COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 41 (2017) 1716–1759

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12452

Characterizing Human Expertise Using Computational Metrics of Feature Diagnosticity in a Pattern Matching Task

Thomas Busey,[a] Dimitar Nikolov,[b] Chen Yu,[c] Brandi Emerick,[a] John Vanderkolk[d]

[a]*Department of Psychological and Brain Sciences, Indiana University*
[b]*School of Informatics, Indiana University*
[c]*Department of Psychological and Brain Sciences & School of Informatics, Indiana University*
[d]*Indiana State Police Laboratory*

Received 7 April 2015; received in revised form 11 March 2016; accepted 13 July 2016

Abstract

Forensic evidence often involves an evaluation of whether two impressions were made by the same source, such as whether a fingerprint from a crime scene has detail in agreement with an impression taken from a suspect. Human experts currently outperform computer-based comparison systems, but the strength of the evidence exemplified by the observed detail in agreement must be evaluated against the possibility that some other individual may have created the crime scene impression. Therefore, the strongest evidence comes from features in agreement that are also not shared with other impressions from other individuals. We characterize the nature of human expertise by applying two extant metrics to the images used in a fingerprint recognition task and use eye gaze data from experts to both tune and validate the models. The Attention via Information Maximization (AIM) model (Bruce & Tsotsos, 2009) quantifies the rarity of regions in the fingerprints to determine diagnosticity for purposes of excluding alternative sources. The CoVar model (Karklin & Lewicki, 2009) captures relationships between low-level features, mimicking properties of the early visual system. Both models produced classification and generalization performance in the 75%–80% range when classifying where experts tend to look. A validation study using regions identified by the AIM model as diagnostic demonstrates that human experts perform better when given regions of high diagnosticity. The computational nature of the metrics may help guard

Correspondence should be sent to Thomas Busey, Department of Psychological and Brain Sciences, Program in Cognitive Science, Indiana University, 1101 E. 10th St, Bloomington, Indiana 47405. E-mail: busey@indiana.edu

To make our work more easily reproducible, we have provided the scripts we used for our data analysis at https://github.iu.edu/busey/FingerprintFeatureRarity. This code contains the data path for going from raw images, to computing representations based on the models, to computing the classifications results and visualization presented here. The code can be run on the test image provided in the repository or on a database of images input to it. The repository does not contain the code for Karklin and Lewicki's CoVar model, which can be obtained from them directly.

against wrongful convictions, as well as provide a quantitative measure of the strength of evidence in casework.

Keywords: Expertise; Categorization; Salience; Forensics; Natural image statistics; Eye tracking; Fingerprints

1. Introduction

Forensic evidence presented in court often takes the form of a source conclusion statement. For example, at a trial a latent print examiner could show two fingerprints, one from a crime scene and the other from the defendant. The examiner might offer an expert opinion that the two impressions were made by the same finger. In support of this conclusion, the examiner may offer diagrams showing the regions in the two impressions that they believe correspond. Despite the compelling nature of such diagrams, it is not enough to find correspondence between two impressions, because when making an individualization claim, the examiner is also stating "the likelihood the impression was made by another (different) source is so remote that it is considered as a practical impossibility"[1] (SWGFAST, 2013). To rule out all other individuals as possible donors of the impression, one would need to demonstrate that these features (at whatever resolution they are represented in the impression) are not likely to appear from the fingers of all others who may have had access to this surface. This reveals another philosophical challenge: There is typically no way to effectively compare the fingerprints of all other individuals who may have had access to this surface.

This challenge places the individualization conclusion on less than firm foundations, as noted by several authors (Cole, 2009; Kaye, 2003). However, this does not mean that fingerprints lack evidentiary value, even if the questioned impression cannot be compared to every individual who may have had access to a surface. Instead, one solution is to rely on features that are the most rare, in that the features are seldom found in a reference database. By inference from face recognition, saying that the suspect had two eyes is almost meaningless for purposes of identification, but saying that a suspect has a heart-shaped mole on the left cheek could be very valuable at a trial. If the goal is to determine whether two impressions could have only have come from one person, an examiner should focus on those features that are specific to this individual and are not shared by others. Knowing which features on an impression tend to be rare would help an examiner make decisions based on the most diagnostic regions of the impression.

But with fingerprints, what is a feature, and how can we define rarity? The challenge with rarity is that it depends both on how a feature is defined, as well as the reference dataset from which these features are drawn. Fingerprints are presumed to be unique, as even identical twins have demonstrably different fingerprints (Srihari, Srinivasan, & Fang, 2008). However, uniqueness is a trivial concept in this context, because each impression is unique, and not all the features of the skin readily transfer to the surface. What is necessary is an approach that characterizes the features that tend to transfer from the skin to

a surface and also tend to be uncommon enough so that only a few individuals may share those features.

This project has two inter-related goals. First, we will measure eye gaze behavior of human experts and use this data in conjunction with the visual details of the fingerprint impressions to build and tune models that attempt to capture elements of human expertise as well as define a feature set. Second, we will use these models to identify whether examiners are using the features that are considered rare and diagnostic by the models. If there are discrepancies, these can be inspected to see whether the model needs to be improved or whether there are diagnostic features that experts are not using. Because neither the models nor the experts are likely to be perfect, and because we do not yet know what constitutes a feature in fingerprints in the absence of a perfect model, we will explore the strengths and weaknesses of the models and the human experts as means to bootstrap the performance of both. Toward this goal, we will use a validation study as well as generalization to new examiners and impressions to demonstrate the degree to which the models do or do not capture human expertise. These will illustrate ways that both the models and humans might be improved.

1.1. Latent print examinations

In most jurisdictions, human examiners typically receive substantial training before performing comparisons between latent prints (typically taken from crime scenes) and tenprints (typically collected during an arrest). Currently, human experts typically outperform computer systems when conducting an examination, and in most U.S. laboratories, computer databases such as the Automated Fingerprint Identification System (AFIS) are traditionally used only to present candidate matches to a print recovered from a crime scene. Due to variation in appearance between impressions, the examiner must decide whether there is enough perceived detail in agreement between two impressions to conclude that they belong to the same category of impressions made by a specific finger. Because this is a local categorization task, it is similar in structure to other visual categorization tasks with a limited number of target categories, such as radiology, security scanning, bird identification, and mushroom hunting. As a result, the techniques described in the present work could be applied to similar visual categorization tasks.

As noted by Dror and colleagues (Dror & Mnookin, 2010; Dror et al., 2011), performance on latent print comparison tasks depends on a number of factors, including decisions about the amount of information that is sufficient for different decisions, as well as interaction with technology and external information about the case. Below we provide a brief summary of these factors because one of the goals of the present work is to demonstrate how the modeling derived from information theory-based approaches can improve the evidentiary values of latent print impressions. These details also specify the nature of the task, which will constrain aspects of the models.

During a comparison, a human examiner will first examine the latent print to determine if they believe the latent to be *of value* for purposes of a comparison. They will also make preliminary marks on the latent print to indicate the locations and orientations of

features they consider diagnostic as part of an analysis stage. The examiner will then compare this print against exemplar impressions collected by a detective or recovered during an automated database search. The results of this comparison stage are typically phrased in terms of an *identification* (implying that the two impressions came from the same finger), an *exclusion* (implying that two different fingers created the two impressions), or *inconclusive* (implying that no determination was possible). During most testimony, the decision is not accompanied by qualifications about confidence or difficulty, and virtually all evidence presented in court is by human experts without reference to a computer match or statistical support from a model of the distributions of features (although this may change; see Neumann et al., 2007, 2006; Neumann, Champod, Yoo, Genessay, & Langenburg, 2015; Neumann, Evett, & Skerrett, 2012; Srihari & Su, 2008; Su & Srihari, 2009).

Latent print examiners make identification or exclusion decisions based on the perceived agreement or disagreement in the detail they observe in both impressions. However, to make a convincing case for identification, the detail in agreement must provide *specificity*. That is, for purposes of identification, not only must features be present in both impressions, but the features must also tend to be absent in impressions made by other sources. This implies that the most diagnostic features are those that are specific to one individual. This specificity is somewhat orthogonal to image clarity, such that a degraded yet highly specific feature may have more utility than a very clear impression that lacks specificity. The features must also have a tendency to be visible when developed on surfaces at crime scenes, which may bias the features toward those that tend to occur multiple times such as the core area (the central region of the print that demonstrates maximum curvature). Features from the edge of the print can vary in appearance from impression to impression depending on the manner in which the finger makes contact with the surface.

Accordingly, a metric must be sensitive to the characteristics of the task and the information that tends to be available. Recent work demonstrates that experts tend to share a common threshold for the determination of sufficiency based on minutiae counts (Ulery, Hicklin, Roberts, & Buscaglia, 2014). In addition, examiners make almost no erroneous identifications (about one in every 1,000 identifications) but have a relatively high rate of inconclusive and erroneous exclusion responses (Ulery, Hicklin, Buscaglia, & Roberts, 2011). There is a relatively high rate of reproducibility among the decisions made by examiners, subject to variability near the decision thresholds (Ulery, Hicklin, Buscaglia, & Roberts, 2012). One conclusion from this work is that experts may rely on a common set of features for purposes of identification or exclusion and therefore have a reasonable degree of expertise that might be used to tune models of this task. However, it has been difficult to determine the exact feature set used by examiners beyond traditional minutiae.

Specificity is relatively difficult to determine, because it is a property of the entire dataset as well as the latent print under consideration, and human examiners must evaluate the diagnosticity of a print relative to the statistics of the reference database. A quantitative metric that describes the base rates of various features could determine the specificity of regions of the latent print, as well as regularize the comparison process.

Recently, fingerprint comparisons have come under fire for a lack of such standardization. A recent report by the National Academies of Science (National Research Council of the National Academies of Science, 2009) pointed to a lack of standards among the community, large variations in training and mentorship, and a number of deficit management policies that may have contributed to past errors. Among their recommendations was a need to standardize the latent print examination process, which in our view could include specific definitions of the visual features used to make identifications or exclusions, along with a method to quantify the probability of observing similar features in a large dataset.

Relatively little is known about the features and process by which examiners conduct examinations. The current technique used by many active examiners to describe the comparison process is known as Analysis, Comparison, Evaluation, and Verification (ACE + V). This has been critiqued as more of a framework that describes the general procedures used by examiners rather than as a specific model because it does not define the thresholds for examiners and it does not specify the features that are to be used (Cole, 2005). An extended specification of the feature set contained in fingerprints has been recently attempted (NIST, 2015), but these characterizations tend to lose the texture details of the print, known as thirdlevel detail (Ashbaugh, 1999; Vanderkolk, 2009). Some of this information may be below the level of visual awareness and therefore difficult to verbalize (Snodgrass, Bernat, & Shevrin, 2004; Vanselst & Merikle, 1993).

An alternative approach that identifies the information used by examiners when conducting latent print examinations is to combine eye gaze data with the natural statistics of fingerprints and the constraints of the task. An accurate characterization of human expertise using quantitative models would create a quantitative specification of the strength of evidence in fingerprint comparisons, because it would characterize the strength of evidence in favor of both same-source and different-source conclusions. The need for such a specification was raised by both the NRC report (National Research Council of the National Academies of Science, 2009) and a working group sponsored by the National Institute of Standards and Technology (Expert Working Group on Human Factors in Latent Print Analysis, 2012) in response to the NRC report.

1.2. Available information in fingerprints

Performance in a comparison task is constrained by the available information in the impressions, as well as the ease of which that information is distinguishable from background noise and robust against distortion. Below we discuss candidate sources of information from fingerprint impressions.

