the minimum summed deviations. In addition, the two sets of eye fixation from two observers may be of unequal length. The Earth Mover metric solves both of these problems by using dynamic programming [27] to find a solution that, although not optimal in a global sense, is very likely to be close to optimal.

Intuitively, this distance metric treats the fixations as small piles of dirt that must be moved from one location to another, and the Earth Mover solution attempts to move the dirt with as little effort as possible. In this case, the fixations are treated as small piles of dirt, normalized by the total number of fixations, and the algorithm moves the dirt from the fixations from one subject to the fixations from another subject for the same image. The fixations are normalized by the total number of fixations, because no two subjects will have an equal number of fixations. The total amount of work required to move the dots from one subject's locations to another is a measure of their overall similarity.

Figure 2 illustrates an intermediate step in the Earth Mover algorithm. We apply a radial Gaussian function of size 41 x 41 pixels with standard deviation of 12.5 pixels to each fixation to acknowledge the spread of information acquisition around each fixation. We then graphically illustrate this pattern of inspection using heat maps, which are shown in Figure 2, where the color indicates the dwell time at that location. If two subjects have very similar fixation patterns, which they would get if they inspected similar features in the fingerprints and therefore their eye visited similar regions on the same fingerprint image, very little work would be required to move one set of fixations from one participant on to the other set of eye fixations generated from the other participant.

Each participant was tested for three sessions, each of which contained one list of 30 pairs of images. For each of these images we calculated the pair-wise similarity of each expert to every other expert, and each novice to every other novice. We were interested in whether the experts or the novices as a group showed more consistency across the image pairs. We had no a priori expectations of what we might find. One might argue that if, in fact, examiners have an implicit set of features that they agree on, they would all look at these regions in the latent print and therefore be more consistent with each other. However, as the images in Figure 1 illustrate, some of these latent print features are quite difficult to visualize. This might have allowed the examiners to inspect different visual features. A third possibility is that novices may have an intuitive sense for what constitutes relevant or diagnostic information in fingerprints. In this case, they may quickly adjust to a style that is similar to those of the experts.



Expert Data *Novice Data*

Figure 2

Examples of heat maps. Color indicates the amount of time spent at each location. The left column contains data from a latent print expert on two pairs of images. The right column contains data from a novice.

The Earth Mover metric is a symmetrical distance function that represents the similarity of two participants, where larger numbers imply more dissimilarity between the two participants. For each trial, we computed the average of all of the distances within the group of all the experts and the average of all of the distance within the novice group. If experts were more consistent than novices, we would expect their distances to be smaller on average than those computed between novices.

Perhaps surprisingly, we found the opposite result. Of the 90 trials in the experiment, experts were more similar to other experts on only 29 trials, whereas on the remaining 61 trials, the novices were more consistent with each other. We computed the mean of the inter-expert distances, which was 48.4 pixels. The mean of the inter-novice distances was smaller at 45.04 pixels. A paired t-test computed across the trials was significant ($t(89) = 2.19$; $p < 0.031$; cohen's $d = 0.464$). We conducted a leave-one-out analysis to determine whether any one participant was responsible for these differences by recomputing the differences between experts and novices by leaving one participant out. We found results consistent with that above on all 12 subjects. Thus, it does not seem to be the case that one outlier participant is responsible for the group differences.

We also restricted the analyses to fixations just on the latent print and found similar results. Experts had smaller distances to other experts on 32 trials, and novices had smaller distances to other novices on 58 trials. The average distance between experts was 38.6 pixels, whereas the average distance between novices was 34.6 pixels, which is significantly different ($t(89)=3.18$; $p<0.01$; cohen's $d=0.67$). However, when we restricted our analyses to just the inked prints, we found no differences between the two groups. Experts had smaller distances to other experts on 48 trials, and novices had smaller distances to other novices on 42 trials. The average distance between experts was 42.5, whereas the average difference between novices was 44.7, which was not significantly different ($t(89) = -1.31$; $p = 0.193$; cohen's $d = -0.28$). Thus, it appears that the group differences are driven mainly by the distribution of latent print fixations. The variability among experts seems to come primarily from the gaze data to the latent prints.

Because we ran each subject in three separate lists, we have an opportunity to address whether the novices are becoming more expertlike as a group. We might find, for example, that novices are more variable as a group on List 3 than on List 1. We did not randomize the order of the lists, and so the results are confounded by the images that were on each list. However, we did separate the data for each list and computed the variability among the experts and the novices separately for each list. For List 1, experts had smaller distances to other experts on 11 trials, and novices had smaller distances to other novices on 19 trials. Experts had slightly larger distances to other experts than novices did to other novices, but this difference was not significant (44.9 vs 42.9 pixels, $t(29)= 0.76$, $p = .45$, cohen's $d = .14$). A similar trend was found for List 2, with experts more consistent as a group on 10 of the 30 trials and the differences trending in the same direction but not significantly different (50.1 vs 47.6 pixels, $t(29)= 0.97$, $p = 0.342$, cohen's $d = .18$). The third list had the largest group differences, with experts more consistent on only 6 trials out of 30. The average distance between experts was 53.3 pixels, and the average distance between novices was 44.7 pixels, a significant difference ($t(29)= 3.18$, $p = 0.003$, cohen's $d = .58$).

