Literature Review

Joe Zemmels

A scientific method must be effective, accessible, and understandable for it to be accepted by the wider scientific community.

In this work, we discuss developments to a particular class of algorithms used to compare cartridge case evidence known as the Congruent Matching Cells (CMC) method. Chapter [#] discusses a modularization of the algorithm into a "pipeline" that enables reproducibility, experimentation, and comprehension. Chapter [#] introduces novel pieces of this pipeline that we demonstrate provide improvements to the current state-of-the-art. Chapter [#] details a suite of diagnostic tools that illuminate the inner-workings of the algorithm and help determine when and why the algorithm does or does not "work" correctly.

Forensic Examinations

A primary goal of a forensic examination is to determine the source of a piece of evidence. This is commonly referred to as the *source identification* problem [citation needed]. A common setting for source identification problems involves obtaining evidence of unknown source from a crime scene and either evidence from a known source or other evidence of unknown source. For example, a fingerprint found at a crime scene may be compared to a suspect's fingerprint or to a fingerprint found at a different crime scene. [cite Ommen and Saunders] refer to the former setting as the *specific source* problem and the latter as the *common source* problem. These two problems differ substantively when it comes quantifying the *probative value* of the evidence, which is outside of the scope of this work [cite SLR papers]. We focus instead on where these two problems agree: measuring the similarity between two pieces of evidence. Specifically, we focus on methods used to measure the similarity between evidence obtained from firearms.

Firearms and Toolmarks Examination

Firearms and toolmarks (F&T) examination involves studying markings or impressions left by a firearm or other tool (e.g., a screwdriver) on a surface [cite Thompson, 2017]. The focus of this work is on a subset of F&T problems that deal with the comparison of cartridge case evidence. A *cartridge case* is the portion of firearm ammunition that encases a projectile (e.g., bullet, shots, or slug) along with the explosive used to propel the projectile through the firearm. When a firearm is discharged, the projectile is propelled down the barrel of the firearm, while the cartridge case is forced towards the back of the barrel. It strikes the back wall, known as the *breech face*, of the barrel with considerable force, thereby imprinting any markings on the breech face onto the cartridge case, creating the so-called *breech face impressions*. These markings have been suggested to be unique to a firearm and are used in forensic examinations to determine whether two cartridge cases have been fired by the same firearm.

During a forensic examination, two pieces of ballistic evidence are placed under a comparison microscope. Comparison microscopes allow for a side-by-side comparison of two objects within the same viewfinder, as seen in [Figure 1]. A pair of breech face images is aligned along the thin black line in the middle of the images. The degree to which these breech face markings can be aligned is used to determine whether the two cartridge cases came from the same source; i.e., were fired from the same firearm. These breech face impressions are considered to be a firearm's unique "fingerprint" left on a cartridge case [cite Thompson, 2017].

A F&T examination typically ends in one of three conclusions: identification, meaning the evidence originated from the same source, exclusion, meaning the evidence did not originate from the same source, or inconclusive, meaning there is insufficient information to conclude identification or exclusion [cite AFTE theory]. Examiners

rarely need to provide quantitative justification for their conclusion. Even for qualitative justifications, it can be difficult to determine what the examiner is actually "looking at" to arrive to their conclusion [cite black vs. white box studies]. [Cite specific conclusions from black vs. white box studies].

Due to the opacity in the decision-making process, examiners have been referred to as "black boxes" in a similar sense to black box algorithms [cite black box algorithm papers]. Their evidentiary conclusions are fundamentally subjective, and there is empirical evidence to suggest that conclusions differ across examiners when presented with the same evidence and even within a single examiner when presented with the same evidence on two different occasions [cite reproducibility and repeatability studies]. This suggests the need to supplement these black box decisions with transparent, objective techniques that quantitatively measure the similarity between pieces of evidence [cite NAS, PCAST]. In this work, we focus on a specific set of techniques used to compare cartridge case evidence.

Forensic Comparison Algorithms

Recent work in many forensic disciplines has focused on the development of algorithms to measure the similarity between pieces of evidence including glass [cite Park and Carriquiry, 2019], handwriting [Crawford, 2020], shoe prints [Park and Carriquiry, 2020], ballistics [Hare et al., 2017; Tai and Eddy, 2018], and toolmarks [Chumbley et al., Krishnan and Hofmann]. These algorithms often result in a numerical, non-binary (dis)similarity score for two pieces of evidence. A non-binary score adds additional nuance to an evidentiary conclusion beyond simply stating whether the evidence did or did not originate from the same source as would be the case in binary classification. For example, the larger the similarity score, the "more similar" the evidence. However, a binary (or ternary, if admitting inconclusives) conclusion must ultimately be reached by an examiner, and there is not yet a consensus on whether a decision rule should be created based solely on results of a comparison algorithm (e.g., defining a score-based decision boundary) or if an examiner should incorporate the similarity score into their own decision-making process [cite the taxonomy of algorithms paper].

