COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science 37 (2013) 731–756

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12029

Temporal Sequences Quantify the Contributions of Individual Fixations in Complex Perceptual Matching Tasks

Thomas Busey,[a] Chen Yu,[a] Dean Wyatte,[b] John Vanderkolk[c]

[a]*Department of Psychological and Brain Sciences, Indiana University*
[b]*Computational Cognitive Neuroscience Lab, University of Colorado*
[c]*Indiana State Police Laboratory*

Received 22 July 2011; received in revised form 14 June 2012; accepted 2 August 2012

Abstract

Perceptual tasks such as object matching, mammogram interpretation, mental rotation, and satellite imagery change detection often require the assignment of correspondences to fuse information across views. We apply techniques developed for machine translation to the gaze data recorded from a complex perceptual matching task modeled after fingerprint examinations. The gaze data provide temporal sequences that the machine translation algorithm uses to estimate the subjects' assumptions of corresponding regions. Our results show that experts and novices have similar surface behavior, such as the number of fixations made or the duration of fixations. However, the approach applied to data from experts is able to identify more corresponding areas between two prints. The fixations that are associated with clusters that map with high probability to corresponding locations on the other print are likely to have greater utility in a visual matching task. These techniques address a fundamental problem in eye tracking research with perceptual matching tasks: Given that the eyes always point somewhere, which fixations are the most informative and therefore are likely to be relevant for the comparison task?

Keywords: Fixations; Machine translation; Expertise; Eye tracking; Fingerprints

1. Introduction

A number of different perceptual tasks require the assignment of correspondences between different regions as part of the solution set. For example, radiologists use both the cranio-caudal and mediolateral-oblique images from a mammogram to provide different views of a potential lesion, and the location of the lesion must be determined in each

Correspondence should be sent to Thomas Busey, Department of Psychological and Brain Sciences, Indiana University, 1101 East 10th Street, Bloomington, IN 47405. E-mail: busey@indiana.edu

x-ray to combine information across the two views (Kopans, 2007, p. 393). In the domain of satellite imagery interpretation, two views of a location may be taken from different angles or using different spectral wavelengths and the analyst must mentally align the views to determine what changes may have occurred in the interim (Seeker & Vachon, 2007). Even perceptual tasks such as mental rotation and viewpoint invariance may involve an alignment step or reference frame translation to derive correspondences between two objects or views (Robertson, Palmer, & Gomez, 1987). In each of these tasks, an alignment stage may precede information integration or source determination.

The present work investigates the role of correspondences in the domain of forensic impressions and focuses on latent/inked fingerprint comparisons as a representative task where the common genesis of two impressions is in question. A forensic examiner will compare a latent print extracted from a crime scene against a set of known standards to determine whether the same finger is the source of both impressions. The known standard may come from a database search, in which case even non-matching candidate prints might tend to share some similarity of visual features with the latent print, because the chances of a close non-match are extremely high when databases have hundreds of millions of prints (Dror & Mnookin, 2010). Thus, the appearance of a single feature in both impressions may not indicate a match. Instead, the relative locations of several features (as well as their appearance) plays an important role in the judgment about whether the two impressions share the same underlying cause.

To determine how the pattern of eye gaze locations contributes to the task of aligning features across different impressions, we recorded gaze data using an eye tracking apparatus during fingerprint comparison tasks. An initial analysis step might be to look at sequential pairs of fixations that are separated by an inter-image saccade to determine those regions that correspond in the mind of the observer. However, there are two problems with this approach, the first of which is detailed by Hayes, Petrov, and Sederberg (2011). They point out that if each object is broken into several regions of interest, the transition matrix from one region to the others has a very limited event horizon, reaching only one step into the future. Extending the transition matrix into multiple steps runs into data scarcity problems because the number of paths grows exponentially large.

The second problem with the paired fixation approach is that non-matching impressions may share similar features, which makes the *configuration* of features more diagnostic. Thus, a forensic examiner may have to fixate several different features to create what the community calls a "target group." This may consist of two or more minutia or other distinctive features, and each feature may require a fixation to place it in visual working memory. This suggests that extended temporal sequences that go beyond paired fixations are required, especially if experts tend to search for multiple features while novices search only for one feature at a time.

To address problems with traditional paired-fixation analyses, we derive extended temporal sequence information from our gaze data to make inferences about the ongoing set of correspondences that subjects are currently investigating. The utility of extended temporal sequences over simple pairs of fixations was recently demonstrated by Hayes et al. (2011), in which the authors used a temporal difference algorithm to infer temporal

sequences of fixations made while completing Raven's Advanced Progressive Matrices. This approach was more successful than any previous eye tracking-based approach at predicting participants' behavioral scores based on eye gaze data and demonstrates the utility of temporal fixation sequences to account for behavioral performance.

Our specific approach relies on techniques derived from machine translation and considers spatial details or regions in each of the two fingerprints as equivalent to words or phrases in two parallel texts that attempt to convey the same meaning in two different languages. The results of the machine translation technique identify correspondences between fingerprints that reveal which fixations most directly contribute to the task of identifying whether the two impressions come from the same source. The novel contribution of this approach is that applications of methods derived for machine translation can provide an estimate of the set of correspondences as revealed by the eye gaze behavior that supports superior identification accuracy by fingerprint experts. Moreover, the approach identifies those fixations that are paired across the two prints even if they are not sequential fixations separated by one saccade across the midline of the display. The output of the analyses determines which fixations likely correspond and therefore may contribute most directly to the comparison task and ultimately the identification decision.

Finally, we note that there have been relatively few studies that have used eye tracking methodologies to study fingerprint examinations, and for the most part these have relied on first-order statistics and none has used extended temporal sequences. One study that focused on fingerprint examinations recorded eye fixations during abbreviated latent print examinations to examine the consistency among experts and novices (Busey et al., 2011). They found that experts were more consistent as a group than novices are as long as viewing times are held constant across the two groups. If experts are given as much time as they like with each print, they have a tendency to become more idiosyncratic as a group. In the field of mammography research, Krupinski et al. (Krupinski, 1996; Kundel, Nodine, Krupinski, & Mello-Thoms, 2008) have used eye tracking to investigate not only what features radiologists rely on when inspecting mammograms but also to suggest cognitive mechanisms such as holistic processing when experts are viewing mammograms. Similar work with chest x-rays demonstrated that dwell times were longer on missed tumors than at other locations, suggesting that errors in radiology are due to identification problems rather than detection problems (Krupinski, Berger, Dallas, & Roehrig, 2003). The field of questioned documents (i.e., handwriting analysis) has also benefited from an eye tracking approach as well, which has helped to delimit the visual features that experts rely on when comparing signatures (Dyer, Found, & Rogers, 2008). In addition, work with chess experts found that experts tended to fixate pieces judged to be relevant more than novices did, and also looked at empty squares more often, which may give them a perceptual encoding advantage (Charness, Reingold, Pomplun, & Stampe, 2001; Reingold, Charness, Pomplun, & Stampe, 2001). All of those applications of eye tracking indicate the promise of using eye movement data in fingerprint examination studies to capture elements of expertise, but none make use of extended temporal sequences.

The central question that motivates this research is: Does the temporal structure of eye fixations allow inferences about the correspondences between the two images that are

identified by participants as they attempt the matching task? The current application asks whether the number and accuracy of these correspondences might change as a function of visual expertise in the domain of fingerprint examination. Although our approach is applied to a fingerprint matching task in our current application, it transcends this particular task and potentially has wide application. We elaborate on potential applications in the Discussion.

2. Method

We tested both fingerprint experts as well as novices. Participants were asked to visually examine pairs of fingerprints and to decide whether the two simultaneously displayed fingerprints match. There was no particular instruction about where they should look during the matching task and that they could freely move their eyes. We recorded eye gaze using a custom video-based eye tracker similar to the design proposed by Babcock and Pelz (2004). Each pair of prints was presented for a fixed duration on each trial. This contrasts with an actual fingerprint examination, which is not time-limited and can take tens of minutes or even hours to complete for difficult prints. However, our statistical analyses require a relatively large number of images to ensure reliable results, and we wanted to gather a complete dataset from each participant.