There is one difference that might explain the greater variability among experts as a group on List 3. Recall that this list had the highest number of nonmatching images (8 out of 30), and this could have made the experts more suspicious as a group when viewing these images. Because experts are more accurate and

therefore more likely to notice nonmatches, the experts may have changed their strategy on List 3 as they began to look for more differences rather than similarities.

There is a concern that the preceding statistics might have been affected by the fact that the individual trials were not independent because the same participants were tested on all the trials. The alternative analysis would be to use the variability among the subjects as a basis for statistical comparison. However, this also builds in dependencies, because each subject contributes to multiple pair-wise distance comparisons. One way to address whether these dependencies are playing a role is to compute the autocorrelation across time. We computed this autocorrelation on the differences between the experts and novices on each trial for fixations on both images, using trial number as the independent variable. The regression function in SPSS reports the Durbin-Watson statistic as a test for correlated residuals. This statistic has a range from 0 to 4, with a midpoint of two, which is consistent with uncorrelated residuals. Our obtained Durbin-Watson statistic was 2.067, which is quite close to 2.0 and demonstrates that our data did not have a strong autocorrelation. There does not appear to be strong dependencies that would invalidate the preceding t-tests.

As discussed earlier, there were large differences in the durations that the two groups chose to spend on the prints, with the experts taking almost 50% longer to view the prints. This may affect the Earth Mover results, because if the experts take more time looking at image detail, they have more opportunity to explore additional parts of the print, and this may lead them to be more variable than novices, at least on List 3. As a partial solution to this problem, we restricted our analyses to just the first 20 seconds of each trial. We still found that experts were more variable as a group than novices. The average distance from one expert to another was 71.45 pixels, whereas the average distance from one novice to another was 56.4 pixels ($t(89)= 5.6$, $p <0.001$, cohen's $d = .593$). As with the full dataset, the differences were strong on the latent print side (46.4 vs 38.2, $t(89)= 5.6$, $p <0.001$, cohen's $d = .586$) but not significant on the inked print side (48.3 vs 49.0, $t(89)= -.39$, $p=.70$, cohen's $d = -.041$).

We repeated this analysis individually for each list. For List 1, there were no differences between experts and novices (experts = 66.4 pixels, novices = 56.7 pixels; $t(29) = 1.48$; $p=.151$, cohen's $d = .270$). However, the data from List 2 did demonstrate larger variability among experts (experts = 79.5 pixels, novices = 57.0

pixels; t(29) = 5.57; p< 0.001, cohen's d = 1.02). List 3 also demonstrated larger variability among experts (experts = 70.4 pixels, novices = 55.6 pixels; t(29) = 3.48; p= 0.002, cohen's d = .635).

The preceding analyses do not directly address the possible reasons for the group differences, because the experts knew that they had the full 60 seconds if necessary, and this may have altered how they chose to spend the first 20 seconds. The only straightforward solution to this issue is to limit both groups to a shorter and fixed duration, which we explore in Experiment 2. However, this duration difference does offer a reasonable explanation for why experts were more variable in Experiment 1 on List 3, however, in that it suggests that experts were more willing to seek out additional visual information before terminating a trial. This leads to more variable fixation locations, as well as longer search times. We also explore two other possible explanations for the differences seen between experts and novices, as discussed next.

Image Reliability at Each Fixation

When discussing these results with latent print examiners, one suggestion put forth was that individual examiners may have different styles or methodologies that would dictate the use of different kinds of information and lead to more variability among experts. Two labels given to the examiners might be "ridgeologists" and "point counters", although such schools of thought may have more overlap than the names imply. Nonetheless, these differences in style may lead the experts away from the higher-quality (but relatively small) regions of the print into regions that have poorer quality but contain potentially more diagnostic information that is particular to the individual examiner's style. To assess this, we computed a rough measure of reliability at each fixation by first estimating the average ridge width across the image, filtering the image by a set of oriented filters at this frequency, and then taking the maximum of all of the filters [28]. Higher image quality is reflected as numbers closer to 1.0, and lower closer to 0.0.

Contrary to these expectations, we found that experts were seeking out slightly higher regions of image quality by this metric, although the difference did not rise to the level of statistical significance. We found the experts had a mean of .49, and the novices had a mean of .48. The 95% confidence interval of the null hypothesis [-0.014; 0.014] included the actual difference

of 0.0112, demonstrating that the difference did not rise to the level of statistical significance. This analysis was performed on just the latent print, although similar findings were also found for the inked prints.

These results demonstrate that the larger variability seen among experts does not arise primarily from their tendency to seek out poorer quality regions of the latent prints, at least to the degree to which we could identify such behavior with our reliability index.

Number of Minutiae Near Each Fixation

A related question to the image reliability hypothesis is that experts might be more or less likely to visit regions that include minutiae. If the novices all tended to gravitate toward a cluster of high-quality inked prints regions and the experts tended to focus instead on nonminutiae information such as ridge orientation and inflection, we might find differences in terms of the number of minutiae each group visits.