Fig. 1 illustrates three patches that may vary in feature rarity.[2] The top row illustrates a patch that lacks classic minutiae (ending ridges and bifurcations), which may limit the specificity of this patch. However, under some circumstances, the ridge element details can provide specificity even in the absence of minutiae, as illustrated by the middle panel. The red and green shaded impressions come from the same finger, and the degree of agreement is clear in the third panel despite the lack of classic minutiae. Thus, even though examiners traditionally describe the features present in latent prints in terms of

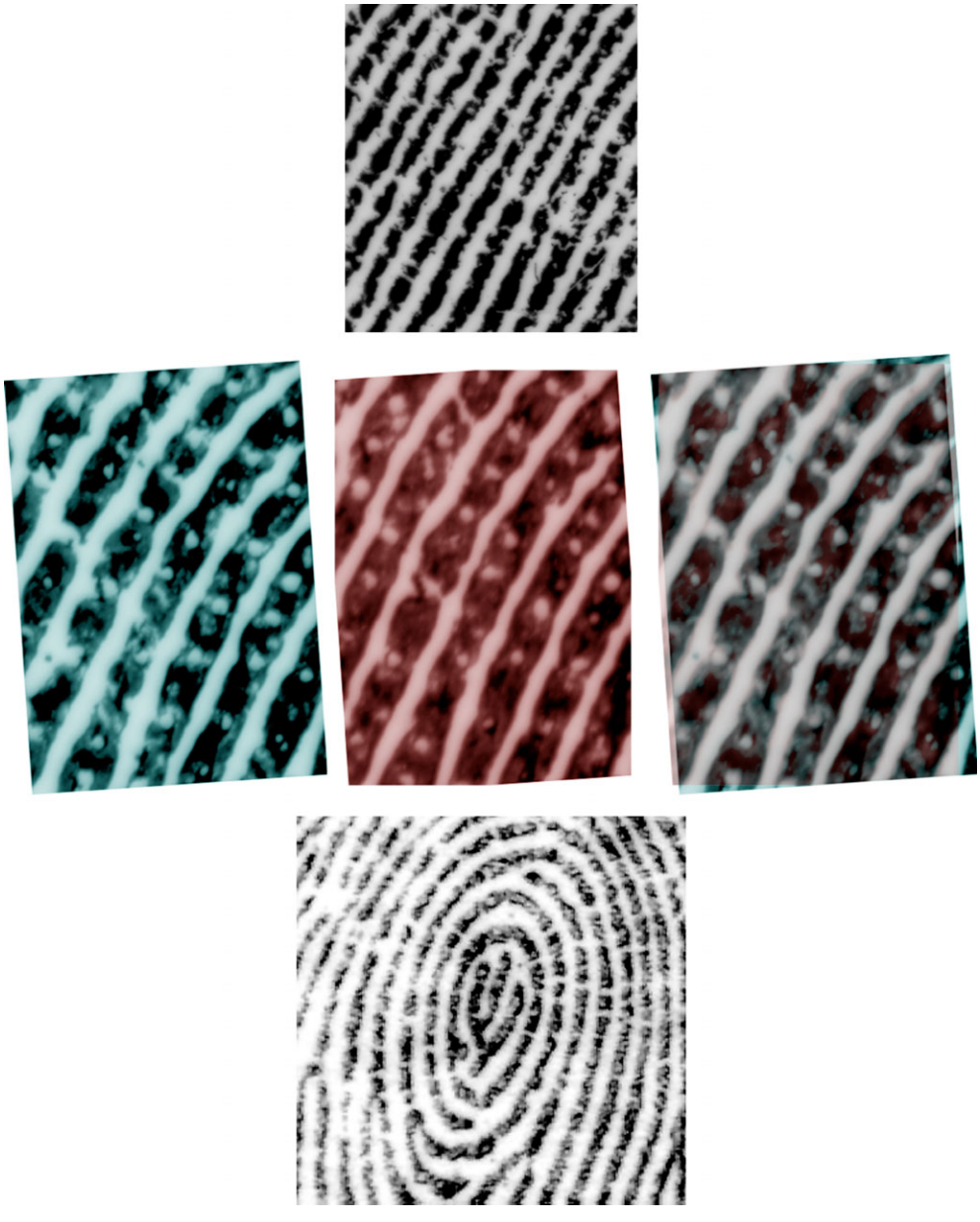

Fig. 1. Three examples of feature specificity or rarity. **Top row**: Although the ridges are fairly clear, the absence of specific features such as ridge elements or minutiae gives this patch relatively little specificity under traditional examination procedures. However, in some cases such an "open field" may prove useful because relatively few candidate prints may have such a large area that is absent of minutiae. **Middle row**: Although this patch lacks minutiae, the clarity of the ridge elements in both left (blue) and middle (red) impressions gives a great deal of specificity as illustrated by the composite image on the right, illustrating that individuating features can exist beyond classic minutiae. **Bottom row**: An extremely uncommon feature that was encountered only once in the 25-year career of a latent print examiner.

classical minutiae, very clear prints have useful detail at the level of individual ridge units. This demonstrates that various spatial scales may play a role in the identification process, from overall ridge flow down to the shapes of the ridge units. However, the ridge element structure may not be visible in degraded latent prints, and examiners may rely on features that span several ridges. The bottom panel of Fig. 1 shows a feature that may be relatively rare (it was noticed during casework by an examiner because it resembles the logo of Indiana University, his alma mater). In over 25 years of casework, the examiner reported seeing this feature only once, and thus this feature may provide a great deal of specificity. Note that the recognition of features may be assisted by linguistic and semantic information as an aid to memory, such as in the example above. This labeling also serves as a form of dimensionality reduction, which is an integral part of both modeling approaches described below (although neither has an explicit linguistic component).

The above examples informally illustrate the different aspects of feature rarity, and demonstrate that rarity may be defined by multiple factors at different scales. In fact, the mere presence of the large open field of ridges that lack minutiae in the top row of Fig. 1 may itself be diagnostic, since such sparse density of minutiae might be rare. Thus, the absence of features could be revealing.

In our discussions with latent print examiners, we found that experts tend to use a variety of different types of features, as well as different strategies to determine which portions of the latent impression reflect ridge detail and which represent visual noise. For example, a dot between two ridges may be quite diagnostic if it was caused by a biomechanical origin. However, if it is simply a spot of lint left behind, its appearance may actually mislead a computer algorithm. The ability to form hypotheses about the manner in which the impression was made may help constrain the interpretation of the visual evidence, and this level of hypothesis generation is not currently embedded in machine comparison algorithms. By using expert eye gaze data to help tune the parameters of a metric, we build in elements of the fingerprint comparison task and overweight particular types of information. However, once the metric is defined, it makes parameter-free predictions for future examiners and images, thus avoiding issues of circularity.

Human experts also describe using information at different spatial scales, which they term "levels of detail." Level 1 detail represents the overall ridge flow and Henry pattern type as such as whorls or loops. Level 2 represents the classic minutiae (ridge endings and bifurcations) as well as possibly mid-level features such as the length of ridges or local curvature. Level 3 detail represents the shapes and appearance of the individual ridge elements, as illustrated in the middle panels of Fig. 1. Although third-level detail is not commonly used in comparisons, Bromage-Griffiths (2011) has conducted recent validation studies on third-level detail to suggest that it may provide some utility for purposes of comparison. Given that information may be available in different spatial scales, one way in which humans may outperform machine identification algorithms is by combining information across spatial scales. Existing computer models tend to rely on level 2 (classical minutiae) features (Champod & Margot, 1995, 1996; Egli, Champod, & Margot, 2007; Fang, Srihari, Srinivasan, & Phatak, 2007; Neumann et al., 2006, 2007; Srihari &

Su, 2008; Su & Srihari, 2008, 2009, 2010), though some more recent proprietary AFIS algorithms may use more pixel-based information.

Human experts may have some ability to select diagnostic features, but they tend to be relatively weak at judging the base rates of different events (Barhillel, 1980). This is especially important when considering a candidate matching print derived from a large database and the problem of close non-matches (impressions that on casual inspection may give the appearance of a match while closer examination indicates an exclusion). As argued by Dror and Mnookin (2010) and modeled by Busey, Silapiruti, and Vanderkolk (2014), larger databases increase the likelihood of finding close non-matching prints. For example, the likelihood of discovering a close non-matching print between a latent print and a print taken from a suspect identified from other evidence such as cell phone records is likely vanishingly small. However, if the full FBI database of 900 million prints were queried using the latent print, each print in the database has the potential to contribute a close non-matching print, and the system is *designed* to discover these close non-matching prints. Thus even if the chance of any one impression in the database creating a close non-match is relatively small, the probability of discovering at least one close non-matching print from the entire database could be quite large. In fact, Busey et al. (2014) argued that databases may have an optimal size, and that larger databases increase the likelihood of false non-matches faster than increasing the likelihood of the suspect being in the database. A quantitative metric that captures the rarity of different types of features could combat this problem, because it would determine the degree of specificity for each similar feature.

1.3. Bootstrapping human expertise and machine-based representations

The goal of this article was to use elements of human expertise to develop a quantitative metric of the information available for latent print examinations. Such a metric would have four potential advantages. First, an appropriately defined metric could capture elements of human expertise to automate portions of the latent print examination task. Second, a metric could combat the problem that close non-matches tend to occur more often when large databases are queried. Third, a metric might help examiners estimate the base rates of particular feature comparisons all while making the mechanics of the matching process observable to both the prosecution and the defense.[3] Finally, a metric could provide a likelihood value that would demonstrate the strength of the evidence, similar to those built using classic minutiae features (Neumann et al., 2006, 2007; Su & Srihari, 2008, 2009).

To develop such a metric, we assume that human experts have developed elements of expertise that do not yet exist in machine identification systems. Note that while machine comparison systems are becoming fairly good with clean prints, at present human examiners outperform the machine identification systems with partial or degraded prints. Human examiners often are asked to work with relatively small patches that lack traditional minutiae, in which case the shape and appearance of individual ridge elements

become important (e.g., see the middle region of the print in the bottom panel of Fig. 1).

Although human experts may not have a full understanding of the statistics and base rates of various kinds of features, the knowledge gained from human experts from eye gaze data may help specify the parameters of a metric that can objectively define both the nature of visual features as well as their rarity. Such a metric could then apply to even novel prints not trained by the system. In this article, we address the construction and validation of such a metric, and other work addresses the application to latent print examinations for purposes of determining strength of evidence. Eye gaze by itself is not perfect: It is subject to sampling error, as well as limitations due to the size of the fovea. In addition, the pattern of eye gaze might differ depending on whether the examiner was looking for evidence to exclude or identify a print as coming from the same source as a comparison print. However, looking for the rarest features is diagnostic for both tasks, and experts describe approaching each comparison as a process of looking for detail in agreement and unexplainable detail in disagreement (unexplainable in the sense that it was not created by distortion or noise). Thus, eye gaze is a reasonable starting point as a means to capture elements of human expertise.

We use eyetracking methods to identify those regions visited by human experts, and then use the visual information derived from the fixated locations to develop a feature space and a resulting metric based on information theory. These are then used to create saliency maps that are then validated against new eyetracking data from novel prints, as well as a validation experiment that illustrates that the regions identified by the model as diagnostic do in fact lead to better identification performance. We will argue that finding close correspondence between the salience maps and the eye gaze of experts would demonstrate that we have captured some elements of human expertise in the quantitative metric, as well as illustrate that experts may gravitate toward the most diagnostic features. We validate this last conclusion using a study that manipulates the information presented to examiners based on the output of the model. Of course, elements of human performance might be suboptimal, in the sense that experts may not be looking at exactly the right features. However, the salience maps are based primarily on the statistics of the reference database, and thus the salience maps may provide a way to improve human performance as well. We will return to the possibility of human expert data and machine-based metrics bootstrapping each other in the Discussion section.

2. Methods

The eyetracking data were collected specifically for this project. We used hardware and software that was designed for portability to collect robust eye gaze data from forensic laboratories (Parada et al., 2015). The eyetracker uses one camera to record the position of the eye relative to the head, and a second camera to record the position of the head relative to the computer monitor that presents the fingerprints. Through calibration procedures afforded by the ExpertEyes software[4] we align the two video streams in time

and establish a set of correspondences between the eye position and the scene camera. This allows us to infer, at each point in time, the visual features that the expert considers to be diagnostic for the fingerprint examination task.

2.1. Participants

We recruited 22 latent print examiners as participants from state or large metropolitan agencies in Indiana, Illinois, and Nevada, as well as at a large conference attended by latent print examiners. They had a mean age of 34.8 years, with a range of 24 to 64 years. We asked that they had a minimum of 2 years of unsupervised casework, and the mean was 7.4 years with a maximum of 33 years of experience. Thirteen were female. We required a visual acuity at our viewing distance of 20/30 vision as measured by a portable Snellen chart. One subject had 20/30 vision, five had 20/25, and the rest were at 20/20. Seven subjects had no ocular correction, two had Lasik surgery, five had contacts, and the rest wore glasses.

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

2.2. Eye tracking recording

Our video-based eyetracker records two video streams at 30 Hz. The eye camera records both the position of the pupil and the position of the corneal reflection, which is the reflection of an infrared LED positioned near the eye camera that provides illumination of the pupil. The position of both of these features is recovered from the eye camera image using the ExpertEyes software, and typically the position of the eye is recorded using the relative location of the two features to control for movement of the eye camera. The scene camera is positioned to record the scene immediately in front of the observer, which in our case is the laptop screen. A series of calibration dots is presented on the monitor at the start and end of the experiment, and the observer is instructed to look at each dot. This establishes a correspondence between the position of the eye relative to the head and a location in the scene camera using a calibration that consists of a second- or third-order polynomial that links a location of the pupil to a unique location in the scene camera.

The scene camera also captures the location of the monitor, and we ultimately need the location of the gaze relative to the images on the monitor. The ExpertEyes software contains a set of routines to track the corners of the monitor. This allows exporting of the gaze data relative to the actual fingerprints. An interim step requires barrel distortion correction (Bouget, 2008), which ensures that linear interpolation between the corners is accurate. The final interpolated values are exported to MATLAB (Mathworks, 2012) for further statistical analyses.