2.1. Datasets

Our results are based on two different datasets that are reflective of the two categories of fingerprint examinations conducted by examiners. Dataset 1 was collected while experts and novices conducted abbreviated fingerprint examinations on latent and inked prints that approximated real casework in appearance. Latent prints recovered from crime scenes are typically only partial representations of the full inked print and are often corrupted by visual noise and have gaps along the ridges. The stimuli were taken from National Institutes of Standards and Technology Special Database 27. Fig. 1 demonstrates a pair of images from this dataset, including the fixations from experts and novices. In this dataset, there were 30 matching pairs and five mismatching pairs. Each trial in this dataset took 20 s and then the subject made an identification decision. Different analyses based on this dataset have been published previously as Experiment 2 of Busey et al. (2011), although the prior analyses addressed consistencies between experts and novices on individual prints, not whether correspondences could be found across prints based on the eye gaze data. The 20 s viewing time was chosen because in prior work with freely viewed images we found that most novices terminated their search after approximately 20 s, while experts tended to take almost twice as long. We did not want our results to be influenced by the fact that experts might take more time, and so we fixed the viewing time to reflect the shortest time in which subjects were typically still working on the print.

Dataset 2 was collected specifically for the analyses described in this article. Prints in this dataset were relatively clean, simulating the biometrically oriented use of fingerprints

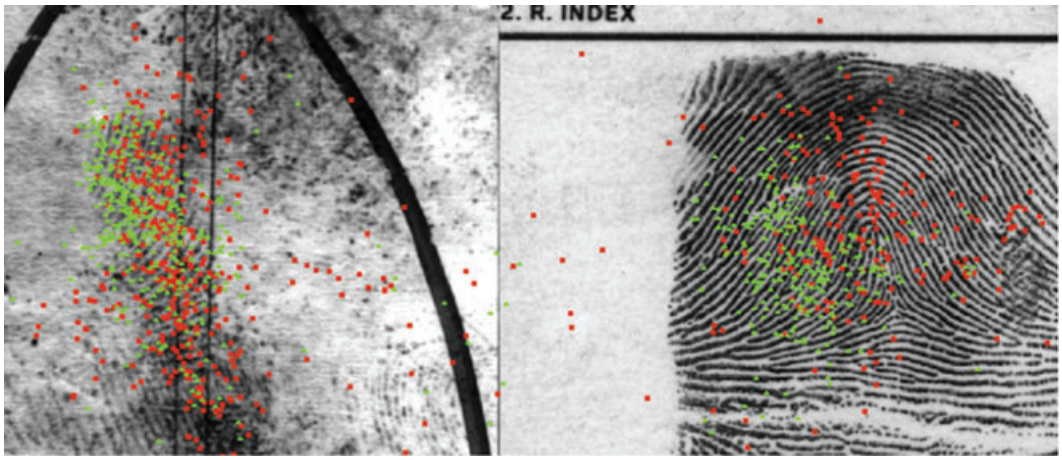

Fig. 1. Latent/inked pair from Dataset 1, with fixations from all experts overplotted as green dots, and fixations from all novices overplotted as red dots. The green dots tend to be clustered in the mid-left portion of the inked print (right image), which corresponds to the area of high detail in the latent print. However, novices have a much wider distribution of fixations, including in regions that have very poor image quality in the latent print.

for identification. In practice, these prints are analyzed for potential matches by a computer and then verified by human experts. The print stimuli in Dataset 2 consisted of pairs of clear images that were scanned at a 1,000-dpi from clear inked prints collected from members of the Bloomington, Indiana community. The dataset consisted of 28 matching pairs and two mismatching pairs. We used more matches than mismatches because the matching trials are more useful for the evaluation of the utility of the machine translation approach. The mismatches were chosen to look very similar, usually by left-right reversing an impression from a finger from one hand to act as a foil for the analogous finger on the other hand of the same individual. Each trial in this dataset consisted of just the first print of the pair displayed for 5 s on the left side of the monitor, followed by just the second print displayed on the right side of the monitor for 5 s, and finally both prints displayed simultaneously for 10 s in their respective locations. Both images together took up the 1,580 $\times$ 759 resolution of the monitor. Because both prints were clear, we included the 5-second pre-exposures of each print individually to encourage participants to view the left image as the "latent" print and the right image as the "inked" print as is typical with traditional latent print examinations. This also allowed the participants to individually look for diagnostic features before looking at both prints simultaneously. Our actual analyses used only the eye gaze data when both images were present simultaneously, and we used only data from the matching prints when evaluating the success of the machine translation approach. Participants were not allowed to interact with the prints to pan or zoom the image, due in part to the limited viewing time. Testing was conducted in dim rooms at illumination levels that allowed for high-contrast images to be displayed on the computer monitor.

In both datasets, a trial ended with a screen clear and a prompt that asked the subject to indicate whether they believe that the two prints came from the same source (an identification), came from different sources (an exclusion), or because 10–20 s is a relatively brief duration, they could indicate "too soon to tell," which could be interpreted as either inconclusive or not yet ready to make a final determination. This last option was added in part because of the conservative nature of examiners (they are reluctant to make decision based on partial or rushed examinations given the gravity of errors in their field) and in part because the three responses could be seen as three levels of confidence along an evidence scale. We will treat these as such when we analyze our data according to signal detection theory below. The terms "identification" and "exclusion" are standard terminology in the latent print community.

Our eye tracker generates gaze data at the sampling rate of 30 Hz, and in total there were approximately 500,000 gaze data points in each dataset. Participants wore a portable head-mounted eye tracker and were seated approximately 60 cm (~24 inches) away from a 21″ LCD monitor. The fingerprint images were presented side-by-side on a 21″ LCD monitor at a resolution of 1,580 × 759 pixels. The monitor itself was set to its native resolution of 1,680 × 1,050 pixels. On our display, 100 pixels subsume about 2.45° of visual arc.

2.2. Participant demographics

There were 12 fingerprint experts and 12 novices in each of the two datasets, giving a total of 48 subjects across both datasets. Experts were recruited at forensic identification conferences and laboratories in Nevada, Illinois, and Indiana, while the novices were members of the Bloomington, Indiana community. The demographics of Dataset 1 are described in Busey et al. (2011). In Dataset 2, experts ranged in age from 32 to 45 years old with a mean age of 36.3 years, while our novices ranged from 18 to 46 with a mean age of 30.9 years. The experts reported a minimum of 3 years of latent print experience, and a max of 13 years, with an average of 7.6 years of unsupervised latent print work. Three novices wore glasses and one wore contacts. One expert wore glasses, three had contacts, and two had Lasik surgery. None of these corrections unduly affect the accuracy of our eye tracker. Five novices were women, and two experts were women. All experts were active latent print examiners in state or large metropolitan crime labs. One expert participant completed both studies. All data were collected under the supervision of the Institutional Review Board (IRB) at Indiana University for purposes of the protection of human subjects.

3. Results

3.1. Behavioral accuracy

The response matrix for the latent/inked comparisons of Dataset 1 and the clean images of Dataset 2 for experts and novices is shown in Table 1. We treat the three

Table 1
Frequency of responses in each category for experts and novices for both datasets

	Dataset 1		
	True Matches		
Participant Category	"Yes"	"Too Soon to Tell"	"No"
Experts	57	247	49
Novices	132	98	130
	True Nonmatches		
Experts	0	16	44
Novices	15	9	36
	Dataset 2		
	True Matches		
Experts	255	78	3
Novices	268	24	44
	True Nonmatches		
Experts	0	2	22
Novices	2	1	21

responses ("match," "too soon to tell," "nonmatch") as three levels of evidence for a match. This allows us to construct a receiver operating characteristic curve for each subject, and then measure the area under this curve. The area under the curve is termed A' (Creelman, 1998). This is a better statistic in this situation than d' because some subjects had zero false alarms and this makes it difficult to compute d' on an individual subject basis. Note that in all statistical tests in the manuscript we set alpha to 0.05 for purposes of null hypothesis significance testing, and report the exact p-value (thresholded at 0.001 for very small probabilities) as well as Cohen's d as a measure of effect size. For nonsignificant or near-significant values we also report the Bayes factor (BF).

For Dataset 1, the mean A' value for experts is .815 and the mean A' for novices is .614. The experts are outperforming novices by this measure, $t(22) = 3.68$; $p < .001$; $d = 1.50$. For Dataset 2, the mean A' value for experts is .982 and the mean A' for novices is .886. The experts are again outperforming novices by this measure, $t(22) = 3.22$; $p = .004$; $d = 1.31$, although even the novices perform quite well at this task.