To answer this question, we processed our inked prints through the Universal Latent Workstation published by the Federal Bureau of Investigation. This program identifies the locations of classical minutiae such as ridge endings and bifurcations or y-branching. We then asked whether experts or novices have a greater number of minutiae near each fixation. We defined "near" as a circle 1.22° (61 pixels) in radius around each fixation and counted the number of minutiae near each fixation. Figure 3 shows one trial with the minutiae and the circles drawn over many of the fixations (not all are displayed for image clarity). We simply compute the average number of minutiae inside the circles for both groups of participants.

Although we did find slightly fewer minutiae near each fixation for the experts relative to the novices, the difference did not rise to the level of a statistical significance. The mean number of minutiae near each fixation for a circle of radius 1.22° for experts was 4.94, whereas for novices the mean was 5.14. The 95% confidence interval constructed from the null hypothesis included the difference of -.12 [-.61; 0.60]. Similar results were found for circles of other sizes. Thus, if there is a difference between the two groups, it is relatively small. Therefore, the tendency to seek out or avoid minutiae in favor of other sources of information does not appear to completely explain the greater variability seen among experts.

Discussion

The Earth Mover statistic demonstrates that the experts as a group were more variable than the novices were as a group. This is contrary to the hypothesis that experts would be more similar to each other if they all rely on the same implicit feature set or standards. We explored two possible reasons (attention to poorer-quality regions in the search for regions that are consistent with a particular style of analysis, and attention to fewer minutiae), neither of which seemed to be major contributors to the differences between the groups. However, the differences in viewing durations between the two groups provided one straightforward explanation for this difference: If experts spend more time, they have more opportunity to become more variable as a group. This alone could explain the greater variability seen with experts. We also explored two possible reasons (attention to poorer-quality regions in the search for regions that are consistent with a particular style of analysis, and attention to fewer minutiae), neither of which seemed to be major contributors to the differences between the groups.

Experiment 2

In Experiment 2, we chose to strictly limit the amount of time that the prints would be available. The mean of the novice viewing durations was about 21 seconds in Experiment 1, and thus we decided to present all image pairs for 20 seconds in Experiment 2. This would allow enough time so that novices (as well as experts, who had even longer viewing durations) would not feel that they had too much time and would therefore stay on task for the entire 20 seconds.

We made several other design changes that improved the accuracy and quantity of the eye gaze data while not substantively altering the nature of the task. A weakness of the Tobii eye tracking system is that it is not portable, and to obtain data from sufficient numbers of participants, we developed a portable eye tracking system that is based on an open-source hardware design [29]. This allowed us to gather data from 12 experts and 12 novices in Experiment 2. We also sought to improve the resolution of the images that were presented, and therefore we switched to a 21″ LCD monitor that allows a much greater 1680 x 1050 image resolution. However, because our testing occurred in the field, we were limited to one 20-minute session per participant, in which we showed a single list of 35 images.

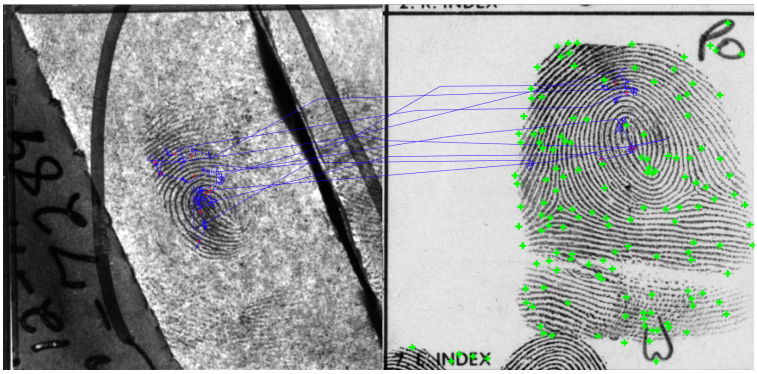


Figure 3

Number of nearby minutiae computation, with the green crosses located on each estimated minutia location. For each fixation, we count the number of minutiae within a given radius (e.g., 25 pixels) of each fixation.

Method

Stimuli

The stimuli for Experiment 2 were also taken from National Institutes of Standards and Technology Special Database 27, but were different than the images used in Experiment 1. We used 35 images, five of which were nonmatches. The nonmatching pairs were considered close nonmatches by our latent print examiner consultant in that they had the same general ridge flow but differed in the exact ridge details.

The images were presented side by side on a 21″ LCD monitor at a resolution of 1580 x 759 pixels. The monitor itself was set to its native resolution of 1680 x 1050 pixels.

Participants

We tested 12 expert and 12 novice participants, including three experts who had participated in Experiment 1 a year earlier. The experts were recruited at forensic identification conferences in Nevada, Illinois, and Indiana, and the novices were members of the Bloomington (Indiana) community. The mean age of the experts was 42.3 years with a range of 25 to 56 years. The mean number of years of experience was 16.3 with a

range of 2 to 29 years of unsupervised latent work. There were five men and seven women. All had self-reported 20/20 vision (corrected or uncorrected). Because some of our recording was done in the field, we did not have an easy way to perform a standardized Snellen chart to measure visual acuity. However, in order to perform the task, subjects had to read text that vertically subtended .19 degrees at the 60 cm viewing distance. This is very similar to the .12 degrees that a 20/20 letter on the Snellen chart subsumes. All subjects reported no difficulty perceiving this text clearly. Four subjects wore glasses and two wore contacts. None had bifocals or graduated contact lenses.

