Behav Res (2015) 47:1377–1392
DOI 10.3758/s13428-014-0550-3

A comparison of scanpath comparison methods

Nicola C. Anderson · Fraser Anderson · Alan Kingstone · Walter F. Bischof

Published online: 25 December 2014
© Psychonomic Society, Inc. 2014

Abstract Interest has flourished in studying both the spatial and temporal aspects of eye movement behavior. This has sparked the development of a large number of new methods to compare scanpaths. In the present work, we present a detailed overview of common scanpath comparison measures. Each of these measures was developed to solve a specific problem, but quantifies different aspects of scanpath behavior and requires different data-processing techniques. To understand these differences, we applied each scanpath comparison method to data from an encoding and recognition experiment and compared their ability to reveal scanpath similarities within and between individuals looking at natural scenes. Results are discussed in terms of the unique aspects of scanpath behavior that the different methods quantify. We conclude by making recommendations for choosing an appropriate scanpath comparison measure.

Keywords Scanpath comparisons · Eye movements · Temporal dynamics of eye movements · Scanpath similarity · Recurrence quantification analysis

When we average over eye movement measures such as fixation counts and durations we ignore that eye movement behavior is a process that unfolds in a particular sequence over time. This sequence is a rich source of information. When Noton and Stark (1971) wrote about their "Scanpath Theory," they noticed that there were only two previous reports (Jeannerod, Gerin, & Perrier, 1968; Yarbus, 1967) that "consider the order in which features are fixated and which, in particular, mention paths followed repetitively from feature to feature of a pattern" (p. 929). Since that time, an explosion of interest has emerged in the particular sequence that the eyes travel over a scene (e.g., Anderson, Bischof, Laidlaw, Risko, & Kingstone, 2013; Brandt & Stark, 1997; Burmester & Mast, 2010; Foulsham et al., 2012; Foulsham & Underwood, 2008; Johansson, Holsanova, & Holmqvist, 2006, 2011; Shepherd, Steckenfinger, Hasson, & Ghazanfar, 2010). Not only has Noton and Stark's Scanpath Theory (1971) been explored in more detail (Foulsham et al., 2012; Foulsham & Underwood, 2008), but sequences of eye movements have been explored when imagining stimuli (Johansson et al., 2006, 2011), when comparing speakers' and listeners' eye movements (D. C. Richardson & Dale, 2005), and when viewing websites repeatedly (Burmester & Mast, 2010). Noton and Stark (1971) were limited to visual inspection to determine whether two scanpaths were similar, but along with the explosion of interest in eye movement sequences came an explosion of methods to analyze such data (e.g., Cristino, Mathôt, Theeuwes, & Gilchrist, 2010; Dewhurst et al., 2012; Henderson, Brockmole, Castelhano, & Mack, 2007).

The purpose of the present work is to bring together the many scanpath comparison methods to both match and differentiate them on a common paradigm and data set. Our experimental paradigm is based closely on that of Foulsham and Underwood (2008). In that work, participants viewed a series of images and were asked to remember them for a later recognition test. Scanpaths were compared within and between individuals and images. Noton and Stark's (1971) Scanpath Theory suggests that, when looking at an image, individuals store both the image features and the gaze sequence used to inspect that image. Thus, the theory predicts that individuals recognizing a

N. C. Anderson (✉)
Department of Cognitive Psychology, Vrije Universiteit Amsterdam, Van der Boechorststraat 1, 1081 BT Amsterdam, The Netherlands
e-mail: nccanderson@gmail.com

F. Anderson · W. F. Bischof
Department of Computing Science, University of Alberta, Edmonton, Canada

A. Kingstone · W. F. Bischof
Department of Psychology, University of British Columbia, Vancouver, Canada

previously seen image follow a scanpath similar to their initial viewing of the image. Foulsham and Underwood (2008) tested this prediction using a computational scanpath comparison technique and found that indeed, a participant viewing the same image twice shows a more similar scanpath than a different participant viewing the same image. Foulsham et al. (2012) expanded their earlier work to include scanpath measures that quantified different scanpath aspects, such as their overall shape or duration similarity. They found that an individual's scanpath was idiosyncratic, that is, more similar within an individual than between individuals. Scanpath similarity was also highest when the same person looked at the same image a second time. Within-participant similarity was highest for the direction of saccades, the position and duration of fixations, and the overall shape of the scanpath. We use these findings as a benchmark with which to evaluate the many available scanpath techniques. We hope that this work provides an informative overview of the available methods that will help researchers select an appropriate measure for use in their own work.

Scanpath comparison methods

We describe the scanpath comparison methods that have been introduced in the literature. In each case, we give a short description, and the reader is advised to consult the original publications for further details. Additional mathematical details are provided in the Appendix.

Edit distance

One successful way for comparing scanpaths is based on the string-edit distance (Bunke, 1992; Levenshtein, 1966; Wagner & Fischer, 1974), which is used to measure the dissimilarity of character strings. In this method, a sequence of transformations (insertions, deletions, and substitutions), is used to transform one string into the other and their similarity is represented as the number of transformation steps between the two strings. This method has been adapted for comparing the similarity of scanpaths (Brandt & Stark, 1997; Foulsham & Kingstone, 2013; Foulsham & Underwood, 2008; Harding & Bloj, 2010; Underwood, Foulsham, & Humphrey, 2009). To achieve this, a grid is overlaid on an image, and each cell in the grid is assigned a unique character. Fixation sequences are then transformed into a sequence of characters by replacing the fixation with the character corresponding to the grid cell the fixation falls in. The dissimilarity of two scanpaths can then be represented by the number of transformations required to convert the string corresponding to the first scanpath to the string corresponding to the second scanpath.

The string-edit distance method was very popular in early scanpath comparison work (e.g., Brandt & Stark, 1997) and has been used subsequently in a variety of experimental contexts (e.g., Harding & Bloj, 2010; Underwood et al., 2009). This is an advantage for researchers wishing to directly compare results to these earlier studies. But the main advantage of the string-edit measure lies in the fact that it captures the intuitive notion of scanpath distance in a simple way. However, several criticisms have been raised against the use of edit distance for scanpath comparisons. First, the grid is defined independently of image content and may be too coarse in regions of interest while being too fine in other regions. Second, two fixations may be considered different even when they are close together, namely if they fall on either side of a grid line. Variants of the string-edit distance have been developed to address these problems. For instance, assigning characters to pre-defined areas of interest allows the researcher to add semantic information to the quantization process (Josephson & Holmes, 2002; West, Haake, Rozanski, & Karn, 2006), but the definition of regions of interest can be time-consuming. In our analysis, we have used a simple grid-based variant of the edit distance.

The string-edit measure has been used on a similar dataset previously in Foulsham and Underwood (2008), where similarity was found to be highest for scanpaths generated from the same person looking at the same image. In addition, in Foulsham et al. (2012), similarity in shape was highest for the same person looking at the same image. Given that the string-edit distance measure is most sensitive to similarities in shape and sequential information, we expect that string edit similarity scores should be quite high between scanpaths of the same person looking at the same image for a second time, convergent with these earlier results.

ScanMatch

Cristino et al. (2010) proposed a generalized scanpath comparison method that addresses many of the deficiencies of the string-edit distance method. Their generalization aligns eye movement sequences based on the Needleman-Wunsch algorithm, which is used in bioinformatics to compare DNA sequences. In their method, scanpaths are spatially and temporally binned and then recoded to create a sequence of letters that retains fixation location, duration, and sequence information. The two character sequences are compared by maximizing the similarity score computed from a substitution matrix, which in turn provides the score for all letter pair substitutions and a gap penalty. Critically, the substitution matrix can encode information about the relationship between specific regions of interest, thus providing the opportunity to include semantic information in the similarity measure.

A major advantage of the ScanMatch method is that it can take into account spatial, temporal, and sequential similarity between scanpaths. In addition, semantic information can be easily added using the substitution matrix. One disadvantage

of this method is that it suffers from the quantization issues inherent in any measure using regions of interest or grids.

Since ScanMatch quantifies spatial, sequential, and duration information together, we expect ScanMatch to do well in revealing within-participant scanpath similarity, and in particular, the strong similarity between observers viewing the same image twice (see Foulsham et al., 2012).

Sample-based measures

Shepherd et al. (2010) introduced several measures for assessing the similarity of two scanpaths. For each of the measures, the scanpaths are first resampled uniformly in time (at 60Hz), and truncated to the shorter length. These measures are sample-based, in that they do not require pre-processing of eye-tracking data into discrete fixation-saccade sequences via saccade and velocity thresholds.

Fixation overlap The first measure, fixation overlap, is defined as the proportion of overlapping samples. Two samples (at time t) overlap if the Euclidean distance between two samples is less than a predefined threshold.

The overlap between two scanpaths yields a similarity measure that is sensitive to temporal and spatial differences between fixation locations. It does not take into account fixation duration, rather it uses the resampling to capture aspects of temporal similarity. Thus, this method preserves temporal ordering but does not account for differences in fixation times. As a result, two scanpaths could have the same spatial positions but different fixation durations, and this method would then evaluate them as not overlapping and therefore being very different. For example, if one scanpath lagged behind another by one fixation but was otherwise spatially overlapped, this method would evaluate the two sequences as very different. One drawback of this method is that it uses an arbitrary, pre-defined radius threshold, with similar disadvantages to the grid-based quantization of string-edit and ScanMatch.

