Executive summary

We developed a prototype system used for physical matching in 2D. The system has two main functions: one is to assist forensic experts in performing physical matching in an objective manner, and the second to collect statistics and build confidence levels regarding the physical matches.

The first function receives input images of broken or torn pieces photographed on distinctive background. The system then extracts the contour of each piece and lets the user choose a segment on one of the contours as a target for matching. The system then runs a search on the other contours, finds potential matches and orders these candidates according to the quality of their fit with the target. Using the statistics from the second function the user can estimate the potential error rates of each candidate matching pair.

The second function takes as an input multiple images of broken or torn pieces and then generates many examples of physical matching that are classified into two populations: the correct matching pairs and the non-matching pairs. The system gathers the frequencies of the matching error for the two populations and presents the results as two histograms. The outputs of the system are several statistical measures describing the quality of matching as a function of the matching error value. The user can find the probabilities of a match being right or wrong for any specific value of matching error and thus is able to compare between matching pairs, and find out which is the best evidence, even among different sizes or materials.

We used the second function to collect statistics for different fracture line lengths of three different materials: silicon, metal coated paper and Perspex. The total error rates

(false negative + false positive) for 1 cm matches were 0.007 for the silicon, 0.37 for the paper and 0.4 for the Perspex.

The statistical results were much lower than initially expected. This is because we used only the 2D fracture lines and not any additional clue commonly used in toolmark comparison such as the 3D nature of some fractures or any existing texture and graphic patterns on the surface or outer border of the pieces to be compared. We also classified the pairs into matches and non-matches and ignored a third classification – "inconclusive". For all these reasons, our results are the entering point to the numerical or quantitative evaluation. Adding the surface details or performing a 3D match, can bring us much closer to the results achieved by toolmark experts.

In presenting results of scientific nature to court, the law system in many countries, demands having a potential error rate for the results. Until today, there wasn't any numerical way to calculate the potential or known error rate, for 2D or 3D comparisons. This project supplies the forensic scientist the essential tools to calculate this crucial information.

The results of this research, experts can express their findings in a more quantitative way. It is in the reach of every scientist, holding this computerized statistics generator, to create a new database for any material. Computing the potential error rate in physical match comparison for objects received in the laboratory becomes possible.

Final progress report

This report has two parts. First we describe the main achievements of this project. Some of the issues were already described in detail in previous reports and are presented here in a shorter version but with the proper perspective of the entire work. Other issues are reported here for the first time, and are fully presented in the second part which also contains discussion and conclusions.

PART I

Background and Introduction:

In forensic literature a physical match is often regarded as conclusive evidence, but the Daubert ruling may change this. Every layman can determine a physical match, and yet the individuality of physical match was never proven scientifically.

All the forensic evidence introduced in Federal courts in the U.S must obey the principles determined by the U.S Supreme Court in the Daubert ruling (Daubert vs. Merrel Dow Pharmaceuticals, INC. Supreme Court of the United States, June 1993). Similar criteria are now becoming mandatory in many western countries.

In brief, the Daubert criteria consist of four independent principles. First, whether or not, the theory or technique in question can be (and has been) tested? Second, was the theory published in peer-reviewed magazines? Third, has the theory gained widespread acceptance within the relevant scientific community? And fourth, are there existing standards controlling the method operation, and the known or potential error rate of the method?

In order to meet the Daubert challenge, one must show that this evidence, and its interpretation, fulfills all the Daubert requirements. However, this all starts from defining the proposed technique, and reporting on its error rate. The goal of this project is to try and develop a method for automatic physical matching, which will enable calculating the uniqueness of the tear and the error rate.

To achieve these goals we followed 13 milestones:

1. Literature review to test matching algorithms

The aim of this part was to check what is known in the scientific community about matching algorithms, contour representations, fracture line characterization and the information content of fracture lines. Here we cite only the most relevant publications; a more detailed review is in report 1. Several groups developed systems that are able to perform matching of pieces. Leitão and Stolfi (2002) developed a multiscale method for the reassembly of two-dimensional fragmented objects. De

Bock and De Smet (2004) presented a practical, semi-automatic approach of performing reconstruction of fragmented 2-D objects. Zhu et al. (2006) described a method for finding candidate matching pairs that is based on the statistics of differences between matching contours. These three groups used similar contour representation known as chain code or curvature curve.

Another important issue studied by Leitão and Stolfi (2002) was the amount of useful information contained in a piece of fragment outline of a given length. Using their estimate of useful information they calculated the probability of two random ceramics fragments, 10.8 mm long being indistinguishable from a true match to be about $1/2^{22} \approx 1/4{,}000{,}000$ (for comparison with our results see PART II, section 6.5).

2. Choice of materials for the experiment and tearing them under controlled conditions, documenting contours of all the fracture lines and building data bases of the torn contours for the materials chosen

This was started in the previous report. It was fully performed here for the three databases of silicon, paper and Perspex. See PART II, section 1.

3. Developing contour representations (rotation invariant)

Building upon the literature we tested several contour representations. The major one we used was the chain code which is a discrete representation of the angle changes along the contour, thus being an approximation to the curvature of the contour line. This easy to compute representation is rotation and translation invariant (i.e. it does not change when the contour is rotated or moved to different locations). On the downside, this representation is computed using derivatives and hence amplifies noise. To overcome this we used smoothing of the chain code.

4. Picture analysis, segmentation of pieces from background to retrieve optimal fracture line

We used a combination of tools including different capturing methods, Photoshop and Matlab to retrieve the fracture lines. The illumination angle during photographing proved critical for viewing the contour on the downside. We tried painting the upper layer in black to increase the contrast, but the separation between the colors was not discreet. A method that proved very effective, revocable and easy to apply was scanning in a scanner with the lid open. We enhanced the contrast between the pieces

and the background by marking the pieces using Photoshop's magic wand tool and painting over the marked regions. Then using our custom software we converted the images to black & white, removed small holes and small grains, segmented the pieces from background and extracted the contour.

5. Applying contour representation to the fracture line

After retrieving the fracture line we applied the chain code representation by first calculating the global angle code which is the direction of the contour relative to a fixed coordinate system, as a function of position along the contour. The local angle code (the chain code) was then calculated by derivation of the global code.

6. Defining correlation coefficient as a function of fracture line length and complexity

We tested several measures of similarity between two vectors of values (each being a local angle representation of a piece of contour): sum of the squares of the difference (SSD), standard deviation of the differences (STD) and two versions of normalized cross-correlation. We found that the normalized SSD measure outperformed the others and therefore we based our matching algorithm on this measure. The values of the similarity measure depended on the length of the fracture line and its complexity in non-trivial ways (see PART II, section 5.5 and 6.4).

7. Performing physical matching between the pieces

This was done by measuring the similarity between one fracture line (the target) and others (the candidates). All similarity measures were calculated between 1D vectors, each being a local angle representation of a piece of contour. The different possible matches were ordered by their similarity with the target and the one with highest similarity measure was marked as the match. The match was then converted back to a 2D contour and was aligned to the original 2D target (see PART II, section 2).

8. Filtering outliers

In some cases the match was not the candidate with the highest similarity measure but one of the others (e.g. the 2nd or 3rd best). We found that checking the similarity on the original 2D contours was much more robust, although it took much longer to compute (see PARTII, section 3). We then decided to build the system using the two

measures. First the system used the 1D measure to quickly find a list of possible matches. Then for each of the candidates the system used the more accurate 2D measure (see PART II, section 2 and 3).

9. Run system for all pictures and sub-pictures

This was started in the previous report for a portion of the silicon database. It was fully performed here for the three databases of silicon, paper and Perspex. See PART II, section 4.2.5.

10. Derive confidence interval statistics (as a function of length and complexity)

This is the major part of this report. See PART II, section 4.

11. Develop a rudimentary user interface

We developed graphical user interfaces (GUIs[1]) for different parts of the system. One GUI, developed for the part that performs the physical matching, is used for selecting the foreground and background of the pieces, and for marking parts of the contours as targets and others as the range for performing candidate search (PART II, section 2). For the second part, the one that gathers the statistics of matching we developed a GUI for marking known matching points along the fracture lines of matching pieces (PART II, section 4.2.3).

12. Fine tune system

We checked that the different parts work together.

13. Summarize conclusions

See PART II, section 6.

[1] Graphical User Interface - abbreviated as GUI is the part of the programs that allows people to interact with the software. It is composed of input elements such as menus and buttons that receives human commands and a visual display of the results of the system that acts according to the user commands.

PART II

1. The research databases

Three materials were used; two of them were described in previous report:

1. Metal coated paper with coating consisting mainly of titanium, calcium, silicon and iron[2]. The paper has a uniform thickness of 0.9 ± 0.01 mm. The tearing process was performed in the production direction. Twenty-four samples were prepared from the metal-coated paper. All samples were cut into 250X50 mm^2 strips.
2. The silicon polymers were self-leveling and prepared on a flat-leveled glass plate. White silicon cast (Epusil RTV 111 + CAT 12 Polymer G'vulot Ltd.), hardness of 40A shor, a relatively uniform thickness of 0.83 ± 0.1 mm, and no distinct direction. And red silicon cast (Epusil RTV 113 + Cat 12 Polymer G'vulot Ltd.), hardness of 58A shor, with varying thickness ranging from 0.85 to 2.16 mm and no distinct direction. The polymers (2 and 3) were self-leveling and prepared on a flat-leveled glass plate.

Twelve samples from each type of silicon sheet were prepared. The thickness of some samples was not even, especially samples made from the red silicon. All sheets were cut into 250X50 mm^2 strips.

3. White Perspex plates 1.9 mm thick were cut in twenty two 7.5X15 cm strips that were broken length wise, in the production direction.

The Third material chosen was Perspex, in order to cover the full range of tearing modules. This material chosen behaves in a totally different way than the first two materials. The silicon behaves as rubber, very elastic so the tearing is not continuous and the paper is ductile, and its tearing is continuous. The Perspex is brittle and during the force applied the Perspex cracks with typical orientations. The silicones had no distinct

[2] The coating composition was determined by Eagle - Micro XRF.

direction while the metal coated paper and the Perspex are manufactured in a rolling process, and therefore have a distinct direction. The Perspex is somewhat similar to the plastic used in the vehicle industry for lamp/light covers and is more common in real case work than the other materials used in the experiment.

Tearing to the paper and the silicon cast were performed with a Zwick 1435 tensile machine (Fig. 1). This is a device used for plastic, paper, wood, rubber and foils. The program application of the machine can be up to 5 kN. Valuable characteristics such as test speed selection and loading force can be controlled. An output was produced of the applied force vs. the tearing advance. The specimen grips were adapted to the test performed and the sample shape. Tearing stress was according to ASTM D 2240 A.
The samples were torn at a set rate of 100 mm/min. Stress was dependent on the tearing rate and the material thickness and qualities; hence it was different for each material torn.
Flat specimen grips were used to prevent the samples from slipping during the tearing process.

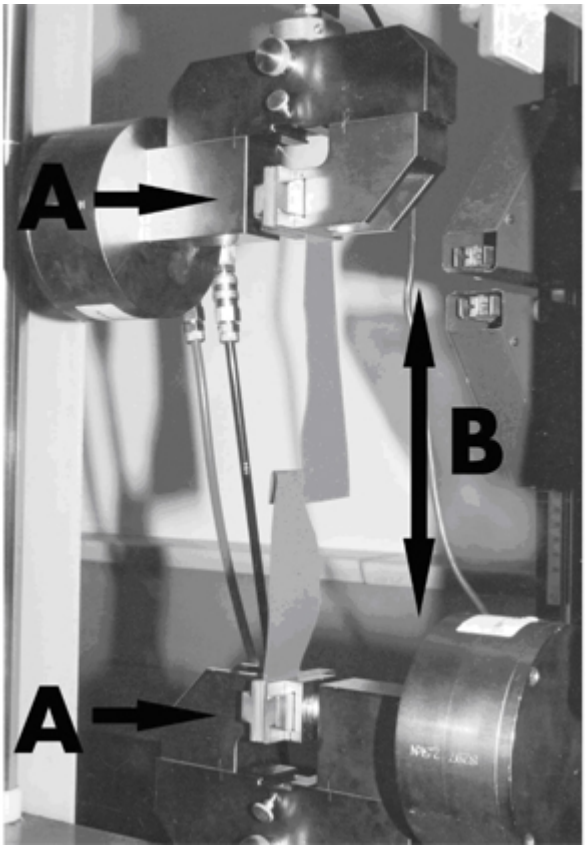


Figure 1. Zwick 1435 tensile machine: A: The specimen grips. B: Movement direction.

The end of each sample was cut 5 cm lengthwise with scissors to assist the clamping by the specimen grips. The cut ends were inserted into the clamps (marked A on Fig. 1) and tightened parallel to the sample, preventing any twisting according to ASTM D 5735-95; hence the clamp movement applied a shearing force perpendicular to the sample (marked B on Fig. 1).

Tensile test was performed to the Perspex with an Instron 6021 machine to break the strips lengthwise. The program application of the machine can be up to 5 kN.

The Perspex samples were torn at a set rate of 10 mm/min. Flat specimen grips were used to prevent the samples from slipping during the elongation process.

The end of each sample was slightly marked lengthwise with a knife to determine the beginning point of the breaking process. The strips were held by clamps on both sides, and the clamps were pulled apart.

All the samples chosen were thin, in order to achieve a nearly two-dimensional tear line, yet the tearing line still had some depth and the illumination conditions were found to be crucial to observe the upper tearing contour. Very oblique lighting at a certain angle revealed the upper contour as a sharp border line, while perpendicular illumination of the sample concealed the height differences, and hence the upper border line did not appear.

Scanning the samples usually produced good results as well with much less effort than photographing. The scanned or photographed pieces were digitally segmented into foreground (the object) and background.

2. Performing physical matching in an objective fashion

Two computerized systems were developed in this project. The first system is designed to help a forensic expert in performing physical matching in an objective fashion and is described in this section. The second system, which is described later, is designed to collect statistics and build confidence levels regarding the physical matches (section 4).

2.1 The first system

In previous reports we developed ways to perform physical matching using a 1D contour representation. Here we describe the GUI for choosing the two sub-segments to be matched. Using the GUI the user marks two areas: a target on one contour and the area for the search of possible matches on the other contour.

The first part of the GUI is related to the previous step of segmentation, the process of separating the pieces from the background (Fig 2). Here the user is asked to choose if the foreground (the piece) is white or black. This is a user decision since it depends on the way the images were prepared in previous stages.

Figure 2. The first part of the GUI is composed of two windows. The current image is presented after its conversion to black & white (BW) image (bottom window). The upper window is a menu in which the user chooses if the object (the piece) is white or black.

Choosing a white or black object is done for the images of the two pieces, the one that contains the designated target and the images with its potential matching fracture line.

Next, the contour is extracted and displayed on another window (Fig 3). Here the user chooses the relevant sub segments on the contour and the results are saved for the matching step.