The mean calibration accuracy is 0.52°, with a standard deviation of 0.18°, a minimum of 0.25°, and a maximum of 1.02°. These values are comparable to those of commercial systems, and more information about the calibration routines are found in Parada et al.

(2015). One degree of visual angle is approximately two ridge widths on the stimuli that we used, and so we have a resolution of approximately 1–2 ridges on the fingerprint impressions. Drift correction routines were implemented between trials to correct for slippage of the eye tracker that would introduce systematic error.

2.3. Stimuli

Our goal is to develop an information metric for relatively clean impressions and then extend it to relatively noisy impressions such as latent prints in future work. However, we want to use relatively difficult comparisons when collecting human data so that the experts will engage the same perceptual mechanisms that they use during casework. For example, if level 3 detail is available in a clean impression but not regularly used during casework, we would not want level 3 detail used to help determine regions of high specificity.

To achieve both goals, we started with relatively clean ten-print impressions as shown in the lower-right panel of Fig. 2. We then created noise patches by sampling from actual latent prints near the latent impression (but excluded regions that had obvious ridge detail). This noise patch was then extended using a texture synthesis algorithm (Portilla & Simoncelli, 2000) where the original impression was used to seed the synthesis algorithm. Multiplying the noise and the fingerprint impression together created the final impression, which is consistent with a subtractive model of adding pigments. This is illustrated in the top row of Fig. 2, and a simulated latent print impression is shown in the lower-left panel of Fig. 2.

The experiment had 39 images, of which 13 had no noise on the simulated latent. Nine of the comparisons were non-mated images, usually created by left-right reversing the opposite hand finger of the same individual. We told the subjects only that many of the prints are matching pairs. We chose this distribution of mated and non-mated impressions because examiners can make exclusion decisions fairly quickly because only one non-explainable difference needs to be found. However, identification requires seeking the most specific regions of the impressions to demonstrate a high degree of certainty in a match.

2.4. Procedure

Subjects were seated in front of a 15″ laptop screen at a comfortable viewing distance. The eyetracker was placed on their head and a brief calibration screen was presented. Subjects were then given 20 s to perform an abbreviated comparison of a simulated latent print and a ten-print that were presented side by side on the computer monitor (see Fig. 2, bottom row). We chose to abbreviate the testing and fix it at 20 s for several reasons. Primarily we wanted examiners to send their gaze to those features that they considered to be most diagnostic. In addition, capping testing at 20 s ensured that all examiners would complete all trials in the allotted 20 min testing interval, and each examiner would spend equal time on all impressions. If we had allowed one subject to spend more time

Fig. 2. **Top row**: Creating simulated latent prints. Left panel—a magnified portion of a tenprint impression. Middle panel—simulated noise generated by sampling noise from an actual latent print near the latent, and then regenerated using a texture synthesis algorithm. Right panel—simulated latent print created by multiplying the noise by the image on a pixel-by-pixel basis. **Bottom row**: Example comparison trial illustrating a simulated latent print with artificial texture noise added to the left print, and a clean tenprint impression on the right. The simulated latent image on the left looks very similar to latent prints, but it has the advantage of allowing access to the ground truth of the ridge detail in the clean version of the print (which is not shown to the participant). A subject would be asked to determine whether the two impressions came from the same source.

on one impression set, this would have overweighted their data for purposes of the information metric. The 20 s testing interval also provides data on more images, thus increasing the range of features available to computational techniques designed to estimate the frequency of features in the database.

After each 20 s exposure duration, we removed the image pair and then asked the examiners to come to one of three conclusions:

"yes - they match"
"no - they don't match"
"too soon to tell"

This last category was necessary because 20 s is likely too short for a reliable conclusion. Recording typically lasted approximately 20 min per subject, which included calibration screens at the start and end of the experiment.

3. Results and discussion

The behavioral results are consistent with a high level of accuracy. To the mated pairs (produced by the same finger), participants made 542 "yes" responses (82.1%), 110 "too soon" responses (16.7%), and 8 "no" responses (1.2%). To the non-mated pairs (produced by different fingers), participants made 2 "yes" responses (1.0%), 31 "too soon" responses (15.7%), and 165 "no" responses (83.3%). To compute one overall measure of accuracy independent of a particular bias, we assume that the three types of responses represent three responses on an evidence continuum, which allows us to compute A′ (Pollack & Norman, 1964), a measure of discriminability that is criterion free (Macmillan & Creelman, 2004) and higher numbers correspond to better performance. It is bounded at 1.0 and typically does not fall below 0.5. The mean A′ value across all participants was 0.97, with five subjects having values of 1.0. The minimum A′ was 0.86, and the standard deviation was 0.04. If the mean A′ value is converted to a measure of sensitivity known as d' (Macmillan & Creelman, 2004) through an inverse normal transformation, the value is 1.91, and the average d′ of the individual subjects is 1.93. This latter value should be viewed as a lower bound, because the five subjects with perfect A′ values must be excluded from this analysis.

All these accuracy values are consistent with prior results (Busey, Yu, Wyatte, & Vanderkolk, 2013; Busey et al., 2011), which find that experts have fairly high accuracy and make relatively few errors relative to novices. Most errors are erroneous exclusions (saying "no" to mated prints) rather than erroneous identifications (saying "yes" to non-mated prints). Given the high level of accuracy, the results reported below do not change significantly if the error trials are excluded.

The raw gaze data from the eyetracker were segmented into fixations and saccades. We have developed our own algorithm of eye fixation finding which uses eye motion to separate fixations (Busey et al., 2013; Parada et al., 2015). First, we perform a running median filter over the data, which takes the median of three consecutive points. This serves to reduce the effect of noise in the pupil estimation. Next, we compute 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 correspond to dramatic eye location changes. This threshold was set to 7.3°/s, which is somewhat lower

than values typically used in the literature, and we chose this in part due to our relatively slow sampling rate (30 Hz) and median filter. However, it is similar to the 20°/s adopted by Sen and Megaw (1984). To avoid spurious brief fixations, we established a minimum duration for a fixation of 67 ms.

This establishes a set of fixations for each participant. To create an information metric based on the available information contained in friction ridge detail, we started by examining those regions visited by experts. For each fixation made by our participants we extracted small regions centered on the fixation, which were used to train the various models on the image content.

The levels of detail described by forensic examiners and illustrated by Fig. 1 demonstrate that the size of the image patches is important. Large image patches tend to focus on more global features, whereas smaller image patches necessarily only contain local features. Because there is no a priori specification of the appropriate spatial scale, we instead explored a range of spatial scales. In fact, we will argue that one reason human examiners are better than machine comparison algorithms is the ability to combine information across spatial scales. One solution is to combine information metrics across spatial scales, something we discuss in a later section. Our image patch sizes ranged from 24 pixels (approximately 1–2 ridge widths) all the way up to 160 pixels (corresponding to about one third of the image, or about 10 ridge widths).

Prior work found much more consistency among experts than novices for fixed-duration stimuli, which is consistent with the hypothesis that experts have a common set of regions or features that they tend to frequent (Busey et al., 2011). How can this consistency of image detail be best represented? We will adopt an approach previously applied to natural images that exploits the dependencies between different pixel locations that the natural scene statistics of fingerprint images provide.

3.1. Grounding eye fixations in image structure

Eye tracking data describe where the subjects moved their eyes, but not necessarily what features or visual information they relied on. As a result, we must use computational techniques to infer the nature of the visual information used by human experts. The basic premise behind dimensionality reduction and machine classification is to derive the basic building blocks or features that are used in the perception and matching of latent prints. We do this by borrowing from the known computational properties of the visual system, which tends to break up the visual scene into individual components and then build back up to a larger, more complete, representation of an object or scene. This requires using a set of elemental features called *basis functions*.

We infer the basis functions using the previously described pixel patches of fingerprint images that are centered on the eye fixations of experts. Our working dataset contains about 22,000 fixations, around which we take pixel patches that are small pixel crops from the larger fingerprint images. These contain 3 to 10 ridges, depending on the size of the patches.

We used independent components analysis (ICA) as developed by Hyvarinen et al. (Hyvarinen, Cristescu, & Oja, 1999; Hyvarinen & Hoyer, 2000), which is similar to the sparse coding techniques developed for images by Olshausen and Field (1996, 1997, 2004). The goal of the ICA approach is to discover the elemental features of a set of image patches, called the basis set. This approach has been widely used in face recognition research (Ding, Mei, Zhang, & Kang, 2001; Kim, Choi, & Yi, 2004; Kim, Choi, Yi, & Turk, 2005; Yang, Gao, Zhang, & Yang, 2005) and may prove useful with friction ridge skin. A basis function is simply a linear combination of pixel locations, which describes how important each pixel is to that basis function. Typically there are fewer basis functions than pixels in the patches. For example, there are 24*24 = 576 pixels in a 24 × 24 pixel patch, and we may have only 100 basis functions. The collection of basis functions creates the basis set.

ICA has two general properties. It looks for a basis set that:

1. Minimizes the mutual information between individual components, such that the information that one basis image tells you about the feature is independent from the information from other basis images.
2. Maximizes the non-Gaussianity of the transformed (projected) data, in the sense that the marginal distributions of the resulting activations tend to be both more peaked and have heavier tails than a typical Gaussian distribution.

The central limit theory says that as you add non-Gaussian sources (like uniform or sparse distributions), the resulting signal looks more and more Gaussian. ICA essentially reverses this process to look for decompositions that provide signals that are as non-Gaussian as possible. These are likely the original signals that were added together to create the mixture signal, which has a distribution that is more Gaussian.

The original use of ICA was to decompose natural images to discover the basis sets (Hyvarinen et al., 1999; Olshausen & Field, 1997). These authors discovered that the resulting components behaved very much like the receptive fields in the early visual system, in the sense that the preferred stimulus for an ICA component is very similar to a preferred stimulus for a neuron in the early visual areas. In addition, when the basis functions are used as filters for the image detail through some operation such as cross-correlation, the resulting activations tend to be sparse such that typically only a few basis functions show strong activation and the rest have fairly weak activation. This is similar to neuronal activity in the brain.

Fig. 3 illustrates a visualization of two basis sets derived from different spatial scales (as specified by the image patch sizes). The gray level of each pixel illustrates the weight associated with this basis function, such that darker regions are determined to be more diagnostic by the metric. The individual basis images tend to reflect variations in spatial scale, as well as curvature.

We conducted some initial machine classification procedures with the raw ICA activations and SVM, which attempted to classify expert fixations from regions that were not fixated by experts. This classification procedure used only the ICA activations of the patches. The results were not compelling, with classification accuracy rarely topping 60%

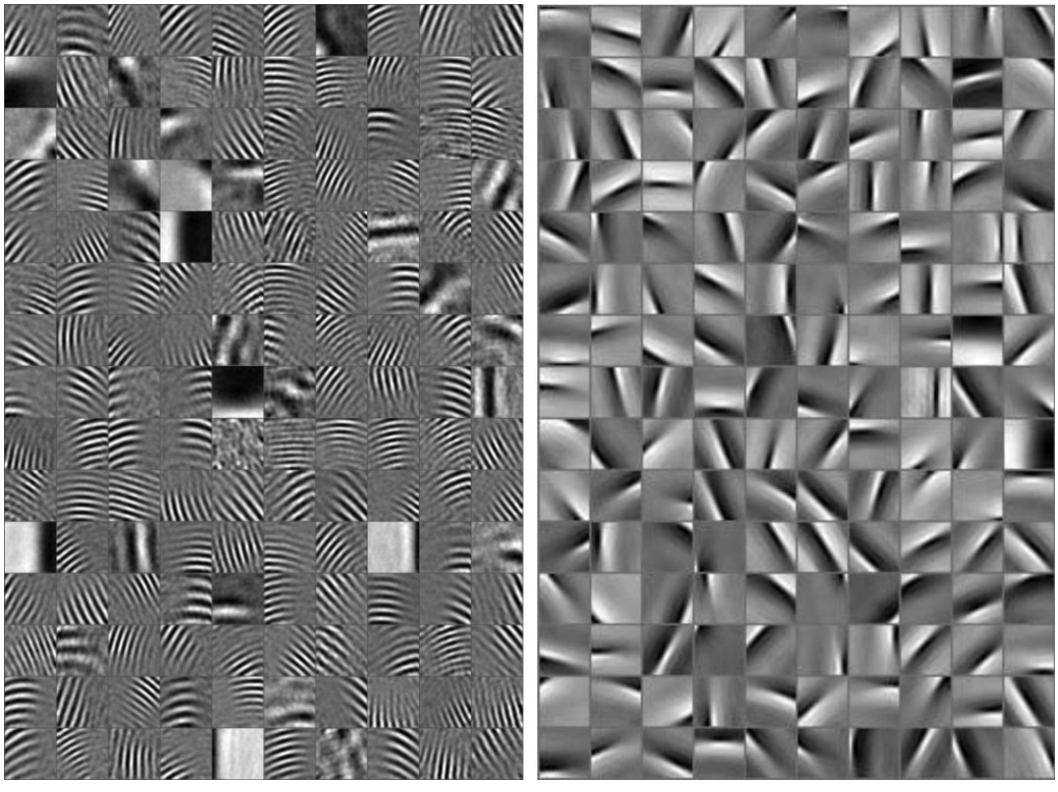

Fig. 3. Independent component analysis (ICA) basis function sets. The grayscale values represent the weights associated with each pixel location for each grid, which also indicate the pattern that produces the strongest activation of a particular basis function. Left panel: basis set constructed using 150 basis functions and 128 × 128 pixel patches. Right panel: basis set constructed using 150 basis functions and 24 × 24 pixel patches. In both cases the size of the basis sets are normalized due to space constraints, but the full basis set retains the dimensionality of the original pixel patches used to construct the set.

where chance is 50%. The linear combination of features may be inappropriate given that the essential elements of the visual system appear to contain important nonlinearities (Riesenhuber & Poggio, 1999). As a result, we explored two extant models that not only perform dimensionality reduction in a principled way, but also provide statistics about feature rarity and diagnosticity. The two techniques have different strengths and limitations, and in some sense the two methods are complementary. Table 1 summarizes the properties, strengths, and weakness of each modeling approach.