To address both sensitivity and response biases for the two groups, we modeled the responses using an equal-variance signal detection model (Macmillan & Creelman, 2005). Fig. 2 illustrates the results of this modeling, which was accomplished using the procedures below.

The signal detection model makes the assumption that the outcome of the decision process can be represented as the amount of evidence along an evidence axis (the abscissa in Fig. 2). The right side of the scale corresponds to evidence that the two prints come from the same finger, while the left side of the scale correspond to a lack of evidence that the two prints come from the same source. In essence, the "signal" in this task is a matching

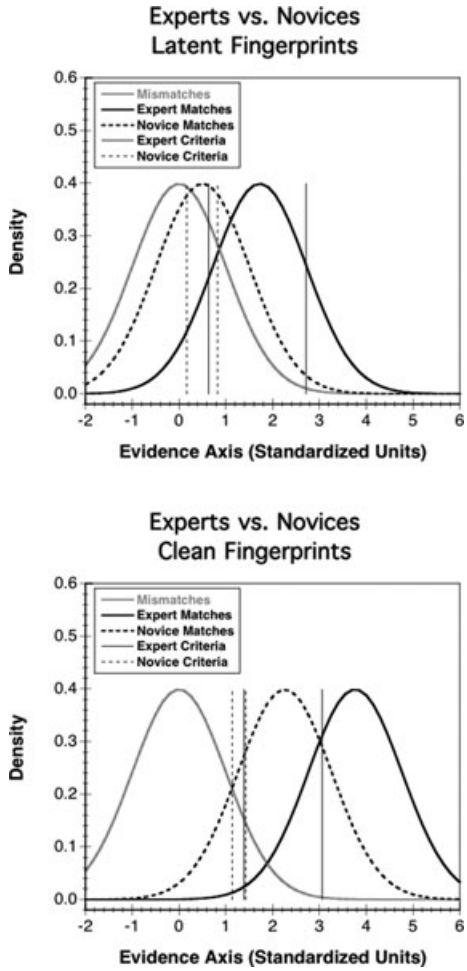

Fig. 2. Signal detection model of responding for dataset 1 (top panel) and dataset 2 (bottom panel). The horizontal axis represents the amount of evidence that the two prints come from the same source, with higher numbers corresponding to more evidence. The gray curves correspond to the mismatch distributions and are axiomatically fixed at a mean of zero for both groups. The dashed lines correspond to the novice match distribution and the two novice criteria. Solid lines are the match distributions for experts, and the two response criteria for experts.

set of prints. Prints of various quality and similarity will differ in their evidence along this axis, and the underlying distributions of evidence values are represented as unit-variance Gaussian distributions. To set the scale, we assign the mis-match distributions to a mean of zero for both experts and novices. The model has three parameters for each group of subjects: the mean of the match distributions (the "signal") and the two decision criteria, for a total of six free parameters to fit eight datapoints that are free to vary (there are 12 total, but the row totals in Table 1 sum to the toal number of matches or mismatches presented to each group).

To make predictions for the proportion of trials that correspond to each cell in Table 1, we assume that if on a given trial the evidence falls below the lower criteria, the model predicts that the observer would report "mismatch." If the evidence falls above the upper criteria, the model predicts that the observer would report "match." Evidence values between the two criteria would lead to a "too soon to tell" response.

The ground truth for the particular trial allows the model to make predictions for the frequency of responding for each of the cells in Table 1. For example, on true non-match trials, evidence above the upper criteria would lead to a false "match" report (a false alarm). For particular values of the mean of the match distribution and values for the two criteria, the model will make predictions for each of the six cells in Table 1 for each group.

The detection model was fit in both Microsoft Excel and Matlab using the assumptions described above using the Solver and fminsearch optimization procedures. We computed the maximum likelihood of the fit during the optimization process, using the simplex algorithm to iteratively discover the best-fitting parameter values. Although this solution is not guaranteed to be global, we find very high correlations between the predicted and actual response proportions. For dataset 1, r^2 between the actual and predicted proportions was .983, and for dataset 2, r^2 between the actual and predicted proportions was .999. These numbers are perhaps not surprising given that there are only four data points and three free parameters for each group. Similar results are found if root mean squared error is minimized.

The results of the model runs are shown in Fig. 2. We pooled the novices together and the experts together and fit the model to the data from both groups. Individual subject fits proved too noisy given that there are only 35 or 39 trials per subject and relatively few mismatching trials. This precludes standard *t*-tests on the parameter values for comparisons. As an alternative we obtained 95% confidence interval estimates around each parameter value by conducting resampling procedures by randomly selecting the data of 12 subjects from each group (with replacement) and fitting the model to the resampled data. This procedure was repeated 1,000 times and the sets of parameters sorted. We then chose the parameter values from the resampling procedures at the 2.5th and 97.5th percentile as our 95% confidence intervals.

We first discuss the parameter values from dataset 1, with 95% confidence intervals following each value in brackets. The mean of the distribution from experts is 1.73 [1.63–1.82] and the mean for novices is 0.50 [0.38–0.63]. The lower criteria for experts is 0.64 [0.55–0.71] and for novices is 0.17 [0.05–0.29]. The upper criteria for experts is 2.72 [2.60–2.90] and for novices is 0.82 [0.17–0.70]. For dataset 2, the mean of the distribution from experts is 3.78 [3.51–4.13] and the mean for novices is 2.26 [2.12–2.42]. The lower criteria for experts is 1.41 [1.13–1.73] and for novices is 1.14 [0.98–1.33]. The upper criteria for experts is 3.08 [2.80–3.42] and for novices is 1.43 [1.26–1.64].

The separation between the match and mismatch groups is usually interpreted as sensitivity, where more separation is equivalent to an improved ability to separate matching from mismatching prints. The experts had greater separation between the matching and mismatching distributions for both experiments by large margins. In addition, in both

datasets the experts tended to be more conservative in their responding, which translates as more likely to give "too soon to tell" responses than novices. This is represented in the model by a greater separation between the two criteria for both experiments.

Taken together, these modeling results demonstrate that experts have fewer errors but are also more likely to stay on the fence with a "too soon to tell" response. This allows them to avoid errors, which prove costly to the novices in terms of the overall separation between the two curves and reduces the overall sensitivity for novices. This result also implies that experts are in some sense choosing to only commit to answers when they are very confident. They are choosing to withhold more responses than the novices do, and this may suggest that one element of expertise is the ability to decide when to make a decision and when one should simply not respond.

The large differences in sensitivity in both datasets between experts and novices demonstrate that there is in fact expertise that is worthy of study. In the next section, we describe the procedures used to collect and analyze moment-by-moment gaze data during visual examination.

3.2. Gaze statistics

The raw eye gaze record consists of eye position recorded at 30 Hz, and we segmented this stream into fixations and saccades in several steps that are based on standard fixation-finding algorithms. All fixation locations are referenced to the fingerprints, not the monitor or scene view of the eye tracker.

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° per second, 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.

We computed low-level gaze statistics to see what separates experts from novices based on the statistical properties of gaze. With the latent/inked prints of Dataset 1, we found no differences in terms of the average duration of each fixation for the two groups, $M_{\text{expert}} = 230.15$ ms, $M_{\text{novice}} = 192.16$ ms; $t(22) = 1.68$; $p = .107$; $d = 0.69$; BF = 1.14. Experts had more fixations on the latent print than novices did, .745 vs. 686; $X^2(1) = 810.18$; $p < .001$, and experts made more saccades within an impression than across impressions than novices did, .852 vs. 780; $X^2(1) = 367.76$; $p < .001$. In addition, experts had smaller saccade lengths on both the latent side, 1.51° vs. 2.35°; $t(22) = -7.60$; $p < .001$, $d = 3.10$, and the inked side, 1.25° vs. 2.68°; $t(22) = -7.28$; $p < .001$, $d = 2.97$. The difference between the two groups in terms of the overall number of fixations did not reach statistical significance, $M_{\text{expert}} = 20.0$, $M_{\text{novice}} = 17.2$; $t(22) = 1.927$; $p = .67$; $d = .79$; BF = 0.81.

