U.S. Department of Justice
Office of Justice Programs
National Institute of Justice

Award Number 2018-R2-CX-0034

Project Title:

Physics and Statistical Models for Physical Match Analysis Utilizing 3D Microscopy of Fracture Surfaces

Principal investigator:

Ashraf F. Bastawros, Professor, Aerospace Engineering, Iowa State University, Ames IA
bastaw@iastate.edu, Tel: 515-294-3039

Submission Date:
October 10, 2023

DUNS: [redacted] **EIN:** [redacted]

Recipient Organization: Iowa State University
1138 Pearson Hall
505 Morrill Road
Ames, IA 50011-2103

Recipient Identification No.: [redacted]

Project/ Grant Period: January 1, 2019 – June 30, 2023

Final Project Report

Signature of Submitting Official:_____
Ms. Marva Ruther, MS, CRA; Senior Award Administrator
Office of Sponsored Programs Administration Iowa State University

Table of Contents

1. SUMMARY OF THE PROJECT

1.1 Project Goals and Objectives

This fundamental and applied research proposal provides the scientific and quantitative paradigms for forensic comparative analysis of fractured and torn metal and plastic objects, utilizing three-dimensional (3D) digital representations of their fractured surfaces and their replicas. An integrated validation study will be conducted on sequentially broken hacksaw blades and their topological replicas. The methodology utilizes 3D spectral analysis of the fracture surface topography, mapped by 3D microscopy (developed under previous NIJ funding). The framework focuses on quantitative statistical measures for the full range of the subclass and individual characteristics of the examined object, and identification of material role on the fracture-feature-characteristic scales. The proposed quantitative forensic comparisons have potential applications across a broad range of fractured materials and/or toolmarks, with diverse textures and mechanical properties.

The proposed framework will assist the examiner by providing analytical and statistical support for his/her decision, by estimating with confidence bounds the probability of a true match, helping to arrive at a quantitative match decision. Our methods will use statistical learning tools trained on databases of both matches and non-matches.

Successful advancement of the proposed technique has the potential to provide a new investigative machine-based analysis with quantified error probabilities that can be applied in performing physical matches for a variety of materials. This research will be conducted in response to the NIJ's expressed need for expanding knowledge underlying forensic science, and in collaboration with forensic scientists working in forensic laboratories.

1.2 Research Questions

Fractured fragments with rough and irregular surfaces, found at crime scenes, are recognized as "being a match" using comparative microscopy and physical pattern analysis. We utilized the surface topography of the fractured fragments, measured by 3D microscopy to provide a quantitative basis for declaring a match with quantified probability and error rates. We set the comparison scale to capture the transition of fracture surface topography to become non-self-affine (surface roughness independent of observation window). At such transition scale, fracture surfaces exhibit unique roughness characteristics; dictated by the intrinsic material properties, microstructure and imperfections, and exposure history to environment and external forces. In the case of the examined class of hardened alloys, which are common in cutlery and tool steel, the identified scale can be related to the microstructure size-scales and is found to be approximately two to three times the grain diameter. This scale closely relates to the characteristic distance necessary for the initiation of cleavage fractures in semi-brittle and hardened metallic alloys. Consequently, the imaging scale required is approximately 20 times the grain-diameter. For each pair of fractures, a k−overlapping images were recorded, with different overlap ratios. The identification of spectral representations for various wavelengths and critical features on the fracture surface was accomplished using the mathematical Fourier Transform. Subsequently, quantitative topological descriptions were devised for the image pairs by performing correlation comparisons on two spectral bands encompassing the limiting fractal scale. These frequency bands are bounded by frequencies corresponding to 2-4 and 4-8 grain diameters. Consequently, each set of fracture pairs under examination yields a total of k−pairs of correlation values. A statistical learning tool was then formulated, employing multivariate statistical analysis methods to classify the fracture pairs based on this collection of k−pairs of topological descriptors. This classification

offers a foundation for establishing the uniqueness of forensic comparisons. The performance of our statistical learning methodology is tested on a robust training data set and validated on a set of thirty-eight different broken pairs of either knives broken in bending, or stainless steel rods with similar grain sizes, broken in tension or bending. The efficacy of the framework under different modes of loading is examined by application to a set of 9-twisted knives to failure. All broken pairs were classified with very high accuracy. The framework lays the foundations for forensic applications with quantitative statistical comparison across a broad range of fractured materials with diverse textures and mechanical properties**.**

1.3 Research Design, Methods, Analytical and Data Analysis Techniques

1.3.1 Introduction

Consider the example of a crime scene where investigators have found the tip of a knife or other tool, which appears to have broken off from the rest of the object. Later, investigators recover a base that appears to topographically match and they wish to show that the two pieces are from the same knife in order to use that evidence later at trial. To this extent, the analyst comparison relies on subjective pattern recognition methodologies. Scientific testimony used in a criminal or civil trial must be "not only relevant but reliable", according to the Supreme Court decision Daubert v. Merrell Dow Pharmaceuticals, Inc (1993). The application of this ruling forced a reconsideration of some previously acceptable forensic evidence and a re-evaluation of the scientific validation of its premises and techniques1. In 2009, The National Academy of Sciences issued a report, "Strengthening Forensic Science in the United States: A Path Forward", which evaluated the state of forensic science and concluded that,

[M]uch forensic evidence—including, for example, bite marks and firearm and tool-mark identification—is introduced in criminal trials without any meaningful scientific validation, determination of error rates, or reliability testing to explain the limits of the discipline.[2]

However, it should be noted that a considerable amount of prior work has been done to provide a quantitative and scientific basis for firearm and toolmark identification, for example, with the consecutive matching striae (CMS) method.[3–5] The report highlighted the need to develop new methods which have meaningful scientific validation and are accompanied by statistical tools to determine error rates and the reliability of the methods. To that end, the American Association for the Advancement of Science has recently published reports on the state of fire investigation[6] and latent fingerprint examination[7].

In the current framework, we focus on fracture matching, the forensic discipline of determining whether two pieces came from the same fractured object. The fracture mechanisms leave surface marks on both surfaces that could be utilized for matching fragments. Current forensic practice for fracture matching visually inspects the complex jagged trajectory of fracture surfaces to recognize a match using comparative microscopy and tactile pattern analysis[8, 9], where macro-features on a pair of fracture fragments are correlated as shown in Figure 1(a, b). Previous research has supported that the observed fracture patterns in metals are unique[10, 11] and that inspection via a microscope of the fracture surfaces by examiners can reliably validate matches[12]. However, this relies on subjective comparison without a statistical foundation, which may be flawed, as the 2009 NAS report argues: "But even with more training and experience using newer techniques, the decision of the toolmark examiner remains a subjective decision based on unarticulated standards and no statistical foundation for estimation of error rates.[2]"

It is therefore desirable to develop more objective methods using quantitative measures that can be validated with less human input for use in a criminal or civil trial. Here we propose to employ the fractal character of the topology of the fracture surfaces and their transition to become non-self-affine (self-affinity means scaling of the roughness with the observation window) with unique roughness characteristic[13] to define the suitable comparison scale and also develop the supporting statistical methods to perform forensic fracture matching using three dimensional (3D) topological imaging of the fracture surface details. The basis for physical matching is the assumption that there is an indefinite number of matches all along the fracture surface. The irregularities of the fracture surfaces are considered to be distinctive and may be exploited to individualize or distinguish correlated pairs of fracture surfaces[8, 14]. For example, the complex jagged trajectory of a macro-crack in the forensic specimen of Figure 1(a) can sometimes be used to recognize a "match" by an examiner or even by a layperson on a jury[8, 14]. However, experience, understanding, and judgment are needed by a forensic expert, to make reliable examination decisions using comparative microscopy and physical pattern match as indicated in Figure 1(b) to identify correlated macroscopic topological features. Indeed, the microscopic details of the non-contiguous crack edges on the observation surface of Figure 1(a, b) cannot always be directly linked to a pair of fracture surfaces, except possibly by a highly experienced examiner. There are many published studies and case reports concerning fracture or pattern matching of different materials such as rubber shoe soles, wood, glass, tape, paper, skin, fishing line, cable, and, most commonly, metal[15–30]. However, at about one tenth the scale of Figure 1(b), the 3D microscopic details imprinted on the topographical fracture surface of Figure 1(c) carry considerable information that could provide a quantitative forensic comparison with higher evidentiary value. Forensically, glass and metal fracture surfaces have been shown to have highly stochastic fracture

branches due to the randomness of the microstructure and grain sizes[10, 31], with limited prior attempts to quantitatively match two measured fracture surface topographies[12, 17]. The fracture surface topography contains many unique features over a wide range of length scales, which generally provide substantial information on damage initiation and propagation.

The material microstructure controls the micro-mechanisms of fracture and the microscopic crack growth path, while the loading direction sets the macroscopic crack trajectory[32]. Mandelbrot et al.[13] first showed the self-affine nature of fractured surfaces and relating its roughness via the exponent characterizing the scale invariant properties 'fractal dimension' to the material resistance to fracture. The self-affine nature of the fracture surface roughness has been experimentally verified for a wide range of materials (metals, ceramics and glasses) and static and dynamic loading conditions[33–37]. A key finding is the variation of such surface descriptors when measured parallel to the crack front and along the direction of propagation [38]. The cut-off length scale of the self-affine behavior was suggested as a unique length scale to characterize the microscale fracture process in ductile [34, 39, 40] and brittle/semi-brittle materials [34, 41, 42].

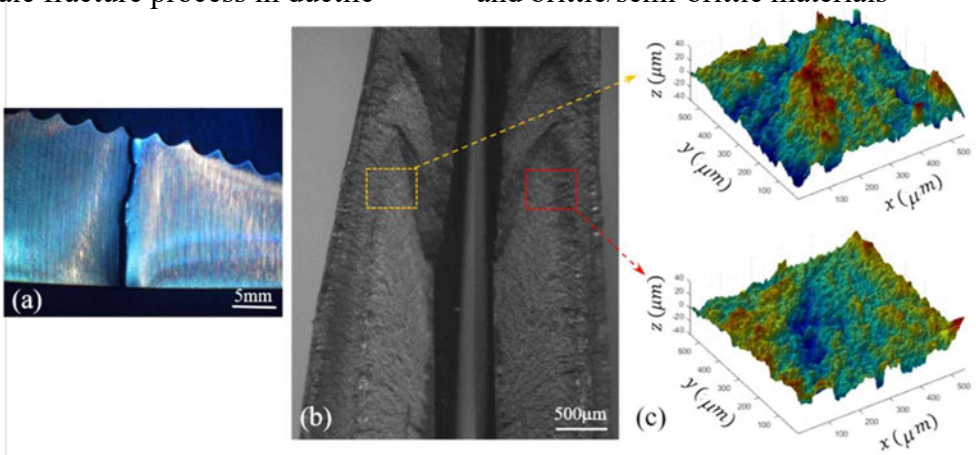

Figure 1: Association of forensic fragments. (a) Visual jigsaw match of the macroscopic crack trajectory at the typical examination scale. (b) Physical pattern match with comparative microscopy, with analyst focusing on macroscopic topological features. (c) 3D representation of fracture surface, showing detailed topographic features at the relevant comparison scale (~ 20 grains), utilized in current work.

We first present an overview of the method and the study objectives. Then we describe the sample generation method and the imaging process used to create training and forensically relevant data sets. We then provide a description of the statistical model, which discriminates the matching fracture surfaces from the non-matching surfaces.

1.3.2 Project Overview and Research Design

Our objective is to find the scale of distinctive features on a fracture surface with unique roughness characteristics, and then create a statistical method, which uses the features to match a pair of fragments in a way that is suitable for use as evidence in a crime scene. Motivated by the observations about the self-affine nature of fracture surfaces, it can be speculated that a randomly propagating crack will exhibit distinctive fracture surface topographical details when observed from a global coordinate that does not recognize the direction of crack propagation. This work explores the existence of such distinction of a randomly generated fracture surface at some relevant length scales. The distinction of these topographical features implies that they can be used to individualize and distinguish the association of paired fracture surfaces. Our approach uses the fact that the microscopic features of the fracture surface in Figure 1(c) possess distinctive or unique attributes at some relevant length scale that arise from the interaction of the propagating crack-tip process-zone and the microstructure details. The corresponding surface roughness analysis is shown in Figure 2 using a height-height correlation function, $\delta h(\delta x) = \sqrt{\left\langle \left(h(x+\delta x) - h(x) \right)^2 \right\rangle_x}$, where the $\langle \cdots \rangle_x$ operator denotes averaging over the x-direction. We see that at the small length scale of less than 10–20 μm, the roughness characteristic is self-affine (i.e. proportional to the analysis window scale). However, at larger length scales (> 50–70μm), this characteristic deviates

and reaches a saturation level, highlighting the individuality of the surface topography at such an observation scale. For the examined class of materials with average grain size of approximately dg = 25–35 μm, this scale is found to be about two grain-diameters. Dauskardt et al.[34] have examined the topography of a wide range of brittle failure of well characterized mild steel at extremely low temperature, and observed two ranges over which the fractal dimension is constant. The first range is 1-10microns corresponding to the cleavage step. This range of cleavage step will be non-unique, as it will be found in all surfaces of the same alloy that exhibit cleavage failure. The second range is of the order of twice to three times the grain size. It is shown that the fractal dimension is constant over a range of the order of twice or three times the grain size range for transgranular cleavage fracture, about twice the grain size range for intergranular fracture, and of the order of the grain size for the quasicleavage fracture[34]. In the field of fracture mechanics, it is postulated that cleavage failure occurs when the local stress ahead of the crack tip exceeds the fracture strength of the material over a characteristic distance, equal to two-grain diameter[34, 45]. This critical scale is required for cleavage crack initiation. However, it is apparent that such critical scale is also embedded in the topography of the fracture surface. When a microcrack is initiated at a hard-particle, it may be arrested if there is insufficient global driving forces to continue crack propagation[46]. Accordingly, the requirement of reaching critical stress over a microstructure critical distance will be maintained for continued crack propagation until the macroscopic crack reach an unstable propagation domain, and thereby set-forth the critical fractal scale on the topology of the fracture surface.

It is important to note that the reported fractographic details are reported for mild steel, examined at extremely low temperature, below the ductile to brittle transition temperature (DTBTT) of (−95ºC), where fracture occurs before general yielding due to slip-induced cleavage.