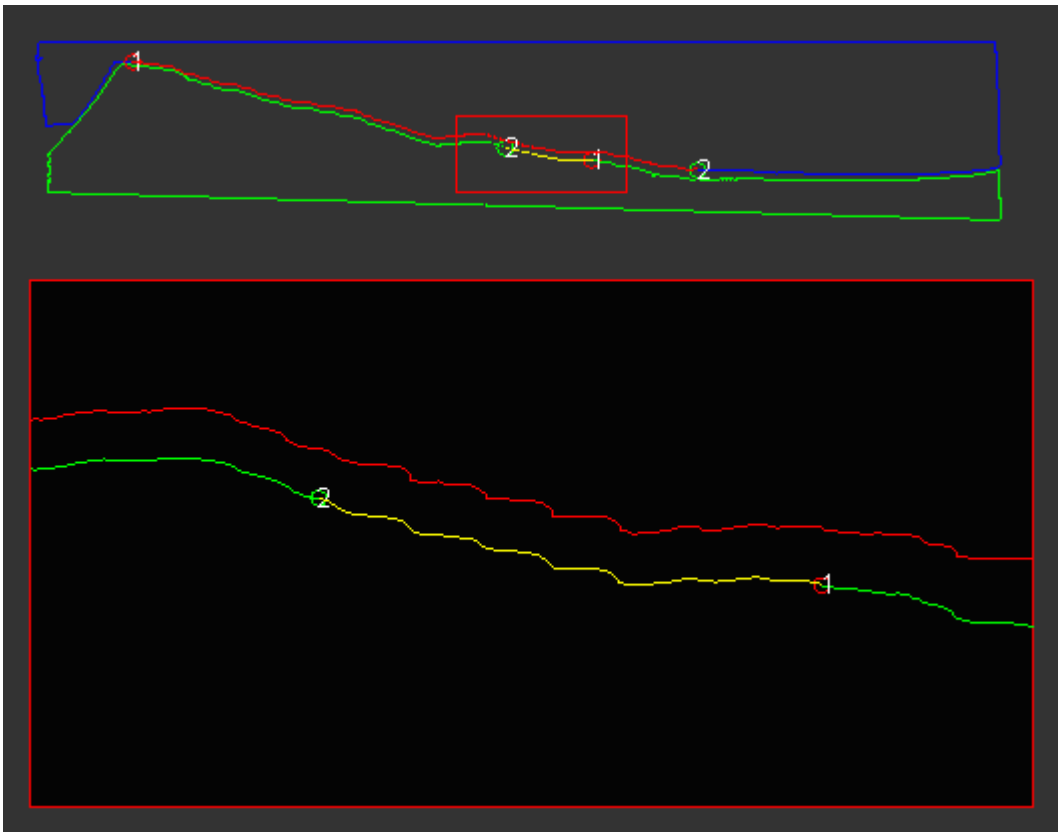


Figure 3. Choosing the target and search area. The two contours are displayed simultaneously in two views: a zoomed out view of the entire contours (top) and a zoomed in view, magnifying a specific feature (bottom). The red rectangle on the top view marks the area that is magnified on the bottom view. The user marks the search area on the top contour using two points (red and green circles). The search area is the red part on top of the blue contour. Setting the borders of the area is done by moving the mouse cursor to the desired location and then pressing the “1” or “2” keys. The same method is used on the bottom contour to mark the target which is displayed as a yellow part on top of the green contour. The user can change the scale of the bottom view (zoom in and out) and move and rotate one contour with respect to the other to be able to finely choose the target and search area.

2.2 The results

The system searches for the best matching candidates using 1D contour coding, as described in previous reports. The results of the search are then displayed (Fig 4).

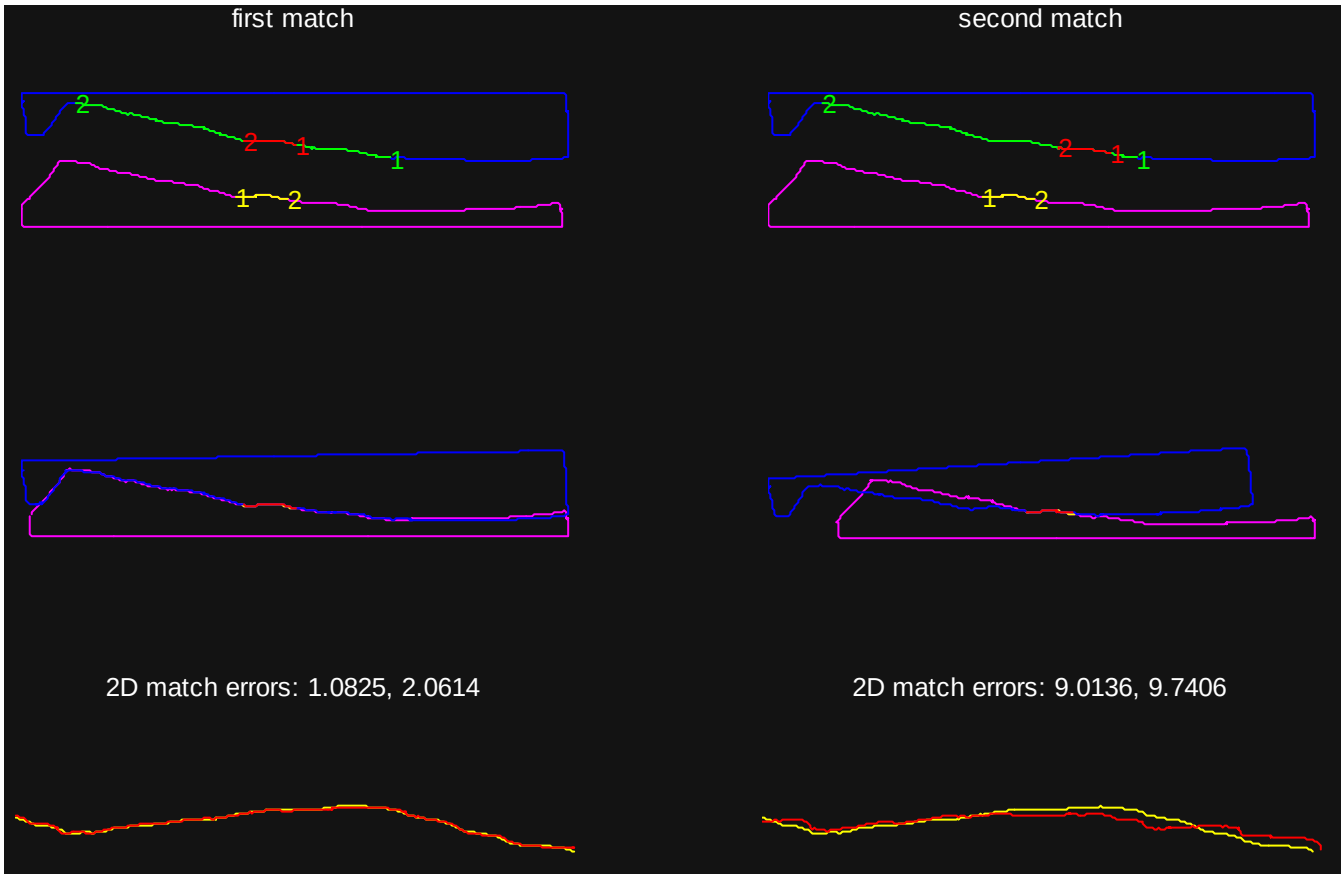


Figure 4. Finding a match. The contours of two pieces are shown in blue (top contour) and magenta (bottom contour). The target is the yellow segment on the bottom contour. Some of the top contour is used as the range for the search (green area). Many candidate segments were generated, all with the same 2D length as the target. The best candidate, i.e. the one with the lowest matching error is shown in red on the top left contour. The second best candidate is shown on the top right contour. Note that the proposed best match and the target have the same position relative to their fracture lines and the 2nd best has a different position. Both candidates are aligned to the target (bottom left and bottom right). The translations and rotations used in aligning the candidates are applied to the entire contours (middle left and middle right). The degree of matching is presented using the 2D error matching values. Two values are calculated for each candidate; the accurate value and the fast to calculate value (see section 3 for explanations).

The results for other candidates are displayed and the user can compare their 2D matching error values. If the statistics for this material and length was already collected, the user is also able to convert the 2D matching error values into matching error rates. This is explained in detail below (sections 4 and 5).

3. Using 2D matching criteria

In a previous report we suggested to increase the robustness of our matching process by using 2D error function. This was successfully applied for two different contexts. In the first, the 2D error was used as an accurate way to check small number of candidates that were found using the faster 1D search. In the second, the 2D error was part of an independent system used to gather the statistics of matching as a function of material and of fracture line length.

In both cases, measuring the 2D error takes the following stages. 1. Registration: the two curves are optimally aligned by translating and rotating one of the curves. 2. The 2D error is calculated as the sum of distances between the two curves.

3.1. Optimal alignment

Note that in this stage a match was already found using the 1D method. Then the target and candidate 1D fracture lines are transformed back to their 2D representation for the 2D alignment method to be applied.

We used the method suggested by Femiani et al. (2004). First, the curves are translated so their centers are aligned. Next, one of the curves is rotated in such a way that the fit between the curves is maximal. The angle of rotation is found analytically by solving for the angle that minimizes the squared difference between the curves. This method works only if the two curves have the same number of points – a condition that is generally not true for our system. We therefore resampled the curves to the same number of points.

3.2. Distance between curves

We compared two methods for finding the distance between curves. The first was a simple sum distances between points of the same indices along the two curves (Fig 5). This method has running time complexity of O(n), where n is the number of point on each of the curves. The second method was the sum of minimal distances between the curves (Fig 5.). Here we find for each point on curve **A** its nearest point on curve **B**, and

for each point on curve **B** its nearest point on curve **A**. The distance between the curves is then defined as the average of the sum of distances from curve **A** to **B** and from curve **B** to **A**. This method has running time complexity of O(n^2), so it is much slower than the first one. However when the number of points is not the same for both curves, the second method is more accurate. Eventually we used the 2nd, more accurate method.

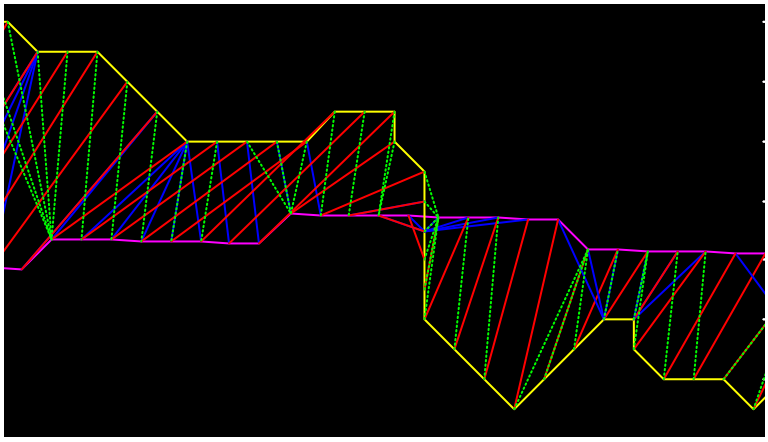


Figure 5. Zoomed in view on two matching curves illustrating the calculation of distance between curves. Method 1: red lines mark the distances between points with equal indices on the yellow and magenta curves. Method 2: green lines mark the distances from each point on the yellow curves, to its nearest point on the magenta curve. Blue lines mark the distances from each point on the magenta curve to its nearest point on the yellow curve. In both methods the result is the sum of all distances divided by the number of points. The result of method 1 is always >= the result of method 2.

4. The statistics of matching

The main object of this project is to develop methods for answering a simple question: given two fracture lines that are suggested to be a matching pair, what is the probability that this match is wrong?

First, we will develop this question, describing accurately and in **greater** detail what statistics are needed for answering the question. Second, we will describe the methods of gathering these statistics and present our results for the three different databases; silicon, paper and Perspex.

4.1. Decision theory

When deciding if a pair of shapes is similar enough to be considered as a match, we should take into account several types of classification options and their implications. If the two shapes are classified as a match this decision may be either right (**hit**[3]) or wrong (**false alarm**). Similarly, if the pair is classified as a non-match, this classification too can be either right (**correct reject**) or wrong (**miss**). Each of these four types of classifications has a probability of occurrence that can be estimated using the distributions of two populations, the **correct matches** and the **non-matches**.

We shall use a schematic example to explain this type of analysis (Fig 6). The left curve is the probability distribution function (PDF) of the matching error for the first group – the correct matches. The right curve is the PDF of the same measure for the group of non-matches. A specific value of the matching error is used as a separation criterion between the two groups (shown as a vertical line Fig 6B,C). Using the areas (light and dark gray) we can calculate the probabilities of hits, misses, false alarms and correct rejects. The decision of what separation criterion to use should be based on the relative significance we set for the different errors. If we can tolerate misses exactly the same as we tolerate false alarms, then the separation criterion will be positioned exactly in the middle of the two PDFs, being the **optimal criterion for separation** between the two groups. If on the other hand, we can not tolerate false alarms, we should move the criterion to the left, until the false alarm area is minimal, unavoidably increasing the probability of misses. The effect of shifting the criterion is shown in Fig 7.

[3] Hit is also known as correct acceptane; false alarm as false positive; and miss as false negative.

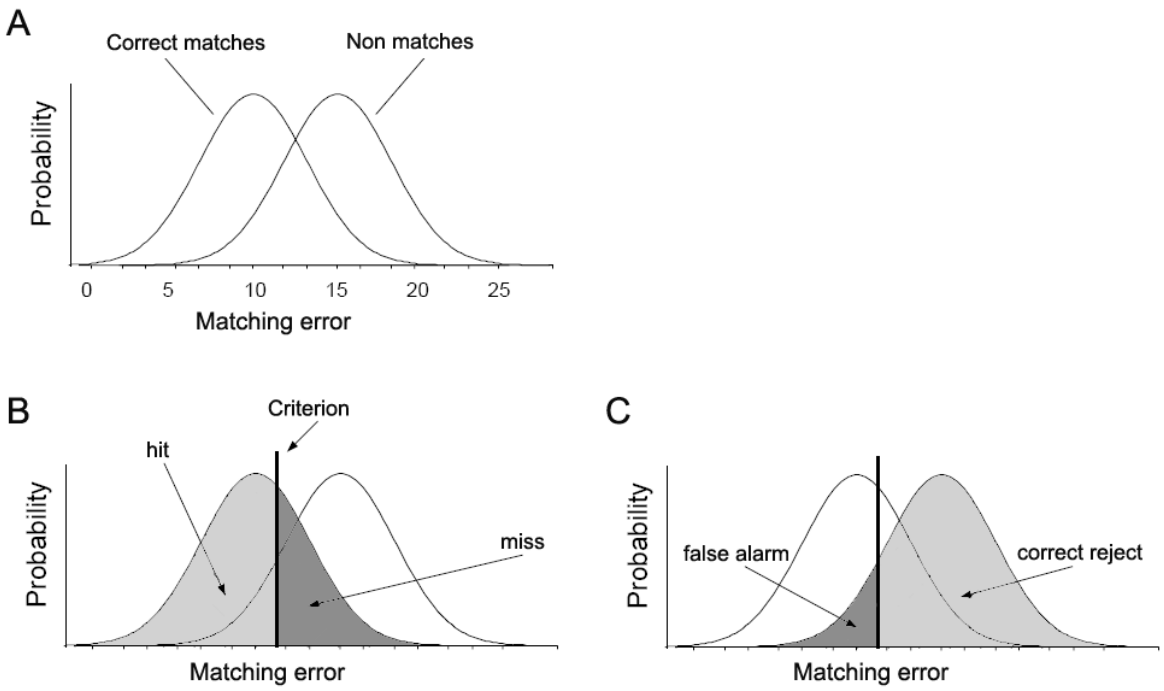


Figure 6. Schematic illustration of the matching error probability density functions for correct matches and for non-matches. *A.* In this example the two PDFs have exactly the same shape with different means. *B.* All pairs of shapes with matching error values to the left of the chosen criterion will be classified as correct matches. Those with matching error values to the right of the criterion will be classified as non-matches. The light gray area is the probability of a **hit**, i.e. a correct classification as a correct match. The dark gray area is the probability of a **miss**, i.e. an erroneous classification as a non-match. C. The dark gray area is the probability of a **false alarm**, i.e. an erroneous classification as a correct match. The light gray area is the probability of a **correct reject**, i.e. a correct classification as a non-match.

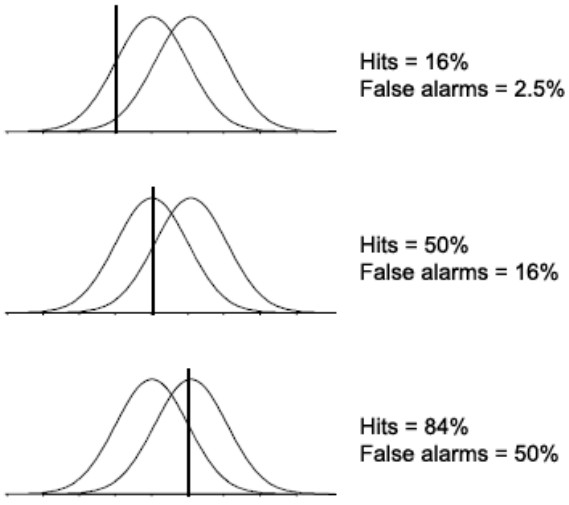


Figure 7. Effect of shifting the criterion. In this example, the two PDFs are not well separated so reducing one type of error, unavoidably increases the other type. Moving the criterion to the left reduces the rate of false alarms to only 2.5% (top), but the hits rate is also low (16%) resulting in a high rate of misses (100-16% = 84%). Such poorly separated PDFs are expected for small length of the matched fracture lines, as shown below in our results.