For Dataset 2, we computed statistics only for the 10 s intervals in which both prints were visible on each trial (Dataset 1 analyses used the entire 20 s recording time). We again found no statistically significant differences in terms of the average duration of each fixation for the two groups, M_{expert} = 176.2 ms, M_{novice} = 174.3 ms; $t(22) = 0.11$; $p = .913$; $d = 0.04$; BF = 3.463. Experts had slightly more fixations on the left print than novices did, .529 vs. 514; $X^2(1) = 29.7$; $p < .01$, and experts made more saccades within an impression than across impressions than novices did, .778 vs. 652; $X^2(1) = 406.5$; $p < .001$. In addition, experts had smaller saccade lengths on both the latent side, 1.59° vs. 2.57°; $t(22) = -5.13$; $p < .001$; $d = 2.09$, and the inked side, 1.47° vs. 2.47°; $t(22) = -7.67$; $p < .001$; $d = 3.13$. There was again no statistically significant difference between the two groups in terms of the overall number of fixations, M_{expert} = 19.0, M_{novice} = 18.1; $t(22) = 1.00$; $p = .33$; $d = .41$; BF = 2.31.

These low-level gaze statistics suggest that experts have a tendency to spend slightly more time on the left print (which was previewed first and is typically viewed as the latent print during actual examinations) and make smaller saccades overall. Similar results are found in mammography where the left image is viewed first and for longer periods (Kundel, Nodine, Conant, & Weinstein, 2007). These results are consistent with experts making smaller eye movements to regions that are closer together prior to making a saccade to the other impression. This suggests that an analysis based on paired fixations separated by an inter-pattern saccade is insufficient to capture the full behavior of experts. Instead, what is required is an algorithm that explores all of the fixations simultaneously as a means to discover the set of correspondences that the participant is currently exploring. We discuss such an algorithm next.

3.3. Machine translation

To consider both the spatial and the temporal elements of the eye tracking data simultaneously, we introduce ideas from machine translation. In a seminal paper, Brown, Stephen, Pietra, Pietra, and Mercer (1994) described how the context of parallel texts could be used to discover a set of probabilistic relations between the words in two languages. The general idea of machine translation is this: Assume that we have parallel texts from two languages, for example, "Harry Potter and the Order of the Phoenix" in both English and French. We would like to infer which words in the two languages share a common meaning. This inference can be done based on statistical information, such as how frequent *egg* in English and *oeuf* in French co-occur together and how frequent *egg* appears without *oeuf*. Intuitively, if a word in English always co-occurs with another word in French and that word in English appears only when the corresponding word in French appears, then those two words are likely to correspond to each other.

Working with this approach, we have developed a similar set of analyses that considers fingerprint regions as "words" and uses the temporal relations among the fixations to these regions to discover correspondences between two fingerprints. Our computational data analysis consists of four steps as illustrated in Fig. 3. First, we use the previously described fixation finding algorithm to partition the continuous time series of raw gaze

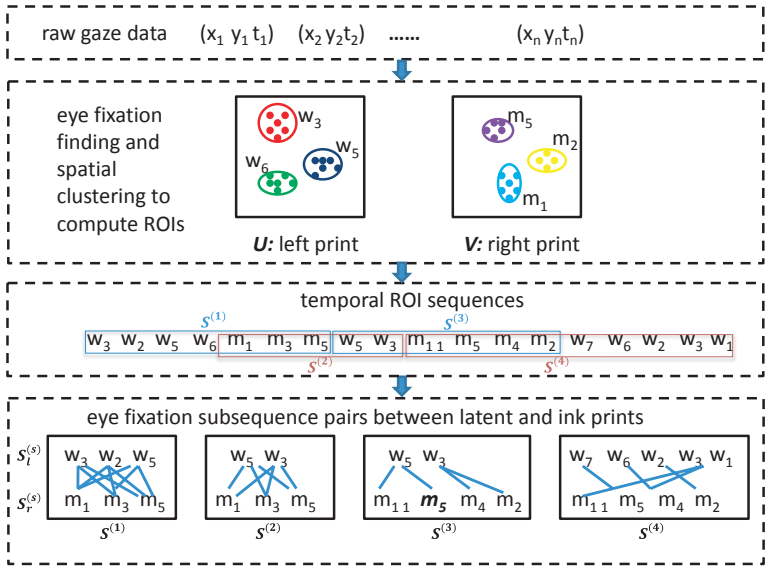

Fig. 3. Steps in Machine Translation analysis. Top box: Raw eye gaze data are segmented into fixations using a variant of velocity-based segmentation. Second box: agglomerative clustering assigns a cluster number to each fixation based on distance measures. Third box: the sequence of cluster numbers on each side is broken into forward sequences (blue boxes) and reverse sequences (red boxes). The $S^{(1)}$, $S^{(2)}$, $S^{(3)}$, and $S^{(4)}$ indicate the four searching instances in this example, which are illustrated with overlapping rectangles. Fourth box: the machine translation algorithm computes the probability of correspondence between each left-side cluster w_g and each right-side cluster m_h. See Appendix for an example illustrating the notation that indexes the location of the ROI m_5 in the third searching instance (italicized).

data into a sequence of eye fixations defined mostly by the speed of eye movements over time (saccade velocity). Second, we use *spatial* clustering to calculate regions-of-interest (ROIs), which are clusters of (x,y) gaze data points based on the spatial distribution of gaze data on the prints. Third, we use the *temporal* information from the ROI sequences to create contextual groupings based on whether the clusters are on the same print or different prints. Finally, we compute the correspondences between ROIs in the inked and latent prints. As a result, we extract the patterns of corresponding areas that subjects examine between two prints. The output of the algorithm is a set of probabilities that indicate the likelihood that a region in one impression corresponds to a region in the second impression. In our case, we consider the temporal information that is bounded by saccades across the midline and discard the ordering of fixations within a temporal sequence because visual working memory may allow an expert to visit different locations in different orders. Essentially our sequences become "bags of fixations" that contain only coarse temporal information generated by midline-crossing saccades.

We will evaluate the success of the machine translation approach in several ways. First, we will identify whether the algorithm can identify more correspondences in the data collected from experts than that collected from novices. Second, we will explore

whether the approach is sensitive to the choice of clustering algorithms or parameters. Finally, for Dataset 2 we will explore whether the correspondences identified for experts or novices are closer to the ground truth corresponding locations.

3.4. Details of the machine translation approach

3.4.1. Temporal fixation finding

We used the same temporal fixation finding algorithm that was used previously to segment the raw datastream for purposes of fixation and saccade analysis. This algorithm is primarily a velocity-based approach that uses spatial proximity only weakly in that it is related to velocity. We found that a wide range of different velocity thresholds gave very similar results, and so the outcomes are not very sensitive to the exact choice of fixation-finding algorithm. The output of this step is a list of fixations and their *x*-*y* centroids. Each fixation consists of typically between 2 and 10 raw gaze points, and thus each fixation duration ranges between 67 and 333 ms.

3.4.2. Spatial clustering

Given the fixation (*x*, *y*) coordinates from Step 1, the next step is to group those gaze fixation points into several clusters—the regions that participants frequently visit. This is a necessary procedure because the data become too sparse for the machine translation algorithm to converge if we treat each individual fixation as a "word." For this step we used hierarchical agglomerative clustering (Jain & Dubes, 1988), which uses spatial distance as the criterion to form clusters of fixations. The basic idea is to treat each eye fixation location as a singleton cluster at the beginning and then successively merge (or *agglomerate*) pairs of clusters until all clusters have been merged into several prototype clusters. We took the centroid of each cluster as an ROI. The criterion used to terminate the clustering algorithm is the minimal distance required to group two data points. In the present study, the distance was set to be 20 pixels, which is equivalent to a visual angle of 0.52°. This produced approximately 175 clusters on the left image for Dataset 1 across both subject groups and 135 clusters on the right image for both groups. For Dataset 2, the left-side images produced about 190 clusters on average and the right side produced about 175 clusters. For the machine translation analysis, we used only prints that actually matched according to ground truth. In a generalization section below we consider alternative clustering algorithms such as *k*-means with similar results.