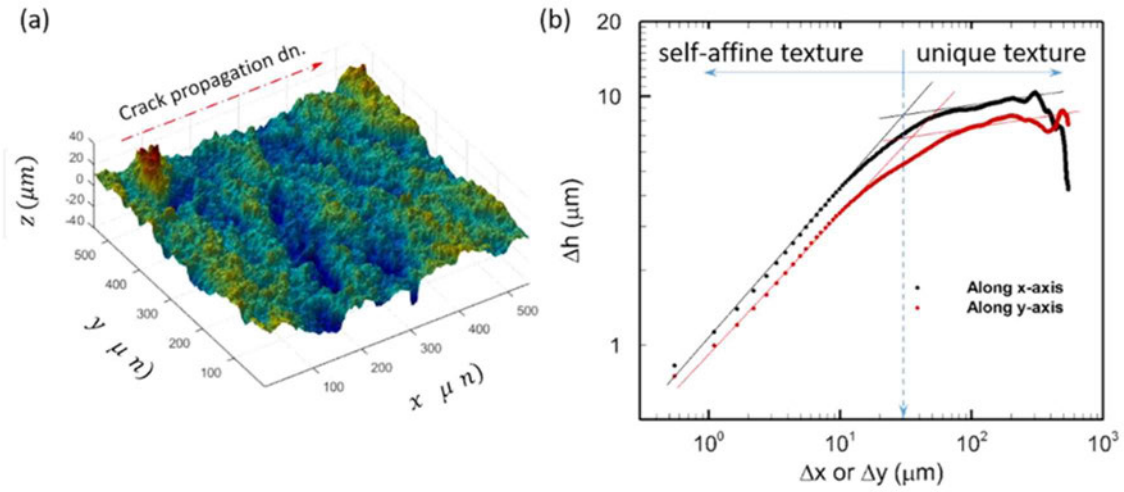

Figure 2: Fracture surface characteristics. (a) 3D surface topography rendering of fracture surface, showing a biased orientation of the low-frequency texture of the fracture surface. The direction of crack propagation is along the x-axis. (b) Height-height correlation variation with the size of the imaging window, showing the domain of the self-affine deformation and the deviation of the fracture surface characteristics at higher length scales (> 50–70μm), which could be used for matching purposes.

For the current examined alloy of AISI 440C stainless steel, a common alloy for cutlery and knives, the alloy has up to 1.2% carbon content in order to make the alloy hard and remains sharp. Such carbon content also shifts the DBTT to be above the room temperature[47]. Accordingly, it is no surprise that the examined alloy in the form of rods of knives at room temperature show similar fractal character to mild steel alloys tested below their DTBTT [34, 45].

Furthermore, it is also important to note that the requirement of local stress ahead of a crack to exceed the fracture stress over a microstructurally significant distance[45] should be viewed in statistical terms. The characteristic dimension represents the location the weakest link for the fracture process to occur[46]. The cleavage fracture process-zone is statistical in nature[48], as a finite volume of the material ahead of the crack tip should include a local defect to nucleate cleavage crack. This very critical argument to show the statistical basis for the individuality of the fracture

surface. Two nominally identical articles from the same material lot might show very different toughness (resistance to fracture) and failure strength values. By extension, we speculate also that such statistical differences will result in different local fracture surface topology because of the statistical randomness of the microscopic spatial location of the critical fracture-triggering particle. Susceptibility of cleavage fracture is sensitive to microstructure (grain size and carbide population), yield strength, stress state (triaxiality), and environment (temperature and radiation). Similarly, the fracture topography will exhibit unique microscopic feature signatures that exist on the entire fracture surface. We extend these parameters to generally include the material microstructure, the intrinsic material resistance to fracture, the direction of the applied load, and the statistical distribution of imperfections within the microstructure. This work utilizes these foundations of the fractal nature of the fracture surface topography and the statistical nature of initiation of cleavage fracture to explore the existence of a critical length scale required for imaging comparison, and the corresponding distinctive or unique attributes of the fracture surface, as well as their applications to forensic comparison of fractured surfaces.

The height-height correlation function at this transition scale captures the uniqueness of the fracture surfaces, so we can use that function's behavior in setting the observation scales (i.e., field of view, FOV, and imaging resolution) for comparing matching and non-matching surfaces to produce a statistical model of each topographical class's behavior for use in classification. This imaging scale should be greater than about 10-times the self-affine limit scale to avert signal aliasing. Further, we can combine multiple observations at different length scales or topographical frequencies (around the self-affine saturation scale and readily segmented in the frequency space) of a single surface into one model in order to improve the ability to discriminate between surfaces of the same class or materials and manufacturing processes (for instance, individualization of a

pry tool from a similar batch of identical tools). The statistical model can produce a likelihood ratio or log-odds ratio of a new set of surfaces belonging to either class, which are common outputs of forensic matching methods[49–55]. The creation of this model can also be used to estimate probabilities of misclassification and compare to the actual rates of misclassification in the test data. Conceptually, this is similar to forensic matching models, which are used in fingerprint identification, and bullet matching. In fingerprint identification, features (minutiae) on the reference print and the latent print are marked and then the pair is given a score based on how well the two match, which may be part of a probabilistic model reporting a likelihood ratio or other probabilistic output[56]. The Congruent Matching Cells approach for matching breech face impressions on cartridge cases in ballistics takes a similar approach: it divides the scanned surfaces into cells and searches for matching cells on the other surface. It then uses this as an input to a statistical model, which outputs a likelihood ratio[57, 58].

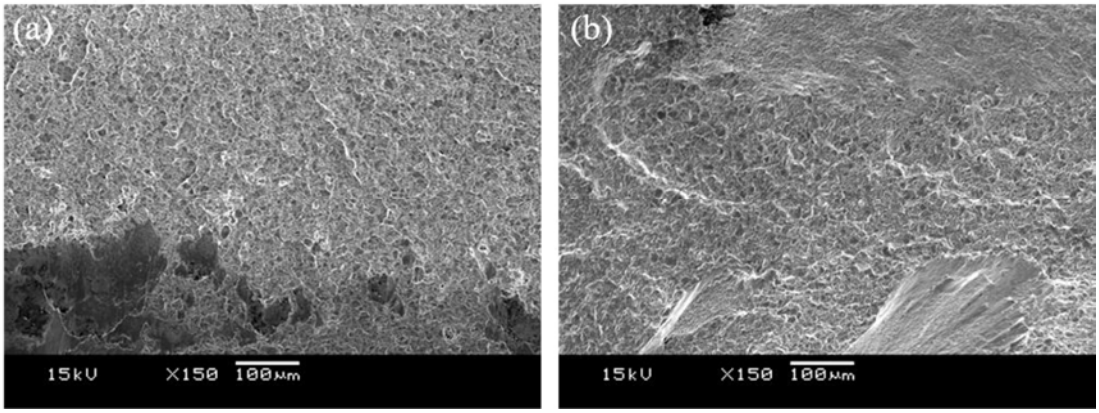

Figure 3: Detailed SEM analysis of a typical fracture surface for (a) bent broken knife broken in bending, showing topological details normal to the imaging plane, and (b) twist broken knife in torsion, showing in-plane swirl textures.

1.3.3 Sample Generation and Imaging

To mimic forensic articles found in a crime scene and that might undergo comparative analysis, we consider two main material classes: sets of rectangular rods of a common tool steel material (SS-440C) fractured under control tension and bending configurations (Figure S.2(b, d)), and sets of knives (Figure S.1) from the same manufacturer (Figure S.1(c, d)), fractured at random employing the fixture shown in Figure S.1(b). Figure 1(a) shows atypical pair of fragments, generated for this study. The average grain size for both groups was approximately dg = 25–35 μm. Four different sets of samples were established with nine specimens in the two sets of knives and ten specimens in the two sets of steel rods. To show the generalization of the approach for modes of loading, an additional set of 9-knives (Figure S.1(e)) was tested by random twisting utilizing the fixture shown in Figure S.1(b). The fracture surface topography would be influenced by a combination of fracture loading modes; that is mixed mode of the tensile mode-I and tearing mode-III loading as shown in Figure 3(b). The scanning electron microscope (SEM) images shows subtle differences between the Modes of loading. Figure 3(a) shows cleaved grains in a direction normal to imaging plane due to the pulling action (mode-I) under bending. Figure 3(b) shows swirl texture due to the combined out of plane tensile (mode-I) and in plane tearing (mode-III) loading. These topographical textures are very different and clearly show the critical role of external loading direction. Further details about sample preparations are given in Section S.1.

For clarity, we refer to the surface attached to the knife handle as the base and the surface from the tip portion of the knife as the tip and apply the same terminology to samples from the rectangular steel rods. The microscopic features of pairs of fracture surfaces were analyzed by a standard non-contact 3D optical interferometer (Zygo-NewView 6300), which provides a height resolution of 20 nm. Utilizing the results of the height-height correlations of Figure 2(b), the

transition scale commences at around 50–70 µm to become non-self-affine and saturate, rendering a required imaging FOV of about 500µm. For the examined material systems, this scale amounts to 2–3 times the grain size (consistent with the fracture process zone for cleavage fracture[45]), and the FOV should cover 20–30 grain-diameters. Accordingly, an optical magnification of 20X is employed, providing a 550µm FOV and 0.55 µm/pixel resolution (Figure 2(a)). Two fragments were aligned for imaging relative to their rectangular edges and their lower right corner. Image mis-registration can greatly affect the correlation estimations between a pair of images. However, the implemented procedure in this work to utilize the spectral (frequency) space is very tolerant to linear mis-registration of up to 20% of the FOV and several degrees of angular miss-registration, further elaborated in S.2.

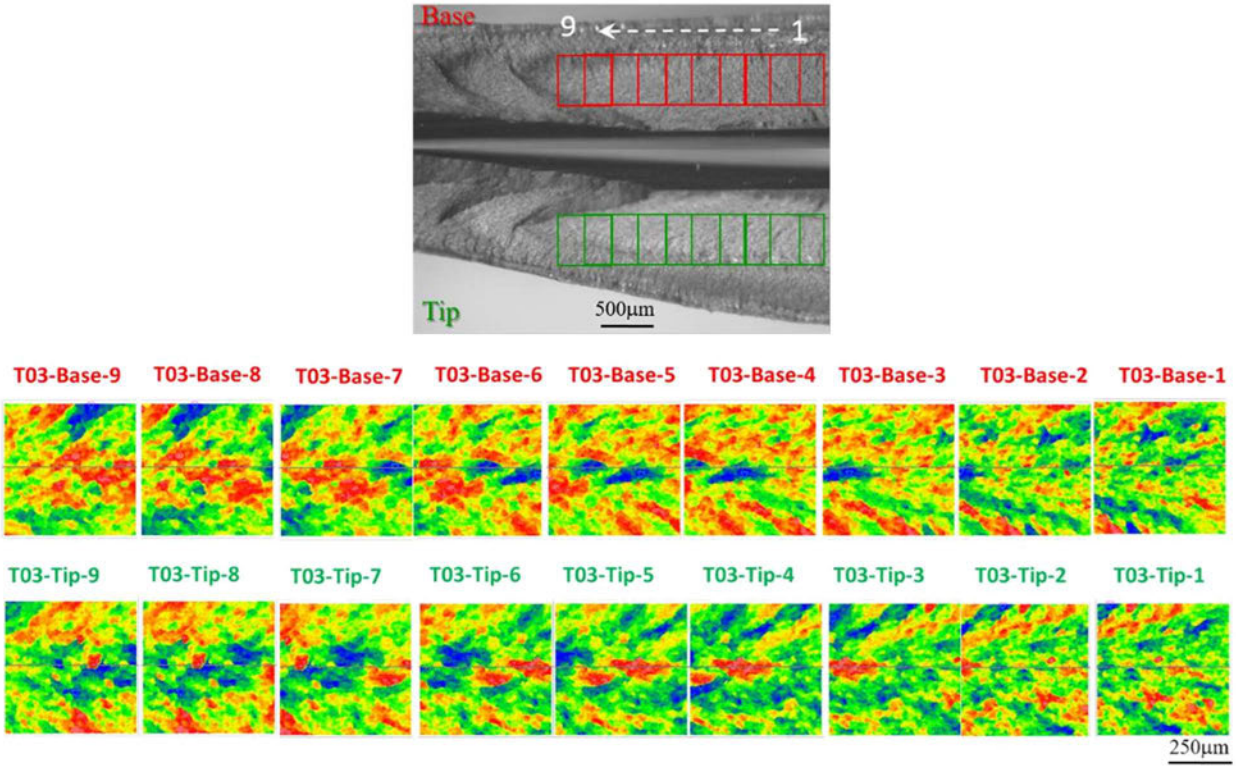

Figure 4: A fracture pair showing the alignment of series of topographical images, relative to a reference coordinate w.r.t. the right edge of the fractured article. A series of k = 9 topographical images with 75% overlap between successive images, rendering three fully independent sequel images on the fracture surface.

A series of k-overlapping surface height 3D topographic maps were acquired from the pairs of fracture surfaces (Figure 4; k = 9), and quantized using Fourier transform based power spectral analysis as summarized in Figure 5 in (a) the image analysis step. The choice of overlap means there are three full independent sequential images on a surface. Multiple overlapping images were needed to overcome problems arising from missing grains between pairs of the fracture surface and/or the special circumstances of complex tortuous path of fracture. The effect of the number of images and overlapping ratio will be further discussed in Section 3. Additionally, having a super-image of stitched FOVs results in mis-registration at the overlapping boundary of the stitched images, leading to an additional interfering frequency within the band of comparison.

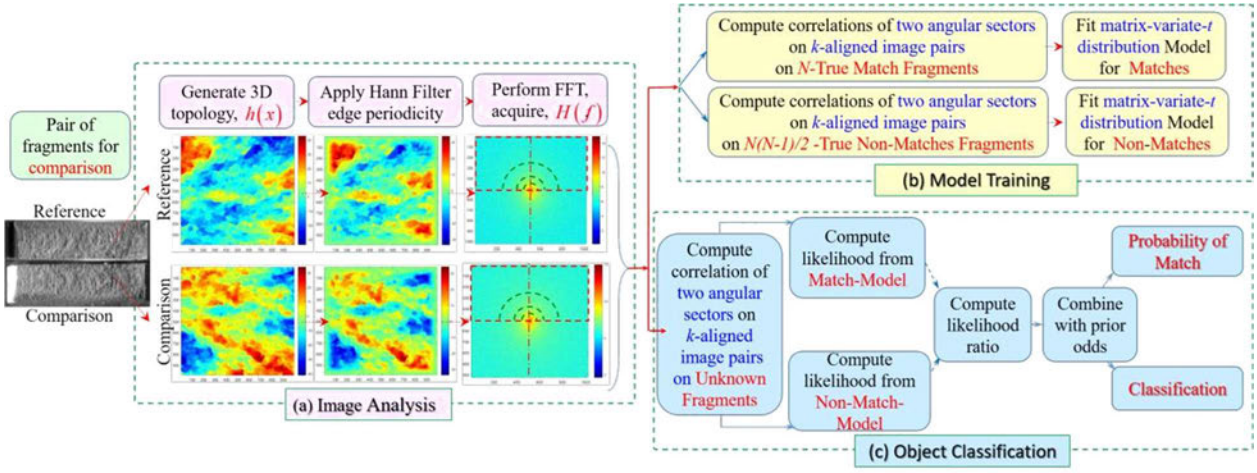

Figure 5: Flow chart showing (a) image spectral analysis, (b) model training, and (c) classification of new objects to provide classification probabilities. For a new field object, an examiner would use (a) to image the object and perform (c) using a model trained in (b) on samples of the same class to guide forensic conclusions.