Fixation overlap is extremely sensitive to differences in absolute timing between two scanpaths, but is slightly less sensitive to differences in position (due to the use of the radius). Given these sensitivities, it is reasonable to expect this measure to perform similarly to the ScanMatch measure, which is also sensitive to the spatial and temporal similarities between two scanpaths.

Temporal correlation Shepherd et al. (2010) also introduced temporal correlation (see also Hasson, Yang, Vallines, Heeger & Rubin, 2008) as a measure of the similarity between scanpaths. For two scanpaths, the temporal correlation is defined as the average of the correlation between their x-coordinates and y-coordinates, respectively.

This measure is very sensitive to temporal and spatial differences between the two scanpaths. The sensitivity to temporal differences can be advantageous when timing is important, e.g., when the stimuli change over time, such as in videos. The correlation measure is also sensitive to small differences in fixation positions, given that there is no spatial quantization of the fixations. A significant advantage of this method is its use of the straightforward and readily interpretable correlation analysis. This measure is more sensitive to similarities in position than the fixation overlap method, while also taking sequential information into account. However, this strong spatial-temporal sensitivity may be less robust to noisy data than other measures that employ a grid or radius.

We expect the temporal correlation measure to be particularly useful in situations and paradigms where precise temporal timing of gaze sequences is crucial.

Gaze shift Shepherd and colleagues' (2010) gaze-shift measure assesses how similar the saccade times and amplitudes are between two scanpaths. Gaze shift is computed as the correlation between the absolute values of the first derivative of each scanpath and is computed in the same manner as the temporal correlation, but using the first derivative instead of the position.

For smoothing and for computing the derivatives of the scanpaths, each scanpath is convolved with the derivative of a Gaussian filter. Gaze shift is sensitive to the amplitude of the saccade as well as to its temporal location. It reflects how similar two scanpaths are in terms of the sequence of large and small saccades. This captures some aspects of a global viewing strategy, as subjects who make small saccades within a localized region would have very different scanpaths than subjects who make large saccades within the entire visible area. This is also useful for comparing dynamic stimuli (e.g., video) to assess how subjects respond to temporal changes in the scene.

The gaze-shift measure quantifies similarity in amplitudes, and might correspond well with the MultiMatch measure that quantifies similarity in scanpath length. In Foulsham et al. (2012), similarity in length was only consistent for the within/between-image comparison. One might expect a similar result for the gaze-shift measure; however, prediction is difficult because it simultaneously quantifies, like the other sample-based measures, temporal similarity.

Linear distance

Mannan, Ruddock, and Wooding (1995) and Mathot, Cristino, Gilchrist, and Theeuwes (2012) analyzed the overall similarity between two scanpaths by computing the linear distances between the fixations in the first scanpath and the nearest neighbor in the second scanpath, as well as the linear distances between the fixations in the second scanpath and the

nearest neighbor in the first scanpath. These distances are averaged and normalized against randomly generated scanpath sequences (see Appendix).

The most significant advantage of the linear distance method is that it does not need to be quantized as in the string-edit method. It simply compares each fixation in one scanpath with the fixations in another in terms of their spatial similarity. However, by comparing only nearest neighbor fixations in terms of distance, this method ignores sequential information. To address some of these issues, Mannan et al.'s (1995) method was modified by Henderson, Brockmole, Castelhano, & Mack (2007) to enforce a one-to-one mapping between two scanpaths, provided that they have the same length. The results for the two methods are very similar (Foulsham & Underwood, 2008), which is likely due to the fact that Mannan et al. average the distances from the first to the second and from the second to the first scanpath, hence clusters of fixations in one scanpath are averaged out. For this reason, we used only Mannan et al.'s (1995) original method in our analyses. Linear distance is a measure that specifically quantifies and compares the fixation positions in two scanpaths, regardless of the order of fixations. Given that scanpath position comparisons have previously revealed an advantage for within-participant similarity (Foulsham et al., 2012), we expect this method to perform well in comparing within- and between-participant scanpath similarity.

MultiMatch

Recently, Jarodzka, Holmqvist, and Nyström (2010), Dewhurst et al. (2012), and Foulsham et al. (2012) introduced the MultiMatch method for comparing scanpaths. The MultiMatch methods consists of five separate measures that capture the similarity between different characteristics of scanpaths, namely shape, direction, length, position, and duration. Computation of each MultiMatch measure begins with scanpath simplification, which involves combining iteratively successive fixations if they are within a given distance or within a given directional threshold of each other. This simplification process aids in reducing the complexity of the scanpaths while preserving their spatial and temporal structure.

Following this simplification, scanpaths are aligned based on their shape using a dynamic programming approach. The alignment is computed by optimizing the vector difference between the scanpaths (note, however, that scanpaths may be aligned on any number of dimensions in MultiMatch). This alignment reduces the comparison's sensitivity to small temporal or spatial temporal variations, and allows the algorithm to find the best possible match between the pair of scanpaths. All subsequent similarity measures are computed on these simplified, aligned scanpaths. The MultiMatch similarity computations presented here follow the implementation described in Dewhurst et al. (2012).

MultiMatch (MM) vector Vector similarity is computed as the vector difference between aligned saccade pairs, normalized by the screen diagonal and averaged over scanpaths. This measure is sensitive to spatial differences in fixation positions without relying on pre-defined quantization. It is a measure of the overall similarity in shape between two fixation-saccade sequences.

MM length Length similarity is computed as the absolute difference in the amplitude of aligned saccade vectors, normalized by the screen diagonal and averaged over scanpaths. This measure is sensitive to saccade amplitude only, not to the direction, location, or the duration of the fixations.

MM direction Direction similarity is computed as the angular difference between aligned saccades, normalized by π and averaged over scanpaths. This measure is sensitive to saccade direction only, but not to amplitude or absolute fixation location.

MM position Position similarity is computed as the Euclidean distances between aligned fixations, normalized by the screen diagonal, and averaged over scanpaths. This measure is sensitive to both saccade amplitudes and directions.

MM duration Duration similarity is computed as the absolute difference in fixation durations of aligned fixations, normalized by the maximum duration and averaged over scanpaths. This measure is insensitive to fixation position or saccade amplitude.

The main advantage of the MultiMatch method is that it provides several measures to choose from for assessing scanpath similarity, and each measure on its own captures a unique component of scanpath similarity. Given the multiplicity of measures, it remains, however, difficult to assess which measure, or which set of measures, is most applicable in a given scenario. Furthermore, because each scanpath is initially simplified it is also not clear how robust each measure is to scanpath variations.

Given that the MultiMatch measures have already been evaluated with a dataset similar to the one generated for the present work, we expect to essentially replicate those earlier results, where saccade direction, fixation position, fixation duration, and shape similarity were found to be higher for within-participant compared to between-participant comparisons.

Cross-recurrence quantification analysis

Recurrence analysis has been used successfully as a tool for describing complex dynamic systems, e.g., for electrocardiograms (Webber Jr & Zbilut, 2005) that are difficult to characterize using standard methods in time-series analysis (e.g., Box, Jenkins, & Reinsel, 2013). It has also been used for

describing the interplay between dynamic systems in cross-recurrence analysis, e.g., for analyzing the postural synchronization of two persons (Shockley, Baker, Richardson, & Fowler, 2007; Shockley, Santana, & Fowler, 2003; Shockley & Turvey, 2005). Richardson, Dale and colleagues have generalized recurrence analysis to categorical data and have used it for analyzing the coordination of gaze patterns between individuals (e.g., Cherubini, Nüssli, & Dillenbourg, 2010; Dale, Kirkham, & Richardson, 2011a; Dale, Warlaumont, & Richardson, 2011b; Richardson & Dale, 2005; Richardson, Dale, & Tomlinson, 2009; Shockley, Richardson, & Dale, 2009). For example, Richardson and Dale (2005) quantified the coordination between a speaker and a listener's eye movements as they viewed actors on a screen. This form of cross-recurrence analysis can provide an overall measure of similarity across two eye movement sequences.

To characterize cross-recurrence patterns, we have developed several measures that are based on the recurrence quantification analysis (RQA) for characterizing gaze patterns of a single observer (Anderson et al., 2013). These measures are introduced briefly below and described in detail in the Appendix.

Consider two fixation sequences $\boldsymbol{f}$ and $\boldsymbol{g}$ that have the same lengths. For sequences of unequal length, the longer sequence is truncated. Within these sequences, any two fixations $\boldsymbol{f}_i$ and $\boldsymbol{g}_j$ are cross-recurrent if they match or are close together, i.e., if their distance is below a given threshold. In the following, we introduce several measures that we have found useful for characterizing cross-recurrent patterns.

Cross-recurrence The cross-recurrence measure of two fixation sequences represents the percentage of cross-recurrent fixations, i.e., the percentage of fixations that match between the two fixation sequences. Cross-recurrence is higher the more spatially similar two fixation sequences are and quantifies their similarity in shape. It is invariant to differences in fixation sequence order as fixations are considered recurrent only if they overlap in position. Given that cross-recurrence quantifies similarity in position, results should be most in line with the linear distance measure and the MultiMatch position measure.