The novices had a mean age of 32.8 years with a range of 21 to 65 and there were five men and seven women. Three subjects wore glasses and two wore contacts, none with bifocals or graduated contacts.

Procedures

Participants were seated approximately 60 cm (~24 inches) away from a 21″ LCD monitor. This implies that each pixel subtended 0.02447 degrees of visual angle, and to convert pixels to degrees of visual angle, simply multiply the number of pixels by this value. Participants wore a head-mounted eye tracker that used two small cameras to monitor the eye and the view of the scene, respectively, according to the hardware proposed by Babcock and Pelz [29]. Both cameras were mounted and specially located on a pair of lightweight safety glasses (Figure 4). One infrared light was located next to the eye camera in order to illuminate the eye properly. This light provided us a constant spot of white light known as the first corneal reflection, which was used for further offline analysis using the ExpertEyes software (an open source application for analyzing eye tracking available at http://code.google.com/p/experteyes/) developed by our research group.

Using this setup, we recorded the video stream from both cameras, which was split later into image sequences. These images were used by the two modules of our software for further temporal alignment, calibration, and gaze estimation. The first module used the images from the eye stream in order to calculate the relationship in time between the pupil and the corneal reflection and fit the eye model. The second module used the images from both streams and the eye model data to synchronize and calibrate both streams.

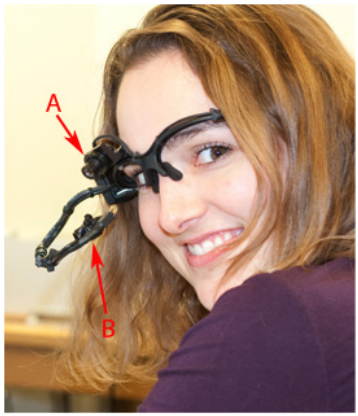


Figure 4

Eye tracker used for Experiment 2. A scene camera (A) monitors the view of the world, and the eye camera (B) monitors the eye position. Calibration procedures provide information about where the observer moves his or her gaze relative to the world.

The ExpertEyes eye tracking system allows the computation of the average error of the eye tracker. The Tobii system reports the average error of 0.5 degrees of visual angle under typical use. We found that our eye tracker produced values in a similar range. The mean error for experts was 0.48 degrees, whereas the mean error for novices was 0.57 degrees. These values were not significantly different for the two groups ($t(22) = 1.76$; $p > 0.05$; cohen's $d = 0.75$). Thus, we have confidence that the eye tracking results from Experiment 2 are comparable in accuracy to those of Experiment 1 and that data from both groups are of comparable accuracy and resolution.

To avoid the confound of different subject groups viewing the images for different durations, in Experiment 2 we limited the viewing durations for all trials to 20 seconds. This number was shorter than the shortest average of the two groups (the Novice group had a mean viewing time of 21.3 seconds in Experiment 1) and ensured that participants would remain on task for the entire trial duration. This change necessitated one other alteration to the procedures. We added an option of "too soon to tell" that allowed the participant to avoid making a forced-choice decision. This was done for two reasons. First, the experts were understandably reluctant to make errors given the potentially severe consequences of an error in casework, especially after

such a brief viewing duration. Second, such a response could be construed as a state of less confidence, and therefore provide additional information about the decision state of the participant. This required some alterations to the way we computed accuracy, but otherwise did not affect our analyses. In fact, because experts make so few erroneous identification errors, this change allowed us to compute an accuracy measure for data that contained no false positive errors, alleviating the correction that was necessary in Experiment 1 to compute d'. We describe these procedures in a subsequent section. Participants responded at the end of each trial by clicking one of three buttons in a dialog window before advancing to the next trial. An early version of our software artificially truncated the accuracy results after 17 trials for six of our experts. However, because our scene camera recorded these mouse clicks in the dialog window, we were able to reconstruct the responses of these subjects for all but 7 trials where the click occurred between video frames. These trials were left blank for the subsequent analyses and represent less than 2% of the overall data for the experts and 1% of the data overall. Our reconstruction procedures were 100% accurate at predicting the responses we did have for each of the six subjects, and, therefore, we are confident in the data we were able to reconstruct.

Fixation Analysis

We developed our own algorithm of eye fixation finding which uses eye motion to separate fixations. First, we performed a running median filter over the data, which takes the median of three consecutive points. This served to reduce the effect of noise in the pupil estimation. Next, we computed the magnitude of velocity at each time point in the data. Finally, we established a velocity threshold to segment the whole continuous stream into several big segments that corresponded to dramatic eye location changes. This threshold was set to 7.3 degrees per second, which is somewhat lower than values typically used in the literature, and we chose this in part because of our relatively slow sampling rate (30 Hz) and median filter. However, it is similar to the 20 degrees/sec adopted by Sen and Megaw [30]. To avoid spurious brief fixations, we established a minimum duration for a fixation of 67 msec. This algorithm provided results that were very similar to the space-based approach used in Experiment 1 with these parameters, as was verified by running this algorithm over the Experiment 1 data and observing a close correspondence between the two approaches.

Results

We describe a similar set of analyses as in Experiment 1.