The two techniques described in subsequent sections were chosen for their ability to encode the statistics of the reference datasets and the ability to reflect feature rarity. We chose these as alternatives to traditional salience models (e.g., the Graph Based Visual Saliency Model [Harel, Koch, & Perona, 2006] and the Itti and Koch model [Itti & Koch, 2000, 2001; Itti, Koch, & Niebur, 1998]) because the fingerprint impressions contain visual features that are very different than natural scenes. As a result, the hand-tuned

Table 1
Properties, strengths, and weaknesses of each modeling approach

	Raw ICA Activations	AIM Model	CoVar Model
Model summary	Image patches centered on fixations are used to extract basis functions using independent projections in pixel space	The model estimates the likelihood of observing an ICA basis activation of a particular value, which is combined across all ICA basis functions. This produces an estimate of the rarity or diagnosticity of each region, relative to the statistics of the entire database	The CoVar model uses a second layer of weightings to capture the residual covariance between ICA basis activations. The activations of this second layer are very good at distinguishing between textures that vary in appearance. However, it does not directly estimate feature rarity
Feature space representation	Activations of individual basis functions, which place each image patch in a *b* dimensional space where *b* is the number of basis functions	Activations of individual basis functions, which place each image patch in a *b* dimensional space where *b* is the number of basis functions	The initial layer uses 500 basis images, and the activations are then reduced to 25 latent nodes through an additional set of weights
Anticipated utility	Rapid calculations support a variety of applications, but poor classification accuracy limits the utility of this approach	This metric is sensitive to the statistics of a database at different spatial scales, which are then combined using logistic regression. Thus, it highlights the regions most diagnostic for purposes of identification	This model captures some of the properties of the initial stages of the human visual system, including areas V1 (the input layer) and areas V2/V3 (the latent layer). Thus, it is likely to characterize those regions that human examiners consider diagnostic. However, it does not directly represent feature rarity based on the statistics of the database
Method of dimensionality reduction	After the initial ICA decomposition, the logistic regression model classifies the two fixation classes as much as possible	The assumption of independence across basis images allows a computation of feature rarity across all basis images. Logistic regression then further combines across different basis image patch sizes and scales	The initial stage acts very much like an ICA decomposition, but the latent layer further reduces the dimensionality to place each image patch in a 25 dimensional space. Logistic regression then separates fixated from random patches based on the activations in this 25 dimensional space

(continued)

Table 1. (continued)

	Raw ICA Activations	AIM Model	CoVar Model
Performance on training sets	64% classification accuracy for fixated versus random patches on trained impressions	75% classification accuracy for fixated versus random patches on trained impressions	81% classification accuracy for fixated versus random patches on trained impressions
Performance on testing sets	62% classification accuracy for fixated versus random patches on novel impressions	74% classification accuracy for fixated versus random patches on novel impressions	80% classification accuracy for fixated versus random patches on novel impressions
Time to process each image	Seconds	Seconds	Minutes to tens of minutes

salience models may not fit the statistics of latent print impressions. However, as we describe next, the two approaches we have adopted naturally reflect the structure of the reference images.

3.2. Attention via information maximization metric

The goal of the attention via information maximization (AIM) metric described below is to develop a quantitative metric of the information content in fingerprints by extending information theory (Shannon, 1948) as formulated by Bruce and Tsotsos (2009). Information theory holds that rare signals are the most informative in applications such as an identification task. For example, if a witness reports that a suspect has light skin, this is relatively uninformative. However, if the witness reports that a suspect has a distinctive scar running right below one eye, this might be a rare feature and therefore more diagnostic for purposes of identification.

We will use the feature set described previously to estimate the likelihood of observing individual features. This will provide a measure of the diagnosticity of the ridge detail at each location in an impression for purposes of individualization, in the sense that it identifies which regions are most rare given a dataset of images. This measure is independent of human examiners, but it depends on the size of the image patches used to construct the basis images, as well as the number of basis images. This first parameter (pixel patch size) determines the spatial scale and therefore the size of the visual features that tend to be represented in the basis set. The second parameter (basis set size) determines the amount of dimensionality reduction that occurs during the construction of the basis set. This second parameter is important because algorithms such as ICA tend to represent those features that are most common. However, information theory posits that features that are *rare* are most relevant for purposes of identification. Thus, it is important not to compress the features set beyond a representation that could still represent rare features with some degree of fidelity.

The individual diagnosticity maps are a function solely of the statistics of the dataset once we specify a pixel patch size and a basis set size. However, experts are likely using more than one spatial scale, and we do not know a priori which basis set size to use because we do not know how much data reduction is optimal. Therefore, we will use data from experts to help determine the appropriate combinations of pixel patch size and basis set size. This poses an initial problem, because human experts may not attend to the most diagnostic features, nor may they fully understand the basis rates of various features.

We will bootstrap our way out of this problem using data from experts to select an appropriate basis set size and pixel patch size, using the following inference steps. We begin with the observation that experts are highly accurate at this task. For example, the sensitivity (d') of fingerprint experts at actual casework was estimated by the current authors to be 2.6 based on data by Ulery et al. (2011). This same study estimates an error rate of one erroneous identification in every thousand comparisons where an identification is made. Human experts may have some knowledge of the relevant features for latent print examinations and use this to distinguish mated pairs from non-mated pairs. Accuracy in our task was similarly high after only 20 s of viewing time, with sensitivity (d') estimated at around 1.9.

Next, we will demonstrate that some combinations of pixel patch size and basis set size produce parameter-free feature rarity maps that agree somewhat with experts' eye gaze. These feature maps are purely a property of the statistics of the database and the visual features of the friction ridge impressions extracted at locations deemed relevant by experts. Finally, we argue that if experts can distinguish between matching and non-matching prints, and they rely on similar features that the model considers to be diagnostic, the AIM metric is capturing some elements of human expertise.

This is not to say that approximating human performance is the end goal of this analysis, because we may find that there are regions that the model determines are diagnostic that humans do not yet look at. In the end, we will find that the interplay between human expertise and the visualizations provided by the AIM metric will help improve both expert performance and the quantitative metric.

To summarize, this analysis will produce a quantitative representation of the information content in friction ridge impressions that is based primarily on the statistical properties of ridge detail, guided somewhat by the use of expert data to determine the best parameterization of the AIM metric.

3.3. Constructing feature rarity maps

The AIM model provides one method of computing feature diagnosticity (Bruce & Tsotsos, 2009). First, we determine the activation of each basis function at each location in the fingerprint. This is done through a process of convolution, and it essentially determines how much a given patch of fingerprint resembles each basis function. Fingerprint patches that are similar to a given basis function produce high activations at that location. The output of this step is a value at each pixel location in the fingerprint that determines how active a given basis function is at that

location ($a_{i,j,k}$). This is repeated for all basis functions in the basis set to produce a *basis activation vector* at each location that represents the activity of each basis function at that location.

Next, we repeat this process for many fingerprints at all possible locations on these impressions rather than just at the fixated locations. We use a reference dataset that is larger than the set used in the current experiment. The reference dataset contained 120 image pairs, and we sample every 5 pixels across each image pair to determine the empirical probability density function for each basis function (B_k). This allows estimation of the *activation distribution* for each basis function B_k. Most of the time the activation will be close to zero, while occasionally it will be quite high if the basis function is a very close visual match to a particular patch of friction ridge skin. This gives the characteristic super-Gaussian distribution that is expected from the ICA algorithm. These empirical density functions are then used to compute the probability of observing each point in an image by combining across the different basis images as described next.

Third, we use these activation distributions to determine the rarity of a particular region of friction ridge skin. Work by Bruce and Tsotsos (2009) demonstrated with visual search tasks that the self-information of a region could be used to estimate the rarity or diagnosticity of that region as formulated by the AIM metric. The self-information is computed as follows. A given location in the friction ridge impression produces a set of activations across all the basis images (essentially how well each basis function matches the patch of skin at that location). We can determine how likely it is to encounter an activation value for a given basis function by looking at the activation distribution estimated from the entire dataset.

Mathematically, this is the Shannon self-information measure (Shannon, 1948):

$$-\log(p(x))$$

where $p(x)$ is the probability of observing an activation value for that particular basis function. The nature of ICA basis functions is such that they tend to be highly active only rarely. The smaller $p(x)$, the larger (in absolute terms) $-\log(p(x))$ will be.

These $-\log(p(x))$ values are summed up over all the different basis function activations. Common features will produce very high values of $p(x)$ and therefore very low values of $-\log(p(x))$. However, rare features will produce activation values that fall in a range that are almost never encountered, and therefore will have a low $p(x)$ value. It is this relation that links self-information with feature rarity. The self-information *SI* computed at location *i,j* is

$$SI_{i,j} = -\sum_{k=1}^{N}(\log(p(B_k \cong a_{i,j,k}))$$

where N is the number of basis images in the basis set, B_k is the distribution of activation values for basis image k derived from the reference dataset, and $a_{i,j,k}$ is the activation of

basis image k at location i,j. The summation of log probabilities is based on the assumption of independence of the activations of the individual basis functions. We justify this assumption due to the nature of independent component analysis, which has a goal of making the activations as independent as possible. In a model described later, we explicitly take advantage of the residual mutual information not captured by the ICA representation.

The self-information value is interpreted as the probability of observing a similar feature in the reference database. To visualize the self-information computed at each pixel, we overlay the images with masks that are similar to heatmaps but instead tend to fade regions that the salience map judges as less diagnostic. As shown in Fig. 4, visible regions are those that are considered to have high self-information and therefore be diagnostic with respect to feature rarity and individualization. Each of the panels in Fig. 4 is derived from a different parameterization of the underlying ICA basis set and illustrates how different parameterizations tend to highlight different regions as diagnostic. All four maps have 150 basis images, but they vary in the pixel patch size that was used to derive the salience map. Smaller pixel patch sizes tend to highlight ridge endings and bifurcations (upper-left panel of Fig. 4), whereas larger patches focus on collections of minutiae that form regions such as the core or in this case an irregular patch in the upper-right portion of the print (upper-right panel of Fig. 4). The lower panels of Fig. 4 illustrate the salience map derived from larger pixel patch sizes, which highlight the core, delta (the triangle regions that typically fall below the core area), and the region above the core as diagnostic regions.

It is worth noting what this representation does *not* highlight. A basic understanding of the way latent prints are captured illustrates that the outer contour of the impression is typically irrelevant for purposes of identification (De Alcaraz-Fossoul, Roberts, Feixat, Hogrebe, & Badia, 2016). The deposition pressure will vary the amount of skin that comes in contact with the surface, and the orientation of the finger may result in variations in the outer contour. For these reasons, most impressions captured under standardized procedures such as an arrest will typically involve a rolling process to capture as much of the side ridge detail as possible. However, latent prints are typically smaller regions of this skin surface, and the edge detail (i.e., where the visible ridges fade due to lack of pressure) is not relevant for purposes of comparison. Because experts rarely look at the edge, this produces an underweighting of the impression edges in the pixel patches used to construct the ICA basis sets. This may prevent the algorithm from highlighting what otherwise could be a very attractive feature for salience models—the edge of the impression. As illustrated in Fig. 4 and in subsequent visualizations, the AIM metric rarely highlights features at the edge of the impression (the dark bands at the top of the impressions on the lower row of Fig. 4 are an artifact that results from edge effects with the larger pixel patch size used to construct the maps). The metric seems sensitive to the statistics of the information selected by examiners, and in some sense the structure of the identification task is built into the construction of the ICA basis sets.