4.2. Gathering the statistics and classifying pairs into correct matches and non-matches

We used the contours of the pieces and the 2D matching method to gather the statistics of the matching values for each of our 3 databases. In order to have enough data for estimating the PDFs, we needed a method for generating many pairs from each of the two groups, the correct matches group and the non-matches group.

4.2.1. The classification problem

When one compares sub-segments of a fracture line to itself, there is no problem in classifying the matched pairs into correct or incorrect. When a candidate piece and the target to be matched to, are the same piece (i.e. extracted from the same position along the contour) the match is correct. In any other case the pair is a non-match. Using this trivial classification method, one can generate enough pairs to estimate the statistics of matching of a fracture line to itself.

But here we are primarily interested in the statistics of matches of pairs of pieces that were torn or broken apart (Fig 4).

This raises the problem of how to decide if a candidate match is indeed a true match. Solving this problem should be done entirely independent of the matching process, so not to bias the statistics of matching. Our system, at this stage of gathering the statistics, can be described as a supervised learning system, where the teacher that is completely outside of the system gives the correct classification for each matching pair, and the learning system, based on thousands of examples, generates a classification criterion. This criterion can be used later in the task of classifying new examples.

Our problem, stated in the terms of supervised learning systems, was to find a way to easily perform the role of the teacher.

4.2.2. Solution to the classification problem

We used a human operator as the teacher that gives the correct classifications. But because of the large number of examples needed (thousands) it was unrealistic for the operator to classify every example. To save time and effort and still be accurate we used a solution based on similarity in positions: in a correct match the target and the candidate should be located in the same place relative to their own fracture lines (Fig 4, left.). These relative positions were estimated by measuring the distances from the candidate and from the target to a set of matching points on each of the contours (Fig 8). The human operator was left with the simple task of marking sets of matching points on every pair of matching pieces. This task was orders of magnitude faster than the classification of all possible pairs.

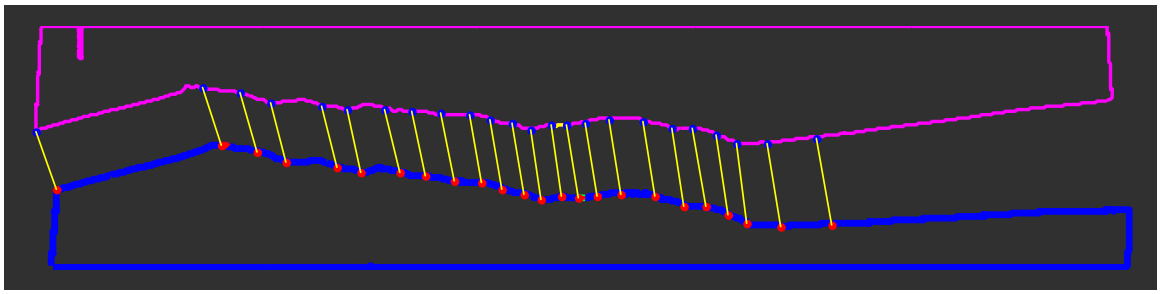


Figure 8. Distinct matching point between the fracture lines of two matching pieces. The matching points are used in the decision of whether a candidate match is a correct match.

4.2.3. Marking the matching points

We developed a graphical user interface (GUI) for marking matching points along the contour lines of two matching pieces (Fig 9).

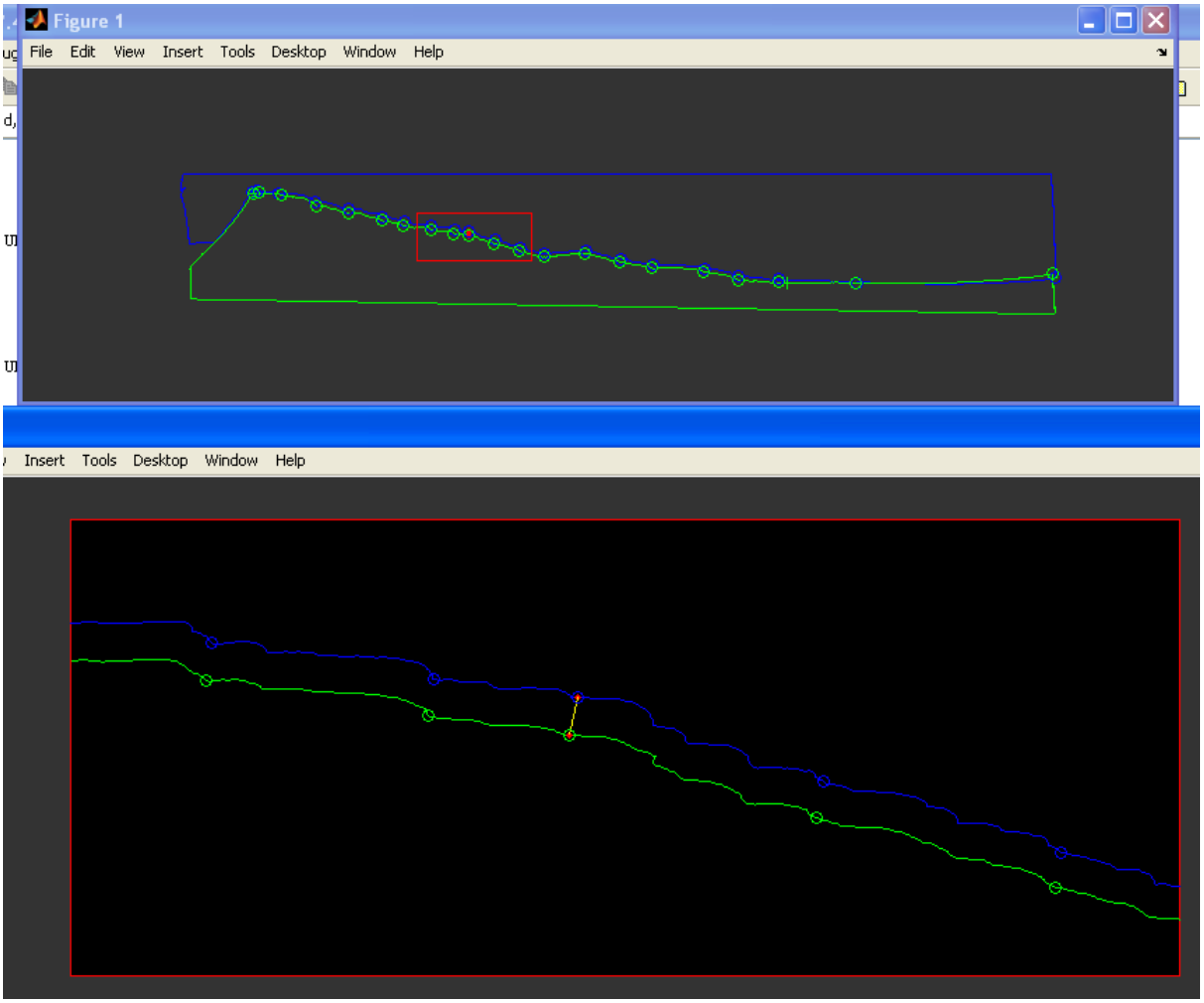


Figure 9. GUI for marking the matching points. The two contours are displayed simultaneously in two windows: a zoomed out view of the entire contours is displayed in the top window and a zoomed in view, magnifying a specific feature is displayed in the bottom window. The red rectangle on the top window marks the area that is magnified on the bottom window. 21 matching points were already marked (blue and green circles). The user can move a cursor (red dots) between the marked points. A yellow line is plotted between the current pair of matching points. The user can move and rotate the lower contour (green) to match it to the upper contour (blue). The user can zoom in to magnify a specific feature and zoom out to have a broader view. The user can add, update and delete points and save the results.

4.2.4. Using the matching points to classify matches

For both the target and the potential match, we measured distances to all matching points along their contours. For a non-match, there was a big difference in the distances (Fig 10) and for a correct match the distances were almost identical (Fig 11).

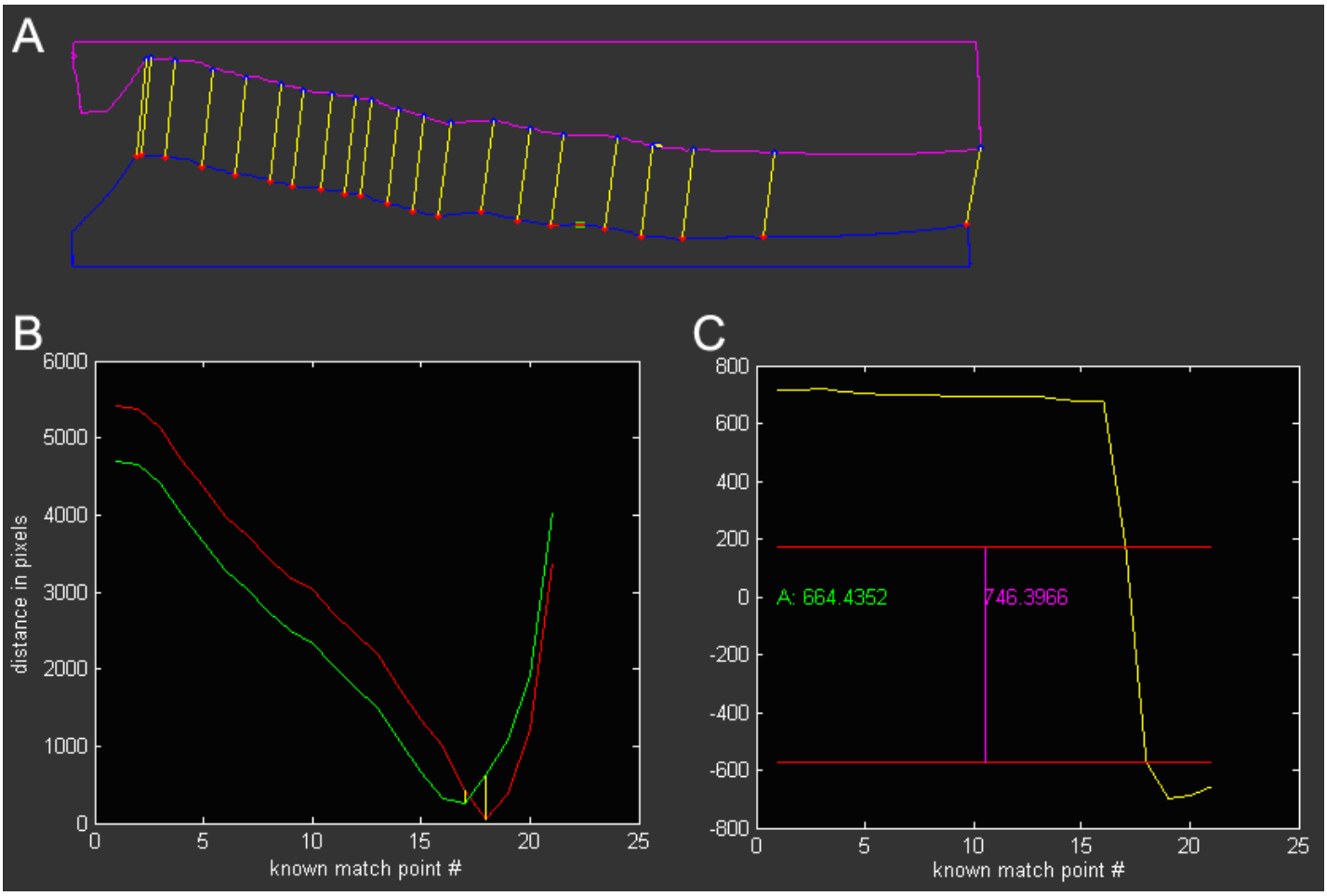


Figure 10. Using the matching points to identify a non-match. *A*. Matching points (blue and red dots connected by yellow lines) marked between two contours. The target (short white bar on the top contour) is situated between matching point 18 and 19 (counting starts from the left). The candidate match (short red and green bar on the bottom contour) is situated between matching points 16 and 17. *B*. Distance of the target from the matching points on the top contour (red curve) and distance of the candidate from the matching points on the bottom contour (green curve). The two vertical yellow lines measure the difference between the red and green curves at the minimum of each curve. *C*. The difference between the red and green curves of B is shown as a yellow curve with a shape typical for a non-match. From matching point 1 to 16, the difference is almost constant (~700 pixels). Then there is a sharp transition to a negative difference (corresponding to the crossing of the red and green curves in B). The average difference (in absolute value) is 664 pixels which is ~1 cm. The magnitude of the small yellow lines from B is shown as red horizontal lines. The difference between the red lines (746 pixels) is termed the ***measure of the mismatch*** and is used to discriminate between correct matches and non-matches during the process of gathering the statistics of matching.

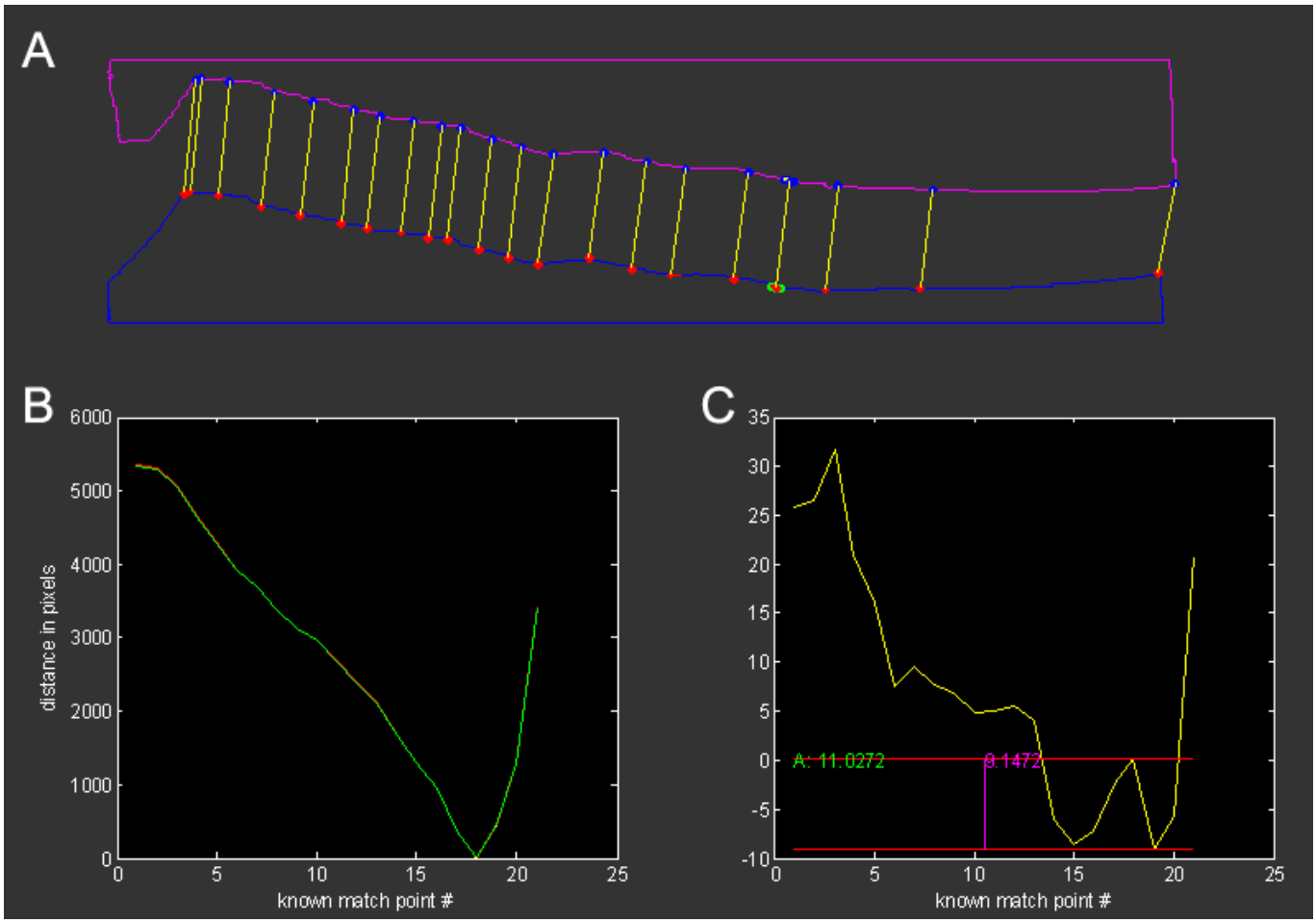


Figure 11. Using the matching points to identify a correct match. As in Fig 10, but for a correct match. *A.* Both target and candidate are positioned very near to matching point number 18. *B.* The distance curve of the target (red) and of the candidate (green) almost overlap. *C*. The difference between the distance curves in B is small (yellow). Its average value is 11 pixels and the measure of mismatch is 9 pixels. Both indicate that the target and candidate are situated at the same relative position; hence it is a correct match.