Fixation Statistics

We again computed fixation durations and saccade length statistics. As with the previous experiment, we found no differences in terms of the average duration of each fixation for the two groups for either the latent or inked prints (all $p>0.5$). We did find that experts spend more time than novices looking at the latent print. Of all the fixations made by experts, 75% were to the latent print, whereas only 69% of fixations made by novices were on the latent ($t(22) = 2.24$; $p<0.05$; cohen's $d = .96$).

We fixed the duration of the stimuli, so we cannot compare the two groups on overall viewing duration. However, we can look at the average length of each saccade. On our display, 100 pixels subsume about 2.6 degrees of visual arc. We found that experts had much shorter saccades than novices both on the latent print side (61.6 vs 96.0 pixels, $F(1,22) = 57.8$, $p<0.01$, cohen's $d = 1.61$) and on the inked print side (51.1 vs 109.4 pixels, $F(1,22) = 53.0$, $p<0.01$, cohen's $d = 1.54$). These results are consistent with experts making smaller eye movements to regions that are closer together.

Behavioral Accuracy

To compute a measure of accuracy, we considered the three response categories ("match", "too soon to tell" and "nonmatch") as a continuum of certainty of a match. By looking at the rate at which each subject made each of these responses to the two image categories (true matches and true nonmatches), we can compute a measure of discriminability that is related to d', called A' (or A-prime) [25]. The three response categories ("match", "too soon to tell" and "nonmatch") are three levels along an evidence axis, and we accumulated the proportion of each response for both true matches and true nonmatches. The cumulative percentage of responses were plotted in a receiver operating characteristic (ROC) graph and A' is simply the area under the ROC curve. No assumptions of normality are required to compute A', although an inverse normal transformation can be used to convert to d' which is unbounded and tends to be more linearly related to underlying psychological dimensions.

A' varies from .5 (guessing) to 1.0 (perfect performance). Because it is bounded at 1.0, it tends not to be a measure that is linear with some underlying psychological dimension such as task difficulty, but it is monotonically related to such dimensions. Thus, higher scores on this measure always imply higher accuracy. In addition, the A' measure will tend to compute accuracy while removing response bias. This is not to say that such response biases are unimportant, but in the present context, we were more interested in accuracy than bias.

A participant may improve accuracy by moving as many trials from the "too soon to tell" category into either the matching or nonmatching category. However, this must be done in such a way that no errors are introduced (i.e., saying "match" to nonmatching stimuli and "nonmatch" to matching stimuli). The experts were quite good at avoiding the first kind of error, known as a false positive. This is very bad in practice since it erroneously identifies a wrong person. Table 1 lists the proportion of responses in each category for experts and novices. As can be seen, experts made no erroneous identifications, while fully 25% of the responses by novices to the nonmatching prints were erroneous identifications. Experts also were somewhat conservative, making many more "too soon to tell" responses than novices did to the True Match stimuli. This is in line with the general concern of our examiners that, although it may be relatively straightforward to conclude in 20 seconds that two prints are not from the same source, they are reluctant to make a positive ID after such a brief duration. This leads to their conservative response bias. Experts demonstrated a missed identification rate of 13.9%, which is almost three times lower than those of the novices (36.1%), as can be seen in the right column of the top section of Table 1.

True Matches

Participant Category	"Yes"	"Too Soon to Tell"	"No"
Experts	16.15%	69.97%	13.88%
Novices	36.67%	27.22%	36.11%
p-value	0.003	0.000	0.000

True Nonmatches

Participant Category	"Yes"	"Too Soon to Tell"	"No"
Experts	0.00%	26.67%	73.33%
Novices	25.00%	15.00%	60.00%
p-value	0.003	0.175	0.181

Table 1

Proportion of responses in each category for experts and novices. The last line of each subtable contains the p-value of the associated t-test comparing the two groups on each response category.

Given the conservative responding of the experts and their reluctance to say "yes" overall, the differences seen in Table 1 might simply be a shift in response criterion. The benefit of using A' as a measure is that it allows for a computation of accuracy that is — in theory — separated from response criterion. Based on this measure, experts have much greater accuracy than novices. Overall, the experts had an average A' value of 0.82, whereas the novices had an average A' of 0.61. This was a significant difference ($t(22) = 3.68$; $p < 0.01$; cohen's $d = 1.57$), and if the values of probability are transformed into d' values via an inverse normal transformation, we see that the experts' accuracy of .94 was almost triple that of the novices, which was .32. Thus, despite evidence of a conservative response strategy, experts were clearly outperforming novices, and this comes mainly because they avoid the very costly erroneous identifications while making fewer correct identifications and fewer missed identifications. The erroneous identification (false positive) errors made only by novices greatly reduce performance on the A' measure, as well as being generally viewed as a worse error than a missed identification by our society. It is of interest to note that this overall increase in accuracy is in evidence despite the fact that the experts made significantly fewer correct identifications, preferring instead to use the "too soon to tell" category significantly more than novices.