Fig. 4. Self-information maps for one impression, processed using basis functions of different sizes. Darker regions are those areas that are considered to have high self-information and therefore be diagnostic with respect to feature rarity and individualization. All maps were constructed using 150 ICA basis functions and illustrate that different patch sizes highlight different kinds of information, ranging from fine detail to relatively large regions. Top left: Map with basis set size of 24 × 24 pixels highlighting local features. Top right: Map with basis set size of 64 × 64 pixels. Bottom left: Map with basis set size of 128 × 128 pixels. Bottom right: Map with basis set size of 160 × 160 pixels highlighting broad regions.

3.4. Validating the AIM metric against expert eye gaze data

The visualizations shown in Fig. 4 demonstrate that the AIM metric is sensitive to both individual features as well as regions of higher complexity such as the core and

delta regions. However, it does not highlight the edges of the impression despite the high contrast that such regions exhibit, and therefore represents some element of the task structure. How well do these visualizations correspond to the behavior of examiners?

Fig. 5 illustrates how close the correspondence between the self-information and eye fixations can be. The dark regions are those the model considers most diagnostic, and the red dots are the fixations from experts. The experts tend to cluster their fixations in those regions the self-information metric deems most diagnostic.

The image pair in Fig. 6 shows close correspondence when a smaller ICA basis set is used. A complete model might benefit from combining self-information diagnosticity maps different spatial scales and basis set sizes. We explore this below.

Fig. 5. Strong predictability for a 160 × 160 pixel patch size basis set with 150 basis functions. The red dots correspond to expert eye fixations, and there is close correspondence between the locations of the dots and the regions identified by the metric as diagnostic. These two images are different impressions from the same finger.

Fig. 6. Smaller numbers of basis functions may pick up different types of features. This is a 128 × 128 pixel patch basis set with only 16 basis images. It appears to be more specific in the regions it identifies.

To evaluate the metric, we use the eye gaze data from 30 images and measure the self-information computed at each fixation. We then compare that to the self-information from an equivalent number of image patches that were not fixated by experts. These regions were chosen by randomly sampling from fingerprint impressions by ensuring that no expert fixation fell within 1 degree of visual angle of the candidate random fixation. Of course, we have no control over where experts move their eyes, so we measured the area that corresponded to regions that could have been assigned to a random fixation to make sure this was not an artificially small proportion of the pixels of each impression. We found that across all the images, an average of 54% of the total image area of the fingerprint detail was available for assignment to the random category (95% CI: 52–56%). Given that this is close to 50%, neither of the two categories (fixated or random) dominated the available area in the fingerprints.

Rather than use a test set of fixations taken from the same images that were used for training, we held out a randomly selected 30% of our images (9 images) for testing. This allows us to argue that our classification results generalize to novel images rather than just to novel fixations from the training images.

We expect that the self-information of regions fixated by experts will be higher than those regions not fixated by experts. We are limited, however, by the fact that we must select an ICA basis set first before computing the self-information. As illustrated by Fig. 4, ICA basis functions produced by image patches of different sizes tend to highlight different levels of ridge detail. Because information is available at multiple scales, we do not know which level of detail the examiners are relying on.

Our solution to this problem is to compute the AIM metric using a variety of basis functions and then use a machine learning procedures (logistic regression and brief

explorations with support vector machines) to discover which basis functions contribute the most to the decision to fixate a particular location. The logistic regression uses the salience maps from a variety of basis set sizes as well as pixel patch sizes to construct a set of weights across all the salience maps. The objective function for the classification algorithm is separating the expert from random fixations as much as possible, with hold-out fixations (and eventually hold-out images and subjects) to test the generalization classification performance. Once found, these weights illustrate the importance of each salience map for separating expert from random fixations.

The classification accuracy values for an individual ICA basis set are found in Fig. 7. Individual ICA basis functions give classification accuracies in the range of 58%–70%, whereas combining several sets together gives a classification accuracy of 75% with only very slight loss of generalization to the testing set (74%). This last result illustrates that the logistic regression does not over-fit the training data. Based on this result, the optimal solution combines information across different spatial scales. The error bars in Fig. 7 represent one standard deviation of the distribution of classification accuracies obtained via resampling procedures.

This method also generalizes readily to new participants and new images. We repeated the above classification training but held out not only a new testing set of images but also a new testing set of participants. The lower panel of Fig. 7 illustrates that classification performance on the training set is similar (~75% classification accuracy) and also has similar generalization accuracy (74%). This approach seems insensitive to the varieties of particular latent print examiners and images used to train the classifier, and it could therefore be expected to generalize to individuals and images outside our testing pool.[5]

On the basis of the logistic regression of the different pixel patch sizes and basis set sizes, we can correctly classify approximately 75% of the fixations made by experts, with very little loss of generality to novel prints because the classification of the testing set is 74%. This is much higher than we saw with the raw ICA activations, and we have good correspondence between the fixations and rarity visualizations in Figs. 5 and 6. We looked at the Area Under the Curve (AOC) values from the receiver operating characteristic curve to make sure that the accuracy values were not determined by one type of error, but found essentially the same data pattern as in Fig. 7: The AOC values ranged from 0.6 to 0.75 with individual pixel patch sizes, and a value of 0.84 with all sets combined and a testing value of 0.83.

One way to address the differences between human performance and the predictions of the AIM model is to address systematic deviations between the two. Fig. 8 illustrates four examples where the salience maps appear to disagree with the eye fixations. The blue-circled regions denote areas that were either fixated by experts and not highlighted by the model (upper left and lower right panels of Fig. 8) or highlighted by the model and not fixated by experts (upper right and lower left panels of Fig. 8). The upper left panel illustrates a set of ridges that contain a number of minutiae and therefore might be considered diagnostic. However, the model highlights these regions as more common in the database than the examiners might believe. The upper right panel highlights a feature that received relatively few fixations, possibly because of its distance to the core. The lower left panel

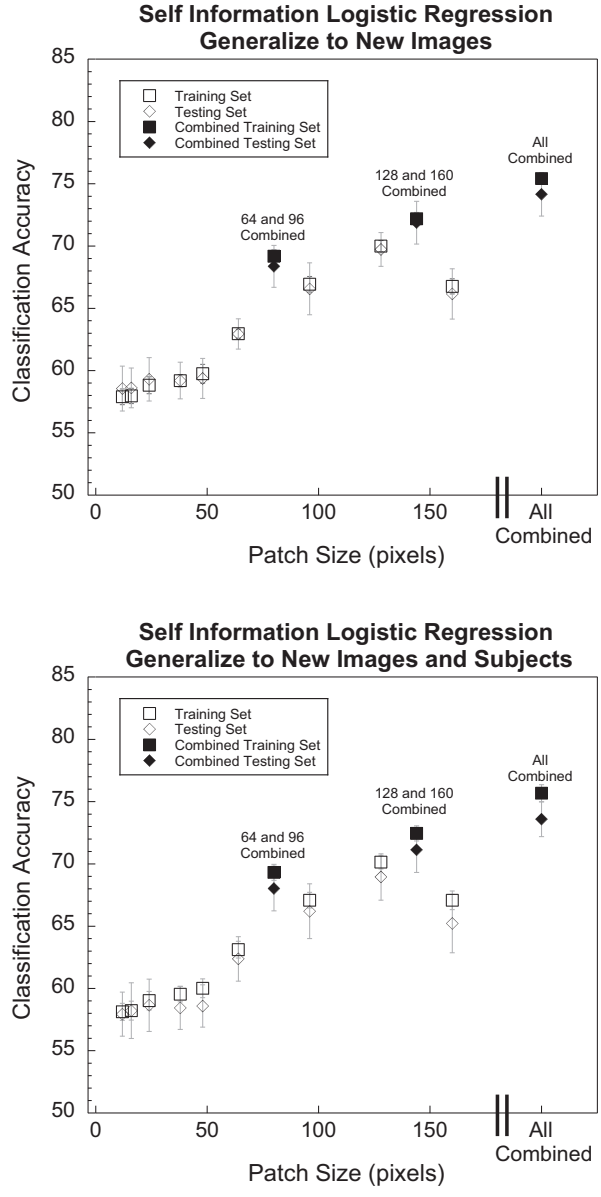

Fig. 7. **Top:** Classification accuracy as a function of pixel patch size for both training and testing sets for the AIM model. Open points are the classification accuracy combined across the different basis set sizes for each patch size. The dark points are the classification accuracy computed by combining across different pixel patch sizes in the logistic regression model. The error bars represent one standard deviation of the classification accuracy distribution across 20 resampling repetitions, which are sometimes smaller than the curve symbols. The best performance (~75% accuracy) is obtained by combining across all pixel patch sizes, with very little loss of generality in the testing set. **Bottom:** As above, but the testing set contains both new images and new subjects. Again, very little loss of classification accuracy is found with the testing set. This illustrates that there is consistency across observers in terms of the regions they choose to fixate.

Fig. 8. Representative examples of poor performance from the AIM model. **Upper Left**: The circle highlights a large number of fixations that the model fails to denote as salient. This may result from the thinning of the ridges, which produces a series of minutiae that examiners deem diagnostic but the model considers relatively common. **Upper Right**: The model highlights a combination of two ridges that separate two ridge endings. This received few fixations from experts, perhaps because of its distance from the core. **Lower Left**: The model highlights the region between the core and the delta likely due to the strong curvature in that area, yet few fixations are directed there, possibly because of the close proximity between the core and delta. **Lower Right**: In contrast, examiners chose to fixate the region between the core and delta, yet the model determined that the region was less diagnostic. The relatively poor image quality may have required more fixations by experts to extract image detail.

illustrates a region between the core and the delta that is highlighted by the model but skipped by examiners, possibly because of the high image quality in the core, delta, and region in between. The lower right panel illustrates that experts will spend more time fixation the region between the core and the delta, but perhaps only if the image quality is poorer in this region.

These deviations between model and gaze data point to a number of improvements to create a more complete model of human eye gaze. Instituting a bias toward the core might improve the model's correspondence with gaze data, as well as accounting for image quality. Human experts may simply need more fixations to extract detail from regions of less quality, and this many not mean that these regions are ultimately considered to be of higher diagnosticity than other regions. Of course, changes to the model depend on the goal of the modeler. If the goal is to account for human performance, a model could include a core bias, but if the goal is to simply identify regions in fingerprints that are the most diagnostic for purposes of comparison, a better approach would be to simply note these discrepancies and encourage examiners to work to overcome the tendency to rely heavily on the core area.

A second way to understand the behavior of the model is to examine the contributions of the various diagnosticity maps to overall classification accuracy for the AIM model. We conducted an analysis of the regression weights in the full model, to identify the contributions of different basis set sizes or pixel patch sizes. The resampling procedures that created the error bars for Fig. 7 also allow us to address the strength of the logistic regression weights for each basis set size and pixel patch size. These weights, when normalized by the standard deviation of the activations produced by each combination of pixel patch size and basis set size, are a measure of the contributions to the overall model. Somewhat surprisingly, none of the pixel patch sizes dominated in this analysis, demonstrating that information at various spatial scales played a role in the final model. In addition, the smaller basis set sizes tended to have higher weights, indicating more contribution to the overall model behavior. Smaller basis sets produce more compression to represent the variability contained in the images. As a result, small basis set sizes tend to filter out idiosyncratic features, which in our case could be visual noise that is not relevant to the comparison task. Larger basis sets will tend to include basis sets that represent this noise. This noise will be ignored by examiners, which may explain why these larger basis sets contribute less to the model when asked to predict human eye gaze.

A limiting factor on accuracy is the inherent noise in eye gaze data (some fixations are more meaningful than others, and the classifier is forced to classify all fixations with equal weight), as well as any error in the estimates of the eye gaze from the eyetracker. However, the AIM metric also assumes independence between the basis function activations. To the degree that this assumption is incorrect, the metric does not take into account the conditional probabilities of observing the activation of one basis function given the activation of a second basis function. These joint activations may miss important elements of the feature space and in an extension below we explore a model that explicitly models the covariance values between basis activations.

3.5. Validation study in a comparison task

The measures of feature diagnosticity provided by the AIM model provide a strong prediction: Regions of fingerprints identified by the model as relatively rare (and therefore highly diagnostic for purposes of comparison) should actually produce greater accuracy in a visual comparison task that is similar to an actual latent print examination. We conducted a behavioral study using 12 latent print examiners who had not previously participated in the eye tracking data collection. We also used a new set of 35 fingerprint pairs that were not previously used to train the AIM model. The images were processed by the AIM model to determine the diagnosticity value at every location in the images. We then collected all the diagnosticity values from pixels that contained ridge detail and divided these values into low and high diagnostic regions by a median split on the AIM diagnosticity values for each image. These were then used to select only those regions that ranked as the best or worst regions for each impression. Fig. 9 illustrates four representative images from the set.