Ballistics Comparison Algorithms

[CMS paper] is a seminal work in developing a more objective technique for comparing ballistic evidence. The authors describe [xyz]

Falling under the category of *pattern evidence*, data collected from ballistics evidence is often pictorial in nature. For example, one may take high-resolution pictures of two cartridge cases and develop algorithms to compare these pictures. As such, many ballistics comparison algorithms rely at least partially on the well-established fields of image processing and computer vision.

Hare et al., Tai and Eddy, Riva papers, Siamese Neural Network paper, Roth(?) polar coordinates paper

Song et al. (2012)

Congruent Matching Cells Methodology

A particular class of algorithms used to compare ballistics evidence is known as the Congruent Matching Cells (CMC) algorithms. Major developments for these algorithms have largely originated from a group of researchers at the National Institute of Standards and Technology (NIST). As these algorithms are of particular import to this work, we present a "timeline" of CMC-related publications here.

[Vorberger et al. (2008)] summarize a study to determine the efficacy a national ballistics identification system. As part of this study, the authors explored automatic methods for comparing various types of ballistics evidence including using the cross correlation function (CCF) to determine the translation and rotation at which two cartridge cases *register* (i.e., are most similar). [Weller et al. (2012)] demonstrate the effectiveness of this method on a set of cartridge cases fired from 10 consecutively manufactured pistol slides in which they conclude [xyz].

Building upon this work, [Song (2013)] proposes partitioning the scan into a grid of *correlation cells* that are individually compared. The assumption underlying this method is that some of these cells will capture

areas of the cartridge case surface that contain identifying markings while other cells will not. By narrowing the focus to only those cells that capture identifying markings, we can more accurately assess the similarity between the regions of the cartridge cases that "matter." The total number of cells that are deemed sufficiently "congruent" between the two cartridge case scans is referred to as the CMC count similarity score. The authors provide a few details of an actual implementation that was tested on the data set of 3D topographical scans from [Fadul et al. (2011)], but these results are presented instead in [Chu et al. (2013)]. Follow-up papers [Song et al. (2014)] and [Tong et al. (2014)] validate this "Congruent Matching Cells" algorithm on 3D topographical and 2D optical images of the [Fadul et al. (2011)] cartridge cases, respectively. Both of these papers demonstrate that the NIST implementation of the CMC method can distinguish between the matching and non-matching comparisons from the [Fadul et al. (2011)] data set, although applying the method to 3D topographies appears to be more effective than 2D optical images. Finally, [Ott et al. (2017)] validate the original CMC method on 3D topographical scans of cartridge cases and firing pin impressions from two firearm proficiency studies [citation needed].

[Tong et al. (2015), Chen et al. (2017), Chen et al. (2018), Tong et al. (2018)] introduce "improvements" to the CMC method. Of particular note is the "High CMC" method presented in [Tong et al. (2015)] and the "Convergence CMC" method presented in [Chen et al. (2017)]. [Chen et al. (2017)] provided a modest sensitivity analysis by comparing the performance of the various CMC methods on four data sets [Fadul et al. (2011), Weller et al. (2012), other two] using the same parameter settings and conclude that the Convergence CMC method performs best.

Of particular note to the contents of this paper is the usage of the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm used to compare firing pin impression scans in [Zhang et al. (2021)]. The authors use the DBSCAN algorithm to map the registration information collected from a firing pin comparison to a binary, match/non-match classification. This method will be discussed in greater detail in [Chapter].

[What's missing from current CMC methodology]

- a well-defined process to validate or experiment with the method
- clarity in how anything is actually implemented
- an indicator of when and why the CMC method "fails" at its intended usage.

While the CMC methodology has been shown to work well in classifying matching and non-matching cartridge case comparisons for a handful of data sets, it is far from being usable in-practice in a forensic or legal setting. This is partially due to

Reproducibility, Comprehensibility, and Approachability of Algorithms

When evidence derived from a scientific method is presented to a judge or jury during court proceedings, there is an ethical imperative that the underlying scientific method be proven effective in its intended usage. The [Daubert] standard legally codifies this notion by requiring a scientific method to satisfy a set of criteria before being considered admissible. These criteria include that the method be generally accepted in the scientific community, that it can be and has been tested, and that it has a known error rate.

An impediment to achieving general acceptance of a scientific method is if the method is difficult to test by the wider scientific community, and therefore difficult to estimate its error rate [cite a paper on difficult to reproduce studies]. This prompts the question: how can we make scientific methods easier to test? Generally, "easier" implies that the method requires few resources or resources that the scientific community has access to. In the case of computational algorithms, virtually everyone has access to a computer and therefore has the means to run (reasonably-sized) algorithms. Additional requirements include code and input data which, if not available, must be produced before an algorithm can be executed. This can be expensive and time-consuming [citation here].