1.3.4 Image Spectral Analysis and Frequency Correlations

From the 3D imaging of the fracture surface, the measured height distribution function $h(x)$ is acquired to define the topography of the fracture surface at every spatial point, x on the fracture

surface of a pair of fragments, shown in Figure 5(a). Each wavelength on the fracture surface has a distribution, in the frequency domain *H(f)*, which is acquired using a Fast Fourier Transform (FFT) operator. For example, grain size has a distribution of frequencies across the spectrum rather than one specific frequency. Similarly, other microscopic fracture features have a range of spectral distributions[59, 60]. For a pair of fractured surfaces, the population of these features contains relevant information about the physical fracture processes present at each length scale (e.g. dimples and voids at the sub-microscale, and river marks at scales of tens of microns).

The spectral space analysis provides a straightforward segmentation of the surface topographical frequency ranges for comparison. After calculating the spectra of each pair of images, each spectrum was divided into multiple radial sectors. The segmented angular sectors for the frequency range (0°, 180°) represents the entire data set, because the amplitude of *H(f)* exhibits inversion symmetry. The spectral array size is proportional to 2^n, as this is a mathematical feature of the FFT. For the image size employed in this work, a spectral array of 1024 by 1024 is acquired, although only the upper half is utilized because of symmetry. The radial segments for comparison in the frequency domain (marked on the FFT spectral representation in Figure 5(a)) are chosen to reflect the physical process scales and the corresponding wavelength, identified from the height-height correlation of Figure 2(b).

To compare the surfaces of two fragments, two-dimensional statistical correlations between their spectra are computed in banded radial frequencies, yielding a similarity measure on each frequency band for the corresponding *k*-pairs of images. As noted earlier, the increments for the bands' frequency are determined by the scale of the image and the material microstructure, covering the transition scale of the height-height correlation of Figure 2(b). A training data set with *N* fracture surfaces is utilized to estimate the correlation distribution among the two selected

frequency bands on all k image pairs for both the population of true matches and true non-matches fracture surfaces. For establishing a statistical match, these distributions, shown in Figure 6, form the basis for our classification and matching process strategy, following the two modules summarized in Figure 5, (b) Model training on an initial data set, and (c) performing classification of new sample(s).

1.3.5 Statistical Model Training/Fitting

A statistical model will be developed to distinguish matching from non-matching fracture surfaces. Employing a training data set, the behavior of the frequency band correlations in the population of matches and non-matches has to be estimated and modeled. The proposed framework provides a separate model for each class (i.e., match and non-match). The model training process, highlighted in Figure 5(b), entails:

1. Choice of controlled and robustly characterized data set of fractured pairs to train the model.
2. Computation of the correlations for the frequency bands for the sets of k images for all N matching and $N(N-1)/2$ non-matching surface pairs.
3. Employing the Fisher's z transformation on the correlation data to stabilize variance[61].
4. Fitting the models using a matrix-variate distribution (as detailed later in this section) to describe the distribution of true matches and true non-matches. The matrix-variate models account for the difference in location of the correlations and account for the covariance of the repeated observations across the surface.

The frequency band correlations for one of the examined data sets (K-1-1) are shown in Figure 6. The proposed method's discrimination ability can be judged from the clear separation of

matching and non-matching surfaces within two separate clusters. The data in this illustration were derived from $N = 9$ base-tip pairs from fractured knives. A series of k = 9 overlapping images were taken from each base and tip fracture surface, resulting in $N \times 2k = 162$ total images (81from the tips and 81 from the bases). Additional details are given in the S.1 for different data sets. In this example, image pairs for when the tip and base surfaces were from the same knife are true matches ($N \times k = 81$ matched-pairs), while those pairs for which the tip and base surfaces were from different knives are true non-matches ($N(N-1) \times k = 648$ unmatched-pairs). Furthermore, there

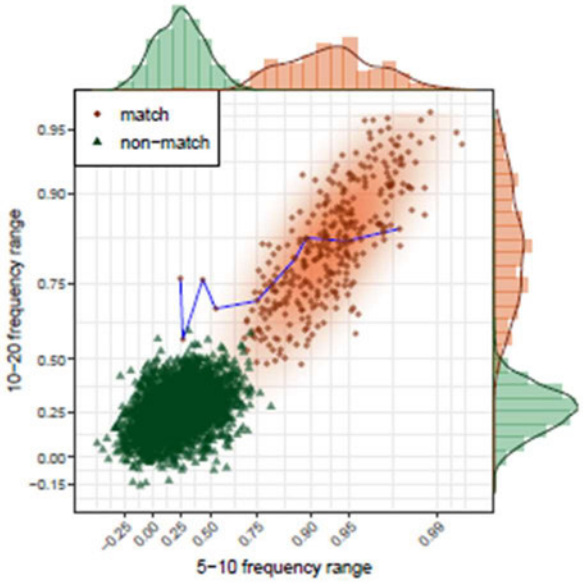

Figure 6: Scatter plot of correlations for 81 matched pairs and 648 non-matched pairs from training set K-1-1 for the 5–10 and 10–20 mm^{-1} frequency ranges on a Fisher-z (nonlinear) axis. We see that true matches and true non-matches are distinguished in this example by features in the 5-10 and 10-20 mm^{-1} frequency ranges. The connected points show the values of nine overlapping images from the same surface, indicating that while some individual images may not distinguish matches from non-matches, taking an ensemble of images from the surface importantly improves the ability to discriminate between the two classes in this data set.

is one image-pair among the true matches in Figure 6 which cannot be distinguished from the true non-matches and three other pairs that are ambiguous. To further improve the discrimination,

This resource was prepared by the author(s) using Federal funds provided by the U.S. Department of Justice. Opinions or points of view expressed are those of the author(s) and do not necessarily reflect the official position or policies of the U.S. Department of Justice.

considering multiple k − observations from the same surface would distinguish it from the nonmatches, since the other observations on that surface are well-separated from the non-matches. In the current framework, we take the information from every pair of images and collectively based on the model, a decision is driven accounting for the fact that the images are not independent, i.e. overlapping and coming from the same fracture surface. The role of imaging repetition or overlap, may improve the signal to noise ratio. Figure 7 summarizes the correlation analysis over several ranges of frequency bands. A clear separation (lower values for the true non-matches and higher values for the true matches) can be observed for the 5–10 and 10–20mm^{-1} frequency-band ranges. Beyond these frequency ranges, there is some overlap, where the true match and the true non-match correlation distributions become less distinct and overlap more.

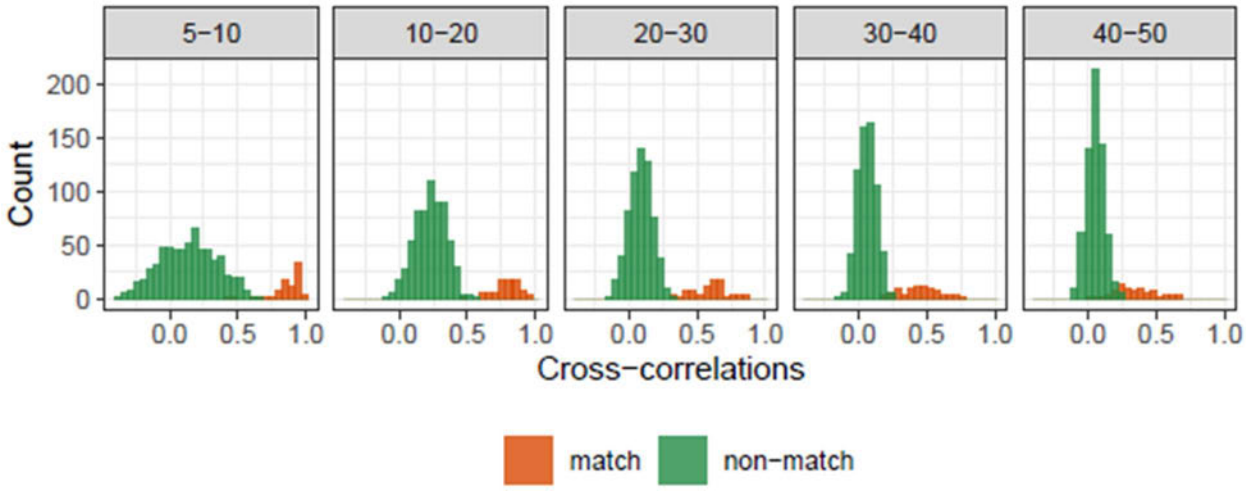

Figure 7: Histograms of correlations of true matches and true non-matches in one set of 9 surfaces with 9 images per surface split by frequency band. Lower frequencies are well-separated but higher frequencies begin to have more substantial overlap.

For the presented data set of $k = 9$ overlapping images for each fracture surface and two (or more) comparison frequencies, each comparison between a pair of fracture surfaces based on

the ensemble of nine images provides a 2 × 9 matrix of correlations. Our model needs to account for the lack of independence in the images from the same specimens. Accordingly, we propose using a matrix-variate distribution[62, 63] to model the densities of the matching and nonmatching populations, and, specifically, a matrix-variate *t* distribution (MxV*t*) because the data for the individual comparisons are approximately elliptically distributed but have heavier tails than a normal distribution. A definition of the distribution is in the Supplement Section S.3 and the density is defined in Equation S1 in the Supplement.

We use matrix-variate distributions to model the relationship between the two frequency bands in each image comparison and across the images being compared in the base and tip pair (e.g. Figure 6). Because of the overlapping-image structure of the data source, our model allows between-image correlations in the matrix-variate model to be related according to an autoregressive model of order 1 (or AR(1)) model (implying that immediately adjacent images can be correlated). The AR(1) model implies that the mean correlations in the two frequency bands remain the same across the images on a surface in the model. The parameters of the model are estimated using an expectation-maximization (EM) algorithm developed for the matrix-variate t distribution[64].

1.3.6 Classification of a New Object

Figure 5(c) summarizes the classification procedure. Suppose the fitted model has been trained on a set of *k*-images per fracture surface, yielding probability density functions f_1 corresponding to the population of true matches and f_2 corresponding to the population of true non-matches. Suppose also that there is a new pair of fracture surfaces that may or may not match. First, the

correlations for the *k*-aligned image pairs in the chosen frequency bands are computed and transformed, yielding a new observation *X*, which is a matrix of observations of correlations with one row for each frequency band and one column for each pair of images—here, a $2 \times k$ matrix. Then, presuming prior probability p of being a true match and prior probability $1-p$ of being a true non-match, we can find, by combining prior probabilities and the match and non-match densities from the model, the posterior probability that the two surfaces match as follows:

$$P(X \text{ is a match}) = \frac{p\, f_1(x)}{p\, f_1(x) - (1-p)\, f_2(x)}.$$

Alternatively, a likelihood ratio (LR) can be produced as $f_1(x) / f_2(x)$, which is a common method in forensic applications[49–55], and these LR results can be incorporated into a framework for expressing the uncertainty about the strength of evidence under different sets of assumptions[65]. The likelihood ratio can be combined with prior odds $(p/(1-p))$ to produce posterior odds:

$$\text{Posterior Odds} = \frac{p}{(1-p)} \times LR$$

with the conversion of odds O to probability P performed by the formula $P = O/(1+O)$. Here, odds and likelihood ratios are employed and reported on the logarithmic scale. Once the posterior odds are obtained, classification decisions can be made under the rules of evidence relevant to each forensic case.

For the purposes of illustrating the method, we are using an equal prior probability of being a match or non-match (i.e., $p = 0.5$ or a log prior ratio of 0). In an actual criminal or civil case, choosing a prior match probability would require carefully considering any other evidence or relevant information previously presented.

1.4 Expected Applicability of the Research

The performed study responds to a need for advanced methods that aid in analysis and comparison of fracture and torn surfaces, employing new tools and algorithms. The study provided a protocol utilizing 3D microscopy and statistical computational methodology to provide a quantitative assessment of fracture surfaces and their replicas for forensic comparison, combined with a statistical expression of the comparison. The performed study enhances the ability of forensic scientists to capture, visualize and analyze fracture patterns, and to express the likelihood of a match between patterns in statistical terms that support the qualitative and quantitative judgment of the examiner. The provided protocol might promulgate a standard operational procedure for fracture surface description, analysis, and comparison that can be used in forensic labs. The current study has the potential to provide forensic laboratories with techniques for:

(i) Examining continuous fracture pieces with relatively few gross macro features in metals and plastics, where visual physical matching may not be possible.

(ii) Providing a wealth of data to establish the foundation of the technique and highlight to prospects and limitation of using surface replicas.

(iii) Prospective extension of the technique for the analysis of torn surfaces (e.g. plastic bags, tapes and fabrics).

Our framework is also general enough to be applied, after suitable modifications to identification of the proper imaging scale, to a broad range of fractured materials and/or toolmarks, with diverse textures and mechanical properties. In doing so, we expect our novel methodology and findings to help forensic scientists and practitioners place forensic decision-making on a firmer scientific footing. This can help formalize the scientific basis for conclusive matching of fragments leading to quantitative and more objective forensic decisions.

2. PARTICIPANTS AND OTHER COLLABORATING ORGANIZATIONS

2.1 Participants at Iowa State University

2.1.1 Principal Investigators

(i) Professor Ashraf Bastawros, Aerospace Engineering, developed the forensic relevant data set, designed and planned the testing matrix for fracture testing, fracture topography imaging and spectral analysis of the fracture surface. Prof. Bastawros was also responsible for the overall steering, management, and progress reporting of the whole project.

(ii) Professors Ranjan Maitra and William Meeker, Department of Statistics, developed the statistical methodology for analysis of the fracture data sets.

(iii) Professor Barbara Lograsso, Department of Mechanical Engineering, oversaw the sample generation, including material selection, heat treatments, and generating randomly fractured sets of samples.