These images were then paired with either different impressions from the same finger or impressions from a different but similar finger (usually by taking the left-right reversal

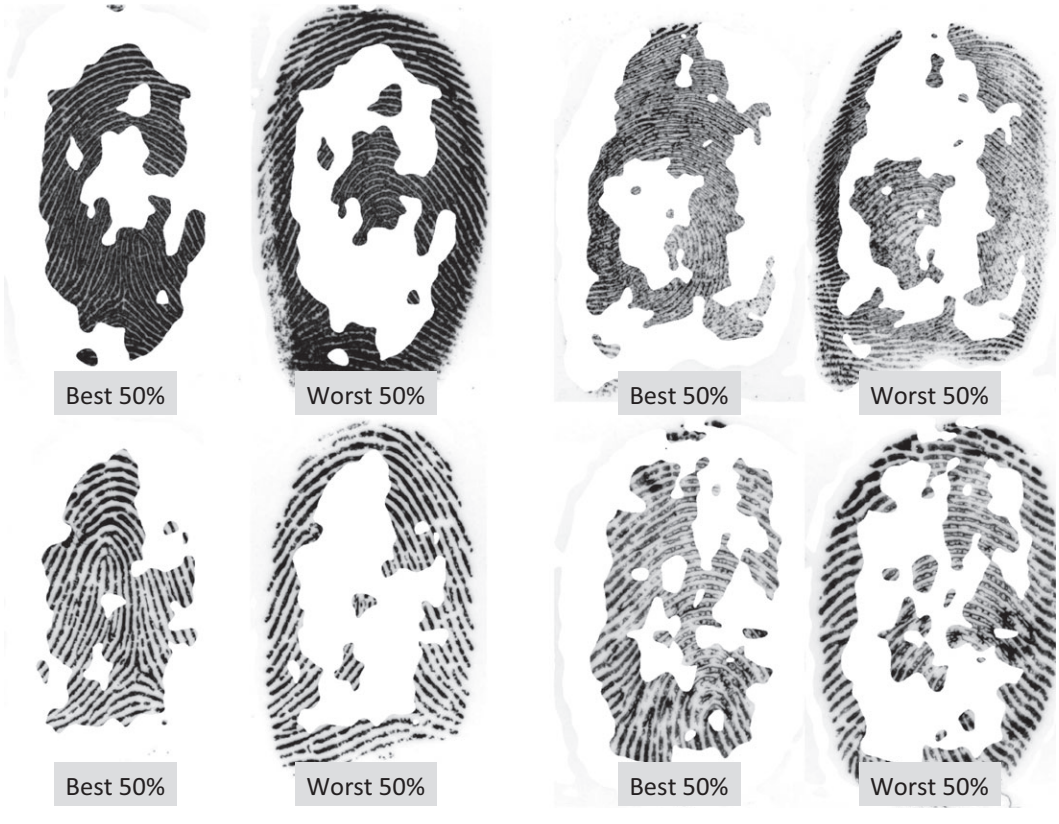

Fig. 9. Example images with regions selected by diagnosticity used in the validation study.

of the same finger on the other hand). Experts conducted a speeded comparison task using the best regions, the worst regions, or the entire impression. The goal of the study was to determine whether accuracy for trials in which the best regions are presented is higher than those trials in which the worst regions are presented. We also included trials in which all information was presented, as a comparison to determine whether the best regions might be similar in performance to the entire image.

3.5.1. Method

3.5.1.1. Stimuli: The diagnosticity of each pixel in 35 new fingerprint impressions was determined from the AIM model by computing individual saliency maps from each combination of basis set size and pixel patch size. These were then weighted using the beta weights from the logistic regression model that combines the saliency maps to produce the best classification performance in Fig. 7 (All Combined). This produces a single measure for each pixel on each of the new images that determines the diagnosticity for that pixel. These diagnosticity maps were then smoothed using a Gaussian low-pass filter with an extent of 64 pixels and a standard deviation of 10 pixels. This avoids very small patches that are smaller than a ridge unit, because the typical distance between ridges in our images is approximately 30 pixels.

The resulting diagnosticity values were collected for only those pixels determined by a human coder to represent actual fingerprint regions, and the median was calculated. New images were created such that only those regions that contained either high or low diagnosticity values were displayed (see Fig. 9). Finally, to encourage examiners to treat the impressions as similar to actual latent prints, we combined the images with noise sampled from an actual latent print. This combination was done using a multiplicative process that mimics the subtractive nature of physical noise (Busey, Swofford, Vanderkolk, & Emerick, 2015). These images were then paired with either a different impression from the same finger or a similar non-mated pair. Fig. 10 illustrates an example stimulus. There were 20 mated pair trials and 10 non-mated pair trials.

3.5.1.2. Participants: The participants were 12 active latent print examiners (8 female) with a mean age of 37.25 (range: 27–59). They had an average of 8.75 years of experience (range: 2–31) conducting unsupervised latent print comparisons in casework. They were recruited from state and federal laboratories in the United States.

3.5.1.3. Procedure: The experiment was conducted online using JavaScript to present the images and PHP/MySQL to save the data. Despite the addition of noise, the images contain a large amount of detail, and prior studies with similar images and unconstrained viewing produced very high accuracy values (Busey et al., 2011, 2013, 2015). Because the goal was to compare among the three types of trials (best regions, worst regions or all regions), we need to have accuracy below ceiling. To limit performance, we chose to present the images for a fixed 15s interval. In prior work we found that 20 s presentations produced relatively high accuracy (Busey et al., 2011), and thus we settled on 15 s as a value that was likely to avoid ceiling effects. Although an actual latent print comparison

Fig. 10. Example trial from the validation study that shows the most diagnostic features as revealed by the AIM model. These are embedded in noise sampled from near an actual latent print to create the simulated latent print on the left. The comparison print is shown on the right.

may take longer than 15 s, the initial search process typically involves comparing a latent print against a long list of images retrieved from a computer database. During this search, each visual comparison can actually take just a few seconds to complete before a print is selected for comparison or rejected, and thus examiners are used to dealing with time-limited comparison circumstances.

The images were randomly assigned to one of the three conditions (Best, Worst, and Full). To improve the experimental power, each image pair was repeated twice, each in a different condition, for a total of 70 experimental trials. Examiners were cautioned that some of the prints would look similar and that they should make a decision based solely on the information presented on each trial. During debriefing, none of the participants commented on the similarity between the images across trials.

Latent print comparisons typically use the response scale of Exclusion, Inconclusive, or Identification. However, examiners might be reluctant to use the Identification response after only a 15 s presentation. Instead, we gave the examiners six response categories: Exclusion, Almost Exclusion, Tending Exclusion, Tending Identification, Almost Identification, and Identification. The simulated latent fingerprint appeared for only 15 s, but the participant could respond at any time before or after the impression disappeared. We intentionally left off the pure neutral category ("Inconclusive") because we did not want examiners to simply give an inconclusive response without giving an indication of which way they were leaning.

3.5.1.4. Results: A single summary score, A' (Pollack & Norman, 1964), was computed for each participant in each of the three conditions using the responses to both the mated

and non-mated pairs. This combines the information from all six response categories into one measure of accuracy without making assumptions about the shape of the underlying distributions. Table 2 illustrates the A′ values for each of the 12 participants for the three conditions.

Because A′ is not linear (it has an upper bound at 1), we used the non-parametric Mann–Whitney test to compare the three conditions. We found that the Worst condition (median = 0.875) produced poorer performance than the Best condition (median = 0.965) (U = 30.5; Critical $U_{\alpha=0.05}$ = 37; $p < .05$; Z = 2.367). However, the Best condition was not reliably different from the Full condition (median = 1.0) (U = 43.5; Critical $U_{\alpha=0.05}$ = 37; $p > .05$; Z = 1.617).

These results demonstrate that experts perform better when given the regions identified by the AIM model as diagnostic for purposes of comparison, and perform at levels that are similar to having the full image. Because these results are collected from entirely new images and participants, this illustrates the generalizability of the AIM metric to identify regions in novel prints that are diagnostic for purposes of comparison. Importantly, this validation experiment breaks the circularity of using elements of the human data to tune the combinations of the saliency maps derived from different spatial scales and basis set sizes. Of course, this does not imply that either humans or the model are perfect, and in the next section we explore a slightly different approach that provides better classification performance for human data, while giving up the ability to make diagnosticity statements about image detail.

3.6 Conclusions from the AIM metric

The AIM metric proposed by Bruce and Tsotsos (2009) uses the ICA decomposition to compute an activation value for each basis function. It then computes the probability of

Table 2
A′ values for all 12 participants in the validation study for the three conditions (Worst, Best, and Full regions)

Worst	Best	Full
0.75	0.87	1.00
0.74	0.90	1.00
0.95	1.00	0.99
0.97	0.95	0.97
0.87	0.94	0.99
0.88	0.98	1.00
0.98	0.85	1.00
0.91	1.00	0.97
0.83	1.00	1.00
0.51	1.00	1.00
0.83	0.92	0.96
0.96	0.99	1.00
0.875	**0.965**	**1.00**

Note. The bottom row in bold represents the median scores of each column.

observing that particular value across all images to determine an overall likelihood of observing a similar patch in the entire dataset. The individual feature maps constructed from one basis set are entirely parameter free, and the maps illustrated in Figs. 4–6 are a function only of the ICA decomposition and the statistics of the reference dataset. When these maps are combined using logistic regression, this produces reasonably high classification of expert fixations in both the training and testing sets, and produces similar generalization performance to new images and observers. These logistic regression models do have free parameters (the weights on the different saliency maps), but they serve mainly to identify which saliency maps contribute most to expert eye gaze data. We argue that the lack of overfitting to the new observers and images, coupled with the results of the validation study, justify the additional free parameters.

Note that the ICA decompositions are a function of the regions visited by experts and therefore may be functions of expertise. However, the statistics of the images are constructed by sampling the entire image and are therefore not a function of expert looking behavior. It is possible that experts are not looking at regions that are most diagnostic, and some interplay between experts and the self-information visualizations might be required to bootstrap performance of both experts and the model. We return to this point in the general discussion.

3.7. Modeling covariance among activations: The CoVar model

The ICA weights used with the AIM metric produced reasonably high classification accuracy for our data, especially if the different pixel patch sizes were combined together to reflect the fact that examiners likely use information at different spatial scales. Central to the AIM metric is the idea that individual basis functions are independent. This allows for the multiplication of probabilities that underlies the self-information computation and gives a statistical measure of feature rarity. This assumption is justified by the fact that the independent component analysis algorithm is *designed* to find components that have independent activations across the entire dataset.

One limitation of this approach is that although basis functions are independent across the entire dataset, they are likely not independent for smaller regions, because similar features that are found close together on an image may strongly activate several different basis functions. This will produce localized covariance in the activations, and there may be important correlations in the basis function activations within similar regions of an image that the AIM model may be missing.

This fact was recognized by Karklin and Lewicki (2009) during an investigation of natural scene statistics. They point out that the earliest stages of the visual system act much like the ICA basis decomposition, with individual neurons sensitive to different orientations and spatial frequency patches at particular locations. However, the visual system must be able to achieve a measure of positional invariance by pooling across similar detectors positioned at slightly different locations. This approach combines the outputs of different detectors that are all tuned to similar features (i.e., same orientation and spatial frequency) but at slightly different spatial positions. This can be done by a second level

of artificial neurons that learns a set of weights on the ICA basis functions to process the *covariance* between the individual basis function activations.

Karklin and Lewicki (2009) proposed a model, called the CoVar model, that could not only discover a set of weights that learned the correlations among the outputs of the basis functions, but also achieved positional and contrast invariance while remaining very sensitive to orientation. This network could have a large number of basis functions (around 500) but have relatively few high-level latent neurons (around 25) that could represent more abstract features by selectively weighting the activations from the input layer. Essentially, the weights on the connections between the input layer (the basis functions) and the latent layer (the high-level neurons) are learned in such a way that they represent a subspace decomposition of the high-level ICA activations that meaningfully represents the natural scene statistics of images.

This particular model seems well suited to fingerprints, because the model naturally represents variations in texture appearance, orientation, and spacing. This model is also a natural progression of the previous ICA approach, because the first layer of the CoVar model is very similar to the ICA decomposition. However, rather than assuming independence between the ICA activations as the AIM metric does, the model explicitly represents these correlations in the second layer of the model. This architecture shares computational principles with the early stages of the visual system, and therefore may be a good candidate as a model to predict what parts of the image are considered diagnostic by experts. However, it does not provide the feature rarity statistics of the AIM metric, so in some ways the CoVar and AIM metrics are complementary and could be used in different settings for different purposes. The next section describes this model and evaluates the accuracy of the model when predicting where new experts will move their eyes on novel fingerprint impressions.

