(Numerical Designation)
2017

This document has been accepted by the Academy Standards Board (ASB) for development as an American National Standard (ANS). For information about ASB and their process please refer to asb.aafs.org. This document is being made available at this stage of the process so that the forensic science community and interested stakeholders can be more fully aware of the efforts and work products of the Organization of Scientific Area Committees for Forensic Science (OSAC). The documents were prepared with input from OSAC Legal Resource Committee, Quality Infrastructure Committee, and Human Factors Committees, as well as the relevant Scientific Area Committee. The content of the documents listed below is subject to change during the standards development process within ASB, and may not represent the contents of the final published standard. All stakeholder groups or individuals, are strongly encouraged to submit technical comments on this draft document during the ASB's open comment period. Technical comments will not be accepted if submitted to the OSAC Scientific Area Committee or Subcommittees.

Standard for Topography Comparison Software for Firearm and Toolmark Analysis

Organization of OSAC Scientific Area Committees for Forensic Science

DRAFT DOCUMENT

(Numerical Designation)
2017

Standard for Topography Comparison Software for Firearm and Toolmark Analysis

Keywords: *3D, Topography, Measurement, Algorithm, Quality Assurance, Firearms, Toolmarks, Identification*

This document specifies the minimum requirements for computer software intended to compare 2D and/or 3D digital representations of toolmarks. It covers necessary conditions for consistent and interpretable comparisons. Software that complies with the specifications of this document can be used for topography analysis and comparison.

Foreword

This standard was proposed by the Firearms and Toolmarks Subcommittee of the Organization of Scientific Area Committees (OSAC) by submitting a request to the American Academy of Forensic Sciences (AAFS) Academy Standards Board (ASB).

This document is part of a series of documents jointly submitted to include:

1. Standard for 3D Measurement Systems and Measurement Quality Control for Firearm and Toolmark Analysis
2. Standard for Topography Comparison Software for Firearm and Toolmark Analysis
3. Standard for Implementation of 3D Technologies in Forensic Laboratories for Firearm and Toolmark Analysis

The purpose of these standards is to ensure that new technologies produce accurate measurements and a validated statistical assessment of the significance of the correspondence. The documents establish performance expectations for new technologies while allowing legacy systems to coexist in the lab. The hardware document specifically refers to 3D scanning hardware and does not apply to legacy 2D type systems. The software document specifies three categories (levels) of software. Legacy systems are Category 0 whereas systems which provide validated statistical measures are Category 2. The implementation document outlines the necessary steps to ensure the proper implementation of 3D technologies.

Acknowledgements

Editor:

Deputy Editor(s):

Drafting Working Group Members:

Ryan Lilien	Cadre Research Labs - Chicago, Illinois
Nicholas Petraco	John Jay College of Criminal Justice - New York, New York
Erich Smith	Firearms/Toolmarks Unit - Federal Bureau of Investigation (FBI) - Quantico, Virginia
Todd Weller	Weller Forensics - Burlingame, California
Xiaoyu Zheng	Engineering Physics Division - National Institute of Standards and Technology (NIST) - Gaithersburg, Maryland

Consensus Group Members:

Table of Contents

1 Scope

This document specifies the minimum requirements for computer software intended to compare 2D or 3D digital representations of toolmarks. It covers necessary conditions for consistent and interpretable comparisons. Software that complies with the specifications of this document can be used for topography analysis and comparison.

Topography analysis and comparison software provides a means of evaluating the similarities and differences between high resolution surface topographies. The aim of such analysis is a quantifiable measure of toolmark topography comparisons to assist an examiner in reaching a conclusion or an assessment of the weight of the evidence regarding common origin. This standard primarily supports firearm and toolmark examination but also can be applied to other measured surface topographies.

Software shall advance through a series of categories. When the requirements of a category are satisfied, the software may be used for the indicated purpose and in the indicated manner. This standard is applicable to all forensic science service providers that provide conclusions regarding toolmark related evidence.

2 Normative References

ISO 25178-72 Geometrical product specifications (GPS) -- Surface texture: Areal -- Part 72: *XML file format x3p*

3 Terms and Definitions

3.1
Comparison Algorithm
A series of computational steps which seeks to assess both the level of geometric similarity (similarity of toolmarks) and the degree of certainty that the observed similarity results from a common origin. A comparison algorithm makes use of a scoring function or similarity score.