2.1.2 Graduate Students

(i) HyeokJae Lee, Current PhD Graduate Student, Aerospace Engineering, generated forensic relevant fracture pairs and supervised the imaging process and spectral analysis.

(ii) Joshua D. Berlinski, Current PhD Graduate Student, Statistics Department, development of the statistical analysis algorithms and performing comparison and statistical decision-making.

(iii) Bishoy Dawood, PhD 2021, Aerospace Engineering, generated forensic relevant fracture pairs and supervised the imaging process and spectral analysis.

(iv) Carlos Llosa, PhD 2022, Statistics Department, development of the statistical analysis algorithms and performing comparison and statistical decision-making.

2.1.3 Undergraduate Students

The following undergraduate students have worked on the project on sample imaging using confocal microscopy and performing spectral analysis of the fractured surface topology:

- Jessica Dwyer (Mechanical Engineering)
- Abdurrahman Moghram (Mechanical Engineering)
- Kaiser Aguirre (Aerospace Engineering)
- Austin Beckhardt (Aerospace Engineering)
- Aidan Furley (Aerospace Engineering)
- Mehmet Sefer (Aerospace Engineering)
- Noah Schumacher (Mechanical Engineering)
- Ryan Schiltz (Material Sciences and Engineering)
- Tyler Chandler (Aerospace Engineering)

2.2. External Collaborators

John Vanderkolk, Retired Forensic Scientist, Indiana State Police Laboratory, 5811 Ellison Road, Fort Wayne, IN 46804. Mr. Vanderkolk reviewed our sample plans and provided input to the forensic perspective. He showed us the protocols used to review such data and comparisons. The PI's communicate with John by email and phone conference calls on a quarterly basis and with semiannual demonstration and discussion.

Lauren K. Claytor, Forensic Scientist Group Supervisor, Virginia Department of Forensic Science, 700 N 5th St, Richmond, VA 23219. Ms. Claytor provided forensic sample set of hacksaws that were analyzied by forensic experts for comparison with the proposed statitical comparartive framework. She has overseen the generation of new samples set under releveant forensic environment**.**

3. OUTCOMES

3.1 Activities/accomplishments

The three major accomplishments have been reached under this grant.

(i) Developed mathematical and algorithm framework to identify imaging scales for different classes of materials with intrinsic length scales (metal grain size) and without intrinsic length scale (amorphous plastics), identifying key variables, probability distributions, and quantification of different sources of variability.

(ii) Performed a systematic study to different fracture scenarios under different loading scenarios (bending, twisting and combination) to achieve realistic forensic data set for fracture comparison.

(iii) Performed a validation study on a fully documented consecutively fractured hacksaw sample set (Claytor and Davis, 2010)[12], provided by our forensic scientist collaborator and utilized for the validation study of fracture match using fracture surfaces topology. Identified key issues of 3D microscopies of surface replicas.

In the next section, we summaries the main finding with additional supplementary materials provided in Appendix S1 to provide additional information about the methods and materials. An R[43]-software package to perform the model fitting and analysis, MixMatrix, and code to reproduce the analysis and figures is available[44].

3.2 Results and Findings

In this work, we attempted to employ the understanding of the fracture topographical details, combined with statistics and machine learning to arrive at a comparison decision. Our focus is to identify the domain of unique individuality through spectral analysis of the fracture surface

topography and provide a quantitative analysis for match probability and the corresponding error rate that is required to be reported [2]. Here, we will discuss some of the developed framework attributes, generalities and limitations.

3.2.1 Imaging scale for comparison

When comparing characteristic features on a fractured surface, identifying the proper magnification and FOV are critical. An optical image obtained by high magnification and a small field of view, will possess a visually indistinguishable characteristic. This is the range where surface roughness shows self-affine or fractal nature as noted on Figure 2(b). In this range, the material intrinsic local fracture mechanism show similar topographical surface features over the fractured surface (e.g. local cleavage steps and river patterns, and/or dimples and voids). On the contrary, employing lower magnifications will result in a lower power of identifying the class-characteristics of the surface. However, we showed that the transition scale of the height-height correlation function captures the uniqueness of the fracture surfaces. We found that this scale is about 2-3 times the average grain size for the class of materials examined here and undergoes cleavage fracture. Interestingly, this scale is consistent with the average the cleavage critical distance for local stresses to reach the fracture critical stress stress[34, 45] for cleavage fracture initiation, and typically extends to 2-3 times the grain size, or around 50–75μm for the tested material system. This critical microstructural size-scale for cleavage crack initiation is stochastic in nature as it statistically encompass the location of the critical fracture-triggering microscopic inclusion or particle[34, 46, 48]. Accordingly, the surface characteristic becomes statistically unique and non-self-affine at a larger scale, where it is eclipsed by the interference of the fracture process zone length scale. This scale sets the (i) the observation FOV to around 10-periods of such scale.

And (ii) the range of wavelengths or frequencies to perform correlations on pairs of fragments. When correlating the frequency bands in the range of 5–20mm^{-1} (i.e. (50–200μm wavelength) full separation and clustering can be clearly observed in Figure 6 for matched and non-match fracture surface pairs. Furthermore, beyond this frequency range, the match and non-match correlations overlaps, as noted on Figure 7. The identified imaging scale (which should be established for each class of materials) coupled with the statistical analysis framework provides a promising quantitative forensic comparison for a wide class of materials. However, the ability to produce accurate 3D topographical representations of the fracture surface is critical for comparisons. Two main issues may pose additional problems for the technique. (1) The class of materials that exhibit cleavage fracture possess a relatively planar fracture surface within several hundred microns. The limitation here is having the imaging depth resolution in the sub-micron range for the entire surface topography range. If the fracture surface exhibits ductile-tearing with large tortuous zigzag-paths and topographical variation in the millimeter-range[66], additional mathematical treatments would be required, similar to comparison of cylindrical surfaces (e.g., cartridge cases)[67, 68]. (2) Surface anomalies including grain fall-out and excessive corrosion will render lower fracture-pair correlations among matches, making the matching and non-matching classes less distinct. In such cases, a larger image set will be required in order to have the same power.

3.2.2 Classification performance

There are two datasets from the knives and two from the steel bars: “K-1-1” is the first set of images from the first set of knives (Figure S.1), and the imaging is independently repeated generating additional sets of images “K-1-2”and “K-1-3” for repeat analysis.“K-2” indicates the other set of knives, whereas “S-1” and “S-2” indicate the two steel bar samples (Figure S.2).

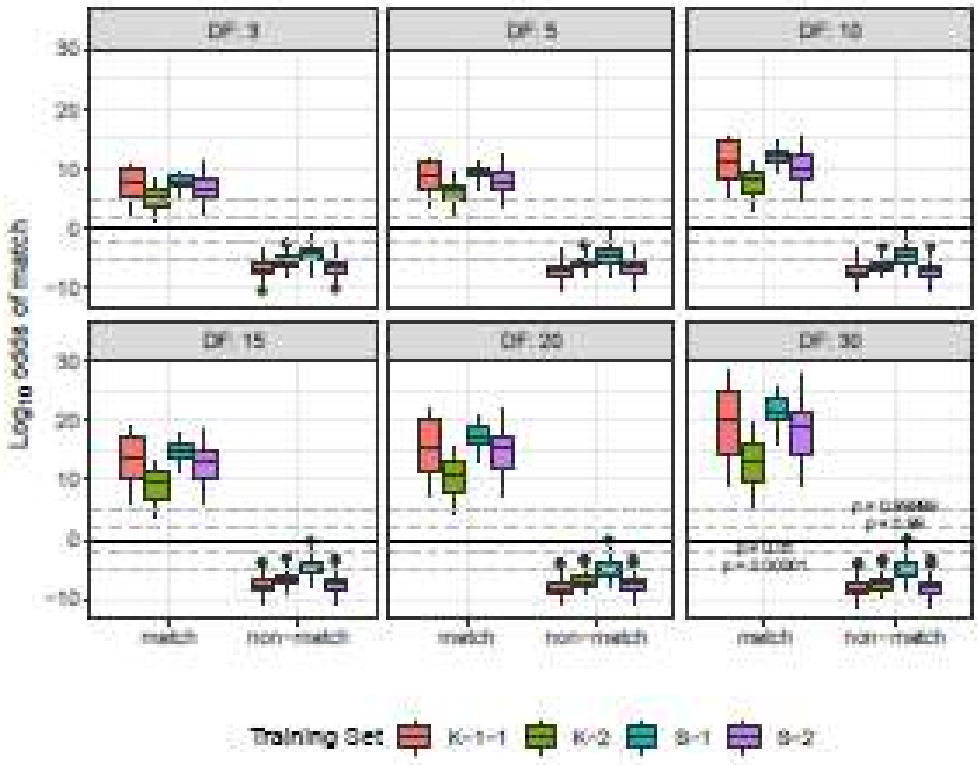

Figure 8: Log-odds of being a match split by training set and true class membership for matrix−t distributions with 3, 5, 10, 15, 20, and 30 degrees of freedom. A log-odds ratio greater than 0 indicates greater odds of being a match than a non-match. The predictions for each training set are on all four sets of surfaces.

Figure 8 shows the classifications obtained by training on each of the four datasets, represented by one of the color boxes, with all 9 images per sample and classifying on all the other sets of surfaces using the matrix-variate t distribution and a common degrees of freedom parameter, $\nu = 3, 5, 10, 15, 20$, and 30, and prior probability of being a match of 0.5 (for example, training on the first set and testing on sets 2, 3, and 4, and continuing the same process with the other sets as the training set). The output (Table S1) given in terms of the log-odds of being a match—log-odds larger than zero ($p = 0.5$) indicate classification as a match. While initially there are no false positives or false negatives, as the degrees of freedom parameter (DF or ν) increases, there is one false positive, though this probability is very close to 0.5 and all of the true positives have a probability close to

1, which suggests using a classification threshold other than 0.5 would yield perfect classification in this set of data. A different threshold can be chosen by selecting a low probability (such as 10^{-4}) as a probability of a false alarm and using the distribution of log-odds of the true non-matches to fix that threshold conservatively by selecting an upper confidence bound of that quantile[69]. Using the upper 95% confidence bound for the threshold at which the false alarm probability based on the distribution of true negatives is 10^{-4} sets the threshold at a probability of 0.8814 for the most conservative training set at the setting of $\nu = 10$, for example, which still results in perfect classification. Additionally, we may consider the probability in the range of $0.5 > P > 0.88$ to bound the range of inconclusive decision.

3.2.3 Reproducibility of results

In order to determine the reproducibility of results for a given sample, we re-imaged one of the knife samples three times and examined the distributions of the true match image correlations in Figure 9. The different re-imaged sets are labeled "K-1-1", "K-1-2", and "K-1-3". The means of the distributions (indicated by the large shapes) are similar and the covariance matrices, visualized using 99% confidence ellipses, are also similar. Using the two-sample Peacock test, a two-dimensional extension of the Kolmogorov-Smirnov test[70, 71], there is no evidence these distributions differ significantly (H_0: distributions are the same for 1 and 2, $p = 0.21$; H_0 for 1 and 3, $p = 0.32$; H_0 for 2 and 3, $p = 0.25$). We conclude that the imaging and analysis processes are reproducible for the analyzed samples.

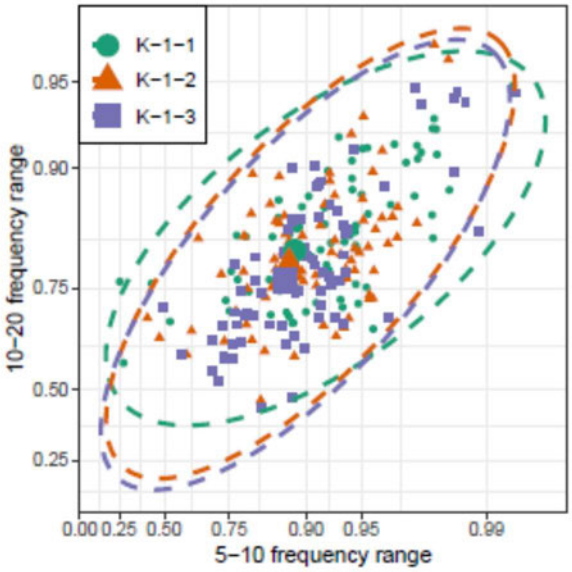

Figure 9: Individual true match correlations for three repetitions of images of the K-1 set of 9 knives and with 9 images per knife. The similarity among the three distributions demonstrates that similar results will be obtained upon re-imaging the same surface, which is important in forensic applications. The large dots indicate the means of the sets, and the ellipses are 99% tolerance ellipses, which provide a representation of the covariance matrices.

3.2.4 Selecting degrees of freedom, DF (ν)

The training sets do not have a sufficient number of observations in both classes to estimate ν in the MxV*t* model. However, the analysis in the previous section indicates it has some influence on the results. We performed a leave-one-out cross validation (LOOCV) procedure to provide guidance about the effects of changing the parameter. For each surface in a training set, a model was trained on the set of observations excluding that surface and tested on the observations using the excluded surface. This was done for $k = 9$ images on training sets S-1 and S-2 and using $k = 5$ images (restricting to the images with only 50% overlap) and $k = 3$ images (restricting to the non-overlapping images) on all four training sets. The procedure was performed only on sets S-1 and S-2 for $k = 9$ because nine surfaces are needed to fit the model and K-1-1 and K-2 have only nine fracture pairs, while S-1 and S-2 have ten fracture pairs. Figure 10 shows the results for $k = 3, 5,$

and 9 respectively. The parameter ν varied from 3 to 30. In all cases, the true matches and true non-matches were perfectly classified using a threshold probability of 0.5 (log-odds of 0). Higher values of ν had more separation between the classes. Using 9 images with 75% overlap had greater separation than 5 images with 50% overlap and greater separation between the identification of true matches. However, given that there is perfect classification in all cases, this finding does not provide much guidance on the selection of ν.