Determinism The determinism measure represents the percentage of cross-recurrent points that form diagonal lines in a recurrence plot and represents the percentage of fixation trajectories common to both fixation sequences. That is, it quantifies the overlap of a specific sequence of fixations, preserving their sequential information (see Fig. 5). An advantage of this measure is that it provides unique information about the type of similarity between two scanpaths. Although two scanpaths may be quite dissimilar in their overall shape or fixation positions, this measure may show whether certain smaller sequences of those scanpaths may be shared.

Laminarity Laminarity is a measure of repeated fixations on a particular region that are common to both scanpaths. Laminarity is closely related to determinism. If both laminarity and determinism are high, then in both scanpaths fixations tend to cluster on one or a few particular locations and remain there across several fixations. If laminarity is high, but determinism is low, then it quantifies the number of locations that were fixated in detail in one of the fixation sequences, but only fixated briefly in the other fixation sequence. It is a measure of the clustering of fixations across two sequences.

Center of recurrence mass Finally, the center of recurrence mass (CORM) is defined as the distance of the center of gravity of recurrences from the main diagonal in a recurrence plot (see Dale et al., 2011b for another method to quantify leading and following in cross-recurrence). The CORM measure indicates the dominant lag of cross-recurrences. Small CORM values indicate that the same fixations in both fixation sequences tend to occur close in time, whereas large CORM values indicate that cross-recurrences tend to occur with either a large positive or negative lag. This is a measure of whether one scanpath may lead (positive lag) or follow (negative lag) its paired scanpath. Their overall similarity in shape or position may be different, but offset, such that one sequence proceeds in a particular trajectory, and the other follows the same trajectory only later on in time (e.g., a few fixations later). In the present work, we use the absolute value of CORM rather than averaging over positive and negative values as we do not have any specific predictions about whether leading or following is more likely to happen in one particular condition or another. Overall, we might predict low corm values for within-image, within-participant comparisons if participants consistently lead or follow a similar scanpath closely (i.e., with a low positive or negative lag) on a later viewing of the same image.

Summary of measurements

All measures presented above are summarized in Table 1 with a list of generalized characteristics. In some methods, e.g., Fixation Overlap, scanpaths are first resampled uniformly in time, for example at 60 Hz. Quantization refers to the way fixations are treated in the different methods: in grid quantization, fixations are discretized on a fixed grid overlaid on the stimuli; in radius quantization, the distance between fixations is thresholded to assess whether two fixations are considered the same or different; and in the direct method, the exact fixation coordinates are used. The MultiMatch measures are used with simplified scanpaths; the other measures use the raw scanpaths. Some methods

Table 1 Overview of scanpath comparison measure properties

Measure	Resampling	Quantization	Simplified?	Truncated?	Preserves temporal ordering?	Target scanpath variable
String edit	No	Grid	No	No	Yes	Position, Sequence
ScanMatch	Yes	Grid	No	No	Yes	Position, Duration, Sequence
Overlap	Yes	Radius	No	Yes	No	Sequence, Position
Correlate	Yes	Direct	No	Yes	Yes	Position, Sequence
Gaze shift	Yes	Direct	No	Yes	Yes	Amplitude, Sequence
Linear distance	No	Direct	No	No	No	Position
MM vector	No	Direct	Yes	No	Yes	Shape
MM direction	No	Direct	Yes	No	Yes	Saccade Direction
MM length	No	Direct	Yes	No	Yes	Saccade Length
MM position	No	Direct	Yes	No	Yes	Position
MM duration	No	Direct	Yes	No	Yes	Duration
Recurrence	No	Radius	No	Yes	No	Position
Determinism	No	Radius	No	Yes	No	Fixation Trajectories
Laminarity	No	Radius	No	Yes	No	Fixation Persistence
Corm	No	Radius	No	Yes	Yes	Leading/Following

require that the two scanpaths to be compared have the same lengths, hence the length of the longer scanpath is truncated to the length of the shorter scanpath. Some measures preserve and use the temporal order of fixations; others do not. Finally, the measures differ with respect to the major characteristic they capture, including, for example, scanpath shape, fixation position and duration, saccade amplitude, and direction.

The present investigation

In the present work, we compared the ability of the scanpath measures reviewed above to reveal scanpath similarities within and between individuals looking at natural scenes. In a paradigm adopted from Foulsham and Underwood (2008) and Foulsham et al. (2012), participants performed a scene encoding and subsequent recognition task. This ensured that participants saw each image in the encoding phase for a second time during recognition, allowing for a comparison of scanpath similarity both within and between subjects and within and between images. The main prediction of these comparisons is that scanpaths are more similar for the same person viewing the same image, than for different people viewing different images. This observation has been verified by a handful of scanpath comparison techniques (see Foulsham & Underwood, 2008; Foulsham et al., 2012). We use it here in order to evaluate the ability of each scanpath comparison method to reveal the uniquely high similarity in the scanpath of the same observer viewing the same image twice, and its ability to capture any singular similarities between observers and images.

Methods

Participants

Twenty-seven participants with normal or corrected-to-normal vision were recruited from the University of British Columbia, and participated for course credit or $5 (Canadian).

Apparatus

Stimuli were presented full-screen on a 19-in monitor operating at a 60-Hz refresh rate. Participants sat 60 cm from the screen with their head restrained in a chin rest. Thus the screen subtended approximately 32.7° × 25.7° of visual angle. Eye movements were recorded using the Eyelink 1000 eyetracker (SR-Research) and participants used a standard keyboard when responses were required.

Stimuli

A total of 36 images at a resolution of 1024 × 768 pixels were selected from a dataset created by Foulsham and Underwood (2008). The images were pictures of buildings, interiors, and landscapes (see Foulsham & Underwood, 2008 for more information). Half of these images were shown during both the encoding and recognition phases of the experiment, while the other half were shown only during the recognition phase to act as "new" images.

Procedure

Participants were seated comfortably with their chin in the chin rest and completed a 9-point eye-tracker calibration

procedure before beginning the experiment. At the start of the experiment, participants were asked to look at the images in preparation for a later memory test. Participants were then given a short break and re-calibrated before they began the recognition phase. In the recognition phase, participants were asked to respond during image presentation with the 'z' key if they had seen the image before and the '/' key if they had not, and the image remained on screen even after response. In both the encoding and recognition phase, images were presented for 10 s and in random order.

Analysis

The data were analyzed using the 15 scanpath similarity measures described in the Introduction. For each measure, we used the parameters recommended by the authors. The relevant parameters are given below.

For the sample-based methods, signals were re-sampled at 60 Hz. For the overlap method, two fixations were considered overlapping if their distance was less than 3.5° visual angle. For the gaze-shift measure, a temporal Gaussian filter with $\sigma =$ 100 ms was used. For the string-edit distance, fixations were discretized on an 8 × 6 grid (approximately $4^{\circ} \times 4^{\circ}$ per grid cell). For the ScanMatch measure, the substitution matrix was used to define similarity in distance between grid locations (as was done in the original manuscript). For the MultiMatch measure, scanpaths were simplified by combining fixations that were closer than 3.5° visual angle or by combining successive saccades whose direction differed by less than 45°. For the recurrence-based measures, two fixations were considered cross-recurrent if their distance was less than 1.9° visual angle. For the determinism and laminarity measures, a minimum line length of 2 was used.

In the scanpath analyses, we were interested in how the two dimensions, participant and image, affected scanpath similarity. Within-participant similarity suggests that the individual creating the scanpath is a factor in scanpath similarity. Within-image similarity suggests that the image itself impacts similarity between scanpaths. The comparisons were made between the same observer viewing the same image (within-participant, within-image similarity); the same observer viewing different images (within-participant, between–image similarity); different observers viewing the same image (between-participant, within-image similarity), and different observers viewing different images (between-participant, between-image similarity). This leads to a 2 (participant; within-participant vs. between-participant) by 2 (image; within-image vs. between-image) repeated measures analysis of variance for each scanpath measure. Comparisons were made such that all possible participant-scanpath pairs were included. In the within-image conditions, encoding scanpaths were compared to recognition scanpaths, and in the between-image conditions, encoding scanpaths were compared to other encoding scanpaths. Note that in the recognition phase, participants responded during image presentation and were subsequently obliged to freely view the image for the entire 10s presentation. It is unknown whether participants employed any sort of unique viewing strategy during this time after response and so we chose not to analyze whether scanpath comparisons resulted from encoding-recognition or encoding-encoding pairs.

Rather than simply compare measures in terms of their ability to detect the main effects and interaction of the analysis of variance, we chose to use effect sizes for making comparisons across measures. For example, if the recurrence measure has a high effect size for the main effect of image, then this measure is sensitive to this factor in the similarity of scanpaths. We used generalized eta squared (η^2_G) for the comparisons (Bakeman, 2005; Olejnik & Algina, 2003; J. T. Richardson, 2011). In repeated-measures designs, η^2_G is comparable across different within-participant and between-participant variables (see J. T. Richardson, 2011, p. 142), but the same is not true for the partial eta squared (η^2_p). Thus η^2_G allows for a direct comparison of the effects of participant and image. Bakeman (2005) suggests that similar guidelines should be used for assessing the size of the effect with η^2_G as is the case for η^2_p , i.e., an effect size of 0.02 is considered as small, 0.13 as medium, and 0.26 as a large effect size.