It is important to note that we merged the data from *all* subjects for each image when computing the clusters. This ensures that if there are systematic differences between the two groups in terms of their eye gaze patterns, it will not produce biases in the clustering algorithm. The results of the clustering are labels assigned to each fixation denoting its cluster membership. In addition, since the clustering results are based on gaze data from all of the participants, we can directly compare gaze patterns from the two groups sharing the same ROI definition. In the Results section, we describe the differences across the two subject groups in terms of how many clusters were actually used by each group. These differ because each subject need not visit each cluster. The exact clustering algorithm is not

critical, and we found similar results to those reported below when we substituted a *k*-means clustering algorithm or used different parameters for the clustering algorithm.

3.4.3. Temporal sequences

The clustering algorithm combined with the temporal sequence of fixations provides a sequence of ROIs extracted from participants' gaze data, some over the left-side image and the rest on the right-side image. Our goal is to calculate correspondences between gazed regions in one image with gazed regions in the other image as participants conducted the matching task. Most often an assumption in machine translation is a sentence-level assignment: We have sentence pairs from two languages that are known to correspond—for example, English and French—and we use this data to infer word correspondences which are unknown. In this case the sentence or paragraph boundaries provide the temporal context (and constraints) from which word correspondences can be derived.

In the fingerprint matching task, we conceptualize the ROIs from one image as words in English, and ROIs on another image as words in French. Based on this conceptualization, the aim here is to find which gazed region in one fingerprint maps to which gazed region in the other fingerprint. To achieve this, we also need to segment the continuous gaze data generated by participants into "sentence" pairs. This is done based on the observation that participants may generate several fixations on the left-side image before crossing to the other image to search for corresponding areas on the right-side image. In light of this, and as shown in the third box of Fig. 3, we first divided a whole sequence into several subsequences by using the midline crossings as breaking points, and then grouped those subsequences into several pairs based on temporal proximity. The outcome of this alignment is a set of fixation sequences that contain fixations from both the left and right images. We call each fixation sequence a *searching instance* as we assume that participants were comparing and matching regions between two prints through those eye fixations on both prints. We used both forward (i.e., left-to-right) as well as reverse (i.e., right-to-left) sequences as input to the machine translation algorithm, although we explored restricting the analyses to just one type of sequence with similar results to using both. The machine translation approach benefits from large datasets, and for the initial analyses we combined the data from all participants to generate a set of correspondences. We also repeated the analysis for individual participants and found similar results for the latent/inked comparisons. These analyses are described in a subsequent section.

3.4.4. Machine translation algorithm

The technical details for the algorithm are found in the Appendix, but we provide a summary here. The goal is to find the probability that a region of interest (ROI) cluster on one print corresponds to a given cluster on the other print. This is done by computing the probability of predicting one set of ROIs in the second half of a searching instance given the ROIs in the first half of the searching instance. This is equivalent to saying: Given this English sentence, what is the probability of observing this French sentence? The advantage of this approach is that the context is provided by the surrounding ROIs and this helps constrain the assignment of correspondences.

The machine translation algorithm iteratively updates the probabilities of correspondence between left- and right-side ROIs on the two images to best predict the observed searching instances. The output consists of the number of times two ROIs co-occurred in the same searching instance, as well as the probability that they correspond. We then apply a threshold to the probability to only consider high-probability pairs as those that correspond, as well as those that have a co-occurring frequency of at least 2, meaning that a participant at least looked at one region in one image and subsequently look at another region in the other region twice. This criteria is necessary because if a ROI in one image co-occurs only once followed by another ROI in the other image, the machine translation algorithm will assign a high correspondence probability to this pair simply because the translation probability between these two infrequent ROIs will be 1. That is, because subjects looked rarely at these ROIs and *only* at these two ROIs, they must have high probability because the sum of all probabilities is 1 for an ROI. For the present purposes we consider these pairs to be less relevant. There are other approaches that might work for "small" correspondences (Och & Ney, 2003), although these are not explored here.

3.5. Evaluation of the machine translation approach

The machine translation algorithm attempts to discover correspondences between two fingerprints using only the temporal sequence of fixations. Although we rely on the centroid of each cluster for visualization purposes, input to the machine translation algorithm does not know anything about space. Its only input is a list of clusters that are visited on each print in the temporally adjacent gaze data, and its output is a probability matrix that describes how likely each cluster on one side corresponds to each cluster on the other side. Fixations associated with clusters that are successfully linked via this approach are likely to have contributed to the identification decision and therefore reflect high-value fixations. Our interpretation places more weight on the fixations associated with clusters that have high-probability correspondences, because the link between clusters can only exist by virtue of the fixations. Note, however, that other fixations may have also contributed to the task yet not be identified by the machine translation algorithm. For example, the examiner's report relying on an "analysis" stage that is focused primarily on the latent print, and then a "comparison" stage where detail from the latent print is sought in the inked print. Fixations associated with the analysis stage may be part of a selection process where features are considered and possibly discarded for purposes of comparison. Machine translation is unlikely to place weight on these fixations unless a match is actively sought on the inked print.

We found that novices generated fixation sequences that were shorter on average than experts. For Dataset 1, the average fixation sequence length was 4.4 clusters, while experts had 6.0 clusters per sequence, $t(34) = -13.6$; $p < .001$, $d = 4.53$. We found significant differences between the two groups both on the latent print, 5.8 vs. 8.2; ($t(34) = -10.5$; $p < .001$; $d = 3.50$, as well as on the inked print, 3.0 vs. 3.7; $t(34) = -6.24$; $p < .001$; $d = 2.07$. Similar results were found for Dataset 2: novices had an average fixation sequence length of 2.9 fixations vs. 4.4 fixations for experts,

$t(29) = -16.2$; $p < .001$, $d = 5.82$. The left-side images showed shorter sequences for experts than novices, 2.9 vs. 4.4; $t(29) = -18.1$; $p < .001$; $d = 6.50$, as did the right side, 2.8 vs. 4.4; $t(29) = -12.8$; $p < .001$; $d = 4.31$. These results are consistent with novices making fewer fixations on each print before crossing over to the other print, which may result from a limited visual working memory for fingerprint detail in the novices.

The machine translation algorithm returns the probability of correspondence between every left-side cluster and every right-side cluster. To assess the strength of these correspondences, we set an arbitrary probability of association threshold of .4 and counted the number of associations that exceed this threshold. This value is reasonable but somewhat arbitrary, and in a later section we explore generalizations of this parameter.

The results of the analysis applied to experts and novices are very clear. Our method produced all of the possible ROI-ROI mappings between fixations on the two images. For Dataset 1 (latent/inked prints), the experts have an average of 14.7 reliable mappings/links found, while the novices have an average of 9.9 links found, $t(34) = 4.51$; $p < .001$, $d = 1.50$. For Dataset 2 with clean prints, we found a similar result. The algorithm found an average of 6.9 links for experts and 4.5 links for novices, $t(29) = 3.18$; $p < .003$; $d = 1.14$. Fig. 4 (from Dataset 1) and Fig. 5 (from Dataset 2) provide a visualization of the results of for both experts and novices for representative images. The lines correspond to strong links identified by the algorithm and the lines are plotted between the centroid of the cluster that was identified as matching by the machine translation solution. These figures and results demonstrate that the temporal dynamics for experts are much better as input to the machine translation algorithm in terms of assigning corresponding links between the two images.

3.6. Generalization studies

We conducted a number of different generalizations studies to ensure that our results are not specific to the particular clustering methods or choice of parameters of the model. To demonstrate that our results are not specific to the choice of the agglomerative clustering algorithm, we repeated the clustering steps with a *k*-means algorithm (Lloyd, 1982). One limitation of the *k*-means approach is that the number of clusters *k* must be pre-specified. However, it does not require a distance parameter like the agglomerative clustering algorithm. We used the agglomerative model with a distance parameter set to 20 (as was used in the main analysis) to choose the number of clusters such that the *k*-means algorithm has the same number of clusters as produced by the agglomerative procedure for a given image. With this preprocessing step, we found an average of 13.6 reliable links for experts and 10.7 reliable links for novices for Dataset 1, which are statistically significantly different, $t(34) = 2.6$; $p = .014$; $d = .87$. We found a similar result for Dataset 2, with an average of 6.1 links for experts and 4.0 for novices, $t(29) = 2.9$; $p = .007$; $d = 1.04$. Thus, the success of the machine translation algorithm does not seem to depend on the choice of clustering algorithm.