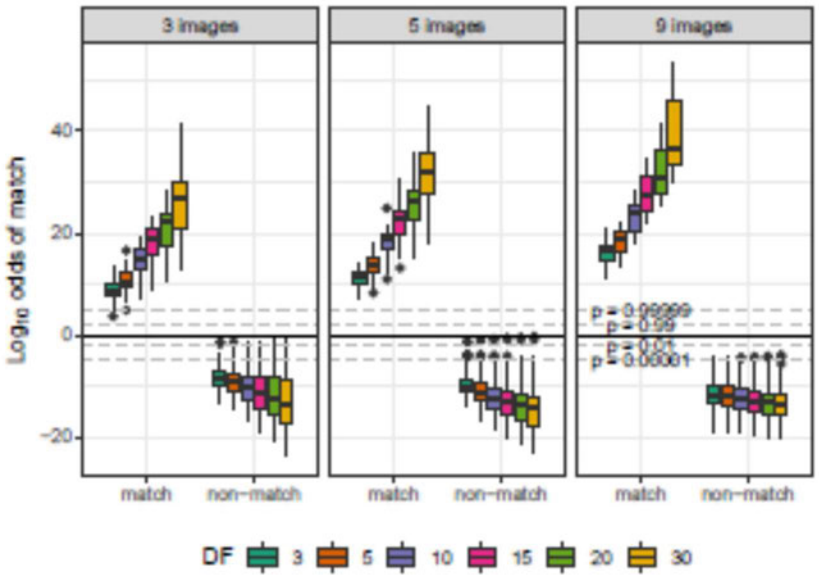

Figure 10: Cross-validation results for models fit using $k = 3$, 5, and 9 images of each surface. The cross-validation was done to provide guidance about the number of images and the choice of DF (ν). There were no false positives or false negatives in this analysis, so it did not provide any conclusive results.

3.2.5 Required number of images for discrimination and model selection

Due to the existence of topographical disturbances in some images (e.g., grains fall out from the fracture surface or substantially large out-of-plane curvature within the range of comparisons), there is not perfect separation between all image pairs for the matches and non-matches. This can be seen in Figure 6 where some image pairs have a correlation coefficient of less than 0.50 for the

two bands of frequency analysis. To mitigate the influence of local topographical disturbances when deciding whether a pair of fragments represent a match or not, multiple observations are needed. To determine how many images are needed to optimize classification performance, we started by training models using all nine images from each base-tip pair in each training set as before. We again used the MxV*t* model with $\nu = 3, 5, 10, 15, 20$, and 30, and then tested them on subsets of consecutive overlapping images of size k, for $k = 2, 3, \ldots 9$ with the model reduced to considering only the selected images and the training set for each model excluded from testing. A summary of the complete results are given in the Supplement Section S.4.

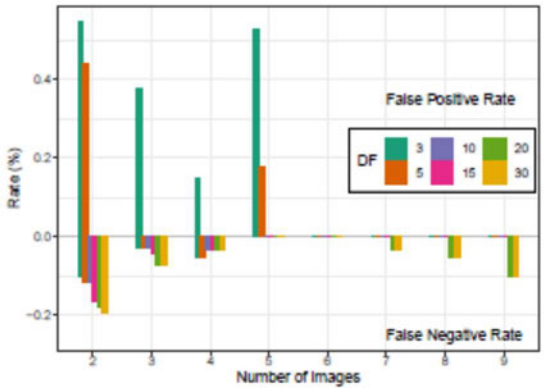

Figure 11: Rates of false positive and false negative classifications (in %) using models trained on the four different sets of surfaces and tested on consecutive subsets of those images for $k = 2, 3, \ldots, 9$.

In Figure 11, models with higher ν have higher false negative rates for all values of k. For values of k over 4, only 20 and 30 DF have false negatives (specifically, they each have one false negative result). Low values of the degrees of freedom parameter have false positives. All of this suggests that choosing a value near $\nu = 10$ and $k \geq 5$ images is sufficient for error-free classification in the sample sets. Figure 12 displays complete results for a model with 10 DF. As k increases, the typical classification results become more separated. However, even with only two images

considered in the test cases for the 10 DF model, the accuracy is very high. The worst case of a false positive is classified with only a probability of 0.8314. The worst case of a false negative is classified with a probability of 0.504. Again, this is the range of match probabilities where an inconclusive match result could be assimilated for $0.5 > P > 0.88$ as noted in Section 3.2.2.

3.2.6 Percentage of overlap between images

Guided by the results of Figure 12, it is apparent that we need at least 5 to 6 images for error-free discrimination in this particular example, and that performance improves with additional images. We reassessed the imaging procedure to gauge the role of the image-overlap ratio. The initial experiment involved imaging surfaces using nine images with 75% overlap between images, which provides three observations for each point on the surface, apart from the edges. However, a similar area can be imaged using 5 images with 50% overlap, which produces two observations of each point on the surface apart from the edges, or using 3 non-overlapping images, which raises the question of whether anything is gained by having an additional third image of the same area and, if so, what level of overlap is best.

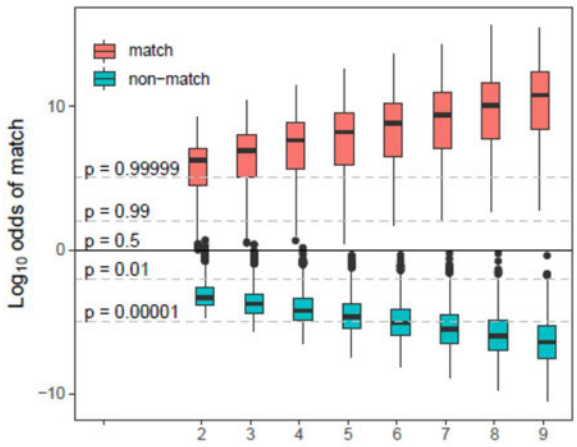

Figure 12: Distributions of the log-odds of a match using models trained on the four different sets of surfaces and tested on subsets of k consecutive images for k = 2, 3, . . . , 9 for a model with 10 degrees of freedom.

We can evaluate this by providing an analysis similar to that done previously: looking at the classification results when restricted to cases with the specified overlap. We train classifiers on the same sets as before, except using 5 images with 50% overlap instead of 9 images with 75% overlap and then test the models on the other sets excluding the set used to train the model by classifying pairs of surfaces using all possible subsets of those images on the surface of sizes 2, 3, 4, and 5. When restricted to the case of 50% overlap, there is only perfect classification when all four or five images are included and the degrees of freedom parameter (ν) is less than 20 (Figure 13). In all cases, there are no false negatives.

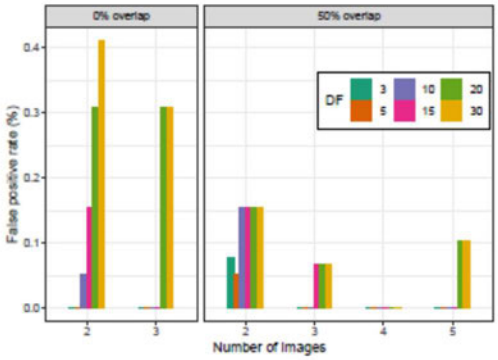

Figure 13: Rates of false positive classifications (in %) using models trained on the four different sets of surfaces using only the images with at most 50% overlap and tested on subsets of k consecutive images for k = 2, 3, 4, 5 and using only the 3 non-overlapping images and tested on subsets of k consecutive images for k = 2, 3. A full summary of the results are in Tables S2 and S3 of the Supplement.

We perform a similar exercise in the case of the non-overlapping images. There are three non-overlapping images per surface which can be used to train the classifiers and the models can then be tested on subsets of those images on each surface of sizes 2 and 3. In the case of non-overlapping images, no model results in perfect classification. The false positives for each model are also shown in Figure 13. There are no false negatives in the classification decisions.

This suggests that, while having more images is generally better, using 5 images with 50% overlap appears to be sufficient if all the images are used. Imaging the entire surface with 50% overlap outperforms imaging the entire surface with 75% overlap in the sense that it works for all of the classes of model. However, if training with 9 images with 75% overlap is possible, testing on new surfaces is feasible with as few as 5 test images with an appropriate choice of the degrees of freedom parameter in the model.

3.2.7 Calibration of output probabilities

The models present the outputs as probabilities; therefore we need to assess how well the probabilities in the models reflect the underlying probabilities in the matching and non-matching populations. Figure 14 displays a calibration plot comparing the output probabilities for all predictions to the empirical proportions in each class with a line drawn by a local regression smoother (LOESS) for each model[72, 73]. These predictions can be compared to the reference line on the plot, $y = x$, to judge the calibration. The true matches correspond with $y = 1$ and the true non-matches with $y = 0$. The vast majority of the model classifications are correct with probabilities of being a match of either < 0.001 for non-matches or > 0.999 for matches. The relative lack of samples in the middle range makes it hard to judge the calibration. The lowest probability of a match among the true matches was 0.3709. Among the various models, the 99th percentile of the predictions for non-matches was, in the worst case, 0.1437. Only outliers overlapped in middle range. We note that our evaluation of the calibration is limited by the sample size in the experiment—with more samples and more observations with match probabilities between 0.1 and 0.9, a better evaluation of the calibration could be made.

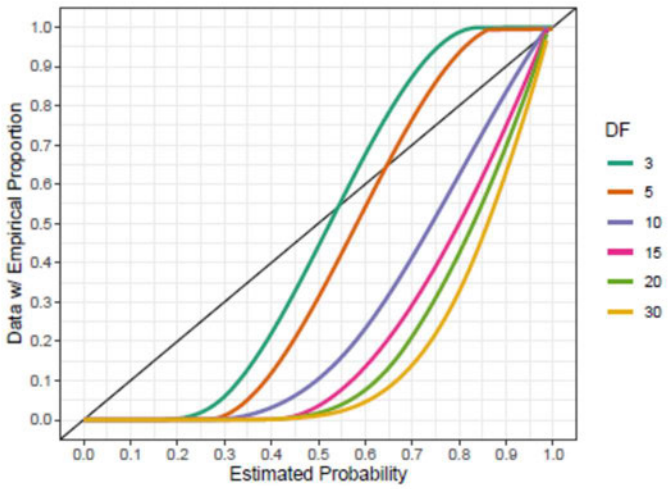

Figure 14: A calibration plot of all the predictions from the various models and the empirical proportions with a LOESS smoother for each model. The lines should be compared to the diagonal reference line. Matches are indicated at $y = 1$ and non-matches are indicated at $y = 0$.

3.2.8 Examining the framework capabilities on a twisted-fracture knife set

All examined sets of fractured articles were tested in tension or bending. This is mode-I cleavage fracture where the crack propagation direction is normal to the loading axis. The fracture surface showed topographical features normal to the fracture surface, similar to those shown in the SEM image of Figure 3(a). However for a general forensic article such as a knife or a pry tool, an edge could be broken due to bending and twisting of the article. This would impose a mixed mode of loading including mode-I opening and mode-III twisting of a crack. To understand the effect of external loading mode on the generality of the proposed analysis framework, a set of nine knives from the same manufacturer similar to the previously used sets were fractured at random using the same fixture shown in Figure S1(b) and forming set of twisted knives shown in Figure S1(e). A typical twisted knife fracture topology is very different at both the macro and micro scales. At the macro scale, the crack trajectory is no longer planer with curvilinear or twisted trajectory as those shown in Figure S1(e). At the micro scale, the SEM image of Figure 3(b) shows twisted topology

in the plane of the crack that is very different from those of mode-I loading. This unique texture would probably further enhance the individuality of the fracture surface. We will attempt to examine the validity of the analysis protocol on such general case of fractured articles. The twisted knife set was imaged using the same procedure discussed in Section 1.3.4 and the same magnification of 20X. However, due to the excessive tortuosity of the crack path, five images (k = 5) with 75% overlap between adjacent images were employed. Using the models previously trained on the four training sets loaded in tension or bending, and restricted to 5 images and setting the degrees of freedom $\nu = 10$.

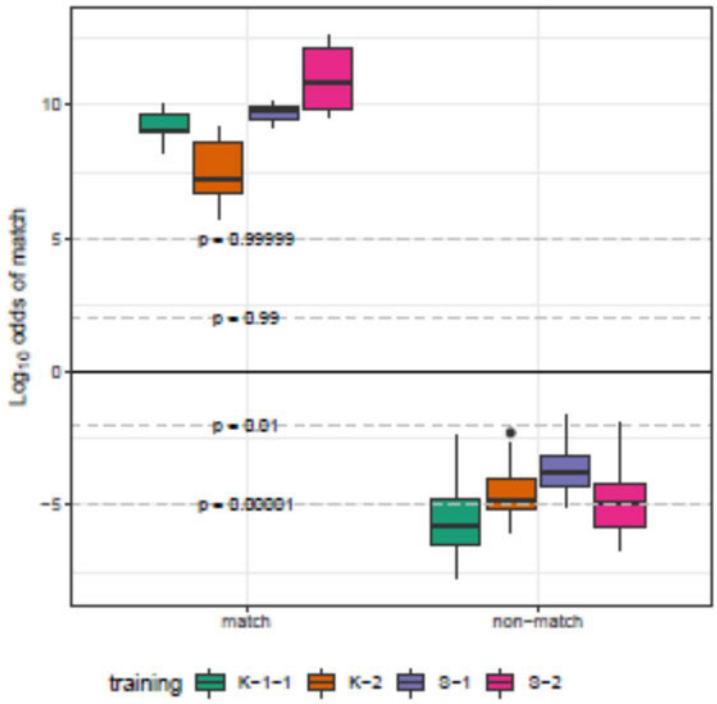

Figure 15: A plot of the classification results for a generality verification data set of nine knives broken by twisting with five images per surface. The models were trained with on different training bending and tensile fracture sets using five images with the degrees of freedom setting $\nu = 10$.

The results for this set, shown in Figure 15, are similar to those obtained in Figure 8 despite the use of a different external loading of mode-I tensile cleavage fracture. The true match cases where

identified with probability exceeding 99.999% and the true non-match were identified with probability not exceeding 0.05% for all the different training sets. This suggests the scale of comparison, derived from the self-affine saturation scale of the fracture surface topology is more general and tied to the microstructure scale (grain size) for hardened tool materials failing by cleavage fracture. This result is far more reaching.