Results

For each scanpath measure, the *F*-ratio and *p*-values for the main effect of participant, the main effect of image, and the participant-by-image interaction are presented in Table 2 and mean scanpath similarity for each measure across the four conditions is presented in Table 3. The remainder of the results will focus on the resulting η^2_G obtained for each term in the analysis of variance. The η^2_G values for the main effect of participant is shown in Fig. 1, for the main effect of image in Fig. 2, and for the participant-by-image interaction in Fig. 3. The means of the significant participant-by-image interactions are shown in Fig. 4, and the corresponding post-hoc comparisons are reported in the participant-by-image paragraph below.

Main effect of participant

Previous work using the same paradigm (Foulsham & Underwood, 2008; Foulsham et al., 2012) has typically revealed that scanpaths are generally more similar if they are from the same individual than if they are from different individuals. In the following, where a main effect of participant was revealed, within-participant similarity was higher than between-participant similarity. In the interest of space, we do not report the similarity means within the text. Of the

Table 2 F-ratio and p values for each scanpath comparison measure

Group	Measure	Participant		Image		Participant x image	
		F	*p*	*F*	*p*	*F*	*p*
Grid-based	String edit	12.55	**0.002**	364.84	**< .001**	11.43	**0.002**
	ScanMatch	53.63	**< .001**	286.80	**< .001**	26.33	**< .001**
Sample-based	Overlap	56.47	**< .001**	236.15	**< .001**	29.80	**< .001**
	Correlate	2.27	0.144	14.71	**0.001**	1.12	0.300
	Gaze shift	0.00	0.953	22.29	**< .001**	1.22	0.279
Direct measures	Linear distance	32.95	**< .001**	201.81	**< .001**	10.99	**0.003**
	MM Vector	14.25	**0.001**	79.84	**< .001**	1.90	0.180
	MM direction	13.17	**0.001**	38.94	**< .001**	2.81	0.106
	MM length	2.83	0.105	25.43	**< .001**	0.68	0.418
	MM position	17.34	**< .001**	150.92	**< .001**	0.70	0.410
	MM duration	12.53	**0.002**	0.31	0.585	0.53	0.472
Recurrence-based	Recurrence	22.20	**< .001**	251.52	**< .001**	22.22	**< .001**
	Determinism	67.99	**< .001**	146.30	**< .001**	35.87	**< .001**
	Laminarity	1.87	0.183	22.98	**< .001**	0.90	0.352
	Corm	0.13	0.718	2.51	0.125	0.31	0.585

Note: Bold values indicate significance

sample-based methods, only the overlap measure reveals the main effect of participant, with a particularly high effect size (η^2_G = 0.23). Note that the overlap method employed a relatively large radius (3.5° visual angle), therefore it is possible that this large effect size is spuriously due to the generous radius size. The correlate and gaze-shift measures do not. Both grid-based measures reveal a significant main effect of participant with ScanMatch showing the highest effect size, (η^2_G = 0.32) and string edit showing a very low effect size (η^2_G = 0.04). All of the direct measures revealed a significant effect of participant with a relatively high effect size for linear distance particularly, (η^2_G = 0.17), with the MultiMatch measures showing rather modest effect sizes. Of the recurrence-based measures, recurrence (η^2_G = 0.14), determinism (η^2_G =

Table 3 Mean scanpath similarity value (with standard deviations in parenthesis) across each condition

Group	Measure	Within participant		Between participant	
		Within image	Between image	Within image	Between image
Grid-based	String edit	28.41 (4.29)	30.16 (4.21)	29.95 (2.62)	31.29 (2.60)
	ScanMatch	0.40 (0.04)	0.33 (0.03)	0.35 (0.02)	0.31 (0.02)
Sample-based	Overlap	0.11 (0.02)	0.07 (0.01)	0.09 (0.01)	0.06 (0.005)
	Correlate	0.03 (0.05)	0.01 (0.01)	0.02 (0.02)	0.005 (0.005)
	Gaze shift	0.46 (0.08)	0.45 (0.07)	0.46 (0.04)	0.44 (0.04)
Direct measures	Linear distance	55.72 (7.71)	42.31 (5.22)	49.13 (3.76)	39.51 (3.68)
	MM vector	0.89 (0.01)	0.88 (0.01)	0.89 (0.01)	0.88 (0.005)
	MM direction	0.78 (0.02)	0.76 (0.02)	0.77 (0.01)	0.76 (0.01)
	MM length	0.91 (0.01)	0.90 (0.01)	0.90 (0.01)	0.90 (0.01)
	MM position	0.82 (0.02)	0.80 (0.02)	0.81 (0.01)	0.79 (0.01)
	MM duration	0.66 (0.05)	0.66 (0.04)	0.64 (0.03	0.64 (0.03)
Recurrence-based	Recurrence	6.24 (2.02)	3.27 (0.71)	4.95 (1.04)	2.82 (0.40)
	Determinism	26.56 (3.92)	17.62 (3.61)	20.69 (1.97)	16.69 (1.21)
	Laminarity	34.76 (7.21)	30.05 (7.65)	32.33 (1.18)	29.10 (0.84)
	Corm	2.04 (3.12)	2.56 (5.09)	1.48 (2.60)	2.66 (2.95)

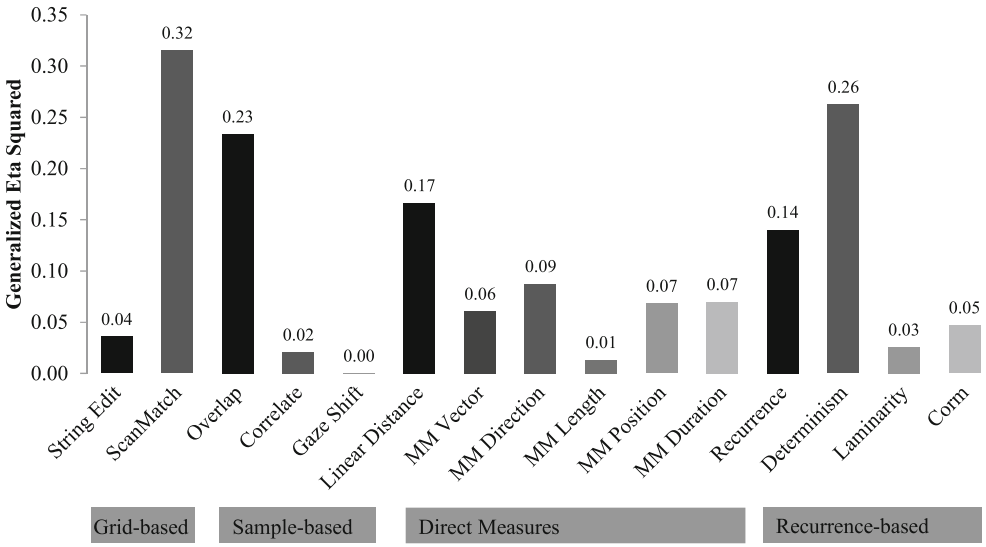

Fig. 1 η^2_G values for detecting the main effect of participant for each of the scanpath comparison measures

0.26) and the corm measure (cell means computed using the absolute value of corm, rather than an average; $\eta^2_G = 0.05$) revealed a main effect of participant, but not laminarity.

Main effect of image

Almost all of the scanpath comparison techniques show a significant main effect of image, with very high effect sizes (see Fig. 2). One notable exception is the MultiMatch duration measure. interestingly, gaze shift, string edit, MultiMatch direction and length, as well as the recurrence-based laminarity and CORM measures show a significant effect of image, but a below-medium effect size.

Participant by image interaction

The interaction between participant and image usually indicates that the similarity between the scanpaths of the same viewer viewing the same image is particularly high relative to the other comparisons (see Table 3 and Fig. 4). Generally η^2_G is relatively low for all of the measures, with the exception of the determinism measure ($\eta^2_G = 0.16$) and the overlap measure ($\eta^2_G = 0.11$).

Post-hoc paired comparisons were performed for all significant interactions, i.e., for overlap, string edit, linear distance, ScanMatch, recurrence and determinism. Cohen's d values are reported for all comparisons.