We next explored the sensitivity of the algorithm to the size of the distance parameter with the clustering algorithm. The original value of 20 (corresponding to 0.52° of visual

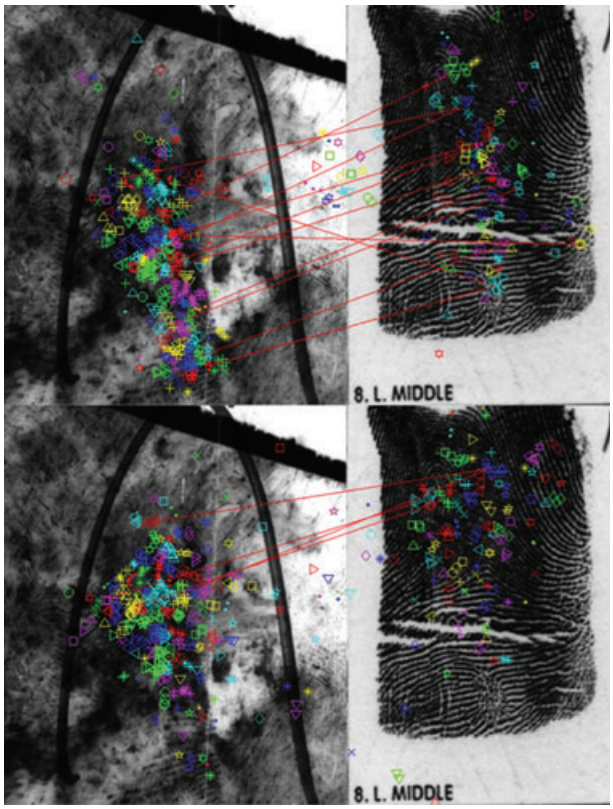

Fig. 4. The corresponding regions from Dataset 1 with inked and latent prints. Top: an example result from experts. Bottom: an example result from novices. The machine translation finds many more reliable links (lines) for experts than for novices.

arc) was increased to 30. With a value of 30, the difference between experts and novices for Dataset 1 drops to trend level significant (13.3 for experts vs. 11.3 for novices; $t(34) = 1.81$; $p = .08$; $d = .60$), although the results from Dataset 2 are consistent with the main findings (11.0 links for experts vs. 8.3 for novices; $t(29) = -3.18$; $p = .003$; $d = 1.14$. Thus, the choice of distance parameter may have some influence on the magnitude of the differences observed between experts and novices.

The machine translation analysis has a threshold parameter for what is considered to be a reliable link, and we set it to a fairly conservative value of .4 (which is typically interpreted as the probability that two clusters correspond). However, to explore more liberal thresholds, we set this value to .1 and .2, which will have the effect of increasing the overall number of links that are considered reliable, at the expense of possibly including erroneous links. For a value of .1, we found more links for experts than novices, 50.0 vs. 32.8; $t(34) = -6.9$; $p < .001$; $d = 2.30$, for Dataset 1, consistent with the original results. Dataset 2 produced a similar pattern, with 18.2 links for experts and 10.7 links for novices, $t(29) = -4.35$; $p < .001$; $d = 1.56$, again consistent with our original findings. Thus,

Fig. 5. The corresponding regions from Dataset 2 with two clean prints. Top: an example result from experts. Bottom: an example result from novices. The lines correspond to the reliable links discovered by the machine translation algorithm.

relaxing the threshold for what is considered reliable seems to increase the differences between experts and novices. Equivalent results are found for thresholds of .2 and .3. Thus, the conclusion that the machine translation approach finds more reliable links for experts than for novices is not sensitive to the choice of threshold value.

Finally, we explored the machine translation approach applied to individual subjects. Recall that the algorithm relies on large collections of data to make links between individual items in each corpus. Traditional text applications use many hundreds of thousands of documents. Our fixation data on each image are limited to only 20 (Dataset 1) or 10 (Dataset 2) s of recording, which leaves relatively little data for each participant. For this reason, we combined the data across subjects in each group for the clustering and machine translation applications described above. Despite the relative scarcity of the data from individual subjects, we explored whether the algorithm could apply to individual subjects. We used the same overall clustering solution derived from the data for all participants, but we used the data only from a single subject.

In recognition of the fact that the results may be less reliable overall due to the limited data, we changed our threshold for what is considered a reliable link from .4 to .3. With

Dataset 1 we have 20 s of recording time per image pair per subject, and this is enough data to allow the algorithm to find significantly more reliable links for experts than for novices. Experts had 22.2 links on average, while novices had only 10.1, $t(22) = 2.65$; $p < .015$; $d = 1.08$. However, the 10 s allowed in Dataset 2 may not provide enough data for the algorithm at the level of individual subjects, 6.7 links for experts, 11 links for novices, $t(22) = -1.1$; $p = .283$; $d = .45$. Similar results are found if the threshold is set to .2, Dataset 1: 31.5 vs. 14.58; $t(22) = 3.48$; $p = .002$; $d = 1.42$; Dataset 2: 15.4 vs. 17; $t(22) = -0.25$; $p = .805$; $d = .10$. These results suggest that under some circumstances the algorithm can be applied to individual subjects, although the duration may have to exceed the 10 s used in Dataset 2.

Each of these generalization analyses demonstrates that the machine translation results are relatively robust to the choice of parameters of the clustering and reliability steps. In addition, the methods apply to individual subjects when sufficient numbers of fixations are included in the dataset.

3.7. Only forward or backward transitions

Latent print examiners report being cautious about the order in which they search the prints, tending to look first at the latent print and then at the inked print. Because the detail in the inked print is usually clearer, there is the danger of seeing detail in the latent that does not exist, based on features observed in the inked print. However, there may be times, such as with exclusions, where "reverse" comparisons are appropriate. Although we do not have sufficient data to address exclusions, we can address whether the temporal sequences from left-right (i.e., forward) or right-left (i.e., backward) comparisons give similar results as when both sequences are used as input to the machine translation algorithm. In all cases below, we found expert/novice differences in the number of reliable links found using parameters of .4 for the threshold and 20 pixels for the clustering algorithm.

For Dataset 1, we found that forward comparisons produced expert/novices differences, 14.7 vs. 9.9; $t(34) = 4.5$; $p < .001$; $d = 1.50$. Backward comparisons produced similar results, 8.5 vs. 3.7; $t(34) = 5.77$; $p < .001$; $d = 1.92$. For Dataset 2, we also found experts with more reliable links for forward, 1.5 vs. 0.6; $t(29) = 2.67$; $p = .012$; $d = .96$, and backward comparisons, 1.0 vs. .7; $t(29) = 3.48$; $p = .002$; $d = 1.25$.

These results suggest that although experts may report using more forward than backward comparisons, there may be enough temporal structure in the order of the fixations for the machine translation algorithm to determine correspondences. Alternatively, experts may simply be using reverse comparisons without being consciously aware of this fact.

3.8. Ground truth accuracy of high-probability correspondences

Are those pairs visually spotted by either experts or novices actually correct correspondences? Did experts do a better job in finding correct correspondences than novices did?

To address these questions, we asked an expert who had not participated in the current studies to independently place corresponding marks on the pairs of clean prints in dataset 2 (the latent/inked pairs from Dataset 1 did not have sufficient clarity to warrant plotting an adequate amount of details for the study). Each markup contained 50–100 marked pairs. We then used a second-order polynomial function to map every point on the left print to a corresponding point in the right print for each print pair. This function tends to smooth any small misplacements in corresponding pairs and provides a right-side match for every left-side pixel and vice versa. A second examiner verified the results of the first.

Having established this ground truth for matching locations, we computed the distance between this true matching location and the location obtained for each correspondence pair discovered by the machine translation algorithm. We found that both groups produce similar deviations. The mean for experts was 53.8 pixels, and the mean for novices was 51.2 pixels. These values were not significantly different, $t(22) = .57$; $p = .58$; $d = 0.23$; BF = 3.04. A deviation of 50 pixels is about 1° of visual angle, which for our images corresponds to about 2 ridge widths in distance. Although we do not see group differences, the spatial accuracy is perhaps surprisingly high given that the algorithm does not know about space directly, the only input being the cluster numbers from the clustered fixations. For comparison, the accuracy of our eye tracker is of similar magnitude (around .5 to 1° of visual angle). These results suggest that the machine translation algorithm is successful in finding spatial matches based solely on the temporal sequences contained in the eye gaze record.