3.7.1. CoVar model training

We trained the CoVar model using similar procedures as the ICA decomposition. We cropped out regions of clear fingerprints near where experts fixated and used these image patches to train the model. The CoVar model is computationally very expensive, and to make the training tractable, we limited the patches to 24 × 24 pixels (about 3 ridge widths). However, in recognition of the observation that spatial scale is an important element of the detail in friction ridges, we used integer multiples of this patch size when extracting patches from our fingerprints, such that we used actual crops ranging from 24 × 24 pixels (Scale 1) to 240 × 240 pixels (Scale 10). As with the ICA/AIM analysis, different spatial scales will likely represent different sources of information. This may range from level 3 detail such as idiosyncratic ridge element shapes, to level 1 detail such as pattern type. The model is very computationally intensive, and so we downsampled all pixel patches to the 24 × 24 pixel patch size, which effectively low-pass filters the larger spatial scales. Thus, all models have 576 input units (24 × 24 pixels) and are parsed by 500 basis functions and 25 latent units that represent the discovered covariances. The model produces a set of 25 activation values for each location it is evaluated at, and

when scales larger than 1 are used, the image content is reduced in size by the appropriate factor to match the scale of the discovered basis functions.

To evaluate the success of the model, we relied on similar procedures as described for the ICA/AIM metric. We computed the activation of the 25 latent neurons at each fixation from the experts. This places each fixation as a point in a 25 dimensional space. We then chose random points on the print that were not near expert fixations and computed the activations of the 25 latent neurons for each random fixation. We then submitted both sets of activations to a logistic regression classifier and a support vector machine, using a held-back set of test fixations from a separate set of images. Note that while it would be possible to apply the AIM metric to the output of the 25 latent neurons, the assumption of independence is likely not met with these activation values and therefore we did not pursue such an option.

3.7.2. Classification results

We conducted logistic regression classification of the CoVar activations using the same expert fixation/random fixation classification procedures used with the AIM metric. Different spatial scales produce different classification accuracy, and these are summarized in Fig. 11. Classification performance improves with larger scales; smaller scales (1–6) produce performance in the 60%–75% range, and larger scales (7–10) produce classification values near 80%. In addition, generalization to new testing images is also in a similar range, indicating that the model does not suffer from over-fitting. The addition of new subjects to the testing set also does not produce a large drop in classification generalization (lower panel of Fig. 11). The latent layer of the CoVar model may capture important covariance information that reflects how humans perceive texture patterns.

3.7.3. Saliency visualization

The logistic regression analysis provides an opportunity to visualize those regions that experts consider to be diagnostic, even for prints that they have never seen before. To visualize the saliency of different regions for purposes of identification, we first compute the activations of the CoVar model at each point in the fingerprint image. This is somewhat computationally expensive, so we instead compute the activations in a grid of every 10th pixel and interpolate between the values. Explorations at finer scales revealed equivalent results at a cost of much more processing time.

Once we have these activations for a particular model and spatial scale, we can then multiply the activations by the weights from the logistic regression and sum up the weighted activations. This weighted sum computes the degree to which an examiner would consider that location particularly diagnostic (at least enough to warrant a fixation).

Representative images are shown in Figs. 12–14. Darker regions in the print are those deemed by the classifier to be more likely to be visited by human experts. The close correspondence between the distribution of fixations (red dots) and the darker regions illustrates how accurate the classifier can be when fitting human eye gaze data. The images

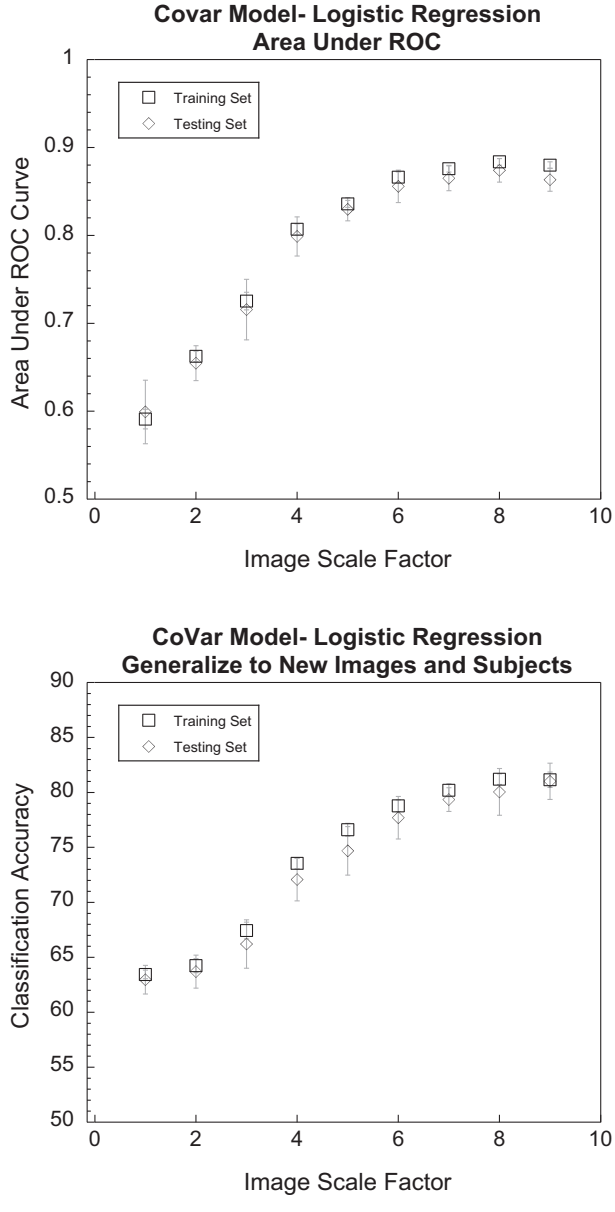

Fig. 11. **Top:** Classification accuracy as a function of spatial scale factor from the CoVar model for both training and testing sets. Classification accuracy improves as the spatial scale increases. Examiners may rely on configural information in addition to simple features. The training and testing classification performance is similar, demonstrating that the model readily generalizes to novel prints that were not used to train the system. Error bars represent one standard deviation of the sampling distribution from 20 resampling repetitions, and in some cases are smaller than the curve symbol. **Bottom:** As above, but the testing set consists of both new images and new subjects. Again we see high classification performance with very minimal loss of generality to the testing set.

Fig. 12. Saliency map showing good correspondence at spatial scale 7 between the fixations and the regions deemed most likely to be associated with expert eye gaze by the logistic model (darker regions). Darker regions are those regions that the logistic model predicts experts will find most diagnostic, and the red dots are fixations from experts. The close correspondence between the two demonstrates that the model accurately captures those features that attract the eye gaze of experts. These are training images.

Fig. 13. Left panel—spatial scale 2. Right panel—same image at spatial scale 6. Different spatial scales represent different information, and both seem to be necessary to capture the distribution of fixations for this print. These are training images.

shown in Fig. 12 illustrate how reliable the classifier predictions can be across different exemplars of the same image.

As with the AIM metric, sometimes multiple spatial scales may be necessary to represent all the information that examiners attend to. An example of this with the CoVar

Fig. 14. Example saliency map for a pair of testing images (i.e., one that was not used to train the logistic regression classifier). Darker regions are those regions that the logistic model predicts experts will find most diagnostic, and the red dots are fixations from experts. The close correspondence between the two demonstrates that the model is able to readily generalize to new fingerprints and predict which regions examiners will visit when they examine the prints.

model is shown in Fig. 13. The left panel is with a small scale, whereas the right side is with a large scale. Both capture fixations, but we may need a combined representation to fully account for human expert performance. We discuss this in a future section.

The logistic regression classifier's performance on the training data is impressive, but to be generally useful, it must generalize to new images. During the training portion of the logistic regression classifier we held back some of our images. This is the best test of generalization: Can the classifier accurately predict where experts will send their eye gaze on images that were not used for training?

The images shown in Fig. 14 illustrate saliency predictions for the model on novel images. These demonstrate that the model can readily generalize to novel images and accurately reflect the fixations for human experts.

Together the classification performance and the saliency maps in Figs. 12–14 illustrate that the CoVar model accurately reflects much of the perceptual mechanisms in human experts that drive their visual performance. Tools based on this algorithm can provide valuable information about human expertise to trainees and even highlight regions that one expert may not have seen but the model identifies as something that experts might look for.

3.7.4. Summary of the CoVar model

The CoVar model achieves higher classification and generalization performance than the AIM model. Its latent modeling of the residual covariances in the ICA activations may be an important element of expertise. It achieves higher classification accuracy and generalization performance than the AIM model at the cost of computational processing time and a loss of the connection to information theory as discussed below.

4. General conclusions

The AIM and CoVar models both represent metrics that reflect elements of human expertise as well as the statistics of the reference databases. The AIM model provides an estimate of feature diagnosticity at a particular scale, and the data from experts helps weight these different scales to produce a metric of the utility of a given patch of an impression. This helps address the close non-match problem, and also characterizes underlying base rates of particular features, which could be more common than the examiner may believe. The CoVar model represents elements of human expertise that involve the covariance of the initial feature detectors, and therefore it provides good estimates of expert looking behavior even on novel prints. On the basis of this, we argue that the CoVar metric embodies elements of human expertise into a computational model, which can be useful for training and automation purposes as described below. However, it does not explicitly rely on the tenants of probability theory and while the weights do represent the statistics of the ridge detail in the training images, it cannot provide a direct measure of feature rarity that the AIM metric computes.

The AIM model has its own advantages. It is computationally much faster, taking seconds per image versus up to 20 min per image to analyze via the CoVar model at even a relatively coarse 10×10 pixel grid. The AIM model might easily take advantage of huge datasets with many hundreds of millions of prints. This will help determine the probabilities of very rare features.

Both models characterize the statistics of the reference databases, but also require human gaze data to determine an appropriate parameterization that defines the feature set. Does this use of human data to tune the models, the results of which are then compared back to human data, represent a form of circular reasoning? Our answer to this challenge is to refer to the very strong generalization performance of both models to both new images and to new observers, and the results of the validation study that demonstrate that humans perform comparison tasks more accurately when using regions that the AIM model designates as diagnostic. However, it remains possible that there is even stronger evidence that the models could discover with the appropriate parameterizations that experts currently do not rely on. We view this as an opportunity to explore the predictions of the models under different parameterizations and work with examiners to highlight new features that they may not have previously considered. For example, the amount of curvature in particular regions, or transitions from convex to concave along a ridge, are not typically described as features by examiners or the extended feature set, but they could be identified by the models as useful for individualizing prints.

The current approach sidesteps the definition of feature much in the same way that the initial stages of the visual system does: by having initial detectors that are sensitive to only the most basic dimensions (edges, orientation, curvature, and phase) and then representing more complex patterns using combinations of the activations produced by these detectors. The continuous nature of the activations produced by these detectors produces a smooth transition between one pattern and another as the weights are adjusted. Despite

this, the models can still assign a diagnosticity value to any particular region or patch given its appearance and the statistics of the reference dataset. This is in contrast to how humans might represent features, where they serve as an important form of communication and compression by humans. For example, the minutiae label clearly denotes some form of ridge ending or the coming together of two ridges despite wide variations in appearance across exemplars of minutiae. However, the current models differ from linguistic labels, as well as extant models of fingerprint comparison, in that they require no preprocessing of the image into individual features such as minutiae. Instead, every location is a feature, and each is evaluated as more or less diagnostic by the model.

This work provides opportunities for automated systems, since neither model requires hand coding of images beyond simply cropping and rotation. This is in contrast to current automated fingerprint information systems (AFIS) that are often manually marked with some assistance from automated minutiae detection algorithms. Either the CoVar or AIM model could be built into AFIS systems to highlight which regions of the searched impression carry more diagnosticity.

This approach can be used to improve performance of latent print examiners in several ways. Knowledge about which regions are useful to experts can be used to train new examiners and to improve performance for current examiners. Trainees could be directed to the most useful regions of prints or a possible route through the print that highlights the most diagnostic regions first. An initial parse of an impression could be used to make a tentative determination about whether prints are of value or not, based on whether there is a sufficient amount of diagnostic information available. Each of these tools would need to be developed and validated to demonstrate that they improve accuracy without affecting the thresholds of examiners.

Both models also have the potential to contribute to the matching portion of the task. Our current modeling addresses the diagnosticity of individual images relative to a reference dataset. However, the models could be extended to look at a pair of images to determine a likelihood that they come from the same source, which is of course the entire purpose of the latent print examination. These forms of quantification of the strength of evidence was noted by both the NRC (National Research Council of the National Academies of Science, 2009) and the National Institute for Standards and Technology as a priority for the field, and we are currently pursuing such extensions.