3.2.9 Assessment of cast replica effectiveness in topological mapping

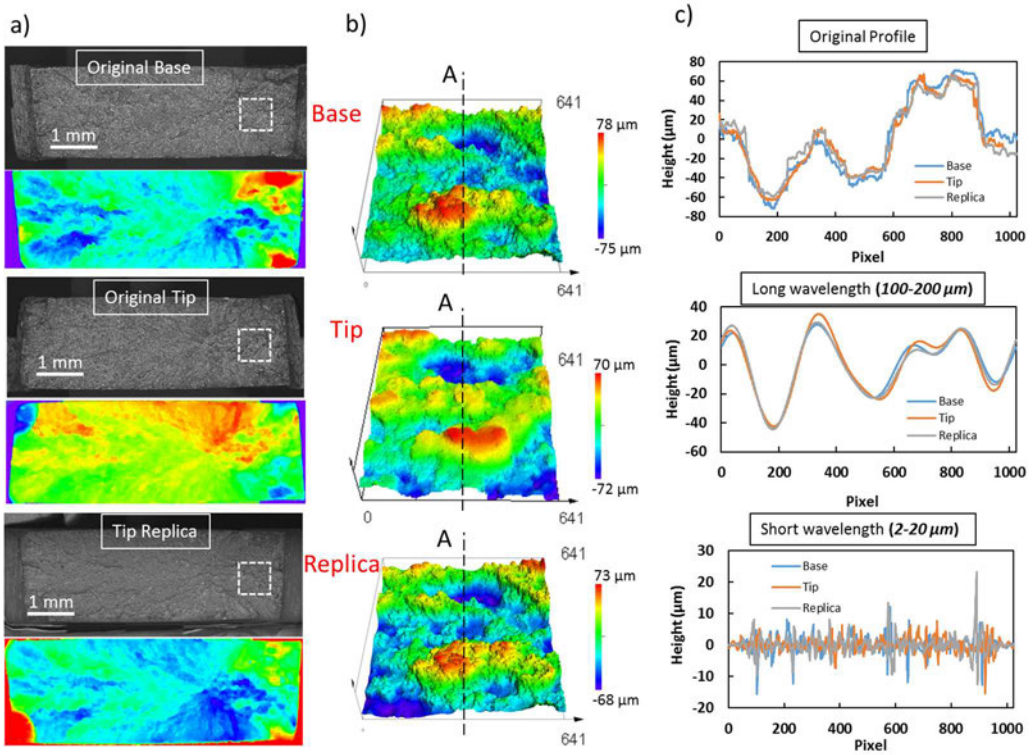

Figure 16: 3D topological analysis of pair of fractured surfaces, Base-Tip pair, and a Tip-replica. (a) Optical micrograph and color rendering of the topological fracture surface. Base and Tip show mirror images with each other, while the Tip and its replica show mirror images with each other (b) 3D topological representation of the fracture surface, utilizing a 640 µm Field of View for the square inset area on the optical image. (c) 2D representation of the height comparison along line A-A showing the original measured height (top) and the corresponding long-wavelength (middle) and short-wavelength (bottom), utilizing spectral decomposition by Fourier Transform Analysis. The low-frequency topological details are the same, while the high-frequency ones are quite different.

Our focus also is on examining the potential of cast replicas, shown in Figure 16, to transfer all the topological details required for the analysis at the proper magnifications. The classification model was applied to three sets of pairs of surfaces: the original Bases with Tips, the Bases with the Tip replicas, and the Tips with their replicas. Figure 17 shows the posterior probabilities of match obtained on the three sets of surface pairs in the log-odds scale. Larger posterior log-odds indicate more evidence that a surface pair is a match, whereas lower log-odds indicate more evidence that a surface pair is a non-match. Utilizing the t-distribution with 5 degrees of freedom provides great confidence in the discrimination power of the proposed comparison and statistical analysis frameworks. The model classified the three cases with ten pairs of true matches and the 90 pairs of true non-matches with zero false negatives and zero false positives. That is, there are a total of 30 pairs of true matches and 270 pairs of true non-matches. The 90 replicas show a high probability of match when compared to the original fractured surfaces. This high accuracy exists for both original Base-Tips, replicas-Bases, replicas-Tips, although the replicas were cast on the surfaces of the tips only. This demonstrates the replicas' ability to capture the relevant features that are important for discrimination. Furthermore, for the true match group, the lowest posterior probability was higher than 0.9996, while the highest posterior probability for the true non-match was less than 0.005. The stark difference between the match and non-match probabilities highlights the strength of using the physical basis of fracture mechanics to guide the imaging procedure and construct the statistical discrimination framework.

Figure 18 summarizes the results for the replication capacity of the silicone replica technique for all the frequency bands in the range of 3-200 mm^{-1} or the corresponding wavelength of 333-5 μm, respectively. The correlation means are shown for each of the comparison bands along with 95% bootstrap confidence intervals for both the matches and non-matches for the

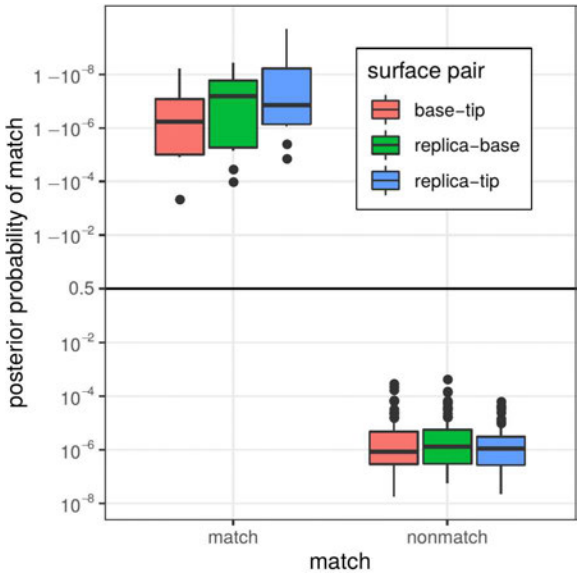

Figure 17: Posterior log-odds of being a match using a model trained on a separate set of images from the same surfaces and tested on the pair of surfaces of base-tip, replica-base, and replica-tip. Higher values indicate stronger evidence of match.

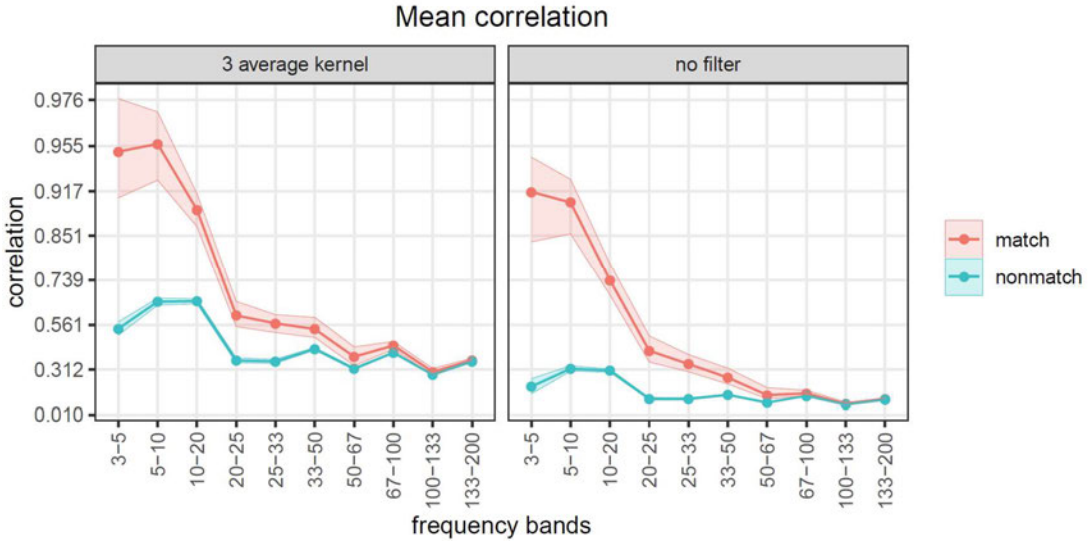

Figure 18: Mean correlations obtained from the comparison between the Tip and its replica along 10 frequency bands and to both matches and non-matches, and for both the filtered and non-filtered FFTs. The matching correlations decrease with increasing frequency bands while still being distinctively different from the non-matches until it reaches frequency bands higher than 100 mm^{-1}.

original FFTs along with the filtered FFTs. Figure 18 shows that with increasing the frequency bands (i.e., reduction of the wavelength in real space), the matching correlations between the Tip and its replica decay in magnitude and spread. Furthermore, the correlations obtained from the filtered FFTs show increased correlations in both the matches and the non-matches. While we are asking too much from the discrimination framework, however Fig. 18(a) shows that the match correlation remains around 0.4-0.5 for a high frequency of 100mm^{-1} or a short wavelength of 10μm. This result gives confidence that replicas can still reliably reproduce wavelengths down to the micron ranges. However, such a range needs additional investigation with a smaller field of view (FOV) and a master surface with well-defined micron range features. One more issue that should be noted at large wavelength, greater than about a fifth of the imaging window size. Figure 18(a) shows a slight drop in correlation at the 3-5mm^{-1} frequency band when compared to the 5-10mm^{-1} frequency band. This is a limitation of the discretization process wherein the resolution per frequency line is 1.56 mm^{-1}, which will provide a very limited number of data points (about 7) in the 3-5mm^{-1} band. A larger FOV at the same magnification would be required to refine the frequency band lines and resolve these long-wavelength limitations.

This analysis provides two significant results. *First*, the analysis supports the developed classification procedure. It shows how replicas are effectively capturing the surface fractures along wavelength topological details in the range of the two frequency-band analyses of 5-10 and 10-20mm$^{-1,}$ which corresponds to 200-100 and 100-50μm wavelengths. The topological features at these length scales are unique and helpful for distinguishing between matches and non-matches. *Second*, the analysis shows the ability of the replica to faithfully replicate fracture features with wavelengths all the way to the 25μm range. For forensic comparison, the replicas are well suited for mapping features of 20 μm and larger. It can be assertively stated that the replicas effectively

distinguish between matches and non-matches in low frequency ranges, and they stop being distinctively different for frequency above 100 mm^{-1}, where the micro-features of the local fracture processes that are common to both the matches and non-matches are compared.

3.2.10 Examining Proficiency Sample Set (Claytor, 2010) [12]

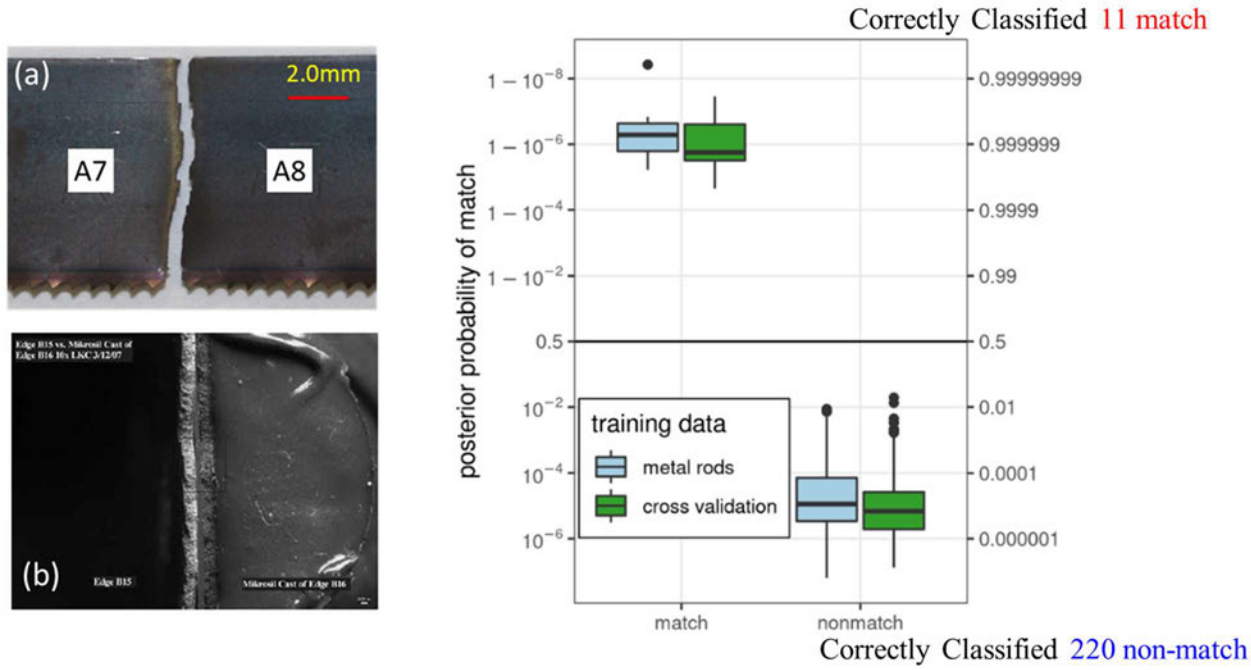

Figure 19: Proficiency sample set, provide by our forensic collaborator[12]. (a) Physical fit of two edges (b) Direct comparison of fractured edge and its replica. (c) Discrimination analysis framework results for posterior probability for 11 pairs of Match and 220 cases of non-match cases.

The fractured surfaces of Figure 19(a) correspond to a validation study performed by Claytor and Davis, 2010,[12] where the proficiency of forensic examiners to correctly associate matching and non-matching fracture pairs was studied. In their study, consecutively manufactured hacksaw blades were fractured in 11 locations each, resulting in 22 fracture surfaces and 231 pairs of fracture surfaces per blade. These fractured hacksaw blades along with laboratory-generated silicon replicas were analyzed by forensic analysists across multiple forensic science centers. Each

test was validated by firearm and toolmark trainees who had completed seven to eight months of toolmark training, including a fracture match proficiency test at the Virginia Department of forensic and who had an average of 12.4 years of qualified experience. Out of the 173 responses received, 157 were correctly classified, indicating a 91% accuracy (true classification rate).

In our study we examined the proficiency of our developed quantitative matching framework by applying the classification method to one of the fractured hacksaw blades used in Claytor and Davis, 2010. [12] This classification algorithm is trained based on cross-validation (such that the material composition is the same for both testing and training datasets), and we will separately train the classification model on a data set of fractured metal rods that are of a different material but with similar fracture topology. In both cases, the quantitative classification frameworks results in 100% accuracy.

In Figure 19(b) we observe a boxplot of the posterior probabilities (in the log-odds scale) that corresponds to the 231 pair surface pairs that result from the 22 fracture surfaces (11 matching pairs and 220 non-matching pairs). Furthermore, in Figure 3 we have boxplots for different training techniques: cross validation training (such that the training and testing datasets correspond to the same hacksaw blade material) and metal rod training data (such that the training and testing datasets correspond to dissimilar materials with similar fracture-surface topology). In all cases the matching and non-matching groups are largely separated, indicating perfect classification. The minimum matching probabilities are 0.99998 for cross-validation training, and 0.999994 for the metal rod training, whereas the maximum non-matching probabilities are 0.019 for cross-validation training and 0.0087 for metal rod training. This implies that any threshold between 0.019 and 0.99998 would result in perfect classification in all cases. Moreover, from Figure 19 we observe that both training methods result in boxplots with exceedingly high overlap, indicating

that both training methods result in remarkably similar classifications. This is a strong result, as it shows that training and testing do not need to occur on the same material but can be used between dissimilar materials if they have similar fracture-surface topologies. This is relevant in scenarios where it is hard or expensive to obtain training data for a specific type of fracture material. In these cases, one can obtain training data from a different material if their fracture-surface topologies contain relevant individualizing features at similar length-scales.