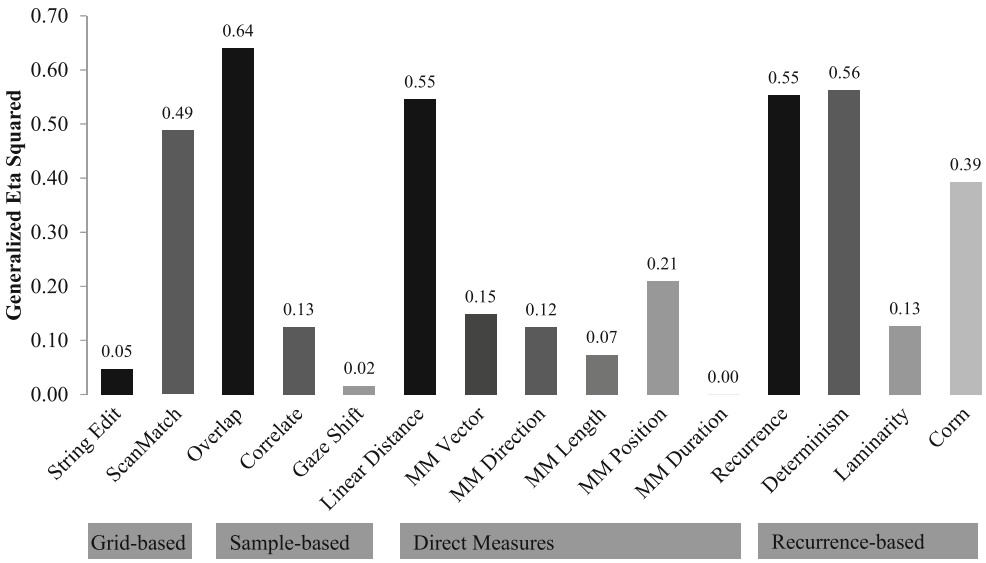

Fig. 2 η^2_G values for detecting the main effect of image for each of the scanpath comparison measures

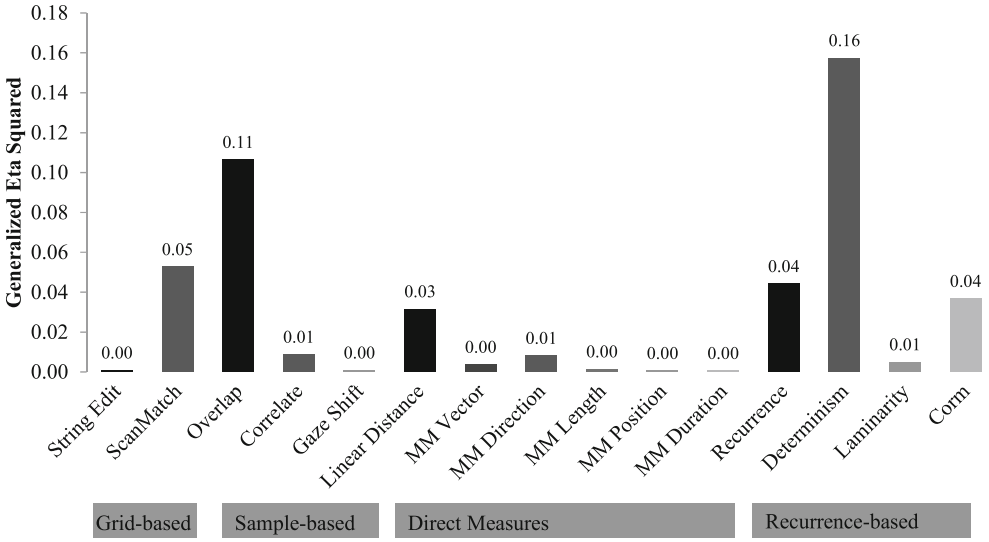

Fig. 3 r^2_G values for detecting the interaction between participant and image for each of the scanpath comparison measures

Overlap Within-participant, within-image similarity ($M = 0.11$) was significantly higher than between-participant, within-image similarity ($M = 0.09$), $t(26) = 7.00$, $p < .001$, $d = 1.37$, within-participant, between-image similarity ($M = 0.07$), $t(26) = 11.45$, $p < .001$, $d = 2.50$, and between-participant, between-image similarity ($M = 0.06$), $t(26) = 12.47$, $p < .001$, $d = 3.02$.

String edit distance Within-participant, within-image similarity ($M = 28.41$, note that in string-edit distance, smaller numbers mean higher similarity) was significantly higher than between-participant, within-image similarity ($M = 29.95$), $t(26) = 3.96$, $p < .001$, $d = 0.43$, within-participant, between-image similarity ($M = 30.16$), $t(26) = 13.10$, $p < .001$, $d = 0.41$, and between-participant, between-image similarity ($M = 31.29$), $t(26) = 7.18$, $p < .001$, $d = 0.81$.

Linear distance Within-participant, within-image similarity ($M = 55.72$) was significantly higher than between-

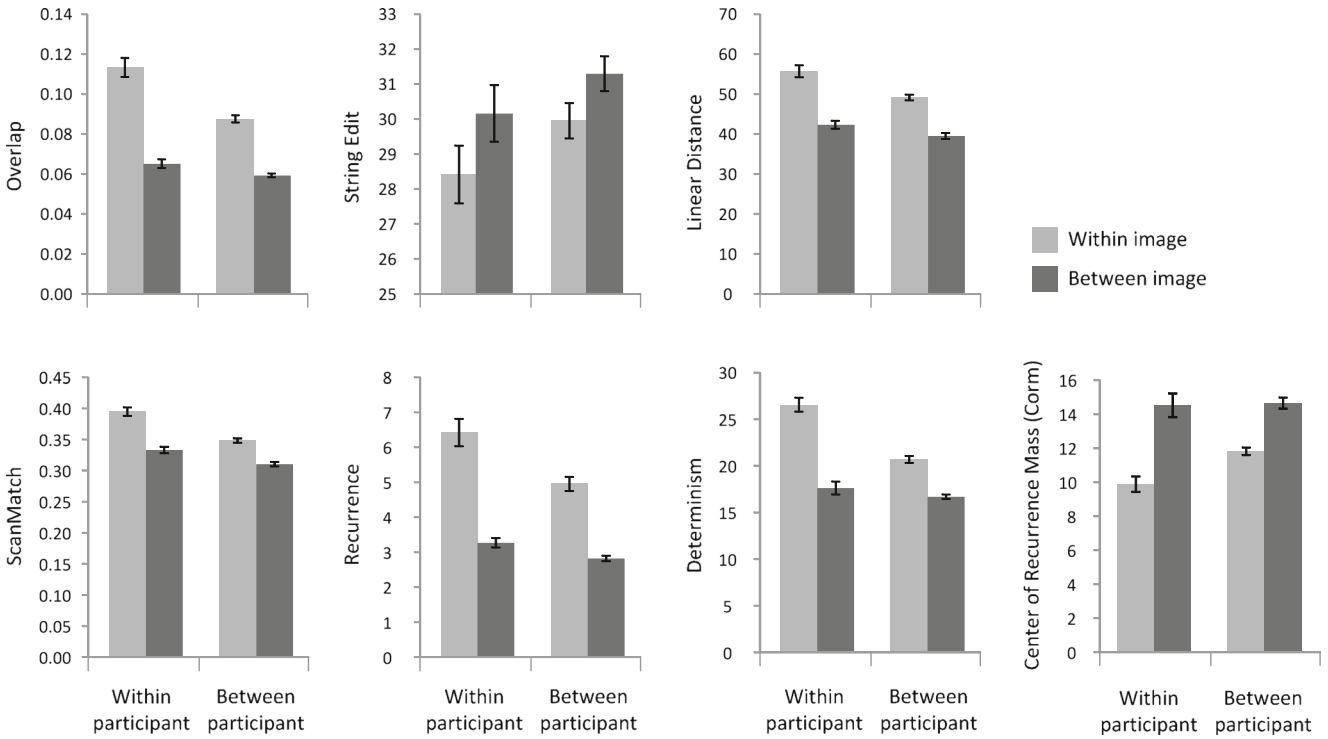

Fig. 4 Interaction between participant and image for the six measures with a significant interaction. Within-image, within-participant similarity is generally higher than all other comparison means. Note that string edit uses a reversed scale compared to the other measures (lower numbers indicate higher similarity)

participant, within-image similarity ($M = 49.13$), $t(26) = 5.57$, $p < .001$, $d = 1.09$, within-participant, between-image similarity ($M = 42.31$), $t(26) = 9.97$, $p < .001$, $d = 2.04$, and between-participant, between-image similarity ($M = 39.51$), $t(26) = 12.33$, $p < .001$, $d = 2.68$.

ScanMatch Within-participant, within-image similarity ($M = 0.39$) was significantly higher than between-participant, within-image similarity ($M = 0.35$), $t(26) = 7.92$, $p < .001$, $d = 1.63$, within-participant, between-image similarity ($M = 0.33$), $t(26) = 11.95$, $p < .001$, $d = 1.96$, and between-participant, between-image similarity ($M = 0.31$), $t(26) = 13.60$, $p < .001$, $d = 2.96$.

Recurrence Within-participant, within-image similarity ($M = 6.42$ %) was significantly higher than between-participant, within-image similarity ($M = 4.95$ %), $t(26) = 5.08$, $p < .001$, $d = 0.97$, within-participant, between-image similarity ($M = 3.27$ %), $t(26) = 11.69$, $p < .001$, $d = 1.96$, and between-participant, between-image similarity ($M = 2.82$ %), $t(26) = 10.71$, $p < .001$, $d = 2.47$.

Determinism Within-participant, within-image similarity ($M = 26.56$ %) was significantly higher than between-participant, within-image similarity ($M = 20.69$ %), $t(26) = 10.21$, $p < .001$, $d = 1.89$, within-participant, between-image similarity ($M = 17.62$ %), $t(26) = 9.85$, $p < .001$, $d = 2.37$, and between-participant, between-image similarity ($M = 16.69$ %), $t(26) = 14.43$, $p < .001$, $d = 3.40$.