We performed this analysis for hundreds of examples of correct matches and non-matches and found a very clear separation between the two groups. Correct matches had a measure of mismatch of less than 35 pixels (and usually much smaller values) and non-matches had values bigger than 150 pixels. To validate this result in another way we deliberately shifted the candidates of correct matches with a known shift of 50 or 25 pixels. We found that the mismatch values for the 50 pixels shift were ~100 and for the 25 pixels shift the values were ~40. We decided that setting a threshold of 40 pixels for the mismatch values is a safe choice. It corresponds to shifts of 25 pixels which is less than 0.4 mm.

To conclude this part, we used a GUI to mark tens or hundreds of matching points for each pair of matching contours. We then calculated the difference between targets and candidates to the matching points and found a clear threshold of 40 for a measure we termed the mismatch value. Using this measure we were able to classify each pair as a correct match or as a non-match. This method of classification was used to generate many matching pairs and gather the statistics of matching, as described next.

4.2.5. Generating the statistics for three materials and several segments lengths

For each of the three databases (silicon, paper and Perspex) we generated hundreds of samples of correct matches and tens of thousands of samples of non-matches. This time consuming task (4-8 hours per database) was repeated several times to gather the statistics of matching for different fracture line lengths. The longer the length of the segments to be matched was, the longer it took to generate enough examples.

The method we used for the generation of correct matches was a bit different from the one used for the generation of the non-matches.

Correct matches: for each pair of matching pieces we sampled K targets along the fracture line. For each target we found its location relative to the already marked matching points. Then, we searched for a candidate on the other fracture line, only around the same relative position. We classified the best candidate as a correct match or as a non-match based on its mismatch value. The results were saved for later analysis.

Non-matches: Let N be the number of pairs of matching pieces. Then the number of pieces is 2*N. For each piece, excluding itself and its match, there are 2*N – 2 other pieces to be checked against. Together there are N * (2*N – 2) = M combinations of pairs of non-matching fracture lines to check. We sampled each of the M combinations to yield a total of tens of thousands of non-matching pairs. The results were saved for later analysis.

4.3. Generating histograms and estimating the PDFs

The results of the previous stages were combined into raw histograms for the two populations of correct matches and of non-matches. Each histogram described the number of occurrences of the different matching error values. We used Matlab's statistics toolbox to estimate the empirical PDF from the raw histograms, the results being histograms with areas normalized to 1 (Fig 12) and non-parametric curves fitted to the histograms (Fig 13). Those normalized histograms and curves were used for the calculation of the empirical probabilities of hits, false alarms, misses and correct rejects, as described next.

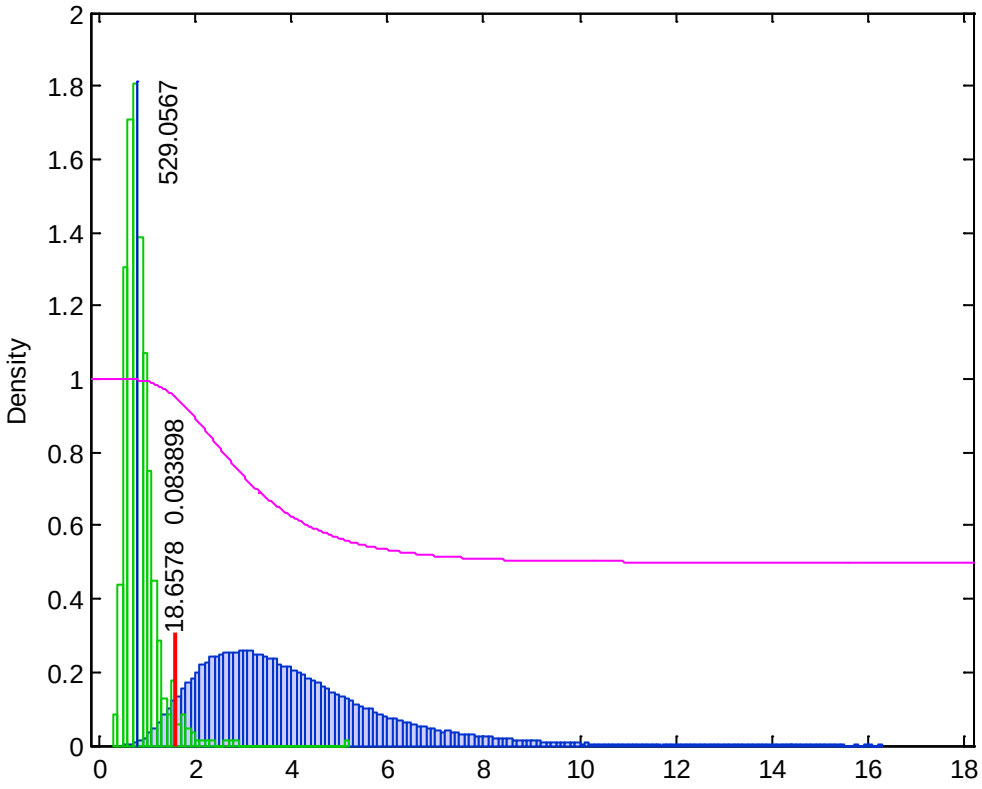


Figure 12. Histograms describing the statistics of matching for 0.25 cm long silicon pieces. Left histogram (green): the distribution of matching error values for the correct matches. Right histogram (blue): the distribution of matching error values for the non-matches. Thick red vertical line is the optimal separation criterion, i.e. the value that separates between the two populations by minimizing the areas of misses and false alarms. Two value are calculated at the separation criterion, the positive likelihood ratio **(hits/false alarms)** = 18.66 and the total error rate **(misses + false alarms)** = 0.084. Thin blue vertical line (around the middle of the green histogram) marks the 50% correct matches' value, i.e. half of the correct matches are left of this line. The positive

likelihood ratio **(hits/false alarms)** at this line = 529. Magenta curve is the positive predictive value (**hits / (hits + false alarms))** as a function of matching error value. It is ~1 for the left side (for very small matching error values) and 0.5 for the right side (for very large matching error values). This value is the empirical probability of match to be a correct match for a given matching error value.

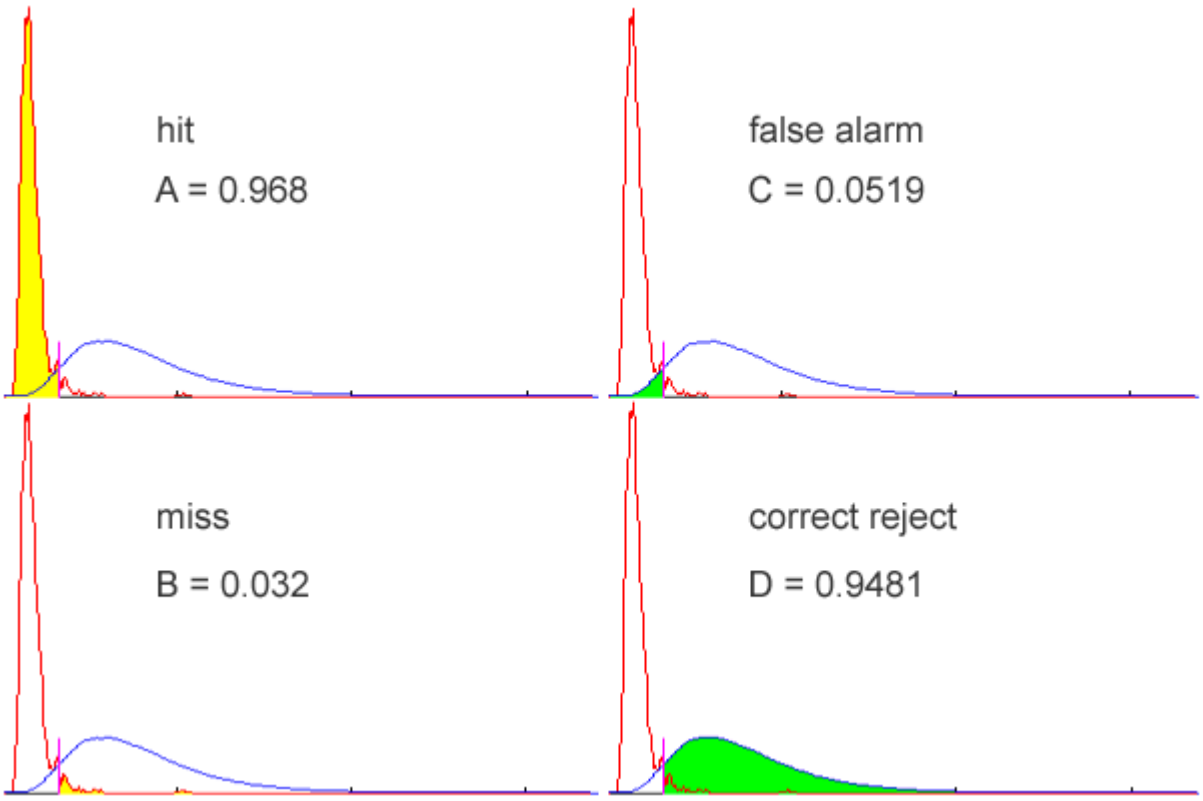


Figure 13. Non parametric fitting to the histograms of Fig 12 (red and blue curves). Areas for the optimal separation criterion: hits = 0.968, misses = 0.032 (1-A), false alarms = 0.0519 and correct rejects = 0.9481 (1-C).

4.4. Calculating the probabilities

4.4.1. Different separation criteria

Given a specific matching error value (on the abscissa – the horizontal axis of the histograms and curves in Fig 12 and 13) as the separation criterion one can calculate the four types of areas. The optimal separation criterion is the value that minimizes the error rates of misses and false alarms. Its value for this example of silicon pieces of 0.25 cm length is 1.6. Using this criterion, we calculated the areas under the graphs (Fig 13). The

probability of classifying a correct match as such (**hit**) is 0.968 and the probability of classifying a non-match as a correct match (**false alarm**) is 0.0519. The ratio of these probabilities (**hit/false alarm**) is the **positive likelihood ratio** and its value for this example is 18.66. In other words, if we use this criterion to separate many examples into matches and non-matches, the chance of a pair that is classified as a match to be classified correctly is 18.66:1. This can be stated also in terms of probabilities, using the **positive predictive value** (**hits / (hits + false alarms)**) giving a value of 0.9491. In other words, using this separation criterion we have a probability of 0.9491 to be correct in classifying pair left of the criterion as correct matches. Another useful measure is the **Bayes risk**. If we assume the same significance to the two types of errors, Bayes risk is the total probability of error for the criterion (**misses + false alarms**) and its value for this example is 0.084.

As stated before, the choice of what criterion to use should be based on the relative significance we attach to the different error rates (Fig 7). Or more precisely, one can use a loss function associated with each type of error and find the criterion that minimizes the total loss.

We used another criterion based only on the histogram of the correct matches, the **50% correct criterion**, in which half of the correct matches are left of this value. The positive likelihood ratio for this criterion was 529 (Fig 12). In other words, if we use this as a separation criterion, pairs with matching error lower than 0.775 (the value of the 50% correct criterion) will be classified as correct matches. These pairs will have a 529:1 chance or a probability of 0.9981 of being classified correctly.

Of course, one can use a criterion that is based on a specific value for any one of the quantities measured. For example, if the minimal tolerated positive likelihood ratio is 100,000 (which is equivalent to positive predictive value of 0.99999) we can search for criterion value with these numbers. In our example (0.25 mm silicon pieces), we do not have such a criterion, since the maximal positive predictive value is 0.99995, equivalent to positive likelihood ratio of 22156 (Appendix 1, 2nd Fig).

4.4.2. The probabilities for a specific matching error value

So far we described how to calculate the probability of a correct classification using a separation criterion. Any pair with matching error value left of the criterion will be classified as a correct match. The positive likelihood ratio and positive predictive value characterize this classification process. But if we want to answer a more specific question: What is the probability of correctly classifying a specific matching pair?

Let us use an example. Suppose we have a silicon matching pair 0.25 cm long, with a matching error value of 1.1. Probability calculation is done by summing the area under the graph between 2 values. So we must use a small range of values around 1.1. We therefore must ask the question again, in the way that it can be answered: What is the probability of correctly classifying a specific matching pair, whose matching error values is between 1.05 to 1.15 (Fig 14)?

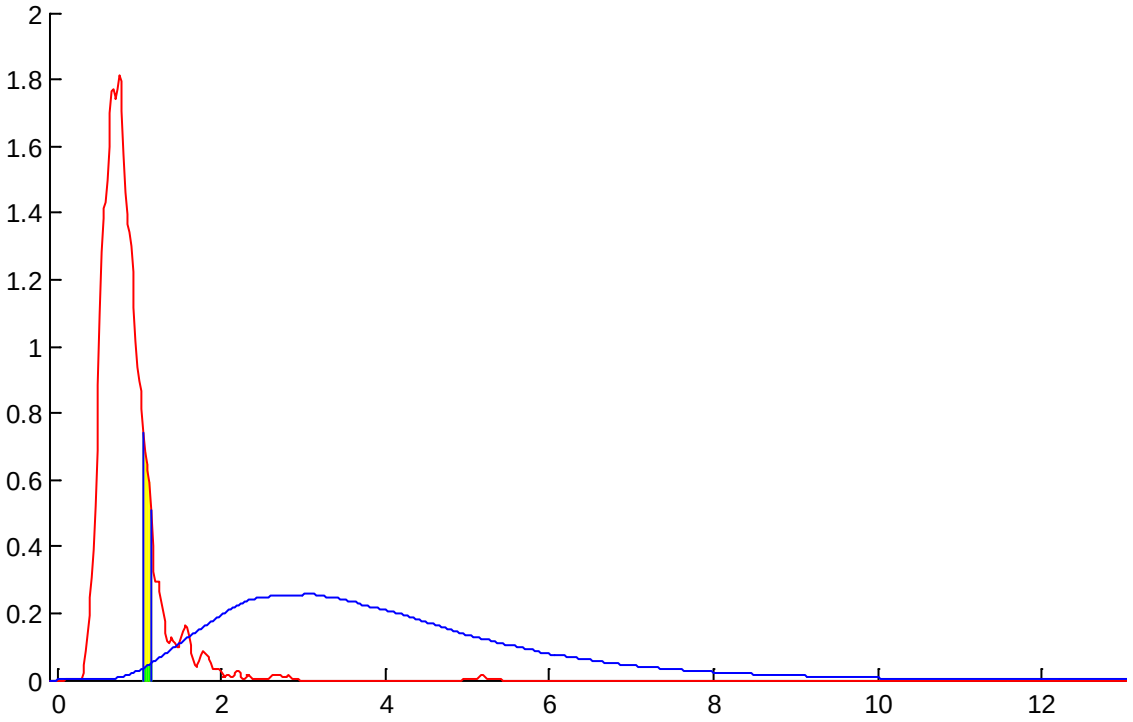


Figure 14. The yellow area (the area under the red curve, including the green area that is dispalyed on top of it) is the probability of a correct match to have matching error values in between 1.05 to 1.15. The green area is the probability of a non-matching pair to be in this range.