3.3 Limitations

Our results suggest that for the class of materials that undergoes cleavage fracture, a single robust training data set under simplified loading conditions would be sufficient to help in discriminating; (i) articles that were exposed to complex external loading (i.e. mixed mode of fracture). (ii) Articles from different classes of materials, but share the same grain size distributions, and (iii) articles with different grain sizes would only require changes of the FOV to include 20-grain diameter and comparison frequency bands corresponding to the 2−4 and 4−8 grain size ranges.

Challenges to this technique arise from high topographical details with a large aspect ratio that might shadow the surrounding details, and might disturb one of the frequency bands.

It is also conceivable to extend these results to glassy metals, polymers and ceramics, that undergoes cleavage and/or brittle or semi-brittle fracture. In such cases, the limits of the fractal scale should be examined and compared to the critical microstructure scale of the fracture surface topology, such as river and herringbone patterns. Though, additional experimental verification are needed for these classes of non-crystalline materials.

4. ARTIFACTS

4.1 Products

1. Thompson, G. Z., Maitra, R., Meeker, W. Q. & Bastawros, A., 2020. Classification with the matrix-variate-t distribution. Journal of Computational and Graphical Statistics, 29:3:668–674. http://doi.org/10.1080/10618600.2019.1696208
2. Bastawros, A., 2021. "*Physics and Statistical Models for Physical Match Analysis Utilizing 3D Microscopy of Fracture Surfaces,*" Invited Talk at the 2021 NIJ R&D Symposium.
3. Bastawros, A., 2021. "*Physical Match Analysis Utilizing 3D Microscopy of Fracture Surfaces,*" Annual Training Meeting, Miami FL, August 22-27, 2021.
4. Thompson G., Dawood B., Yu T., Lograsso B., Vanderkolk J., Meeker, W., Maitra, R., and Bastawros, A.F. 2021, Fracture Mechanics-Based Quantitative Matching of Forensic Evidence Fragments. Under review, Nature Communications; arXiv: https://arxiv.org/abs/2106.04809
5. Bishoy Dawood, Carlos Llosa-Vite, Geoffrey Z. Thompson, Barbara K. Lograsso, Lauren K. Claytor, John Vanderkolk, William Meeker, Ranjan Maitra, and Ashraf Bastawros, 2022. *Quantitative Matching of Forensic Evidence Fragments Utilizing 3D Microscopy Analysis of Fracture Surface Replicas*, Journal of Forensic Sciences 67(3): 899-910, 2022. https://doi.org/10.1111/1556-4029.15012
6. Bastawros, A., 2021. "*Physical Match Analysis Utilizing 3D Microscopy of Fractured Surface,*" Poster session, 2021 NIJ R&D Symposium.

4.2 Data Sets

1. An R[43]-software package to perform the model fitting and analysis MixMatrix, is available[44]. A GitHub repository containing the data and the code to reproduce the figures and analysis for the submitted paper to Nature Communications, at https://github.com/gzt/fracturematching.

2. TopologyMatcheR; an R-package software and data set archival at GitHub; https://github.com/carlos-llosa/TopologyMatcheR for Dawood, JOFs 2022 paper for identifications of 10-pairs of fractured articles and their casted replicas. Five sets of text-files ".CSV-formatted" are added as reduced input data sets for the correlation coefficients for the range of examined frequencies of 5-200mm^{-1} for each of the six image-pairs of each fracture surface pairs. These are the input data, which can be used with the program. These file data files include:

   (i) A3-corrs.csv Correlation coefficients for a set of 10 fractured pairs of metallic articles of similar grain size range. This is a master calibration set, acquired by a well-trained operator and used for training the comparison algorithm. It has 600 entry; 60 matches (10-pairs x 6-images per pair) and 540 unmatched entry (for every matched pair of surfaces, there are 9-other possibilities of unmatched combination, resulting in 10 pairs x 9 combinations x 6images per pair).

   (ii) base-tip-corrs.csv Correlation coefficients for pair of fractured base-tip surfaces.

   (iii) replica-base-corrs.csv Correlation coefficients for pair of fractured base and replica of the tip surfaces.

   (iv) replica-tip-corrs.csv Correlation coefficients for pair of fractured tip and its replica surfaces.

   (v) 3crossavg-replica-tip-corrs.csv Correlation coefficients for fractured tip and its replica surfaces after applying the average kernel on the original FFT spectra of each surface.

3. Berlinski, J. et al., 2023. "*Matching pairs of bending fractured samples*," an archive of the entire data set for 10 pairs of bent fracture samples/ Each fracture pairs has six pairs of overlapping images. doi:10.25380/iastate.23947302.v1

https://iastate.figshare.com/articles/dataset/Matching_pairs_of_bending_fractured_samples/23947302

4. Berlinski, J. et al., 2023. "*Matching pairs of tensile fractured samples*," an archive of the entire data set for 10 pairs of tensile fracture samples/ Each fracture pairs has six pairs of overlapping images. doi:10.25380/iastate.23929215.v1

https://iastate.figshare.com/articles/dataset/Matching_pairs_of_tensile_fractured_samples/23929215

4.3 Dissemination Activities

1. Bastawros, A., 2023. "Physical Match Analysis Utilizing 3D Microscopy of Fractured Surface," Seminar and discussion, Virginia Department of Forensic Science, March 21, 2023. Dr. Bastawros spent over two hours discussing his technique, development framework and findings. Dr. Bastawros fielded questions from 10 forensic scientists from across the State of Virginia.

2. Bastawros, A. and Maitra, R., 2022. Developed a full session on "Collaborative case study: Engineering and statistical models for error quantifications in comparative 3D microscopy for physical match analysis," during the Conference on "Statistical Methods in Imaging 2022," VANDERBILT UNIVERSITY MEDICAL CENTER, May 25-27, 2022 https://www.vumc.org/biostatistics/SMI/Maitra

   Three Consecutive talks were given by the PIs on:

   (a) Fracture matching: A historical perspective and a path forward in forensic science, Lauren K. Claytor, Virginia Department of Forensic Science.

   (b) Fracture matching: Role of fracture mechanics on setting the comparison scales, Ashraf F. Bastawros, Iowa State University.

   (c) Quantitative matching of forensic evidence fragments using fracture mechanics and statistical learning, Geoffrey Thompson, Indiana University

5. APPENDICES

S.1 Details on Sample Generation and Imaging

Two main material classes are considered: two sets of nine single serrated edged knives from the same manufacturer (Chicago Cutlery), and two sets of ten rectangular (0.25" wide, and 1/16" thick) rods of a common tool steel material (SS-440C) cut from the same metal sheet to minimize any variability from the manufacturer. The knives were fractured at random using a controlled bend fixture shown in Figure S1(a). A set of fractured pairs of knives is shown in Figure S1(b). The two sets of the tool steel rods were loaded under either controlled tensile loading at 1*mm/min* displacement rate (Figure S2(a)) or controlled bending loading at 1.5 *mm/min* displacement rate (Figure S2(c)) until fracture. The pairs of tool steel samples fractured by tension and bending are shown in Figure S2(b, d), respectively. The average grain size for both groups was approximately $dg = 25–35\mu m$. For clarity, we refer to the surface attached to the knife handle as the base and the surface from the top portion of the knife as the tip. The same terminology was applied to the tool steel samples as well.

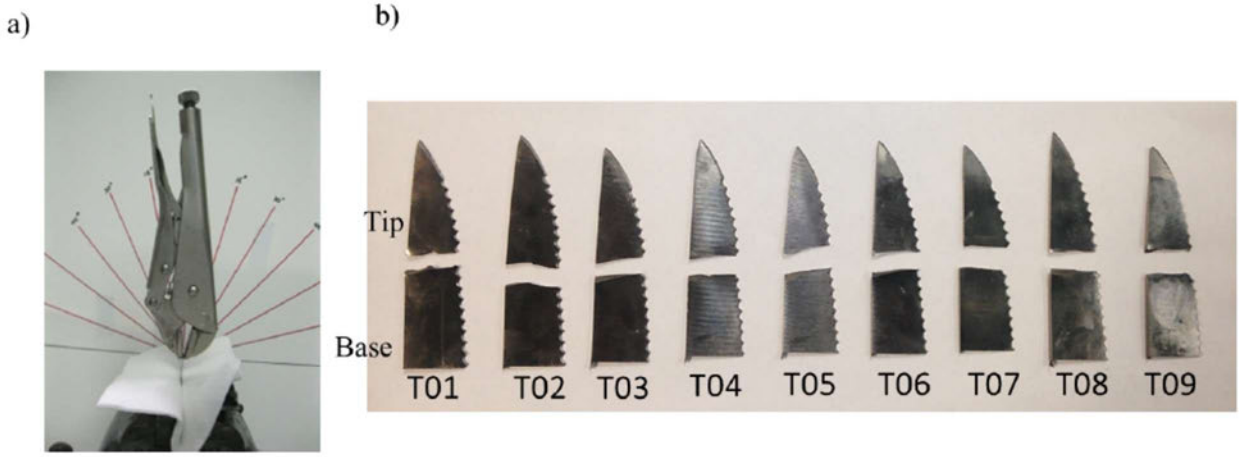

Figure S1. Knife-breaking protocol. (a) Loading fixture. (b) Pairs of knives from the same manufacturer fractured by bending.

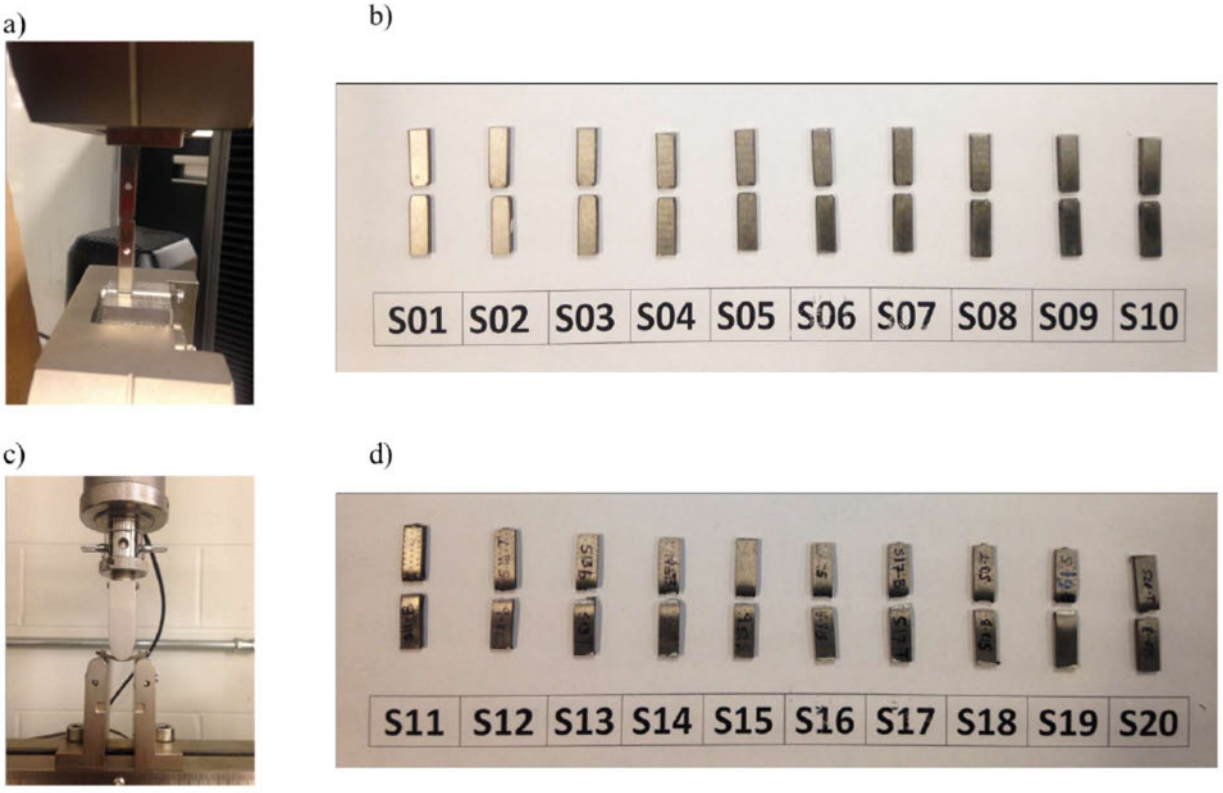

Figure S2. Sample generation protocol. (a) sample fracture under a controlled tensile loading, (b) pairs of steel samples fractured by tensile loading, (c) sample fracture under three point bending loading, (d) pairs of steel samples fractured under bending.

S.2 Imaging, Processing and Alignments

The microscopic features of pairs of fracture surfaces were aligned and analyzed by a standard non-contact 3D optical interferometer (Zygo-NewView 6300), which provides a height resolution of 20 *nm*. At the utilized magnification of 20X, defined from the scale of the transition from the affine deformation (Figure 2(b)), a spatial inter-point resolution of $0.55\mu m$ was achieved to record topological surface heights similar to Figure 2(a). A series of $k-$ overlapping surface height 3D topographic maps were acquired for each pair of fracture fragments as shown in Figure 3. After acquiring each topological image, standard digital image processing were applied through a 3×3

or 5 × 5 averaging window kernel to (a) fill missing data, and (b) remove digital noise that is above two standard deviation from the mean of the window. A mathematical Hann filter with 10% edge smoothing ratio was applied to the each topological image to provide a periodic boundary for the image edges. The FFT operator was applied to each image to generate the corresponding frequency-space representation and the image spectral contents. Correlations of the selected frequency bands were carried out for the model development and decision-making corresponding to those shown in Figure 4.

Alignments of pairs of images for the comparative analysis is a critical issue for the generality of the technique, especially when there is a lack of clear reference to start the 3D topological imaging from. The case of circular object like a bullet casing would require additional angular alignment of the images, either physically or computationally before taking the FFT spectra for the comparative analysis. For the rectangular objects or knives with clear edges studied in this work, the specimen edges naturally rendered a reference coordinate. We utilized the sample edge as our reference and aligned the edges with the imaging field of view of the microscope. Then we moved a predetermined distance in the planner $x-$ and $y-$coordinates to start the imaging sequence. In general, one may identify three types of miss-registration that can greatly affect the correlation estimations between a pair of images. Some of the alignment issues that affect the utilization of virtual comparison microscopy are:

1. Planner misregistration in the $x-$ and $y-$coordinates. The planar alignment is critical when comparing the images in their spatial (real) space, wherein an auto correlation must be utilized to adjust the planar miss-registration. However, the implemented procedure in this work to utilize the spectral (frequency) space is tolerant to planar miss-registration of up to 20% of the

field of view. Actually, we have examined different planar miss-registration up to 400% of the FOV, and we found that the correlation will degrade by about 10% for up to 50% shift for the FOV.