Notes

1. This language is recommended by the Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST), and it represents the current consensus body for fingerprint examinations. However, the work of this group has been subsumed by Organization of Scientific Area Committees (OSAC) Friction Ridge Subcommittee (FRS), which may recommend different language.
2. One note on the concept of rarity: With infinite precision of measurement, virtually all objects are equivalently rare to the point of being unique. Even the same finger is

changing over time through skin cells sloughing off, and so the concept of "sameness" requires some thought in a forensic context. However, any form of description below perfect measurement will necessarily lose information, thus creating apparent similarities between impressions created by different objects depending on what information is lost in the deposition, development, and measuring processes. Our approach creates a feature description that in the process also dimensionalizes the feature space, thus creating the opportunity to calculate a rarity measure that is the product of the probabilities of observing particular values along each dimension. Our use of the term "rarity" should be interpreted in a probabilistic sense, as the probability of observing a particular combination of feature detector outputs.

3. The closed and proprietary nature of existing search algorithms has raised questions about the ability of a defendant to face their accuser under situations where the initial accusation comes from a database search (Wexler, 2015).
4. ExpertEyes may be found at https://code.google.com/p/experteyes/
5. We explored alternative classification metrics such as support vector machines (SVMs) and found that while we achieved slightly higher classification on the training set (up to 83% on the training set with all pixel patches combined) the generalization performance dropped to 74% on the testing set, which is similar to the logistic regression. This is perhaps not surprising since we observed the same over-fitting in our previous approach when only using ICA basis functions, without self-information, as a representation.

References

Ashbaugh, D. R. (1999). *Quantitative-qualitative friction ridge analysis: An introduction to basic and advanced ridgeology*. Boca Raton, FL: CRC Press.

Barhillel, M. (1980). The base-rate fallacy in probability judgments. *Acta Psychologica*, *44*(3), 211–233. doi:10.1016/0001-6918(80)90046-3

Bouget, J. Y. (2008). The MATLAB camera calibration toolkit. Available at http://www.vision.caltech.edu/bouguetj/calib_doc/. Accessed October 12, 2015.

Bromage-Griffiths, A. (2011). Investigation of the reproducibility of third-level characteristics. *Journal of Forensic Identification*, *61*(2), 171.

Bruce, N. D. B., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, *9* (3), 1–24. doi:10.1167/9.3.5

Busey, T., Silapiruti, A., & Vanderkolk, J. (2014). The relation between sensitivity, similar non-matches and database size in fingerprint database searches. *Law, Probability and Risk.*, *13*(2), 151–168. doi:10.1093/lpr/mgu002

Busey, T., Swofford, H. J., Vanderkolk, J., & Emerick, B. (2015). The impact of fatigue on latent print examinations as revealed by behavioral and eye gaze testing. *Forensic Science International*, *251*, 202–208. doi:10.1016/j.forsciint.2015.03.028

Busey, T., Yu, C., Wyatte, D., & Vanderkolk, J. (2013). Temporal sequences quantify the contributions of individual fixations in complex perceptual matching tasks. *Cogn Sci*, *37*(4), 731–756. doi:10.1111/cogs.12029

Busey, T., Yu, C., Wyatte, D., Vanderkolk, J. R., Parada, F. J., & Akavipat, R. (2011). Consistency and variability among latent print examiners as revealed by eye tracking methodologies. *Journal of Forensic Identification*, *61*(1), 60–91.

Champod, C., & Margot, P. A. (1995). Computer assisted analysis of minutiae occurrences on fingerprints. Paper presented at the International Symposium on Fingerprint Detection and Identification, Ne'urim, Israel.

Champod, C., & Margot, P. (1996). Analysis of minutiae occurrences on fingerprints—the search for non-combined minutiae. Paper presented at the Meeting of the International Association of Forensic Sciences (IAFS), Tokyo, Japan.

Cole, S. A. (2005). More than zero: Accounting for error in latent fingerprint identification. *Journal of Criminal Law & Criminology*, *95*(3), 985–1078.

Cole, S. A. (2009). Forensics without uniqueness, conclusions without individualization: The new epistemology of forensic identification. *Law, Probability and Risk*, *8*(3), 233–255.

De Alcaraz-Fossoul, J., Roberts, K. A., Feixat, C. B., Hogrebe, G. G., & Badia, M. G. (2016). Fingermark ridge drift. *Forensic Science International*, *258*, 26–31.

Ding, P. L., Mei, J. F., Zhang, L. M., & Kang, X. L. (2001). Research of automatic face recognition based on ICA. *Journal of Infrared and Millimeter Waves*, *20*(5), 361–364.

Dror, I. E., Champod, C., Langenburg, G., Charlton, D., Hunt, H., & Rosenthal, R. (2011). Cognitive issues in fingerprint analysis: Inter-and intra-expert consistency and the effect of a "target" comparison. *Forensic Science International*, *208*(1-3), 10–17.

Dror, I. E., & Mnookin, J. L. (2010). The use of technology in human expert domains: Challenges and risks arising from the use of automated fingerprint identification systems in forensic science. *Law, Probability and Risk*, *9*(1), 47.

Egli, N. M., Champod, C., & Margot, P. (2007). Evidence evaluation in fingerprint comparison and automated fingerprint identification systems: Modelling within finger variability. *Forensic Science International*, *167* (2–3), 189–195. doi:10.1016/J.Forsciint.2006.06.054

Expert Working Group on Human Factors in Latent Print Analysis. (2012). Latent print examination and human factors: Improving the practice through a systems approach: The report of the Expert Working Group on Human Factors in Latent Print Analysis. Washington, DC: NIST NIJ, National Institute of Justice.

Fang, G., Srihari, S. N., Srinivasan, H., & Phatak, P. (2007). Use of ridge points in partial fingerprint matching. Paper presented at the Defense and Security Symposium.

Harel, J., Koch, C., & Perona, P. (2006). Graph-based visual saliency. Paper presented at the Advances in neural inzrformation processing systems.

Hyvarinen, A., Cristescu, R., & Oja, E. (1999). A fast algorithm for estimating overcomplete ICA bases for image windows. Paper presented at the Neural Networks, 1999. IJCNN'99. International Joint Conference.

Hyvarinen, A., & Hoyer, P. (2000). Emergence of phase- and shift-invariant features by decomposition of natural images into independent feature subspaces. *Neural Computation*, *12*(7), 1705–1720.

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10–12), 1489–1506. doi:10.1016/S0042-6989(99)00163-7

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203. doi:10.1038/35058500

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259.

Karklin, Y., & Lewicki, M. S. (2009). Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, *457*(7225), 83–U85. doi:10.1038/Nature07481

Kaye, D. H. (2003). Questioning a courtroom proof of the uniqueness of fingerprints. *International Statistical Review*, *71*(3), 521–533. doi:10.1111/j.1751-5823.2003.tb00209.x

Kim, J., Choi, J., & Yi, J. (2004). Face recognition based on locally salient ICA information. *Biometric Authentication, Proceedings*, *3087*, 1–9.

Kim, J., Choi, J., Yi, J., & Turk, M. (2005). Effective representation using ICA for face recognition robust to local distortion and partial occlusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(12), 1977–1981.

Macmillan, N. A., & Creelman, C. D. (2004). *Detection theory: A user's guide*. Mahwah, NJ: Psychology Press.

Mathworks, I. (2012). *MATLAB*. Natick, MA: Mathworks Inc.

National Research Council of the National Academies of Science (2009). *Strengthening forensic science in the united states: A path forward*. Washington DC: National Academies of Science.

Neumann, C., Champod, C., Puch-Solis, R., Egli, N., Anthonioz, A., & Bromage-Griffiths, A. (2007). Computation of likelihood ratios in fingerprint identification for configurations of any number of minutiae. *Journal of Forensic Sciences*, *52*(1), 54–64. doi:10.1111/J.1556-4029.2006.00327.X

Neumann, C., Champod, C., Puch-Solis, R., Egli, N., Anthonioz, A., Meuwly, D., & Bromage-Griffiths, A. (2006). Computation of likelihood ratios in fingerprint identification for configurations of three minutiae. *Journal of Forensic Sciences*, *51*(6), 1255–1266. doi:10.1111/j.1556-4029.2006.00266.x

Neumann, C., Champod, C., Yoo, M., Genessay, T., & Langenburg, G. (2015). Quantifying the weight of fingerprint evidence through the spatial relationship, directions and types of minutiae observed on fingermarks. *Forensic Science International*, *248*, 154–171. doi:10.1016/j.forsciint.2015.01.007

Neumann, C., Evett, I. W., & Skerrett, J. (2012). Quantifying the weight of evidence from a forensic fingerprint comparison: A new paradigm. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *175*(2), 371–415. doi:10.1111/j.1467-985X.2011.01027.x

NIST. (2015). NIST Latent Fingerprint Homepage. Available at http://fingerprint.nist.gov/latent/. Accessed April 5, 2016.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*(6583), 607–609. doi:10.1038/381607a0

Olshausen, B. A., & Field, D. J. (1997). Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision research*, *37*(23), 3311–3325.

Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current opinion in neurobiology*, *14*(4), 481–487. doi:10.1016/j.conb.2004.07.007

Parada, F. J., Wyatte, D., Yu, C., Akavipat, R., Emerick, B., & Busey, T. (2015). ExpertEyes: Open-source, high-definition eyetracking. *Behavioral Research*, *47*(1), 73–84.

Pollack, I., & Norman, D. A. (1964). A non-parametric analysis of recognition experiments. *Psychonomic Science*, *1*(5), 125–126.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision*, *40*(1), 49–71.

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, *2*(11), 1019–1025. doi:10.1038/14819

Sen, T., & Megaw, T. (1984). The effects of task variables and prolonged performance on saccadic eye movement parameters. In A. G. Gale, & F. Johnson (Eds.), *Theoretical and applied aspects of eye movement research* (pp. 103–111). Amsterdam: Elsevier.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal*, *27*(379–423), 623–656.

Snodgrass, M., Bernat, E., & Shevrin, H. (2004). Unconscious perception at the objective detection threshold exists. *Perception & Psychophysics*, *66*(5), 888–895.

Srihari, S. N., Srinivasan, H., & Fang, G. (2008). Discriminability of fingerprints of twins. *Journal of Forensic Identification*, *58*(1), 109–127.

Srihari, S., & Su, C. (2008). Computational methods for determining individuality. Paper presented at the Proceedings of the 2nd international workshop on Computational Forensics, Washington, DC. http://link.springer.com/chapter/10.1007%2F978-3-540-85303-9_2. Accessed April 5, 2016.

Su, C., & Srihari, S. N. (2008). Generative models for fingerprint individuality using ridge models. Paper presented at the International Conference on Pattern Recognition, Tampa, FL.

Su, C., & Srihari, S. N. (2009). Probability of random correspondence for fingerprints. In *International Workshop on Computational Forensics* (pp. 55–66). Berlin: Springer.

Su, C., & Srihari, S. (2010). Evaluation of rarity of fingerprints in forensics. Paper presented at the Advances in Neural Information Processing Systems

SWGFAST. (2013). Document #10 standards for examining friction ridge impressions and resulting conclusions (latent/tenprint), Scientific Working Group on Friction Ridge Analysis, Study and Technology (SWGFAST). Available at http://www.swgfast.org. Accessed April 5, 2016.

Ulery, B. T., Hicklin, R. A., Buscaglia, J., & Roberts, M. A. (2011). Accuracy and reliability of forensic latent fingerprint decisions. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(19), 7733–7738. doi:10.1073/Pnas.1018707108

Ulery, B. T., Hicklin, R. A., Buscaglia, J., & Roberts, M. A. (2012). Repeatability and reproducibility of decisions by latent fingerprint examiners. *PLoS ONE*, *7*(3), 1–12. doi:10.1371/journal.pone.0032800

Ulery, B. T., Hicklin, R. A., Roberts, M. A., & Buscaglia, J. (2014). Measuring what latent fingerprint examiners consider sufficient information for individualization determinations. *PLoS ONE*, *9*(11), 1–16. doi:10.1371/journal.pone.0110179

Vanderkolk, J. R. (2009). *Forensic comparative science: Qualitative, quantitative source determination of unique impressions, images, and objects*. Burlington, MA: Elsevier.

Vanselst, M., & Merikle, P. M. (1993). Perception below the objective threshold. *Consciousness and Cognition*, *2*(3), 194–203.

Wexler, R. (2015). *Convicted by code*. *Slate*. Available at http://www.slate.com/blogs/future_tense/2015/10/06/defendants_should_be_able_to_inspect_software_code_used_in_forensics.html

Yang, J., Gao, X. M., Zhang, D., & Yang, J. Y. (2005). Kernel ICA: An alternative formulation and its application to face recognition. *Pattern Recognition*, *38* (10), 1784–1787. doi:10.1016/J.Patcog.2005.01.023