The probability of a correct match to have this range of values is 0.0561. The probability of a non-match to have this range of values is 0.0039. The ratio between the probabilities is 14.35. Therefore the chance of a pair in this range to be a correct match is 14.35 larger than being a non-match. Or in probability terms, it has a probability of 0.93 of being a correct pair.

5. Results – The statistics of matching

5.1. Silicon fracture lines

We ran our system for 6 different fracture line lengths: 0.125, 0.25, 0.5, 1, 2 and 3 cm.

The results are summarized in table 1 and Fig 15. The histograms are presented in Appendix 1.

Length [cm]	Length [pixels]	# of non-matches	# of correct matches	Total error rate	Optimal separation criterion	50% correct criterion	Positive predictive value at the optimal separation criterion	Positive predictive value at the 50% correct criterion	Positive likelihood ratio at the optimal separation criterion	Positive likelihood ratio at the 50% correct criterion
0.125	84	180520	1453	0.22690	1.05	0.6351	0.8657	0.963900	6.44	26.7
0.250	167	177608	842	0.08380	1.60	0.7757	0.9491	0.998100	18.66	529
0.500	333	275585	828	0.01850	2.07	0.8986	0.9942	0.999994	172	180186
1.000	667	155022	765	0.00690	3.80	1.1111	0.9959	0.999999	242	5.4578e+010
2.000	1334	63201	531	0.00036	5.10	2.2467	0.9997	1.000000	3370	2.2468e+015
3.000	2001	43633	425	0.00016	6.71	2.1405	0.9999	1.000000	11692	7.0461e+016

Table 1. Statistical results for matching silicon pieces of 6 different lengths.

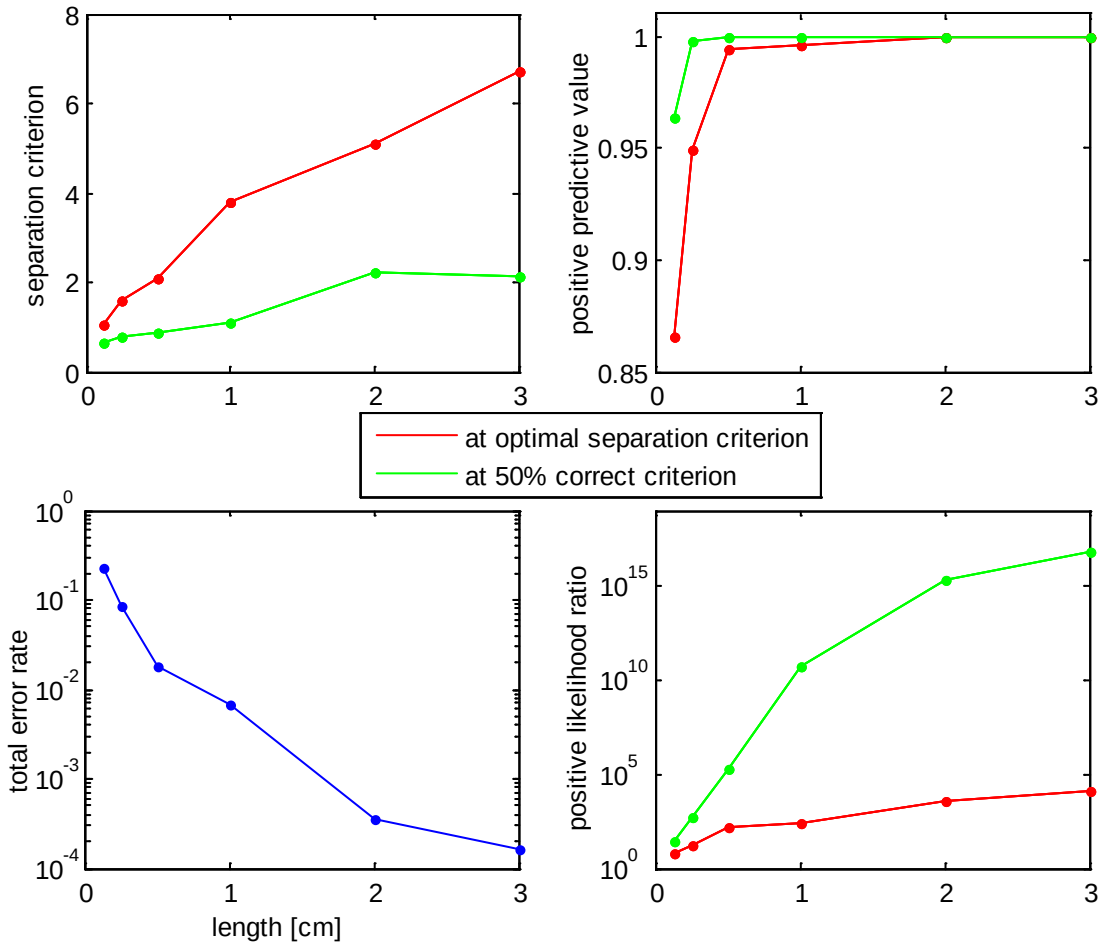


Figure 15. Results for the silicon pieces, based on table 1. Top graphs have a linear y axis scale. Bottom graphs have a logarithmic y axis scale.

5.2. Paper fracture lines

We ran our system for 5 different fracture line lengths: 0.5, 1, 2, 3 and 4 cm.

The results are summarized in table 2 and Fig 16. The histograms are presented in Appendix 2.

Length [cm]	Length [pixels]	# of non-matches	# of correct matches	Total error rate	Optimal separation criterion	50% correct criterion	Positive predictive value at the optimal separation criterion	Positive predictive value at the 50% correct criterion	Positive likelihood ratio at the optimal separation criterion	Positive likelihood at the 50% correct criterion
0.5	150	58787	1006	0.4138	3.87	2.8121	0.8012	0.935972	4.03	14.6
1.0	300	54911	886	0.3771	5.65	3.9699	0.8100	0.958333	4.26	23
2.0	600	45440	1183	0.2594	8.50	5.0973	0.8346	0.987008	5.05	76
3.0	900	32977	720	0.1368	9.10	5.4465	0.9253	0.998102	12.38	526
4.0	1200	12257	517	0.0846	10.16	6.2327	0.9484	0.999176	18.38	1213

Table 2. Statistical results for matching paper pieces of 5 different lengths.

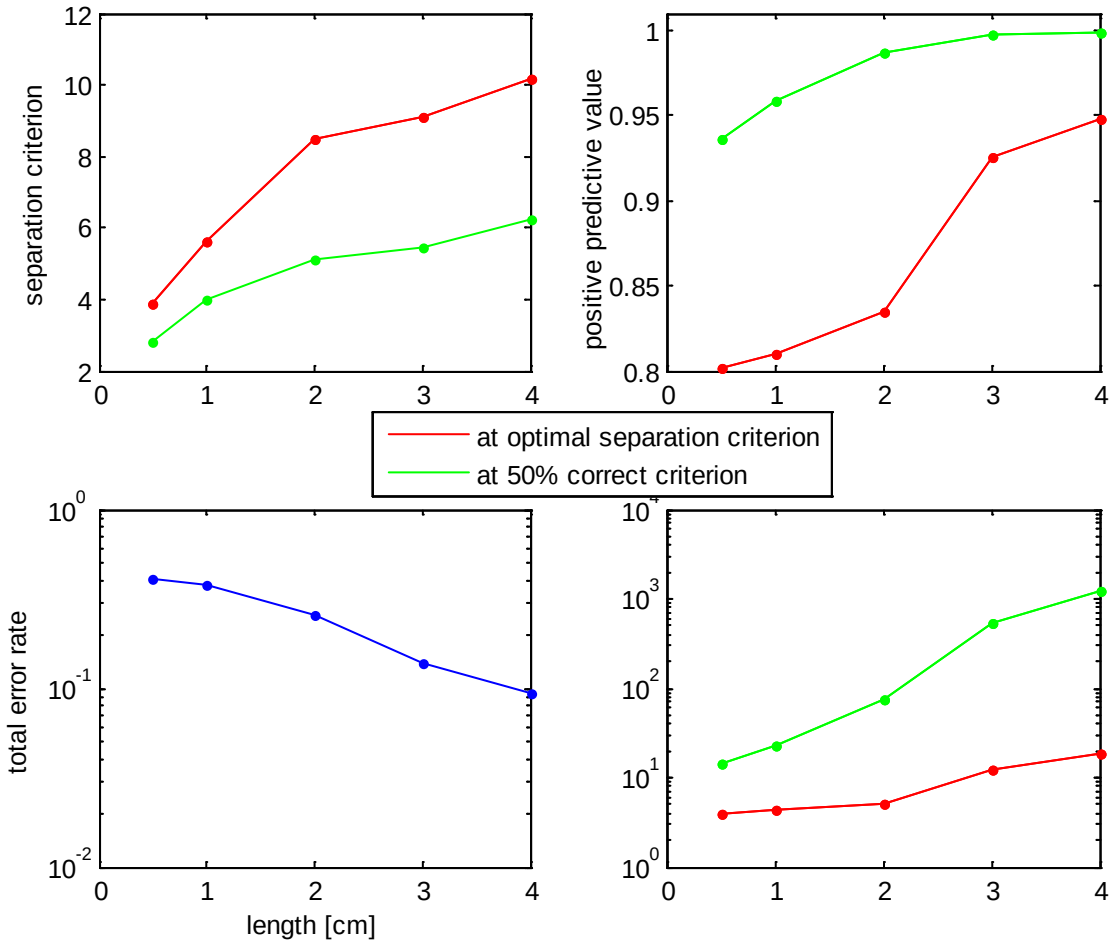


Figure 16. Results for the paper pieces, based on table 2. Top graphs have a linear y axis scale. Bottom graphs have a logarithmic y axis scale.

5.3. Perspex fracture lines

We ran our system for 3 different fracture line lengths: 1, 2 and 5.333 cm.

The results are summarized in table 3 and Fig 17. The histograms are presented in Appendix 3.

Length [cm]	Length [pixels]	# of non-matches	# of correct matches	Total error rate	Optimal separation criterion	50% correct criterion	Positive predictive value at the optimal separation criterion	Positive predictive value at the 50% correct criterion	Positive likelihood ratio at the optimal separation criterion	Positive likelihood at the 50% correct criterion
1.0	225	130008	1085	0.4030	0.66	0.4342	0.8487	0.960089	5.61	24.06
2.0	450	116040	894	0.1666	1.58	0.6178	0.8832	0.989823	7.56	97.26
5.33	1200	132802	676	0.0405	3.04	1.1603	0.9783	0.998781	45.03	819

Table 2. Statistical results for matching Perspex pieces of 3 different lengths.

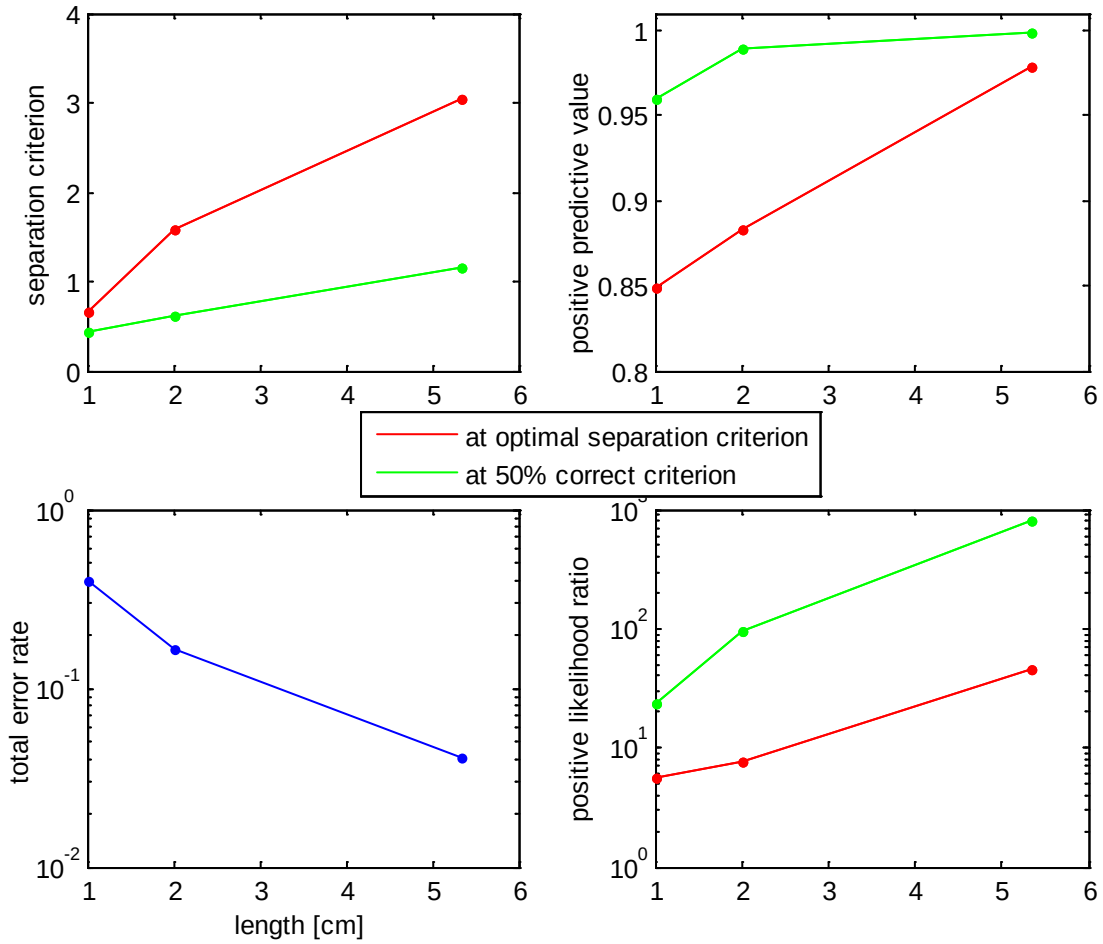


Figure 17. Results for the Perspex pieces, based on table 3. Top graphs have a linear y axis scale. Bottom graphs have a logarithmic y axis scale.

5.4 Comparing the results for different materials

There are some clear differences between the statistics of the silicon fracture lines and the statistics of the two other materials (Fig 18).