Corm Within-participant similarity ($M = 9.87$) is significantly lower than between-participant, within-image similarity ($M = 11.82$), $t(26) = 4.88$, $p < .001$, $d = 1.05$, within-participant, between-image similarity ($M = 14.52$), $t(26) = 6.74$, $p < .001$, $d = 1.53$, and between-participant, between-image similarity ($M = 14.64$), $t(26) = 10.30$, $p < .001$, $d = 2.35$.

Discussion

In the present work, we performed multiple types of scanpath comparisons on the scanpaths generated from a foundational encoding and recognition experiment. Specifically, we compared sample-based measures (Shepherd et al., 2010), grid-based measures (Bunke, 1992; Cristino et al., 2010; Levenshtein, 1966; Wagner & Fischer, 1974), the direct linear distance measure (Mannan et al., 1995), the measures computed from the MultiMatch algorithms (Dewhurst et al., 2012), and recurrence-based measures (D. C. Richardson & Dale, 2005). This design allowed for the comparison of scanpath similarity across images and participants, focusing on the relative contribution of the individual generating the scanpath and the influence of the stimulus itself on resulting similarities. In addition, this study was designed for the more general comparison of the methods used to compute scanpath similarity, which we compared by computing generalized eta squared. In the following, we review the contributions of each measure to the understanding of scanpath similarity in terms of participant and image similarities and then discuss the performance of each group of measures, with particular emphasis on the type of information they quantify.

Our results support previous work using a similar paradigm where the same person looking at the same image is more likely to have a more similar scanpath (Foulsham & Underwood, 2008; Foulsham et al., 2012). This was borne out in the number of measures that revealed a significant interaction between participant and image, where the within-image, within-participant comparison had the highest similarity scores (see Fig. 4 and Foulsham et al., 2012, Fig. 2). Interestingly, because each scanpath comparison technique differs in the characteristics that it captures, the different scanpath comparison measures may be revealing in terms of the aspects of scanpath similarity that result from within-participant or within-image similarity.

The quantification methods that were most notable in revealing within-participant idiosyncrasies were those that quantified similarity in shape and position: overlap, linear distance, ScanMatch and recurrence. Interestingly, and similar to previous work (Foulsham et al., 2012), measures that also quantified some aspect of sequential order such as ScanMatch and determinism were particularly discriminatory. Beyond the overall similarity in shape and sequential order, the determinism measure revealed that scanpaths are repeated by particular individuals on a more local level, where a person may have an idiosyncratic strategy of repeating particular short sequences of scanpaths in a consistent manner, regardless of the image they are viewing. Taken together, individual differences in scanning behavior reveal themselves in terms of overall scanpath shape and to some extent, the order in which this shape unfolds.

The image had a very strong influence on the similarity of scanpaths. Most measures were able to detect a main effect of image and the resulting η^2_G values were quite high. Most notably, those measures that quantify shape, position and sequential order were again very discriminatory: overlap, linear distance, ScanMatch, recurrence, and determinism. Interestingly, most of the measures that performed well on the main effect of image were those with configurable grid or radius sizes. Although we used author-recommended or previously used radius/grid sizes, an interesting avenue of future research would be to examine to what extent the present results may vary with respect to the radius/grid size chosen.

Within-image, within-participant similarity is the special case where the similarity arises from scanpaths created by the same individual looking at the same image. The overlap and determinism measures showed the highest η^2_G values for

detecting the interaction between image and participant. Follow-up contrasts indeed revealed that this arose from the within-image, within-participant scanpaths being more similar than the other comparisons. Overlap quantifies both the spatial and temporal overlap between fixation sequences that have been resampled to 60Hz. This means that scanpaths were quite similar in both where they landed and how long they stayed there, with particular sensitivity to the duration in which the eyes were at any given position. Determinism quantifies the repetition of small segments of fixation trajectories that occur in a particular order. This means that individuals, when viewing the same image again, reinstate local patterns of their scanpath behavior. For example, if a participant looked at a book, then at a vase on a shelf, upon viewing that image again, that short book-to-vase sequence was likely to be repeated. An example of this behavior in the present work is shown in Fig. 5. The significant interaction in the CORM measure is interesting, because it indicates that when participants view the same image for the second time, their pattern of cross-recurrences occurs close in timing to the earlier viewing. This provides converging evidence that both spatial and sequential aspects of the scanpath are preserved upon a second viewing.

Relation to previous work

In the present work we broadly replicated Foulsham and Underwood (2008), where scanpaths were compared using the string-edit distance algorithm. That data was also recently used in the evaluation of the MultiMatch methods (Foulsham et al., 2012). Scanpath similarity was generally highest for the same participant looking at the same image again (arising from the significant interaction between image and participant) than for different participants looking at the same image. We did not fully replicate these findings in our experiment when using the MultiMatch measures. For example, we did not observe a main effect of image for the duration measure nor an interaction between image and participant for the vector, direction and position measures. One reason for these differences may be the difference in viewing time. We used a viewing time of 10 s for both encoding and recognition trials compared to the 3 s used in Foulsham and Underwood (2008). To better understand the difference, we re-analyzed our MultiMatch data using the first 3 s of viewing time. Now only the interaction between image and participant for the direction measure was significant, $F(1, 26) = 6.36$, $MSE = .001$, $p < .05$, suggesting that viewing strategies change over time as scene information is acquired. Importantly, when looking at the paired-samples t-tests between conditions for the MultiMatch measures in our data, the particular comparison between the same person looking at the same image, and different participants looking at the same image mirrored the paired comparisons of Foulsham et al., (2012) for this contrast except for the length measure (vector, direction, position, and duration t's$(26) > 2.9$, p's $< .01$).

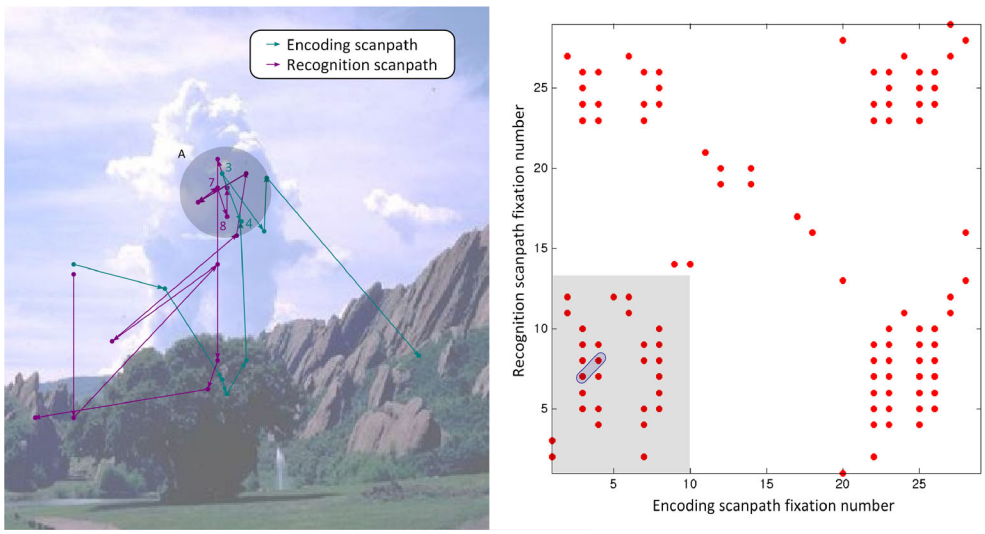

Fig. 5 Left panel: A sequence of eye movements from a subset of fixations in an encoding and recognition trial. The sequences of fixations in the shaded area A in the left panel create patterns of determinism (diagonal lines) in the cross-recurrence plot. Right panel: A cross-recurrence plot of the fixations in an encoding and recognition trial from the same participant (within-image, within-participant scanpaths). The red dots represent recurrences, where fixation locations between the scanpaths fall within 1.9 degrees visual angle of each other. The grey shaded area is the section of the plot that is created from the subset of fixations shown in the left panel, the blue shaded region is an example of a diagonal line of recurrences that create a deterministic sequence. It is the result of the sequence of fixations 3 and 4 in the encoding scanpath being repeated at fixations 7 and 8 of the recognition scanpath. The percentage of recurrent points in the cross-recurrence plot that forms these diagonal lines is the Determinism measure used in the present work

Choice of radius/grid size

In the present work, we were concerned with how the measures compared when they were computed using author-recommended or previously-used parameters. However, the choice of radius size or granularity of grid is extremely important. As radius size increases, so does the chance that spatially sensitive scanpath measures are considered similar. As radius is reduced, the chances of an overlap are reduced. The Overlap method was an extremely successful scanpath comparer in this dataset, but uses a much larger radius compared to the recurrence-based measures, which also performed quite well. This could reflect the fact that the Overlap measure is reflecting some overlap in time that the recurrence-based measures are not, but it could also mean that the radius was simply bigger and therefore, overlaps in space were more likely to be found. It is our speculation that there may be a fine line between real scanpath similarity and spurious similarity due to radius size. One interesting avenue for future investigation would be to run scanpath comparisons on multiple radius and grid sizes and compare the results along with effect sizes. The general recommendation, based on previous work (Anderson et al., 2013; Dewhurst et al., 2012) is to choose a sensible radius size that reflects roughly the size of region of foveal or parafoveal vision. However, how the choice of radius and grid size impacts scanpath quantification and comparison is another fruitful avenue of future research.