2. Angular misregistration within the plane observation (fracture plane). This is more critical when considering a series of overlapping $k-$ images, wherein the last image-pair might have much less of an overlap due to angular separation. For our work, this was not an issue as we were studying fragments with good references for alignments. Beyond the presented work in this manuscript, we are implementing an aliment technique using the frequency spectra on angular trajectories within the two comparison bands to correct for angular misregistration that would maximize the correlation coefficients for an image pair. For a set of $k-$ image pairs, the miss-registration angle is optimized to maximize the correlation coefficient over the entire $k-$ image pairs on a pair of fracture surfaces. This process is conducted for pairs of fragments, regardless of their classification as true match or true non-match.
3. Fracture plane tilt: This is the case with a very tortuous fracture surface, especially for semi-brittle or ductile articles. In such case of tortuous path, the topological details of the surface will have high aspect ratio that might eclipse/shadow other features. Furthermore, if large plasticity exists on the surface and the surface topology changes within several hundred microns with non-planer mean, the 3D topological imaging process would be much harder and additional mathematical treatment would be required, similar to comparison of cylindrical surface (e.g. cartridge cases).

6. REFERENCES

1. Fradella, H. O. & Fogary, A. L. The impact of Daubert on forensic science. Pepperdine Law Review 31, 323 (2003−2004).

2. National Academy of Sciences (NAS). Strengthening Forensic Science in the United States: A Path Forward (The National Academies Press, Washington, DC, 2009). URL https://www.nap.edu/catalog/12589/strengthening-forensic-science-in-the-united-states-a-path-forward.

3. Biasotti, A. A. A statistical study of the individual characteristics of fired bullets. Journal of Forensic Sciences 4, 34 (1959).

4. Uchiyama, T. The probability of corresponding striae in toolmarks. AFTE Journal 24, 273-290 (1992).

5. Miller, J. & McLean, M. Criteria for identification of toolmarks. AFTE Journal 30, 15−61 (1998).

6. Almirall, J., Arkes, H., Lentini, J., Mowrer, F. & Pawliszyn, J. Forensic science assessments: a quality and gap analysis−fire examination (American Association for the Advancement of Science, Washington, DC, 2017).

7. Thompson, W., Black, J., Jain, A. & Kadane, J. Forensic science assessments: a quality and gap analysis−latent fingerprint examination (American Association for the Advancement of Science, Washington, DC, 2017).

8. Vanderkolk, J. Forensic Comparative Science: Qualitative Quantitative Source Determination of Unique Impressions, Images, and Objects (Academic Press, Cambridge, MA, 2009). URL https://www.xarg.org/ref/a/B074CCCQJC/.

9. Van Dijk, T. & Sheldon, P. The Practice of Crime Scene Investigation (International Forensic Science and Investigation Book 10) (CRC Press, Boca Raton, Florida, 2004). URL https: //www.xarg.org/ref/a/B00UVAK1NA/.

10. Katterwe, H. W. Fracture matching and repetitive experiments: a contribution of validation. AFTE JOURNAL 37, 229 (2005).

11. Miller, J. & Kong, H. Metal fractures: Matching and non-matching patterns. AFTE Journal 38, 133−165 (2006).

12. Claytor, L. K. & Davis, A. L. A validation of fracture matching through the microscopic examination of the fractured surfaces of hacksaw blades. AFTE JOURNAL 42, 323 (2010).

13. Mandelbrot, B. B., Passoja, D. E. & Paullay, A. J. Fractal character of fracture surfaces of metals. Nature 308, 721−722 (1984). URL https://doi.org/10.1038/308721a0.

14. Van Dijk, T. & Sheldon, P. Physical comparative evidence. In The Practice Of Crime Scene Investigation, 393−418 (CRC Press, 2004).

15. Klein, A., Nedivi, L. & Silverwater, H. Physical match of fragmented bullets. Journal of Forensic Sciences 45, 722−727 (2000).

16. Walsh, K., Gummer, T. & Buckleton, J. Matching vehicle parts back to the vehicle. AFTE Journal 26, 287−289 (1994).

17. Matricardi, V. R., Clarke, M. S. & DeRonja, F. S. The comparison of broken surfaces: A scanning electron microscopic study. Journal of Forensic Sciences 20, 507−523 (1975).

18. McKinstry, E. A. Fracture match − a case study. AFTE Journal 30, 343−344 (1998).

19. Verbeke, D. J. An indirect identification. AFTE Journal 7, 18−19 (1975).

20. Townshend, D. Identification of fracture marks. AFTE Journal 8, 74−75 (1976).

21. Dillon, D. J. Comparisons of extrusion striae to individualize evidence. AFTE Journal 8, 69−70 (1976).

22. Karim, G. A pattern-fit identification of severed exhaust tailpipe sections in a homicide case. AFTE Journal 36, 65−66 (2004).

23. Smith, E. D. Bullet and fragment identified through impression mark. AFTE Journal 36, 243 (2004).

24. Katterwe, H., Goebel, R. & Gross, K. D. The comparison scanning electron microscope within the field of forensic science. AFTE Journal 15, 141−146 (1983).

25. Goebel, R., Gross, K. D., Katterwe, H. & Kammrath, W. The comparison scanning electron microscope: First experiments in forensic application. AFTE Journal 15, 47−55 (1983).

26. Moran, B. Physical match/toolmark identification involving rubber shoe sole fragments. AFTE Journal 16, 126−128 (1984).

27. Rawls, D. A rare identification of glass. AFTE Journal 20, 154−156 (1988).

28. Hathaway, R. A. Physical wood match of a broken pool cue stick. AFTE Journal 26, 185−186 (1994).

29. Zheng, X. et al. Applications of surface metrology in firearm identification. Surface Topography: Metrology and Properties 2, 014012 (2014). URL https://doi.org/10.1088%2F2051-672x%2F2%2F1%2F014012.

30. Petraco, N. D. K. et al. Addressing the National Academy of Sciences′ challenge: a method for statistical pattern comparison of striated tool marks. Journal of Forensic Sciences 57, 900−911 (2012). URL https://doi.org/10.1111/j.1556-4029.2012.02115.x.

31. Katterwe, H., Goebel, R. & Grooss, K. The comparison scanning electron microscope withinthe field of forensic science. Scanning Electron Microscopy 1982, 499−504 (1982).

32. Anderson, T. L. Fracture Mechanics: Fundamentals and Applications (Academic Press, 2017).

33. Underwood, E. & Banerj, K. Fractals in fractography. Material Sciences and Engineering 80, 1−14 (1986).

34. Dauskardt, R., Haubensak, F. & Ritchie, R. On the interpretation of the fractal character of fracture surfaces. Acta Metall. Mater. 38, 143−159 (1990).

35. Cherepanov, G. P., Balankin, A. S. & Ivanova, V. S. Fractal fracture mechanics—a review. Engineering Fracture Mechanics 51, 997−1033 (1995). URL https://www.sciencedirect.com/science/article/pii/001379449400323A.

36. Bouchaud, E. Scaling properties of cracks. J. Phys. Condensed Matter 9, 4319−4344 (1997).

37. Charkaluk, E., Bigerelle, M. & Iost, A. Fractals and fracture. Engineering Fracture Mechanics 61, 119−139 (1998).

38. Ponson, L., Bonamy, D. & Bouchaud, E. Two-dimensional scaling properties of experimental fracture surfaces. Physical Review Letters 96, 035506−1−4 (2006).

39. Srivastava, A. et al. Effect of inclusion density on ductile fracture toughness and roughness. Journal of the Mechanics and Physics of Solids 63, 62−79 (2014).

40. Yavas, D. & Bastawros, A. F. Correlating interfacial fracture toughness to surface roughness in polymer-based interfaces. Journal of Materials Research 36, 2779—-2791 (2021).

41. Bonamy, D., Ponson, L., Prades, S., Bouchaud, E. & Guillot, C. Scaling exponents for fracture surfaces in homogeneous glass and glassy ceramics. Physical Review Letters 97, 135504 (2006).

42. Morel, S., Bonamy, D., Ponson, L. & Bouchaud, E. Transient damage spreading and anomalous scaling in mortar crack surfaces. Physical Review E 78, 016112 (2008).

43. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2018). URL https://www.R-project.org/.

44. Thompson, G. Z. MixMatrix: Classification with Matrix Variate Normal and t Distributions (2020). Http://github.com/gzt/MixMatrix/, https://gzt.github.io/MixMatrix/.

45. Ritchie, R., Knott, J. & Rice, J. On the relationship between critical tensile stress and fracture toughness in mild steel. Journal of the Mechanics and Physics of Solids 21, 395-410 (1973).

46. Lin, T., Evans, A. & Ritchie, R. Statistical model of brittle fracture by transgranular cleavage. Journal of the Mechanics and Physics of Solids 34, 477-496 (1986).

47. Armstrong, T. & Warner, L. Low-temperature transition of normalized carbon-manganese steels. In Symposium on Impact Testing. ASTM International (1956). URL https://api.semanticscholar.org/CorpusID:137632596.

48. Curry, D. & Knott, J. Effects of microstructure on cleavage fracture stress in steel. Metal Science 12, 511-514 (1978).

49. Aitken, C. G. & Taroni, F. Statistics and the Evaluation of Evidence for Forensic Scientists

(John Wiley & Sons, Ltd, 2004). URL https://doi.org/10.1002/0470011238.

50. Meester, R. Why the effect of prior odds should accompany the likelihood ratio when reporting DNA evidence. Law, Probability and Risk 3, 51–62 (2004). URL https://doi.org/10.1093/lpr/3.1.51.

51. de Keijser, J. & Elffers, H. Understanding of forensic expert reports by judges, defense lawyersand forensic professionals. Psychology, Crime & Law 18, 191–207 (2012). URL https://doi.org/10.1080/10683161003736744.

52. Martire, K., Kemp, R., Sayle, M. & Newell, B. On the interpretation of likelihood ratios in forensic science evidence: Presentation formats and the weak evidence effect. Forensic Science International 240, 61–68 (2014). URL https://doi.org/10.1016/j.forsciint.2014.04.005.

53. Zadora, G., Martyna, A., Ramos, D. & Aitken, C. Likelihood Ratio Models for Classification Problems (John Wiley & Sons Ltd, 2013). URL https://doi.org/10.1002/9781118763155.

54. Taroni, F., Biedermann, A., Bozza, S., Garbolino, P. & Aitken, C. Bayesian Networks for Probabilistic Inference and Decision Analysis in Forensic Science (John Wiley & Sons, Ltd, 2014). URL https://doi.org/10.1002/9781118914762.

55. Dawood, B. et al. Quantitative matching of forensic evidence fragments utilizing 3d microscopy analysis of fracture surface replicas. Journal of Forensic Sciences 67, 899–910 (2022).

56. Champod, C., Lennard, C., Margot, P. & Stoilovic, M. Fingerprints and Other Ridge Skin Impressions, Second Edition, chap. 2.7 (CRC Press, 2016). URL https://doi.org/10.1201/b20423.

57. Song, J. Proposed "NIST ballistics identification system (NBIS)" based on 3d topography measurements on correlation cells. AFTE Journal 45, 184–193 (2013).

58. Chen, Z., Song, J., Chu, W., Tong, M. & Zhao, X. A normalized congruent matching area method for the correlation of breech face impression images. Journal of Research of the National Institute of Standards and Technology 123 (2018). URL https://doi.org/10.6028/jres.123.015.

59. Kobayashi, T. & Shockey, D. A. Fracture surface topography analysis (FRASTA)-development, accomplishments, and future applications. Engineering Fracture Mechanics 77, 2370–2384 (2010). URL https://doi.org/10.1016/j.engfracmech.2010.05.016.

60. Jacobs, T. D. B., Junge, T. & Pastewka, L. Quantitative characterization of surface topography using spectral analysis. Surface Topography: Metrology and Properties 5, 013001(2017). URL https://iopscience.iop.org/article/10.1088/2051-672X/aa51f8.

61. Fisher, R. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika 10, 507–521 (1915).

62. Gupta, A. & Nagar, D. Matrix Variate Distributions, vol. 104 (CRC Press, 2018).

63. Iranmanesh, A., Arashi, M. & Tabatabaey, S. On conditional applications of matrix variate normal distribution. Iranian Journal of Mathematical Sciences and Informatics 5, 33–43 (2010). URL http://ijmsi.ir/article-1-139-en.html.

64. Thompson, G. Z., Maitra, R., Meeker, W. Q. & Bastawros, A. F. Classification with the matrixvariate-*t* distribution. Journal of Computational and Graphical Statistics 29, 668–674 (2020). URL https://doi.org/10.1080/10618600.2019.1696208.

65. Lund, S. P. & Iyer, H. Likelihood ratio as weight of forensic evidence: A closer look. Journal of Research of the National Institute of Standards and Technology 122 (2017). URL https: //doi.org/10.6028/jres.122.027.

66. Beachem, C. & Yoder, G. Elastic−plastic fracture by homogeneous microvoid coalescence tearing along alternating shear planes. Metallurgical Transactions 4A, 1145—-1153 (1973).

67. Duez, P., Weller, T., Brubaker, M., Hockensmith-II, R. E. & Lilien, R. Development and validation of a virtual examination tool for firearm forensics. Journal of Forensic Sciences 63, 1069−1084 (2018).

68. Chapnick, C. et al. Results of the 3d virtual comparison microscopy error rate (vcmer) study for firearm forensics. Journal of Forensic Sciences 66, 557−570 (2020).

69. Meeker, W. Q., Hahn, G. J. & Escobar, L. A. Statistical Intervals: a Guide for Practitioners and Researchers (John Wiley & Sons, 2017), second edn.

70. Peacock, J. A. Two-dimensional goodness-of-fit testing in astronomy. Monthly Notices of the Royal Astronomical Society 202, 615−627 (1983). URL https://doi.org/10.1093/mnras/202.3.615.

71. Xiao, Y. A fast algorithm for two-dimensional Kolmogorov-Smirnov two sample tests. Computational Statistics & Data Analysis 105, 53−58 (2017). URL https://doi.org/10.1016/j.csda.2016.07.014.

72. Cleveland, W. S. Robust locally weighted regression and smoothing scatterplots. Journal of the American Statistical Association 74, 829−836 (1979). URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1979.10481038.

73. Austin, P. C. & Steyerberg, E. W. Graphical assessment of internal and external calibration of logistic regression models by using loess smoothers. Statistics in Medicine 33, 517−535 (2014). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.5941.