The finding of similar performance between experts and novices on ground truth despite large differences in the number of reliable links found is interesting in part because it is counter-intuitive. Suppose, for example, that our novices could hold fewer visual items in memory. Previous data suggest that the visual working memory span is longer for experts than for novices (Busey & Vanderkolk, 2005; see also Curby & Gauthier, 2007). If this were the case, experts and novices may be equally accurate at identifying corresponding points (especially in the relatively clean prints), but the experts can identify multiple corresponding points at a time while the novices may be limited to a point-by-point comparison. This would make the data from experts more useful for machine translation because the greater working memory could support longer "sentences" (fixation strings). We do not have any direct evidence for greater working memory for visual targets for our experts, but this is certainly consistent with the general perceptual expertise literature on visual working memory (Charness et al., 2001).

3.9. Summary of results

The converging evidence across both datasets and several different analyses is that the machine translation algorithm can use the data from both experts and novices to successfully identify matching regions in two fingerprints. The results suggest that experts can find more of those pairs than novices. In the process, the approach also reveals which eye fixations most meaningfully contribute to the matching task. Given large datasets of eye

movement data, our method can successfully extract meaningful patterns that are not apparent from simple fixation or saccade statistics.

4. Discussion

A pervasive (and commonly ignored) problem in eye tracking research is that the eyes are always pointing somewhere, but the researcher has no direct way of inferring whether the subject considers this to be a meaningful location for a particular task. In fact, the subject may simply be staring off into space. The present result considers the entire corpus of fixations in the context of the matching task to assign correspondences between fixations on each stimulus. Fixations that are assigned a high probability of matching to a location on the other stimulus could be considered as higher-quality fixations with respect to the comparison task, since they imply not only that the subject was performing the task but also was successful in his or her attempt to find a match. In principle, the subject need not even be aware that a match was found; the machine translation algorithm could discover a correspondence by inference.

What do these analyses reveal about the nature of expertise in fingerprint examiners? Although our examinations were time-limited, they had the match/non-match structure of a real examination. Fingerprint examinations provide a nice example of perceptual learning in a context of very restricted visual input. Fingerprints are in some ways similar to faces in that they are all built out of the same set of visual features. The texture may be even less relevant for fingerprints than faces, because the impression development process (e.g., black powder vs. cyanoacrylate fuming vs. ninhydrin) can create impressions with very different surface characteristics. This may force examiners to rely on more configural approaches which would require a pattern of eye fixations that is sequential and regular to identify configurations of features. In this regard, similar processes may be at work in the fingerprints and faces (Busey & Vanderkolk, 2005).

The shorter saccades observed with experts are consistent with a "chunking" strategy in which several features are placed into working memory (Pomplun et al., 2001). Individual features such as single minutiae may not be by themselves particularly diagnostic, but combinations of minutiae may be more rare and therefore more diagnostic. Such a chunking strategy would prove problematic for any analysis technique that relies on pairs of fixations separated by a single saccade that spans both images. Indeed, the order of items within an image need not be the same across the two images, which makes our "bag of fixations" approach more applicable.

Although these results are applied specifically to latent print examiners and fingerprints, we believe that the findings generalize far beyond this particular subject group and task. Indeed, there are many disciplines in which visual comparisons are required to look for similarities and differences. For example, satellite imagery taken at multiple points in time must be inspected for differences that may reflect human activity. Other forensic disciplines such as tool marks, firearms, and footwear are all based on similar principles of visual comparison. Indeed, Pomplun et al. (2001) noted the strengths of the comparative

visual search approach as a paradigm to investigate the fundamental processes underlying vision in complex displays.

The specific utility of analyzing extended temporal sequences in complex visual displays over simple pairs of fixations was recently demonstrated by Hayes et al. (2011). They applied a temporal difference algorithm to eye movements recorded while participants completed the Raven's Advanced Progressive Matrices. The temporal difference algorithm extends the temporal sequences into more than the next step by including, but discounting, future fixations. This creates a successor representation (Dayan, 1993) in which future fixations beyond the next step still contribute, and the algorithm computes the expected number of future fixations of each region of interest given that the subject fixated on one particular region of interest. The authors used a dimensionality reduction approach based on the set of successor representations to predict approximately half of the variance on the final score on the Raven's Matrix across subjects. This approach was more successful than any previous eye-tracking-based approach at predicting behavioral scores based on eye gaze data.

The present approach uses machine translation to assign weights to individual fixations that express the degree of correspondences with fixations in the other image. Fixations that have high probability values as determined by the machine translation approach likely represent important or diagnostic fixations for purposes of comparison. This approach could be extended to include clusters of clusters in a hierarchical representation, much like phrases are built up from words in the text domain. Such clusters of clusters would be consistent with a cognitive strategy that combines several individual features into a single chunk with increased task relevance. For tasks that require a matching decision, such as the fingerprint examination task used in our experiments, this chunking strategy might promote a mental representation that is more diagnostic than one formed using individual features. We are currently exploring these extensions using modified techniques that are designed to more directly address the existence of hierarchical representations.

The most important contribution of the machine translation analyses is the demonstration that the temporal sequence in which locations are visited plays an important role in expertise. The algorithm itself is more successful at identifying correspondences in print pairs viewed by experts because of the temporal information present in the sequence of fixations. We should note that the entire endeavor could have failed spectacularly. Eye gaze data can be quite noisy for the reasons noted above, and the proportion of task-relevant fixations may be quite low if participants engage in a relatively random search process. The success of this approach is made more impressive by the fact that the machine translation algorithm knows *nothing* about the spatial locations of clusters. It only receives a list of labels of clusters that are visited on each side, and a pair of "sentences" is simply a list of the left- and right-side fixations. In fact, the algorithm does not even use the order within a sequence. That such a system could be successful based on such sparse input is a testament to the structure of the information contained in the temporal sequences and the strength of the original Brown et al. (1994) algorithm.

Acknowledgments

This research was supported by National Institute of Justice grants NIJ 2005-MU-BX-K076 and NIJ 2009-DN-BX-K226. The authors would like to thank Bethany Schneider, Francisco Parada, and Ruj Akavipat for their help with data collection and analysis.

References

Babcock, J. S., & Pelz, J. (2004). Building a lightweight eyetracking headgear. Paper presented at the ETRA 2004: Eye Tracking Research and Applications Symposium.

Brown, P. F., Stephen, A., Pietra, D., Pietra, V. J. D., & Mercer, R. L. (1994). The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, *19*, 263–311.

Busey, T. A., & Vanderkolk, J. R. (2005). Behavioral and electrophysiological evidence for configural processing in fingerprint experts. *Vision Research*, *45*(4), 431–448.

Busey, T. A., 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.

Charness, N., Reingold, E. M., Pomplun, M., & Stampe, D. M. (2001). The perceptual aspect of skilled performance in chess: Evidence from eye movements. *Memory & Cognition*, *29*(8), 1146–1152.

Creelman, C. D. (1998). Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers. *Contemporary Psychology*, *43*(12), 840–841.

Curby, K. M., & Gauthier, I. (2007). A visual short-term memory advantage for faces. *Psychonomic Bulletin & Review*, *14*(4), 620–628.

Dayan, P. (1993). Improving generalization for temporal difference learning – The successor representation. *Neural Computation*, *5*(4), 613–624.

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*, 47–67.

Dyer, A. G., Found, B., & Rogers, D. (2008). An insight into forensic document examiner expertise for discriminating between forged and disguised signatures. *Journal of Forensic Sciences*, *53*(5), 1154–1159. doi:10.1111/j.1556-4029.2008.00794.x.

Hayes, T. R., Petrov, A. A., & Sederberg, P. B. (2011). A novel method for analyzing sequential eye movements reveals strategic influence on Raven's Advanced Progressive Matrices. *Journal of Vision*, *11* (10), 10. doi:10.1167/11.10.10.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Englewood Cliffs, NJ: Prentice Hall.

Kopans, D. B. (2007). *Breast imaging* (3rd ed.). Baltimore, MD: Lippincott Williams & Wilkins.

Krupinski, E. A. (1996). Visual scanning patterns of radiologists searching mammograms. *Academic Radiology*, *3*(2), 137–144.