Earth Mover Analysis

As with Experiment 1, we computed the Earth Mover distance for each trial (including each of the individual presentations within each trial) between each expert and every other expert, as well as between each novice and every other novice. We found that under these testing conditions, experts showed much more consistency than the novices. Experts had smaller distances to other experts on 28 trials, and novices had smaller distances to other novices on only 7 trials. Experts had an average distance to other experts of 121.3 pixels, whereas the average distance of novices to other novices was 144.3 pixels ($F(1,68) = 15.7$; $p<0.001$, cohen's $d = 0.96$). When the analysis is restricted to just the latent print, we found a similar difference (77.53 vs 95.37 pixels, $F(1,68)=14.5$; $p<0.01$, cohen's $d = 0.92$), as well as with the inked prints (92.2 vs 117.1 pixels, $F(1,68)=20.3$, $p<0.01$, cohen's $d = 1.09$). Thus, under a variety of analyses, the experts demonstrated more consistency than novices.

The images in Figure 5 provide a representative visualization of why experts may show more consistency than novices. The latent print in this pair is a very difficult distorted impression with the addition of visual noise. The fixations from experts (plotted as dark triangles) show a clear clustering on the inked print that corresponds to the region in the latent print that has the greatest clarity. The novices have a much wider distribution of fixations, including in areas with apparent poor ridge detail. These results are consistent with the idea that experts have a clear sense of what constitutes high-quality information in latent prints. They tend to move their eyes to this region in the latent print, and then move their gaze to corresponding locations in the inked print.

Number of Minutiae Near Each Fixation

We again processed each pair of images using the Universal Latent Workstation and computed the number of nearby minutiae within a 1.22° (50 pixel) radius of each fixation for both experts and novices, although the results are consistent across a wide range of radii. Recall that in Experiment 1 we found no difference between the two groups, which may have resulted from the fact that the poor quality of the latent print dictated what information was usable.

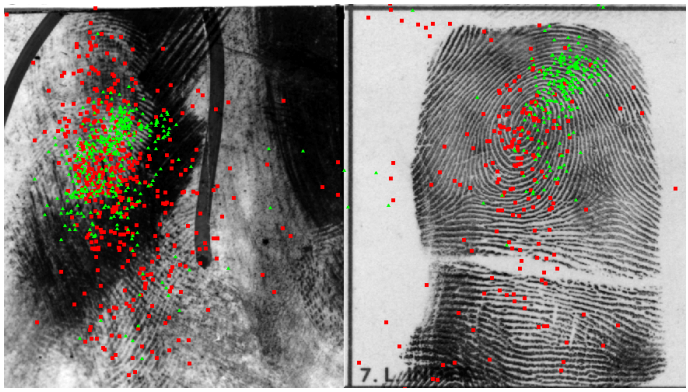


Figure 5

Latent and inked pair from Experiment 2, with fixations from all experts overplotted as green triangles, and fixations from all novices overplotted as red squares. The green triangles tend to be clustered in the upper-right portion of the inked print (right image), which corresponds to the area of high detail in the latent print. However, novices have a much wider distribution of fixations, including in regions that have very poor image quality.

We again found no difference between the two groups in terms of the number of minutiae visited. Experts had an average of 1.24 minutiae near each fixation, and novices had an average of 1.29. This difference is not significant because the 95% confidence interval based on the null hypothesis of no difference between the two groups is [-0.23, 0.21] and includes the actual difference of -0.045.

Image Reliability at Each Fixation

We looked at the image reliability at each fixation and found a small but significant difference between the two groups. For the latent prints, we found that experts were actually seeking out slightly higher regions of image quality by this metric. We found that the experts have a mean of .614 and the novices have a mean of .612. The 95% confidence interval on the difference [0.0004, 0.0050] does not include zero, demonstrating this difference is small but statistically reliable. Thus, for the latent print, experts were choosing to view regions that tended to be of higher image quality.

Interestingly, the reverse is true for the inked print. Experts have a mean of .703 by this metric, and novices have a mean of .754, and the confidence interval on the difference [-0.0711, -0.0315] does not include zero. This may result from the fact that the experts tended to look in the regions of the inked print that correspond to the clear areas in the latent print, whereas novices may look at a broader set of areas, including regions that were of higher quality in the inked print, not realizing that these were of little value if there is no matching region in the latent print.

Discussion

Several results of Experiment 2 stand in contrast to those of Experiment 1. First, experts as a group now show more consistency than novices, which is consistent with the hypothesis that experts share a common understanding of which regions of the latent print are more informative. They then move their eye gaze to the corresponding locations on the inked prints. This leads to slightly higher estimates of image reliability in the latent prints relative to novices, but the matching constraint in the inked print may mean that they were forced to look at regions of the inked print that have relatively poorer image quality in order to find correspondence. Novices may look to regions of the inked print that have higher quality, not realizing that there is no matching region in the latent print.

General Discussion

Consistency among experts is an important issue for the legal system, because we value a body of experts who generally agree upon a set of procedures and methodologies that is shared by the community. However, if multiple experts contribute to a decision, there may be value in each examiner acquiring different sources of information when making an independent decision. The results of List 3 in Experiment 1 demonstrate that when given relatively unconstrained viewing time, experts will look at images longer than novices, and as a result may tend to be more variable as a group simply because they have more opportunity to explore a wider set of areas. Note that this candidate explanation is not directly testable given the design of Experiment 1, but the fixed and shorter viewing durations of Experiment 2 demonstrated quite clearly that once the confound of different viewing durations is removed, experts as a group will tend to be much more consistent than novices.