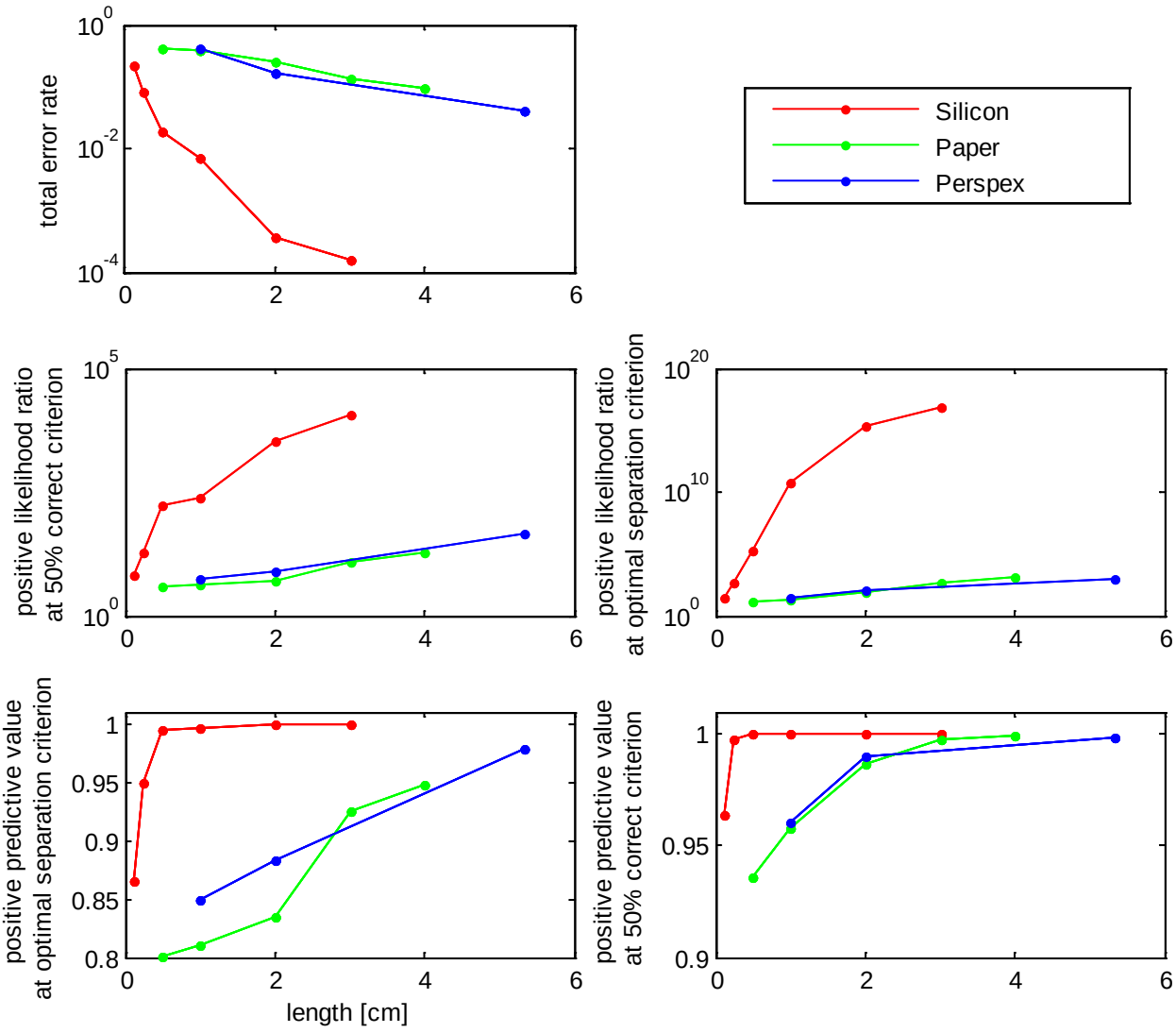


Figure 18. The results for the different materials presented together. Top and middle graphs have logarithmic y axis scale. Bottom graphs have linear y axis scale.

The silicon pieces are easier to match than the other two materials, as evident by every measure presented here (Tables 1-3, Fig 18). Matching the silicon pieces has much lower total error rates and much higher positive predictive values and positive likelihood ratios. This means that for a given error rate, one can find a correct match for much smaller silicon pieces than for the other two materials.

The paper and Perspex pieces have very similar statistics. However, since they have different physical characteristics the similarity in their statistics calls for an explanation. The paper fracture lines are curved and twisted in contrast to the very straight lines and simple curves of the Perspex fracture lines (Fig 19).

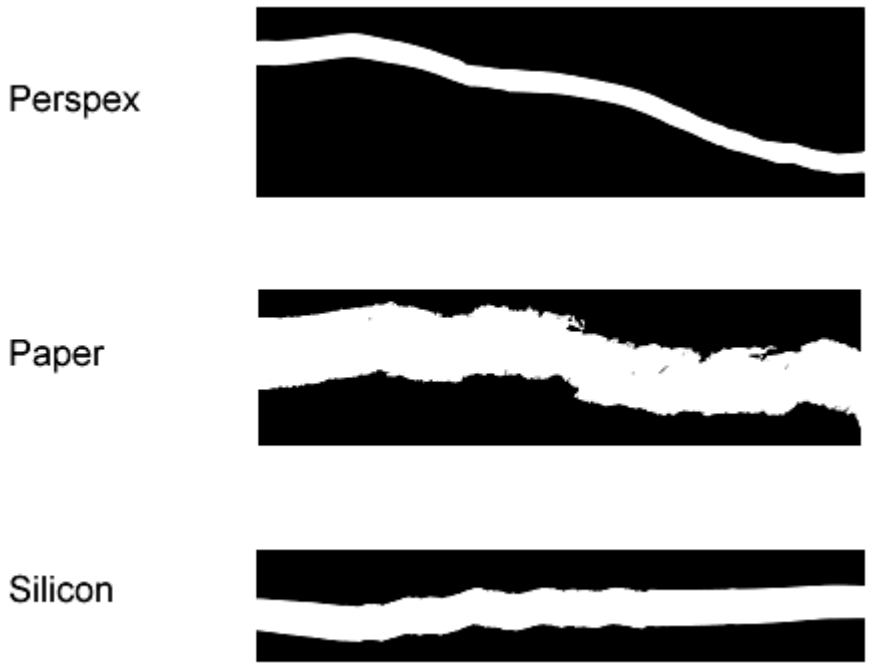


Figure 19. Typical fracture lines of the three materials. The Perspex fracture lines are straight or slightly curved, with almost no details. The two pieces match almost exactly both in overall shape and in fine details. The paper (composed of paper and metal layers) fracture lines are very detailed, with a lot of turns and twists. The overall features of the two halves are similar but they do not match well in fine details. The silicon fracture lines have intermediate complexity and the two halves match very well, especially in fine details.

In terms of the amount of information in a fracture line, the paper pieces have much higher values than the Perspex pieces. On the other hand, the matching paper pieces are not very similar while Perspex matching pieces are almost identical (Fig 19). The low degree of similarity of matching paper pieces has two sources, one that is directly related to material properties and the other that is more a matter of the algorithms we use. The first source is material tear and deformations of the paper and metal layers of the pieces.

The second is errors in the segmentation of figure from ground. Some improvements are expected if a different segmentation algorithm will be employed, but it will probably not be as good as the segmentation of the silicon and Perspex pieces.

To conclude, there are two opposing trends in these materials. The fracture lines of the paper pieces have a lot of information but also a lot of noise. The Perspex fracture lines have low amount of information and very little noise. The combined effect of these two trends is that both databases show similar statistics.

The superior results of matching silicon pieces can be also explained in these terms. The silicon fracture lines have enough information to ensure unique matches and a low amount of noise that does not ruin the similarity of matching pairs.
In conclusion, the property that may well predict the quality of matching is not the amount of information of a typical fracture line but the mutual information between the two matching pieces, or how well one side of the fracture can predict the other side.

5.5 The statistical results of the system developed as apposed to results achieved in real cases

Surprisingly, the statistical results were much lower than initially expected or that usually concluded by toolmark experts. In most cases of toolmark comparison, the fracture surface is not two-dimensional, but three dimensional. In addition, almost always the fracture is in a lattice of visual information, which raises the level of certainty significantly. This information includes graphic patterns existing on the surface or the outer border of the pieces to be compared. This information can rarely be quantified, and it did not get any representation in our research.
In our research we related to the two dimensional fracture contours only, and ignored any assisting information. The results of our research are the entering point to the numerical or quantities evaluation. If the chances of getting a two dimensional fracture line are so low- adding the surface details or performing a three dimensional match, can bring us much closer to the results achieved by toolmark experts.

In our research, the results were divided in only two categories: matches and non-matches. In the forensic practice, another category exists – inconclusive. Some of our results would have fallen in this category; hence the statistics for a "match" would be higher.

5.6. The quality of matching as a function of fracture line length

The first obvious result is that longer pieces give better results (Fig 18). The total error decreases with length in all three materials. The positive predictive values and the positive likelihood ratios increase with length for all three materials.

The second result is the increase in matching error values with length in all cases, as seen in the different histograms. An outcome of this is the increase in criterion values with length (Tables 1-3, Fig. 12-14). How can we compare results of different lengths? We suggest using the histograms and the statistical measures as discussed below (section 6.4).

6. Summary and discussion

We developed two prototype systems used for physical matching.

The first is a system to assist forensic experts in performing objective physical matching in 2D and the second system to collect statistics and build confidence levels regarding the physical matches.

6.1. First system – performing physical matching in an objective fashion

The inputs to the system are two digital images of the two pieces to be matched. This could be obtained either by a digital camera or by a scanner. Scanning the samples usually produced results as good as photographing with much less effort. The photographic conditions should be such that the segmentation of pieces from background should be easily performed by a simple threshold on the gray levels (e.g. using Photoshop's magic wand tool). Each image should contain only one piece.

The system segments the pieces from the background of the images and extracts the contour of each piece. The user chooses a target segment on one of the contours, and an

area for candidate search on the other contour. The system runs a search and order the potential matches according to their matching error.

The outputs of the system are several sub-segments along the matching fracture line which are the best potential matches to the target. Each candidate has a matching error value that characterizes the quality of the fit with the target.

The user can review the results and estimate the potential error rates using the statistics on that material (results of the second system).

6.2. Second system - collecting statistics and building confidence levels regarding the physical matches

The inputs to the system are 2*N digital images with the same specifications as for the inputs of the first system. The images should come as N pairs, each containing the images of two matching pieces. Another input is the fracture length in cm for which the statistics will be gathered.

The system generates many examples that are classified into two populations, the correct matching pairs and the non-matching pairs. The system gathers the frequencies of the matching error for the two populations into two histograms.

The outputs of the system are several statistical measures describing the quality of matching as a function of the matching error value. 1. Best separation criterion for the two populations. 2. Error rates as a function of matching error value. 3. Positive predictive value and positive likelihood ratio as a function of matching error value.

The user can find the probabilities of a being right or wrong for any specific value of matching error and thus is able to compare between matching pairs, and find out which is the best evidence, even among different sizes and materials.

6.3. Performing the match

We used two major methods for finding the best match. The 1D method was described in detail in previous reports. It has the benefit of being fast and therefore can be used to scan large number of pieces in search of potential matches. The 2D method was much more accurate but also much slower. When the results of the two methods were compared we

found that pairs with low 1D error had also low 2D error, but the order among them was not always the same for the two methods. In our physical matching system we combined the two methods. The initial search is done by the 1D method. It narrows down the number of potential matches from thousands to tens of pairs. The 2D method is then used to select the best match from a small number of good candidates found by the 1D method.

6.4. Comparing matching results for different fracture line lengths

The matching error and separation criterions increase with length (see section 5.5). This means that if we want to compare the error of matching for two individual pairs of pieces, each with a different length, we should normalize the matching error values according to the length. A possible normalization is to divide the matching error values by the optimal separation criterion of each length. However a better way for comparing pairs with different lengths is to avoid using the matching error values directly, but rather use the statistical results such as the positive predictive values and the positive likelihood ratios that characterizes pairs with given matching error values.

An example will illustrate this point. Suppose we have two pairs of silicon pieces, one with fracture line length of 2 cm and matching error value of 4, and the other with fracture line length of 0.5 cm and matching error value of 1.4. Checking the histograms and statistics for 0.5 cm and for 2 cm (Appendix 1) we find that a 0.5 cm pair with matching error in a small range around 1.4 has a positive predictive value of 0.993 and a 2 cm pair with matching error in a small range around 4 has a positive predictive value of 0.999. We can therefore conclude that the 2 cm pair have a higher chance of being a correct match and therefore is better evidence.

In a similar way we can even compare the degree of matching for pairs made of different materials. All we need is the statistics for the specific materials and specific lengths.

This of course raises the important questions of extensibility. First and easier to answer is the question of using our results to estimate the error rates for a length we do not have in our analysis. If this new length is in the range of lengths we checked, we believe it is safe

to interpolate our results. However, this is not really needed because in a few hours of computer work we can generate the statistics for the new length we are interested in.

The more difficult question is whether we can estimate the error rates for materials we did not check. We believe that this may be possible in the future, after performing some more research, following these steps. 1. Develop ways to characterize populations of fracture lines of different materials. 2. Develop ways to estimate the similarity between populations of fracture lines of different materials. 3. Check if materials that have similar population characteristics share also similar statistics and if so find the confidence levels of such estimation technique.

6.5. The amount of information in fracture lines

Leitão and Stolfi (2005) estimated the information content of fracture lines using a sample of 50 pairs of matching ceramics pieces. They calculated the amount of information for pieces 10.8 mm long to be 22.35 bits. This means that the probability of two random fragments, 10.8 mm long being indistinguishable from a true match will be about $1/2^{22} \approx 1/4{,}000{,}000$.

How does this number compare to our results?

The estimate of Leitão and Stolfi is an average over all matching pairs. To have a comparable estimate we did the following. First we calculated the ratio between the areas of correct matches group and non-matches group for all values of matching errors using small bin sizes (see section 4.4.2. and Fig. 14). Then we averaged over these values and found that the averaged positive likelihood ratio for 1 cm fracture lines of that criterion was 2200 for the silicon pieces. Converting the numbers to bits of useful information, the silicon piece of 1 cm had an average of ~11 bits.

This is a rough comparison since there were differences between our work and Leitão and Stolfi (2005) in the way the fracture lines were extracted, and because of the different statistical measures we used. For example, Leitão and Stolfi based their analysis only on the best 10.8 mm long parts of their matching pairs. They did not sample uniformly the entire population of matching pairs, and also omitted completely the non-matching pairs from their analysis. They did check the information content of non-matching pairs but that was only used as a control. Therefore using their numbers to estimate the probability

of correct classification of matching pairs, is expected to yield a higher estimate then our method. Trying to mimic some of their choices, let us calculate the average over the better half of the population – those that are left of the 50% criterion. In that case we get an averaged positive likelihood of 1.2e11, comparable to ~36.8 bits.
To conclude, comparing the statistics of matching for ceramics and silicon pieces in a reliable way should be done by applying the same method to the two materials.

Stone (2004) used a much simpler and entirely theoretical model to estimate the number of possible fracture lines shapes for brittle metals. His model was based on the assumptions that the fracture line changes direction randomly in 2 directions, every arbitrary unit of length. The number of shapes that can be generated using such a model is 2^n, where n is the length. This can be generalized into k possible turn directions yielding k^n possible shapes. Obviously this is an upper limit on the number of actual shapes. Stone (2004) used the assumption that the consecutive turns are statistically independent. If there is some dependency, the number of possible shapes will be reduced. The interdependency between consecutive turns should be estimated experimentally. We did not do that systematically and can only comment on the nature of the materials we checked. In all three materials we found characteristic shapes that are repeated along the fracture lines, a phenomena that could lead to the assumption that the turns in the fracture lines are statistically dependant, and do not behave randomly. A second point that Stone's model makes impractical for real statistical estimates on matching is that the model considers the fracture line on its own. It does not take into account that any real matching pairs of fracture lines will not be exactly the same.
However we can try and compute how distant is this upper limit from our results. Our silicon fracture lines of 1 cm were 667 pixels long. We did not quantify the minimal turning length along these contours but let us use a conservative estimate of 10 pixels, giving 67 turns per cm. Using 3 possible directions per turn (again a conservative estimate) we get $3^{67} \approx$ 9e31 possible shapes. Using this number as an estimate on the rate of possible errors is obviously wrong, it is many orders of magnitude larger than our experimental result of 1/2200 and even of the 50% criterion value of 1.2e11.

7. Conclusions.

Our analysis was based on the experimental statistics of matching. We estimated the probability distribution functions (PDFs) of matching error values for correct matches and for non-matches. We applied this analysis for different fracture line lengths and three different materials. Eventually, we were able to calculate error rates much more reliably than previous estimates. With the results of this research, the expert can express his findings in a more numerical way. The Daubert criteria that demands knowing "the potential or known error rate" can now be fulfilled.

During our research we created databases from three different materials. It is in the reach of every scientist, holding this computerized statistics generator, to create a new database for any material. Computing the potential error rate for objects received in the laboratory for physical match comparison becomes possible..

References

De Bock J. and De Smet P. Semi-automatic reconstruction of fragmented 2D objects. *American Academy of Forensic Sciences, 56th Annual Meeting (AAFS2004),* Dallas, Texas, USA. Feb. 16-21, 2004.

Femiani J., Razdan A. and Farin G. Curve Shapes: Comparison and Alignment. *IEEE PAMI* (November 2004).

Leitão H.C.G. and Stolfi J. Information contents of fracture lines, Technical report – IC-99-24 , Institute of Computing(IC), Unicamp, Brazil. In English, 15 pages, 1999.

Leitão H.C.G. and Stolfi J. A multi-scale method for the re-assembly of two dimensional fragmented objects. *IEEE - Transactions of Pattern Analysis and Machine Intelligence*. 24(9), pp 1239-1251, 2002.