Comprehensibility = understanding what the algorithm is doing. Approachability = being able to use or change the algorithm yourself.

Baggerly and Coombes (2009) *Deriving Chemosensitivity from Cell Lines: Forensic Bioinformatics and Reproducible Research in High-Throughput Biology*

Donoho (2017) *50 Years of Data Science*

tidyverse functionality

National Academies of Sciences, Engineering, and Medicine (2019) *Reproducibility and Replicability in Science*

In cartridge case evidence, the Congruent Matching Cells (CMC) methods are one class of algorithms used to measure the similarity between two cartridge cases. Numerous authors have demonstrated the ability of the CMC methods to effectively distinguish between matching and non-matching pairs of cartridge cases. However, to-date only conceptual descriptions of the CMC methods, along with results derived from an internal implementation of the described algorithm, have been published for the wider scientific community. These published descriptions and results demonstrate that the authors' implementation of the CMC methods work as intended, yet fail to ensure others can reproduce or develop upon the published work without having to create their own implementation. By "reproduce," we mean *computational reproducibility* as defined by the National Academy of Science, Engineering, and Medicine:

> Definition here

For a method to be widely accepted by the scientific community, results must be reproducible (by others).

By definition, algorithms are repeatable assuming the same data are provided as input and a seed is set for any internal random number generation. That is to say, if the exact same procedure is performed on the same data on two separate occasions, then the results will be the same. However, reproducibility is still in-question for many forensic comparison algorithms. By reproducibility, we mean *computational reproducibility* as defined by the [National Academy of Sciences, Engineering, and Medicine]: "obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis." As we argue in [Chapter], many published forensic comparison algorithms fail to provide sufficient detail or resources to satisfy one or more of these criteria and, therefore, are not reproducible. In-short: if the code and data exist, then you should share it. Using a specific cartridge case comparison algorithm as an example, we detail a development process by which algorithms can not only satisfy reproducibility, but can also be more easily understood and accessed by the wider scientific community.

In Chapter 6, we introduce a taxonomy to classify various levels of computational reproducibility. We argue that currently-published versions of the CMC method are provided in the form of *conceptual descriptions* while our implementation, paired with data that are open-source on the National Ballistics Toolmark Database [cite NBTRD], satisfies the definition of computational reproducibility provided by the NASEM [cite NASEM].

Beyond considerations of reproducibility, which is a core tenet of science, are considerations of comprehensibility and approachability of an algorithm. [Discuss how pipeline aids in comprehension, experimentation].

For data discussion: cite Vorberger et al. (2007), NBTRD, and TopMatch

[Carpentry R package, gapminder reproducibility seminar] https://swcarpentry.github.io/r-novice-gapminder/

Visual Diagnostics

Forensic examiners often provide expert testimony in court cases. In these situations, examiners need to be able to explain the process by which they reached an evidentiary conclusion to the fact finders of the case; namely the judge or jury. [Talk about explanation of testimony historically – look for papers].

As algorithms are more often used in evidentiary conclusions, the technical know-how needed to not only understand how the algorithm works but also be able to explain it to lay-people has increased. In some cases, the creators of the algorithm have been willing to sit as expert witnesses [check if this is the proper term] [cite strmix testimony]. [However, this will not be as viable as algorithms are used more often in more cases. There are two potential solutions to this problem. The first is to educate examiners on the technical details of the algorithm. This, however, requires resources to create and disseminate the educational materials and examiners' willingness to learn these details. Even if examiners were willing to learn about the algorithm,

many algorithms require technical knowledge that may be out of the scope of an examiner's standard training. Even if examiners were to be educated to have the technical know-how needed to understand the algorithm, many algorithms, specifically commonly-used machine learning algorithms, are currently "unexplainable" in the sense that they rely on derived features that are not human-interpretable [cite unexplainable machine learning algorithms].]

Because of these hurdles, educating examiners on the use of highly technical algorithms seems currently implausible. An alternative solution, is to develop algorithms from the ground-up to be intuitive for examiners to understand and explain. One aspect of explainability is being able to diagnose when and why an algorithm succeeds or fails at its intended usage [cite a paper detailing tenets of explainable machine learning algorithms].]

We use the visual diagnostic tools discussed in Chapter [5] to develop a set of features. By definition, these features are human-interpretable unlike, for exmaple, features that are calculated in the convolution layer of a convolutional neural network. The interpretability of these features imply that they can be explained to forensic examiners or lay-people. This will make it easier to introduce such methods into forensic labs and court rooms.