The statistics of saccades and fixations reveal other group differences that are indicative of the strategies adopted by experts. Experts spent more time on the latent prints than novices and tended to have shorter saccades. This is consistent with the reported strategy described by experts in which they first look at the latent print for high-clarity regions and then place collections of target features in working memory prior to looking for correspondence with the detail on the inked print.

Kundel and Nodine [31] showed that eye movements can be combined with stored memories to build up a perceptual representation of a picture. Thus in Experiment 2, experts might be able to use concepts learned during training to dictate which regions of the latent print may correspond to high-quality visual features in the latent print. Further support for this hypothesis comes from research that shows that language and labels shape concepts and percepts [32]. Experts have a definite language used to describe the characteristics of fingerprints. Thus, this language could shape the relative importance of fingerprint regions and make them psychologically salient, which in turn leads to agreement among experts in terms of where they direct their gaze.

We would like to point out that these experiments were intended to answer specific questions about intersubject variability and consistency, which require each participant to inspect relatively large numbers of image pairs in a brief time. Caution should be exercised when generalizing these results to full-blown, latentprint and inked print casework, where the examiner has much more time to inspect each print pair. However, we believe that these results tap into whatever skills and strategies experts have developed as part of their training and experience with wide varieties of qualities and quantities of details throughout latent and standard prints.

Still unanswered is the question of what or how many features experts rely on when matching fingerprints — a research question that cannot be answered by using surveys or questionnaires, but we can infer this expertise from eye movement data. This will require data from many more participants and is an ongoing research topic in our laboratory. Overall, our results show the promise of this eye tracking approach, which has the potential to apply perceptual and cognitive principles to fingerprint practice. We show that moment-by-moment eye movement data generated by experts are an informative resource that we can use to infer their underlying cognitive states and cognitive processes when conducting an examination. Meanwhile, as a first study of this

type, our efforts here also pointed to the technical challenge in this venue – given a huge amount of fine-grained eye movement data, how to effectively derive meaning results from such data. This is one primary focus of our future research – developing and using more sophisticated data analysis methods to extract more interesting results from eye movement data.

Final Thoughts

The notion of variability and consistency among experts is bound to provoke strong feelings among both supporters and critics of latent print comparisons. It is clear that latent prints typically contain multiple types of information, and it could be argued that this variability may be worrisome if different experts use different features. Note that the finding of greater variability among experts need not invalidate the latent print examination practice, but it would foster a healthy debate about the varieties of techniques and approaches currently employed by the practitioners. Anecdotally, there exist several different aspects of examinations that might be taught, such as point counting, point parachuting (looking for points without regard to their configuration), ridgeology, ridge running, holistic, qualitative quantitative, or variations of these. Thus there may already exist the bases for variability among experts, although data from a larger group of experts would be required to see evidence of these approaches in the eye tracking results. It may be reassuring that at least among our experts we see greater consistency among our experts than the novices in the well-controlled conditions of Experiment 2. However, it should be noted that under these conditions, the experts were unwilling to commit to positive identifications for the most part. These are issues that the latent print community will have to grapple with, and we hope that this article will provide the data necessary to continue the discussion about what constitutes the best practices in latent print examinations.

Acknowledgment

This research was supported by grants #2005-MU-BX-K076 and #2009-DN-BX-K226 from the National Institute of Justice.

For more information, please contact:

Tom Busey
Department of Psychology
Indiana University
Bloomington, IN, 47405
busey@indiana.edu

References

1. Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council. *Strengthening Forensic Science in the United States: A Path Forward*. National Academies Press: Washington, DC, 2009.
2. Snodgrass, M.; Bernat, E.; Shevrin, H. Unconscious Perception at the Objective Detection Threshold Exists. *Perception & Psychophysics* **2004**, *66* (5), 888–895.
3. Van Selst, M.; Merikle, P. M. Perception Below the Objective Threshold? *Consciousness and Cognition* **1993**, *2* (3), 194–203.
4. Dror, I. E.; Charlton, D. Why Experts Make Errors. *J. For. Ident*. **2006**, *56* (4), 600–616.
5. Dror, I. E.; Charlton, D.; Péron, A. E. Contextual Information Renders Experts Vulnerable to Making Erroneous Identifications. *For. Sci. Int*. **2006**, *156* (1), 74–78.
6. Dror, I. E.; Péron, A. E.; Hind, S.; Charlton, D. When Emotions Get the Better of Us: the Effect of Contextual Top-down Processing on Matching Fingerprints. *Appl. Cognitive Psychology* **2005**, *19* (6), 799–809.
7. Krupinski, E. A. Visual Scanning Patterns of Radiologists Searching Mammograms. *Acad. Radiol*. **1996**, *3* (2), 137–144.
8. Krupinski, E. A.; Nishikawa, R. M. Comparison of Eye Position Versus Computer Identified Microcalcification Clusters on Mammograms. *Med. Phys*. **1997**, *24* (1), 17–23.
9. Kundel, H. L.; Nodine, C. F.; Krupinski, E. A.; Mello-Thoms, C. Using Gaze-tracking Data and Mixture Distribution Analysis to Support a Holistic Model for the Detection of Cancers on Mammograms. *Acad. Radiol*. **2008**, *15* (7), 881–886.
10. Manning, D.; Ethell, S.; Crawford, T. An Eye-tracking AFROC Study of the Influence of Experience and Training on Chest X-ray Interpretation. In *Medical Imaging 2003: Image Perception, Observer Performance, and Technology Assessment*; Chakraborty, D. P., Krupinski, E. A., Eds.; Proceedings of SPIE: Bellingham, WA, *5034*, 257–266.
11. Kundel, H. L.; Nodine, C. F.; Conant, E. F.; Weinstein, S. P. Holistic Component of Image Perception in Mammogram Interpretation: Gaze-tracking Study. *Radiology* **2007**, *242* (2), 396–402.
12. Kundel, H. L.; Nodine, C. F.; Carmody, D. Visual Scanning, Pattern Recognition and Decision-making in Pulmonary Nodule Detection. *Invest Radiol*. **1978**, *13* (3), 175–181.
13. Dyer, A. G.; Found, B.; Rogers, D. Visual Attention and Expertise for Forensic Signature Analysis. *J. For. Sci*. **2006**, *51* (6), 1397–1404.