Leitão H.C.G. and Stolfi J. Measuring the Information Content of Fracture Lines**,** *International Journal of Computer Vision*. 65(3), pp 163–174, 2005.

Stone R.S. A Probabilistic model of Fractures in Brittle Metals. *AFTE Journal*, Vol 36(4). 2004.

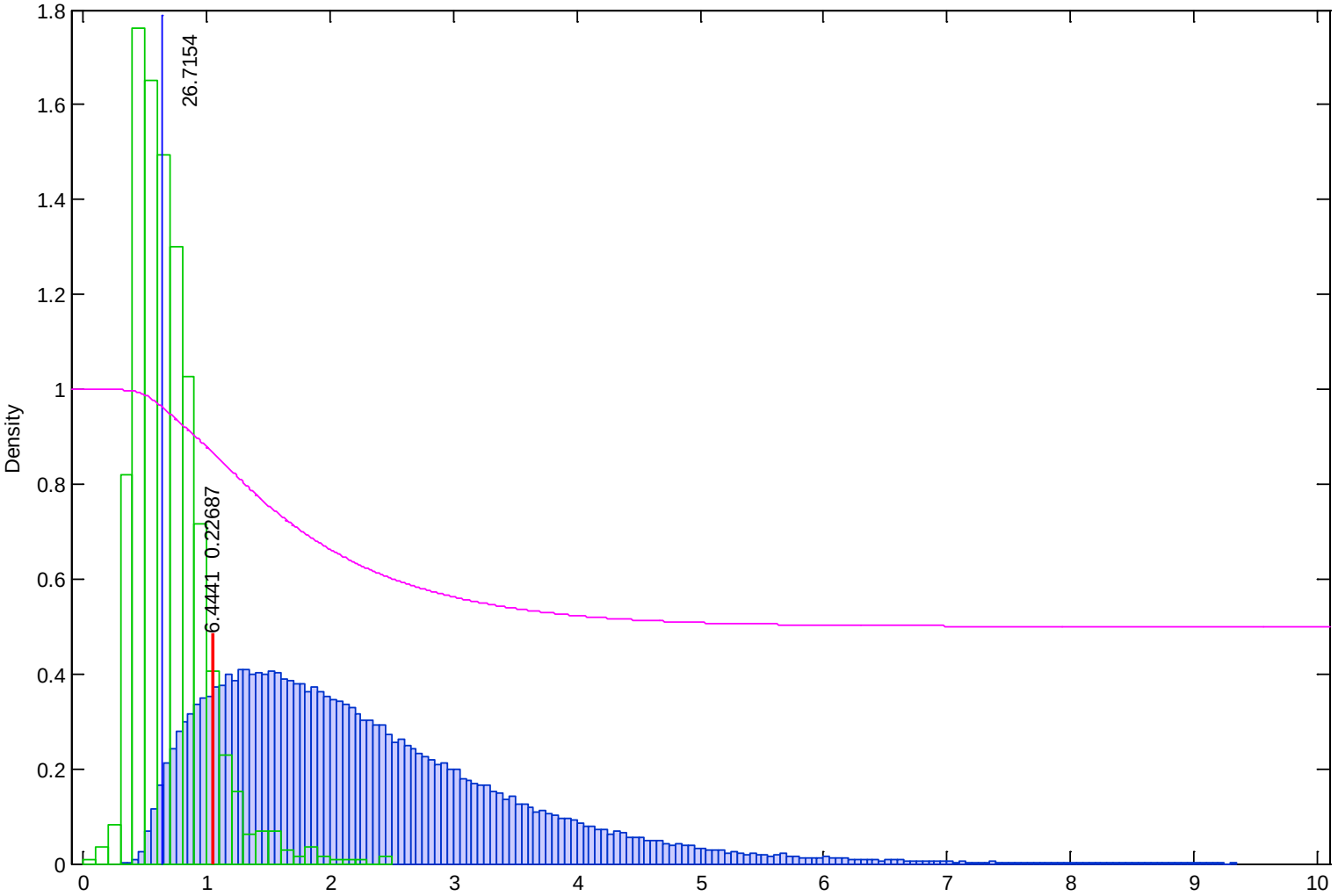
Zhu L., Zhou Z., Zhang J. and Hu D. A Partial Curve Matching Method for Automatic Reassembly of 2D Fragments. *International Conference on Intelligent Computing (ICIC 2006), Kunming**, China*. 2006.

Appendix 1
Histograms for the silicon matching experiments

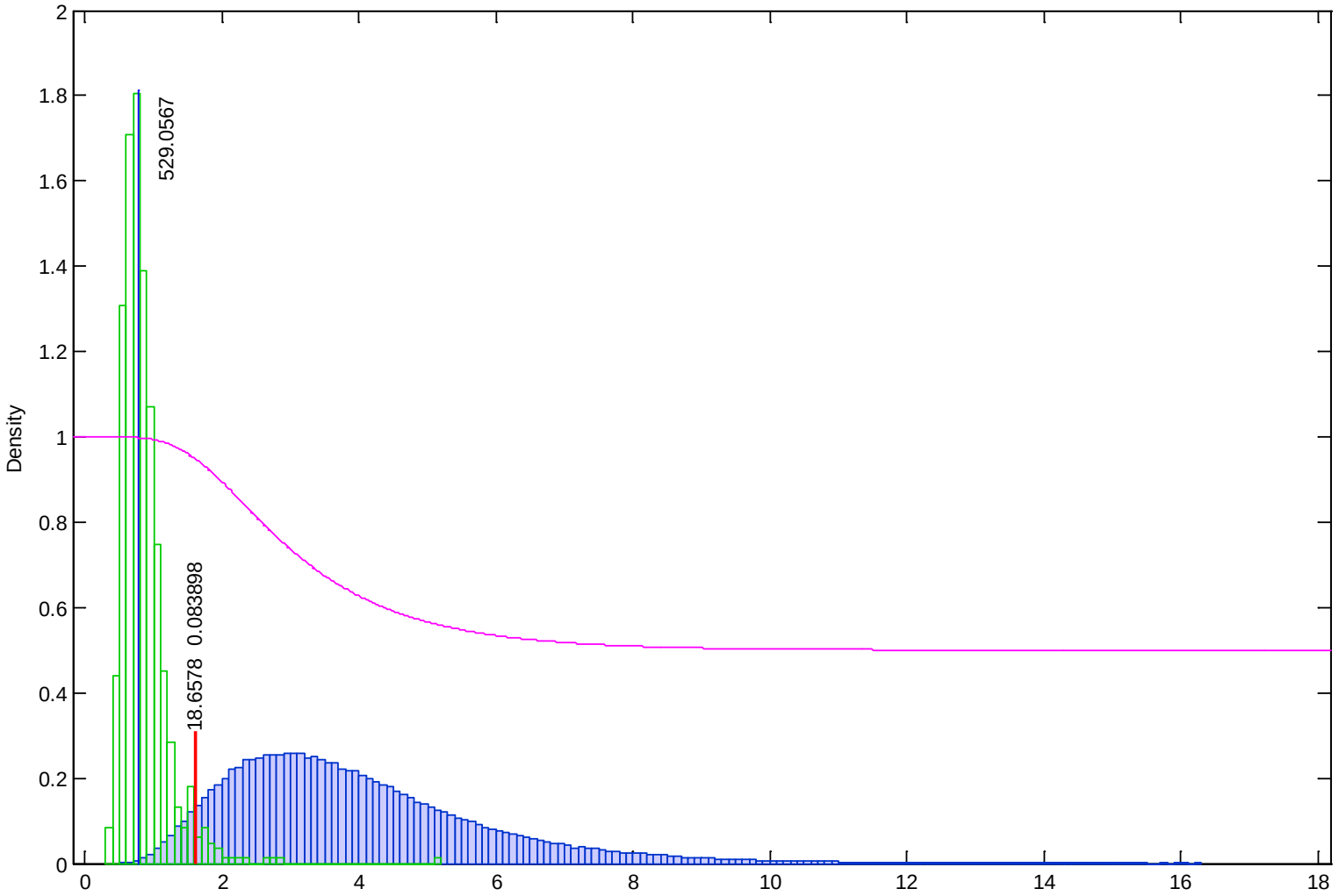
Common legend for all figures in the appendices:

Left histogram (green): the distribution of matching error values for the correct matches. Right histogram (blue): the distribution of matching error values for the non-matches. Thick red vertical line is the optimal separation criterion, i.e. the value that separates between the two populations by minimizing the error rates (of misses and false alarms). Two value are calculated at the separation criterion, the positive likelihood ratio **(hits/false alarms)** is displayed just above the line and the total error rate **(misses + false alarms)** is displayed a bit higher. Thin blue vertical line (around the middle of the green histogram) marks the 50% correct matches' value, i.e. half of the correct matches are left of this line. The number displayed next to it is the positive likelihood ratio **(hits/false alarms)** at this line. Magenta curve is the positive predictive value (**hits / (hits + false alarms))** as a function of matching error value. It is ~1 for the left side (for very small matching error values) and 0.5 for the right side (for very large matching error values). This value is the empirical probability of pair to be a correct match for a given matching error value.