Krupinski, E. A., Berger, W. G., Dallas, W. J., & Roehrig, H. (2003). Searching for nodules: What features attract attention and influence detection? *Academic Radiology*, *10*(8), 861–868. doi:10.1016/s1076-6332 (03)00055-2.

Kundel, H. L., Nodine, C. F., Conant, E. F., & Weinstein, S. P. (2007). Holistic component of image perception in mammogram interpretation: Gaze-tracking study. *Radiology*, *242*(2), 396–402.

Kundel, H. L., Nodine, C. F., Krupinski, E. A., & Mello-Thoms, C. (2008). Using gaze-tracking data and mixture distribution analysis to support a holistic model for the detection of cancers on mammograms. *Academic Radiology*, *15*(7), 881–886. doi:10.1016/j.acra.2008.01.023.

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. doi:10.1109/tit.1982.1056489.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Och, F. J., & Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational Linguistics*, *29*(1), 19–51.

Pomplun, M., Sichelschmidt, L., Wagner, K., Clermont, T., Rickheit, G., & Ritter, H. (2001). Comparative visual search: A difference that makes a difference. *Cognitive Science*, *25*(1), 3–36.

Reingold, E. M., Charness, N., Pomplun, M., & Stampe, D. M. (2001). Visual span in expert chess players: Evidence from eye movements. *Psychological Science*, *12*(1), 48–55.

Robertson, L. C., Palmer, S. E., & Gomez, L. M. (1987). Reference frames in mental rotation. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *13*(3), 368–379.

Seeker, J., & Vachon, P. W. (2007). Exploitation of multi-temporal SAR and EO satellite imagery for geospatial intelligence. Paper presented at the Information Fusion, 2007, Quebec, Que.

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.

Appendix: Machine translation details

The formal definition of the machine translation is as follows. Suppose we have one ROI set (as determined from the clustering algorithm) from the left image $U = \{w_1, w_2, \ldots, w_G\}$ and the other ROI set from the right image $V = \{m_1, m_2, \ldots, m_H\}$, where G is the number of ROIs in left print and H is the number of ROIs in the right print. Let T be the number of temporal sequences (searching instances, both forward and backward) as defined in Step 3 in the main text. Define the data from a searching index as $S^{(S)}$ which contains two subsequences $S_l^{(S)}$ and $S_r^{(S)}$ for the left and right subsequences that contain the ROIs from the left and right image, respectively. The superscript (s) indexes the searching instances and ranges from 1 to T. In the example in Fig. 3, T is 4. All gaze data are in a dataset $\chi = \{(S_l^{(S)}, S_r^{(S)}), 1 \le s \le T\}$, where each subsequence $S_l^{(S)}$ consists of $l^{(s)}$ left-side ROIs $w_{u^{(s)}(1)}, w_{u^{(s)}(2)}, \ldots\ldots, w_{u^{(s)}(l^{(s)})}$ associated with searching index (s) and each $u^{(s)}$ can be selected from 1 to G. In the example in Fig. 3, $l^{(s)}$ is 5 for searching instance 4. We use $u^{(s)}(i)$ to index the ROIs for each subsequence. Similarly, the corresponding gaze sequence on the other print $S_r^{(S)}$ includes $r^{(s)}$ possible right-side ROIs $m_{v^{(s)}(1)}, m_{v^{(s)}(2)}, \ldots\ldots, w_{v^{(s)}(r^{(s)})}$ and the value of $v^{(s)}(j)$ is in the range from 1 to H. In the example in Fig. 3, there are four subsequences (bottom row of Fig. 3) that provide data to determine which ROI in one image should be mapped with one or several co-occurring ROIs in the other image. In the third searching instance (s), if $j = 2$, $v^{(3)}(2)$ equals 5, which is italicized for easy reference. So $m_{v^{(s)}(j)}$ where $s = 3$ and $j = 2$ is linked to the location (x-y coordinate) of the 5th ROI on the right hand image.

The computational challenge here is to build several one-to-one mappings from many-to-many possible mappings within multiple searching instances as not all of the ROIs (generated by participants) within an instance can reliably map to the other ROIs on the other image. We suggest that to determine which ROI in one image goes to which ROI

in the other image, a good solution should not consider just the mapping of a single ROI-ROI pair, but instead we should estimate all these possible mappings simultaneously. Thus, we attempt to estimate the mapping probabilities of all of these pairs so that the best overall mapping is achieved. In doing so, the constraints across multiple searching instances and the constraints across different ROI-ROI pairs are jointly considered in a general system which attempts to discover the best ROI-to-ROI mappings based on the overall statistical regularities in the whole eye fixation sequence.

Formally, given a dataset χ, we use the machine translation method proposed in (Brown et al., 1994) to maximize the likelihood of generating/predicting one set of ROIs from one image given a set of ROIs from the other image:

$$\begin{aligned} P(S_m^{(1)}, S_m^{(2)}, \ldots, S_m^{(T)} | S_w^{(1)}, S_w^{(2)}, \ldots, S_w^{(T)}) &= \prod_{s=1}^{(T)} \sum_a p(S_m^{(s)}, a | S_w^{(s)}) \\ &= \prod_{s=1}^{T} \frac{\in}{(r_s+1)l_s} \prod_{j=1}^{ls} \sum_{i=0}^{r_s} p(m_{v_{(j)}^{(s)}} | w_{u_{(i)}^{(s)}}) \end{aligned} \tag{1}$$

where the alignment a indicates which ROI in one image is aligned with which ROI in the other image. $p(m_{v_{(j)}^{(s)}} | w_{u_{(i)}^{(s)}})$ is the mapping probability for a ROI-ROI pair that are part of subsequences $v^{(s)}$ and $u^{(s)}$ at positions j and i in the subsequences, and ε is a small constant. This is the probability that we compare against a threshold to determine reliable links. Note that these are not probabilities between two ROIs in individual sequences, but are the global probabilities between the two ROIs, that are in this case indexed by the particular searching instance.

To maximize the above likelihood function, a new variable $c(m_h | w_g, S_w^{(s)}, S_m^{(s)})$ is introduced which represents the expected number of times that any particular ROI w_g in one subsequence $S_{w^{(s)}}$ generates any specific ROI m_h in the other subsequence $S_{m^{(s)}}$:

$$\begin{aligned} c(m_h | w_g, S_w^{(s)}, S_m^{(s)}) = &\frac{p(m_h | w_{u_{(i)}^{(s)}})}{p(m_h | w_u^{(s)_{(1)}}) + \cdots + p(m_h | w_u^{(s)_{(r)^{(s)}}})} \\ &\times \sum_{j=1}^{l} \delta(m_h, v(j)) \sum_{i=1}^{r} \delta(w_g, u(i)) \end{aligned} \tag{2}$$

where $\delta = 1$ when both of its arguments are the same and equal to zero otherwise. The second part in Eq. (2) counts the number of co-occurring times of w_g and m_h. The first part assigns a weight to this count by considering it across all the other ROIs in the same

searching instance. By introducing this new variable, the computation of the derivative of the likelihood function with respect to the mapping probability $p(m_h \mid w_g)$ results in:

$$p(m_h|w_g) = \frac{\sum_{s=1}^{S} c(m_h|w_g, S_w^{(s)}, S_m^{(s)})}{\sum_{h=1}^{H} \sum_{s=1}^{S} c(m_h|w_g, S_w^{(s)}, S_m^{(s)})} \tag{3}$$

As shown in Algorithm 1, the method sets an initial $p(m_h \mid w_g)$ based on the co-occurrence statistics, and then successively compute the occurrences of all ROI-ROI pairs $c(m_h|w_g, S_w^{(s)}, S_m^{(s)})$ using Eq. (2) and the mapping probabilities using Eq. (3). In this way, the method runs multiple times and allows for re-estimating ROI-ROI mapping probabilities. The algorithm iteratively estimates the mapping probabilities, maximizing the likelihood function until a convergence criterion is met. A detailed technical description can be found in Brown et al. (1994).

Algorithm 1 Estimating ROI-ROI mapping probabilities
Assign initial values for $p(m_m \mid w_n)$ based on co-occurrence statistics.
repeat
    E-step: Compute the counts for all ROI-ROI pairs using Eq. (2).
    M-step: Re-estimate the mapping probabilities using Eq. (3).
until the mapping probabilities converge.