Choice of scanpath measure

The findings of the present investigation reveal that there are several issues to keep in mind when choosing a particular scanpath measure. First, the choice must be hypothesis-driven. If, for example, the aim of an experiment is to determine whether two scanpaths are similar in terms of their duration, then the sample-based methods may be preferred; however if sequential order is not important, then a good choice would be to use the MultiMatch duration measure. If fixation sequence is paramount, good options are the ScanMatch, correlate and determinism methods that quantify differences in sequence. The reason why the sample-based methods were appropriate in their original context (Shepherd et al., 2010) was partially because the participants were watching movies. In this scenario, eye movements may be much more strongly driven by the stimulus (and therefore scanpaths will be more similar both within and between participants; Dorr, Martinetz, Gegenfurtner, and Barth, 2010), requiring the employment of more sensitive measures. In addition, spatial similarity at different time points is actually less informative because the interesting points in the video are more likely to move or change. One important characteristic of the ScanMatch method is that it is possible to specify particular relationships between areas of interest in a scene, such that overlaps between these regions are given a stronger weight. In this manner, it is possible to test a prediction that relates specific scene items to scanpath sequences. For example, in a social attention experiment, one might predict that observers will move their eyes between the people in an image in a specific sequence, or between the eyes and mouth in a picture of a face. Using ScanMatch, it is possible to test this prediction directly using areas of interest. For a detailed discussion of hypothesis-driven techniques for selecting a scanpath measure, see Dewhurst et al. (2012).

Second, it is important to keep in mind the various requirements of each method. For example, some methods, such as the recurrence-based measures, require that the scanpaths be trimmed to the shorter of the two scanpath lengths. This may result in a loss of some data. Certain methods, such as the sample-based methods and ScanMatch may require the data to be resampled. Thus, fixations and saccades are no longer relevant units. For example, the sample-based methods work directly on the sample-level data from the eye-tracker which are then re-sampled to 60 Hz. In comparing the resulting scanpaths, the specifics of fixations and saccades are lost. Some methods encourage the simplification of data, although this is not a requirement. For example, in our implementation of MultiMatch, angular differences between saccades smaller than 45° were collapsed. Simplification is typically performed in order to increase processing speed. This may be desirable when many comparisons need to be computed (especially for the computation-intensive scanpath comparison measures), but is by no means absolutely necessary. Investigating the effect of simplification on subsequent scanpath comparisons is a fruitful avenue of future research.

Finally, a very important consideration in choosing a scanpath comparison measure is the form of quantization the method requires. Some of the methods, such as string edit and ScanMatch require the dividing of the image into grids. This can potentially be a problem because nearby fixations may be classified into different grid cells. However, there are situations where this may be desirable. Grid-based quantization allows for pre-specifying interest areas. For example, in ScanMatch, as mentioned previously, it is possible to define and quantify specific relationships between these areas of interest. The overlap method and recurrence-based methods require the specification of a radius (although recent work with the recurrence measures is aimed toward direct comparisons). When quantization is direct (i.e., without using a radius or grid), eye movement behaviors are directly compared across scanpaths. This is a significant advantage of the MultiMatch and sample-based methods.

Although, as outlined above, the choice of scanpath measure is highly dependent on the research question, we have a few general recommendations. Scanpath comparison has evolved over many decades of research in eye movement behavior, but perhaps the most striking improvements have occurred most recently. ScanMatch is a remarkable improvement on more simple methods such as string-edit and linear distance and, as mentioned previously, is one of the few methods that naturally includes semantic information as part of the scanpath similarity score. MultiMatch is an excellent example of a method that is robust, easy to use, highly intuitive and freely available. Although cross-recurrence has only recently been developed for eye movements, we feel that it provides exciting opportunities to understand eye movement behavior beyond a single similarity score. The results from many of the cross-recurrence measures, such as determinism, for example, are simple percentages of fixations that overlap in a trial and are thus directly interpretable. These most recent contributions represent, in our opinion, the state-of-the-art in scanpath comparison techniques. However, these and many other scanpath comparison techniques are still under active development.

Resources

One impressive aspect of the methods considered in the present work is that they are freely available, either upon request from the authors or directly online. Some measures are even accompanied by excellent user interfaces along with tutorials for their use (most notably, ScanMatch and MultiMatch), reducing the barrier to using these methods. Below is a list of URL's where, as of this writing, those methods that are available online can be found.

- String edit: A general version has been implemented in many programming languages at Rosetta code: http://rosettacode.org/wiki/Levenshtein_distance
- ScanMatch: Matlab code can be found here: http://seis.bris.ac.uk/~psidg/ScanMatch/ and a version is already implemented as part of a more general eye movement analysis package, GazeParser (Soho, 2013): http://gazeparser.sourceforge.net/index.html
- Linear distance: An updated version of this analysis written in Python by Mathot, Cristino, Gilchrist, and Theeuwes (2012), can be downloaded here: http://www.cogsci.nl/eyenalysis.html
- MultiMatch: Matlab code for MultiMatch can be found here: http://wiki.humlab.lu.se/dokuwiki/doku.php?id=public:useful_links#scanpath_comparison
- Cross-recurrence analysis: Tutorials and code for cross-recurrence and other analysis techniques can be found here: http://ecem2013.eye-movements.org/workshops/eye-movements-space-and-time

Conclusion

In the present work, we compared some commonly available scanpath comparison measures by providing an overview of the each measure and then applying it to a single data set. We compared the measures based on the aspects of scanpaths that they quantify and on how well they performed in revealing differences in scanning behavior across participants and images. This analysis provides a framework with which other researchers can use to determine the most suitable comparison technique that is most appropriate for their application. We hope that this overview and these results will help those interested in studying both the spatial and sequential aspects of eye movement behavior navigate the myriad of scanpath comparison measures.

Appendix

Temporal correlation

Shepherd et al. (2010) introduced temporal correlation as a scanpath similarity measure. The scanpaths are first resampled uniformly in time, e.g., at 60 Hz, and truncated to the shorter of the two paths. The temporal correlation is defined as the average T_c of the correlation between the x-coordinates and between the y-coordinates of two scanpaths $\boldsymbol{f}$ and $\boldsymbol{g}$:

$$T_c = \left(corr(f_x, g_x) + corr\left(f_y, g_y\right)\right)/2$$

Gaze shift

Shepherd et al.'s (2010) gaze shift measure is computed as the correlation between the absolute values of the first derivative of each scanpath. The first derivative is computed by convolving each scanpath with the derivative of a Gaussian filter. Gaze shift is then computed in the same manner as the temporal correlation, but using the first derivative instead of the position.

$$G = \left(corr(|\mathrm{f'_x}|, |\mathrm{g'_x}|) + \mathrm{corr}\left(|\mathrm{f'_y}|, |\mathrm{g'_y}|\right)\right)/2$$

Linear distance

Given two scanpaths $\boldsymbol{f}$ and $\boldsymbol{g}$ with n_1 and n_2 fixations, Mannan, Ruddock, and Wooding (1996), compute the distance d_{1i} between the ith fixation $\boldsymbol{f}_i$ and its nearest neighbor fixation in $\boldsymbol{g}$ and the distance d_{2j} between the jth fixation $\boldsymbol{g}_j$ with its nearest neighbor fixation in $\boldsymbol{f}$. Then the similarity S of the scanpaths $\boldsymbol{f}$ and $\boldsymbol{g}$ is defined as

$$S = 100\left(1 - \frac{D}{D_r}\right)$$

With

$$D^2 = \left(n_1 \sum_{j=1}^{n_2} d_{2j}^2 + n_2 \sum_{i=1}^{n_1} d_{1i}^2 \right) / \left(2n_1 n_2 \left(w^2 + h^2 \right) \right)$$

and w and h the width and height of the images. D_r is the same as D but with randomly generated scanpaths.

Cross-recurrence analysis

Consider two fixation sequences $\boldsymbol{f}_i$, $i = 1,\ldots,N$, with $\boldsymbol{f}_i = <x_i, y_i>$ and $\boldsymbol{g}_i$, $i = 1,\ldots,N$, with $\boldsymbol{g}_i = <x_i, y_i>$. For fixation sequences of unequal length, the long sequence is truncated. Two fixations $\boldsymbol{f}_i$ and $\boldsymbol{g}_j$ are cross-recurrent if they are close together, i.e., we define the cross-recurrence of two fixations c_{ij} as

$$c_{ij} = \begin{cases} 1, & d\left(f_i, g_j\right) \leq \rho \\ 0, & otherwise \end{cases}$$

where d is the Euclidian distance, and ρ is a given threshold.