14. Charness, N.; Reingold, E. M.; Pomplun, M.; Stampe, D. M. The Perceptual Aspect of Skilled Performance in Chess: Evidence from Eye Movements. *Mem. Cognit.* **2001**, *29* (8), 1146–1152.
15. Busey, T. A.; Vanderkolk, J. R. Behavioral and Electrophysiological Evidence for Configural Processing in Fingerprint Experts. *Vision Res.* **2005**, *45* (4), 431–448.
16. Bond, G. D. Deception Detection Expertise. *Law Hum Behav.* **2008**, *32* (4), 339–351.
17. Chi, M. Two Approaches to the Study of Experts' Characteristics. In *The Cambridge Handbook of Expertise and Expert Performance*; Ericsson, K. A., Charness, N., Feltovich, P. J., Hoffman, R. R., Eds.; Cambridge University Press: Cambridge, 2006; pp 21–30.
18. Abernethy, B.; Russell, D. G. The Relationship between Expertise and Visual-Search Strategy in a Racquet Sport. *Hum. Movement Sci.* **1987**, *6* (4), 283–319.
19. Schyns, P. G.; Rodet, L. Categorization Creates Functional Features. *J. Experimental Psychology: Learning, Memory, and Cognition*, **1997**, *23* (3), 681–696.
20. Shiffrin, R. M.; Lightfoot, N. Perceptual Learning of Alphanumeric-Like Characters. *Psych. of Learning and Motivation* **1997**, *36*, 45–81.
21. Mello-Thoms, C.; Ganott, M.; Sumkin, J.; Hakim, C.; Britton, C.; Wallace, L.; Hardesty, L. Different Search Patterns and Similar Decision Outcomes: How Can Experts Agree in the Decisions They Make When Reading Digital Mammograms? In *Digital Mammography: Lecture Notes in Computer Science*, Vol. 5116/2008, Springer Berlin: Heidelberg, 2008; pp 212–219.
22. Rayner, K. Eye Movements in Reading and Information Processing: 20 Years of Research. *Psychol. Bull.* **1998**, *124* (3), 372–422.
23. Nodine, C. F.; Mello-Thoms, C.; Kundel, H. L.; Weinstein, S. P. Time Course of Perception and Decision Making During Mammographic Interpretation. *Am. J. Roentgenol.* **2002**, *179* (4), 917–923.
24. Kundel, H. L.; Nodine, C. F. Interpreting Chest Radiographs without Visual Search. *Radiology* **1975**, *116* (3), 527–532.
25. Creelman, C. D. Signal Detection Theory and Roc Analysis in Psychology and Diagnostics: Collected Papers. Contemporary Psychology, **1998**, *43* (12), 840–841.
26. Rubner, Y.; Tomasi, C.; Guibas, L. J. The Earth Mover's Distance as a Metric for Image Retrieval. *Int. J. Computer Vision*, **2000**, *40* (2), 99–121.

27. Bellman, R. Dynamic Programming and a New Formalism in the Theory of Integral Equations. *Proceedings of the National Academy of Sciences of the United States of America*, **1955**, *41* (1), 31–34.
28. Kovesi, P. D. MATLAB and Octave Functions for Computer Vision and Image Processing. www.csse.uwa.edu.au/~pk/research/matlabfns/ (accessed May 2008).
29. Babcock, J. S.; Pelz, J. Building a Lightweight Eyetracking Headgear. Paper presented at the ETRA 2004: Eye Tracking Research and Applications Symposium, 109–113.
30. Sen, T.; Megaw, T. The Effects of Task Variables and Prolonged Performance on Saccadic Eye Movement Parameters. In *Theoretical and Applied Aspects of Eye Movement Research*; Gale, A. G., Johnson, F., Eds.; Elsevier: Amsterdam, 1984; pp 103–111.
31. Kundel, H. L.; Nodine, C. F. A Visual Concept Shapes Image Perception. *Radiology* **1983**, *146* (2), 363–368.
32. Lupyan, G.; Rakison, D. H.; McClelland, J. L. Language Is Not Just for Talking - Redundant Labels Facilitate Learning of Novel Categories. *Psychol. Sci.* **2007**, *18* (12), 1077–1083.