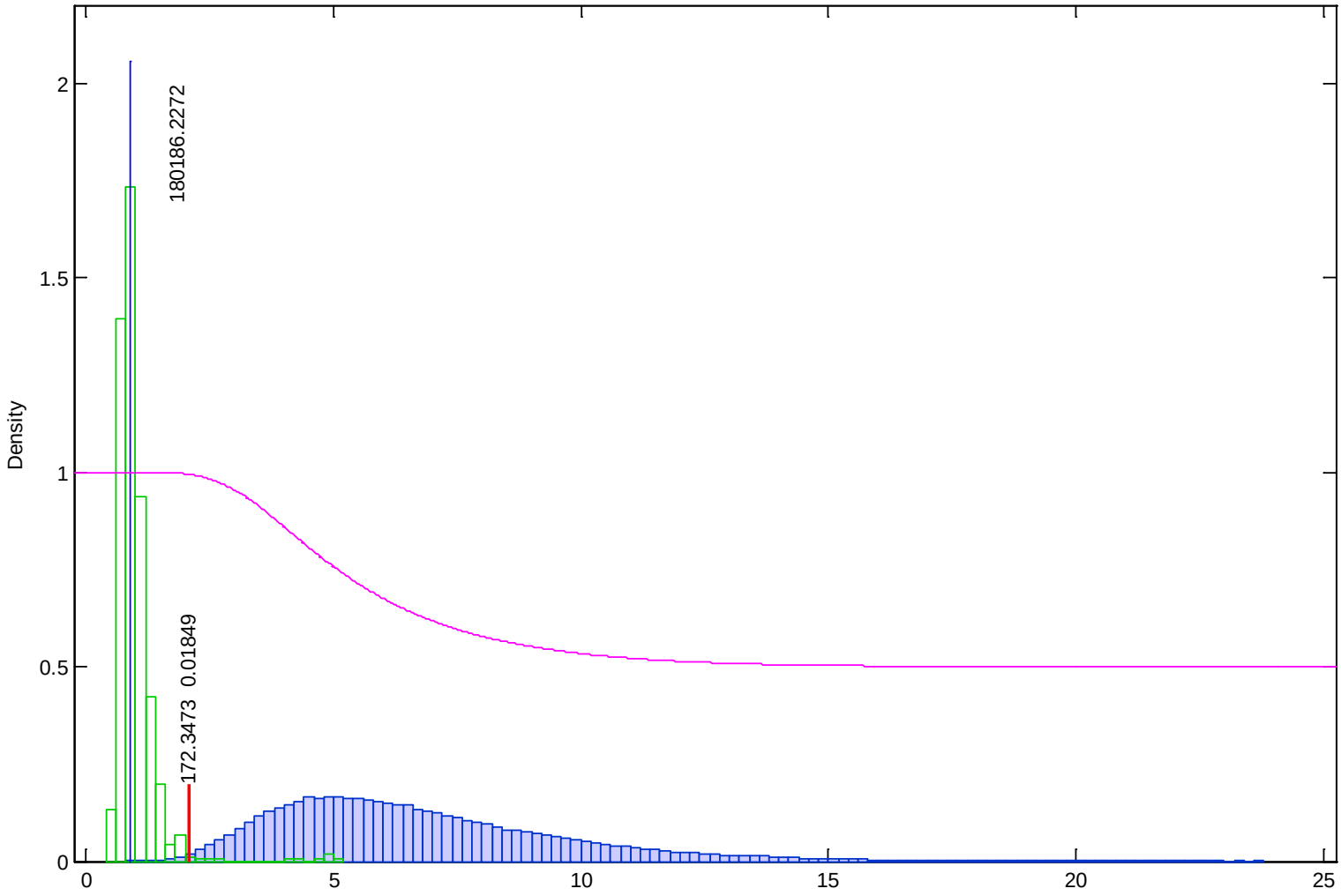
0.125 cm or 84 pixels



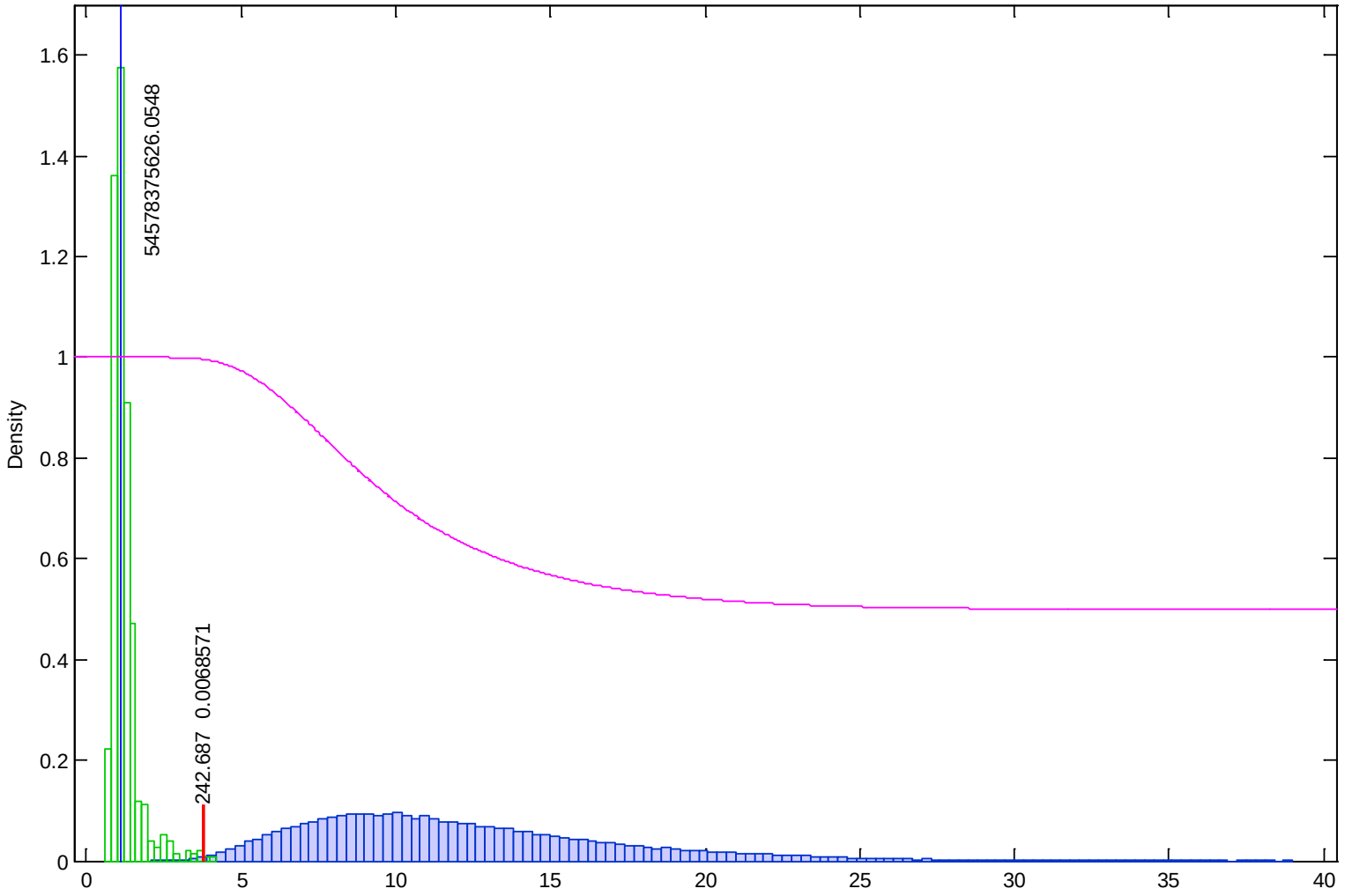
0.25 cm or 167 pixels



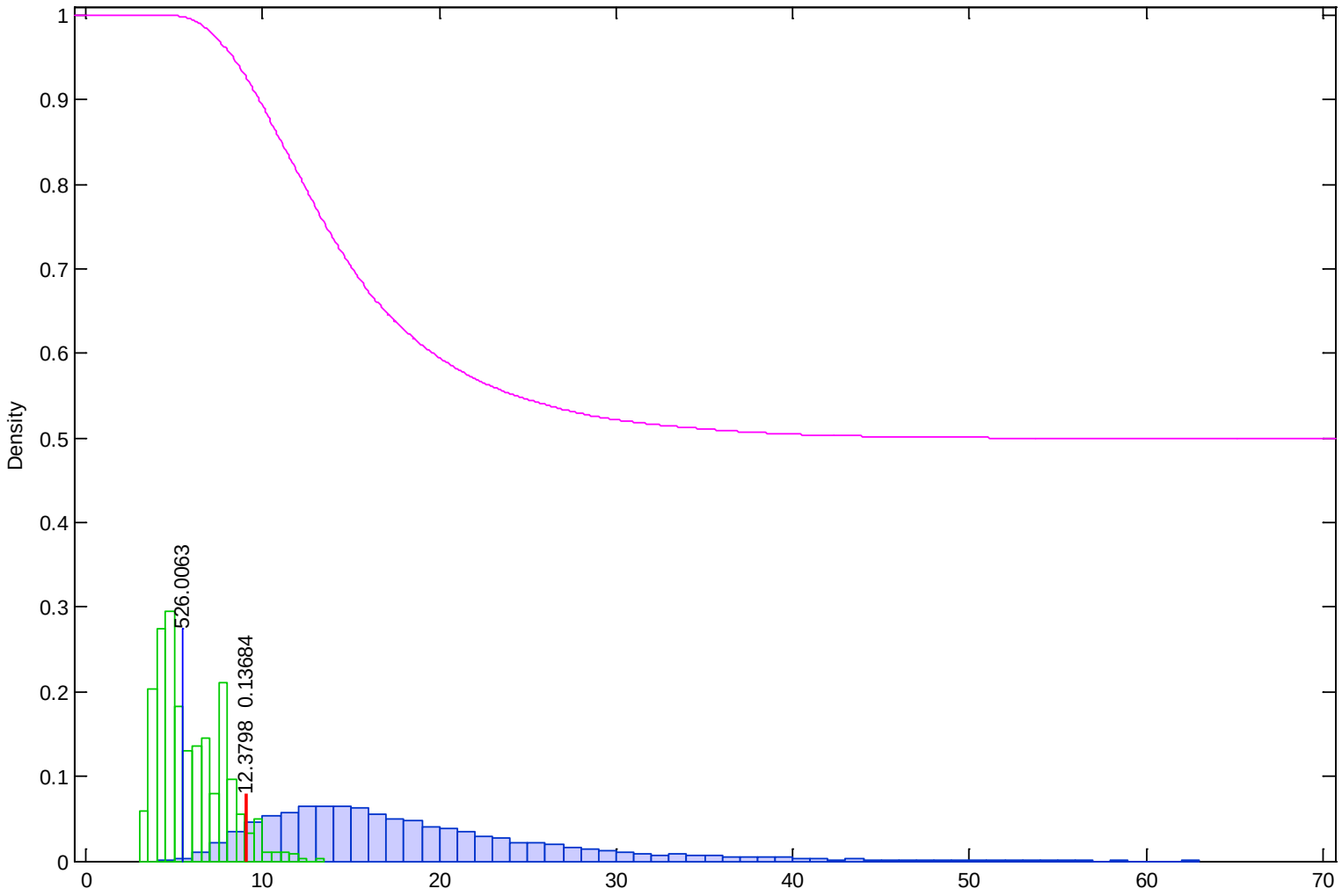
0.5 cm or 333 pixels



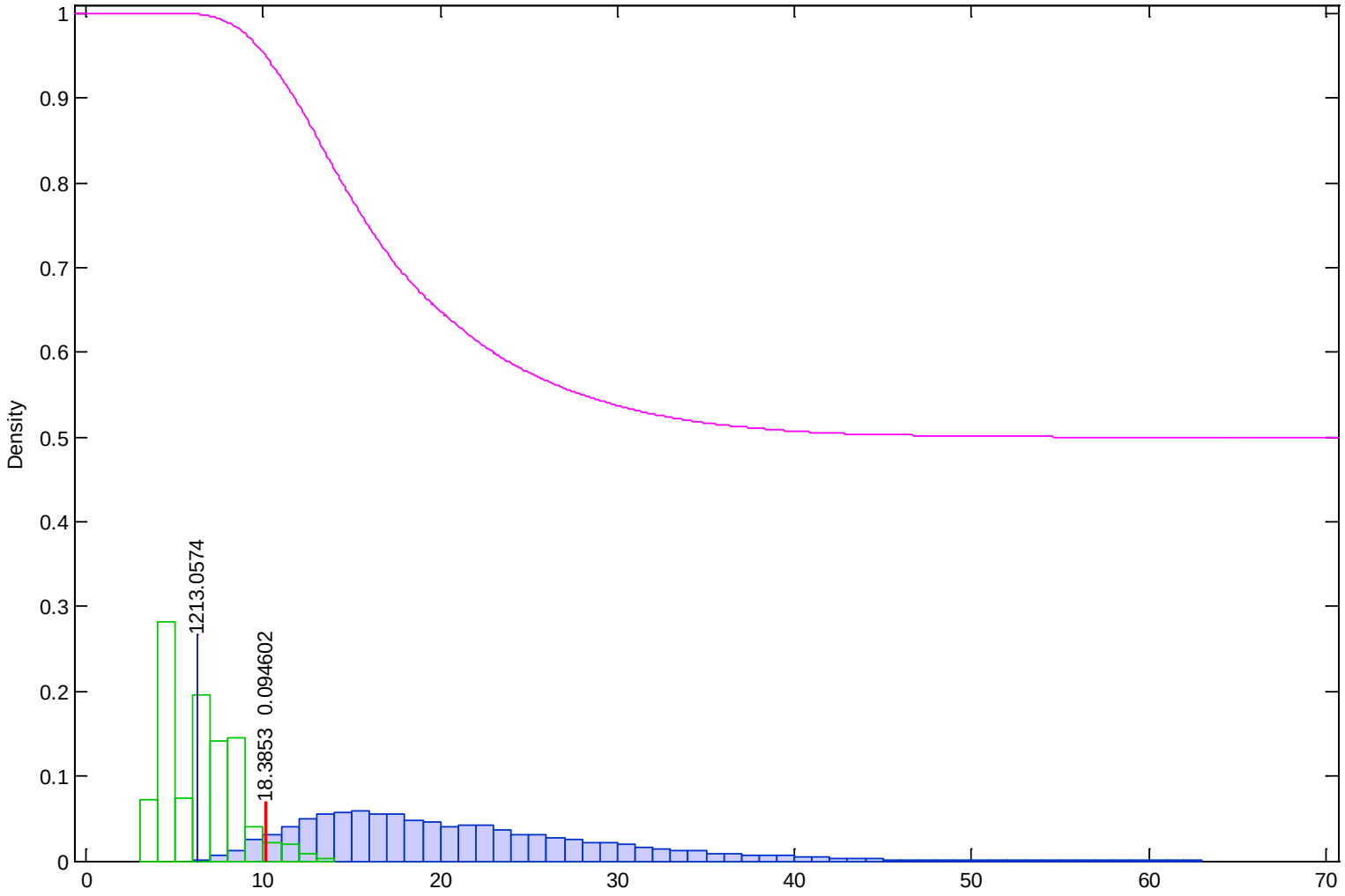
1 cm or 667 pixels



3 cm or 900 pixels



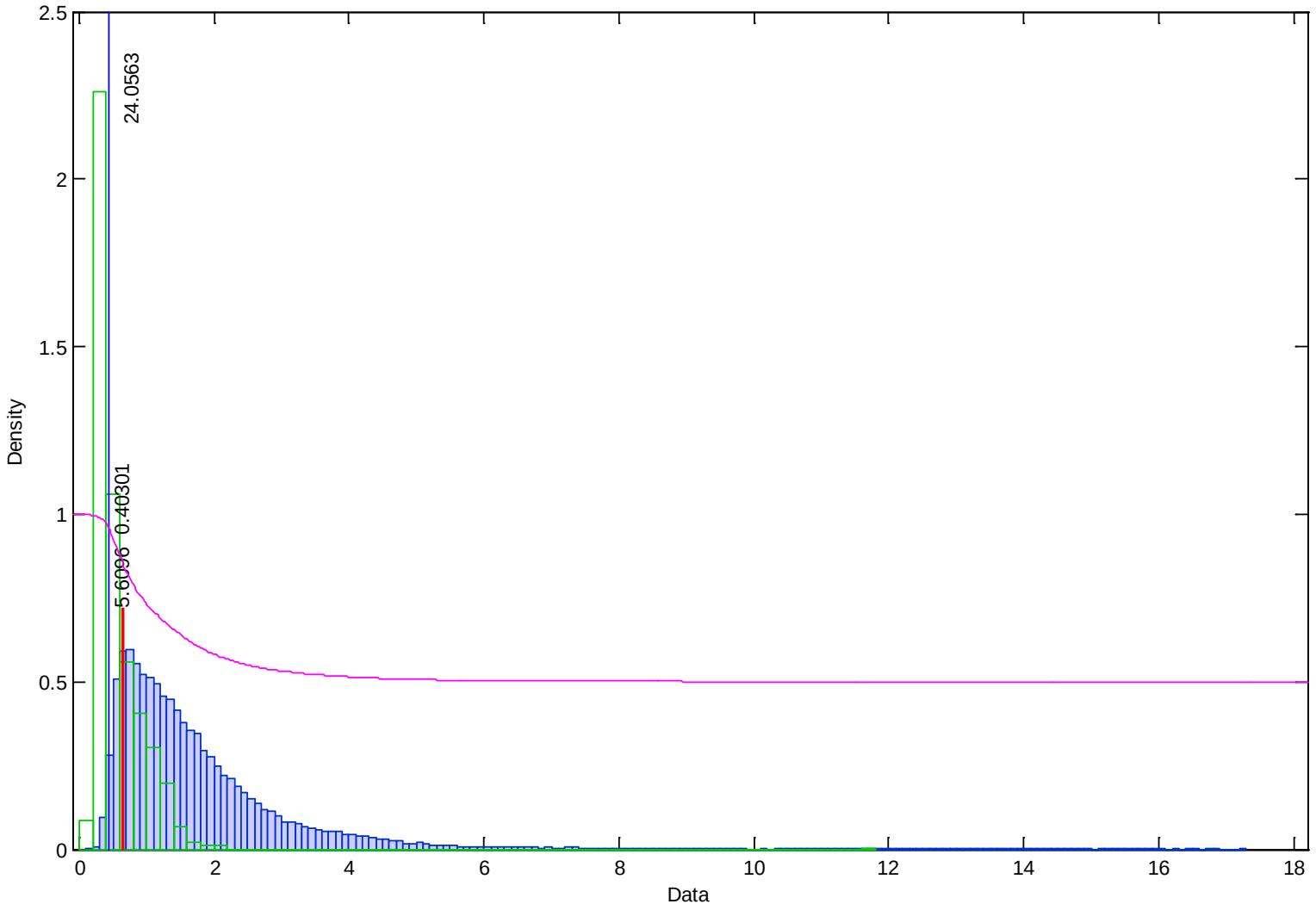
4 cm or 1200 pixels



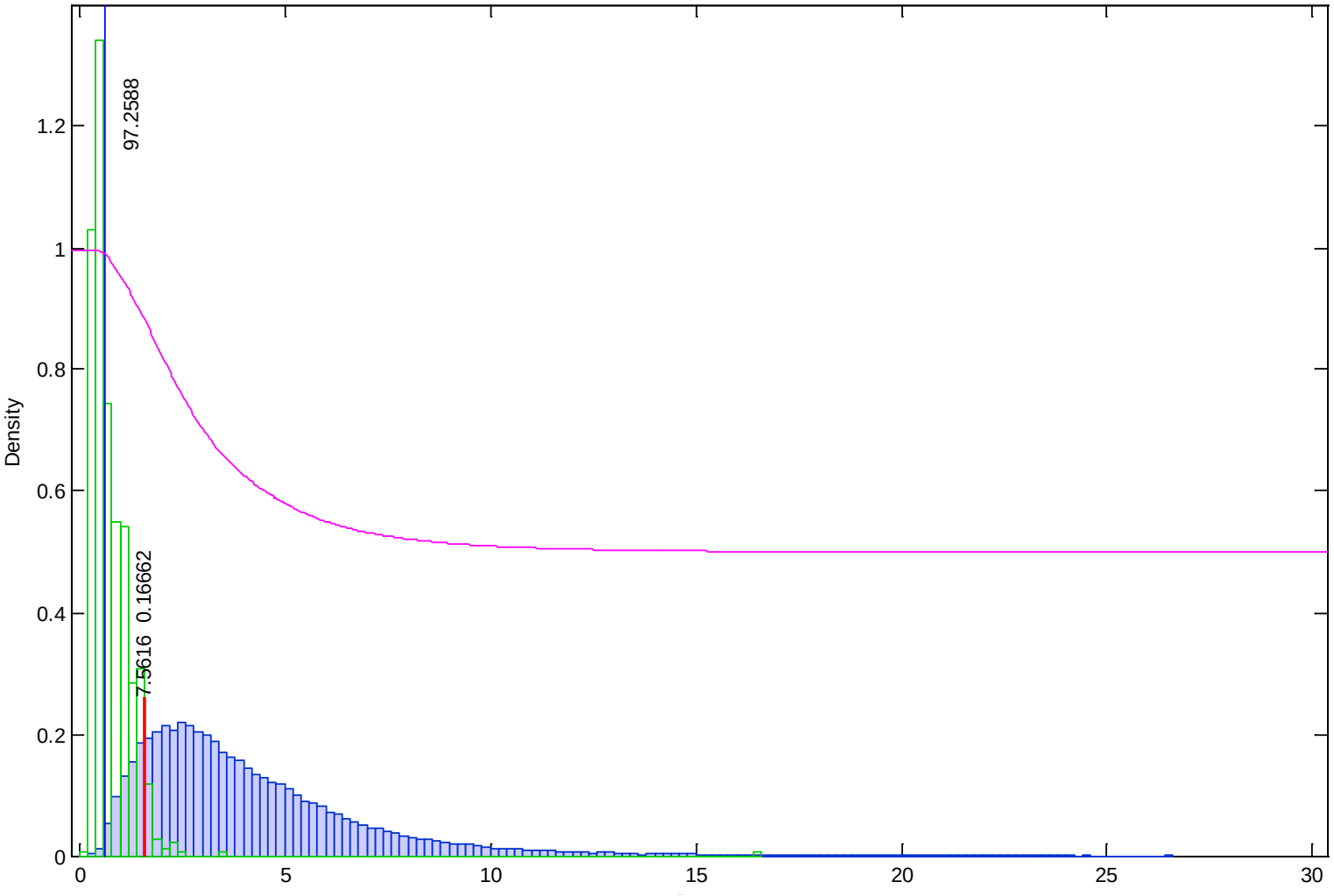
Appendix 3

Histograms for the Perspex matching experiments

1 cm or 225 pixels



2 cm or 450 pixels



5.333 cm or 1200 pixels