Several measures can be used for characterizing cross-recurrence patterns. The measures are extensions of the recurrence measures introduced by Anderson et al. (2013). Let C be the sum of recurrences, i.e., $C = \sum_{i=1}^{N} \sum_{j=1}^{N} c_{ij}$. Further, let D_L be the set of diagonal, H_L the set of horizontal, and V_L the set of vertical lines in the cross-recurrence matrix, all with a length of at least L, and let $|\cdot|$ denote cardinality.

Cross-recurrence The cross-recurrence measure of two fixation sequences is defined as

$$\mathrm{REC} = 100 \cdot \frac{C}{N^2}$$

It represents the percentage of cross-recurrent fixations, i.e., the percentage of fixations that match (are close) between the two fixation sequences.

Determinism The determinism measure is defined as

$$DET = 100 \cdot \frac{|D_L|}{C}$$

It measures the percentage of cross-recurrent points that form diagonal lines and represents the percentage of fixation trajectories common to both fixation sequences. That is, it quantifies the overlap of a specific sequence of fixations, preserving the sequential information. The minimum line length of diagonal line elements was set to $L = 2$.

Laminarity The laminarity measure is defined as

$$\mathrm{LAM} = 100 \cdot \frac{|H_L| + |V_L|}{2C}$$

Laminarity represents locations that were fixated in detail in one of the fixation sequences, but only fixated briefly in the other fixation sequence. Again, we set the minimum line lengths of vertical and horizontal lines to $L = 2$.

Center of recurrence mass Finally, the center of recurrence mass (CORM) is defined as the distance of the center of gravity from the main diagonal, normalized such that the maximum possible value is 100.

$$CORM = 100 \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} (j-i) c_{ij}}{(N-1)C}$$

The CORM measure indicates the dominant lag of cross-recurrences. Small CORM values indicate that the same fixations in both fixation sequences tend to occur close in time, whereas large CORM values indicate that cross-recurrences tend to occur with either a large positive or negative lag.

References

Anderson, N. C., Bischof, W. F., Laidlaw, K. E., Risko, E. F., & Kingstone, A. (2013). Recurrence quantification analysis of eye movements. *Behavior research methods, 45*(3), 842–856.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior research methods, 37*(3), 379–384.

Box, G. E., Jenkins, G. M., & Reinsel, G. C. (2013). Time series analysis: forecasting and control: John Wiley & Sons.

Brandt, S. A., & Stark, L. W. (1997). Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience, 9*(1), 27–38.

Bunke, U. (1992). Relative Index Theory. *Journal of Functional Analysis, 105*(1), 63–76.

Burmester, M., & Mast, M. (2010). Repeated web page visits and the scanpath theory: A recurrent pattern detection approach. *Journal of Eye Movement Research, 3*(4), 5.

Cherubini, M., Nüssli, M., & Dillenbourg, P. (2010). This is it!: Indicating and looking in collaborative work at distance. *Journal of Eye Movement Research, 3*(3), 1–20.

Cristino, F., Mathôt, S., Theeuwes, J., & Gilchrist, I. D. (2010). ScanMatch: A novel method for comparing fixation sequences. *Behavior Research Methods, 42*(3), 692–700.

Dale, R., Kirkham, N. Z., & Richardson, D. C. (2011a). The dynamics of reference and shared visual attention. *Interfaces Between Language And Cognition, 103*.

Dale, R., Warlaumont, A. S., & Richardson, D. C. (2011b). Nominal cross recurrence as a generalized lag sequential analysis for behavioral streams. *International Journal of Bifurcation and Chaos, 21*(04), 1153–1161.

Dewhurst, R., Nystrom, M., Jarodzka, H., Foulsham, T., Johansson, R., & Holmqvist, K. (2012). It depends on how you look at it: scanpath

comparison in multiple dimensions with MultiMatch, a vector-based approach. *Behav Res Methods, 44*(4), 1079–1100.

Dorr, M., Martinetz, T., Gegenfurtner, K. R., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision, 10*(10), 28.

Foulsham, T., & Kingstone, A. (2013). Fixation-dependent memory for natural scenes: an experimental test of scanpath theory. *J Exp Psychol Gen, 142*(1), 41–56.

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision, 8*(2).

Foulsham, T., Dewhurst, R., Nystrom, M., Jarodzka, H., Johansson, R., Underwood, G., et al. (2012). Comparing scanpaths during scene encoding and recognition: A multi-dimensional approach. *Journal of Eye Movement Research, 5*(4).

Harding, G., & Bloj, M. (2010). Real and predicted influence of image manipulations on eye movements during scene recognition. *Journal of Vision, 10*(2).

Hasson, U., Yang, E., Vallines, I., Heeger, H. J., & Rubin, N. (2008). A hierarchy of temporal receptive windows in human cortex. *Journal of Neuroscience, 28*(10), 2539–2550

Henderson, J. M., Brockmole, J. R., Castelhano, M. S., & Mack, M. (2007). Visual saliency does not account for eye movements during visual earch in real-world scenes. *Eye movements: A window on mind and brain, 537–562*.

Jarodzka, H., Holmqvist, K., & Nyström, M. (2010). A vector-based, multidimensional scanpath similarity measure. Paper presented at the Proceedings of the 2010 Symposium on Eye-Tracking Research and Applications.

Jeannerod, M., Gerin, P., & Perrier, J. (1968). Déplacements et fixations du regard dans l'exploration libre d'une scène visuelle. *Vision Research, 8*(1), 81–97.

Johansson, R., Holsanova, J., & Holmqvist, K. (2006). Pictures and spoken descriptions elicit similar eye movements during mental imagery, both in light and in complete darkness. *Cognitive Science, 30*(6), 1053–1079.

Johansson, R., Holsanova, J., & Holmqvist, K. (2011). The dispersion of eye movements during visual imagery is related to individual differences in spatial imagery ability. Paper presented at the Expanding the space of cognitive science: Proceedings of the 33rd Annual Meeting of the Cognitive Science Society.

Josephson, S., & Holmes, M. E. (2002). Attention to repeated images on the World-Wide Web: Another look at scanpath theory. *Behavior Research Methods, Instruments, & Computers, 34*(4), 539–548.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions and reversals. Paper presented at the Soviet physics doklady.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1995). Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-D images. *Spatial Vision, 9*(3), 363–386.

Mannan, S. K., Ruddock, K. H., & Wooding, D. S. (1996). The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision, 10*(3), 165–188.

Mathot, S., Cristino, F., Gilchrist, I. D., & Theeuwes, J. (2012). A simple way to estimate similarity between pairs of eye movement sequences. *Journal of Eye Movement Research, 5*(1).

Noton, D., & Stark, L. (1971). Scanpaths in Saccadic Eye Movements While Viewing and Recognizing Patterns. *Vision Research, 11*(9), 929–942.

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological methods, 8*(4), 434.

Richardson, J. T. (2011). Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review, 6*(2), 135–147.

Richardson, D. C., & Dale, R. (2005). Looking to understand: The coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension. *Cognitive science, 29*(6), 1045–1060.

Richardson, D. C., Dale, R., & Tomlinson, J. M. (2009). Conversation, gaze coordination, and beliefs about visual context. *Cognitive Science, 33*(8), 1468–1482.

Shepherd, S. V., Steckenfinger, S. A., Hasson, U., & Ghazanfar, A. A. (2010). Human-monkey gaze correlations reveal convergent and divergent patterns of movie viewing. *Current Biology, 20*(7), 649–656.

Shockley, K., & Turvey, M. T. (2005). Encoding and retrieval during bimanual rhythmic coordination. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(5), 980.

Shockley, K., Santana, M.-V., & Fowler, C. A. (2003). Mutual interpersonal postural constraints are involved in cooperative conversation. *Journal of Experimental Psychology: Human Perception and Performance, 29*(2), 326.

Shockley, K., Baker, A. A., Richardson, M. J., & Fowler, C. A. (2007). Articulatory constraints on interpersonal postural coordination. *Journal of Experimental Psychology: Human Perception and Performance, 33*(1), 201.

Shockley, K., Richardson, D. C., & Dale, R. (2009). Conversation and coordinative structures. *Topics in Cognitive Science, 1*(2), 305–319.

Sogo, H. (2013). GazeParser: an open-source and multiplatform library for low-cost eye tracking and analysis. *Behavior Reserch Methods, 45,* 684–695. doi:10.3758/s13428-012-0286-x

Underwood, G., Foulsham, T., & Humphrey, K. (2009). Saliency and scan patterns in the inspection of real-world scenes: Eye movements during encoding and recognition. *Visual Cognition, 17*(6–7), 812–834.

Wagner, R. A., & Fischer, M. J. (1974). The string-to-string correction problem. *Journal of the ACM (JACM), 21*(1), 168–173.

Webber, C. L., Jr., & Zbilut, J. P. (2005). Recurrence quantification analysis of nonlinear dynamical systems. *Tutorials in contemporary nonlinear methods for the behavioral sciences, 26–94*.

West, J. M., Haake, A. R., Rozanski, E. P., & Karn, K. S. (2006). eyePatterns: software for identifying patterns and similarities across fixation sequences. Paper presented at the Proceedings of the 2006 symposium on Eye tracking research & applications.

Yarbus, A. L. (1967). Eye Movements and Vision: Springer-Verlag US.