3.2
comparison software
Software that implements a comparison algorithm and may also include database, search, and visualization functionality.

3.3
comma-separated value file
CSV
A simple file format for tabular data where individual values are separated by a comma (or other designated delimiter).

3.4
frequency
The rate at which an event occurs.

3.5
heightmap
A three-dimensional topographic data set consisting of surface points (*x*, *y*, *z*) where each dimension is a coordinate measured in standard units (e.g., micrometers). See topographic data.

3.6
interpretable scoring function
A Category 1 scoring function. The score is explainable and has quantifiable meaning. See Section 4.2.5.

3.7
noncompliant 3D measurement
A 3D Topographic measurement made using hardware that does not comply with the "Standard for 3D Measurement Systems and Measurement Quality Control for Firearm and Toolmark Analysis“ document. The accuracy and precision of these measurements may be uncertain and untraceable.

3.8
rank-score only scoring function
A Category 0 scoring function. See section 4.2.4. A rank-score only scoring function is either non-quantified, inconsistent, or non-explainable.

3.9
probability
A quantified measure between zero and one indicating how probable or likely it is that an event will or has occurred. In the frequentist interpretation, probability is based on the rates at which events occur. In the Bayesian interpretation, probability reflects a degree of belief. On this scale, zero indicates impossibility and one indicates absolute certainty.

3.10
scoring function (or similarity score)
The mathematical core of a comparison algorithm which takes two input topographies and generates one or more outputs that quantify the comparison based on geometric similarity. Scoring functions are either rank-score only, interpretable, or statistically validated.

3.11
statistically validated scoring function
A Category 2 scoring function. See section 4.2.6.

3.12
topography data
A measurement of an object’s surface geometry. Topographic data may be one, two, or three dimensional.

3.13
topography data: 1D data
1D topography data is also known as a Linear Profile. A linear profile with n points shall be represented as a function of a single coordinate where at each point along a single axis (x_i) (where i=1,...,n) there is a measured height (z_i). Both x_i and z_i are measured in standard units (e.g., micrometers). An example of a linear profile is a cross-section through a striated toolmark (e.g.,

bullet land area); where x_i is a spatial position measured in micrometers and z_i is the corresponding height of the striation profile measured in micrometers.

3.14
topography data: 2D data
2D topography data is also known as a Planar Image. An n-by-m planar image I shall be represented as a function of two coordinates where at each point (x_i, y_j) (where i=1,...,n; j=1,...,m) there is a measured surface color or intensity denoted $I(x_i, y_j)$. An example of 2D topographic data is an image taken through a comparison microscope; where each point $I(x_i, y_j)$ is the RGB (red, green, blue) color value measured at the specified (x_i, y_j) position. The measured color or intensity is a function of the surface geometry and the environmental conditions (e.g., light position). Although reference scales may be included in the collected image, the points (x_i, y_j) may or may not be measured in standard units (e.g., micrometers).

3.15
topography data: 3D data
3D topography data is also known as a Heightmap. An n-by-m heightmap H shall be represented as a function of two coordinates where at each point (x_i, y_j) (where i=1,...,n; j=1,...,m) there is a measured surface height $z_{i,j} = H(x_i, y_j)$. All three coordinates x_i, y_j, and $z_{i,j}$ are measured in standard units (e.g., micrometers). The surface H is a 1-to-1 representation of the actual object. An example of 3D topographic data is a primer surface measured using a confocal microscope; where each point $H(x_i, y_j)$ represents the surface height (in micrometers) measured at the specified (x_i, y_j) position. 3D data captured on hardware compliant with the “Standard for 3D Measurement Systems and Measurement Quality Control for Firearm and Toolmark Analysis” document can be used in virtual comparison microscopy.

3.16
XML 3D Surface Profile
X3P
X3P is an open file format for the exchange of three dimensional surface topography data in standard units. Details are specified in the ISO Final Draft International Standard (FDIS) 25178-72.

4 Requirements

4.1 Data and Format Requirements (Mandatory)

4.1.1
Comparison Software shall accept as input one or more of the following topographies: a Linear Profile (1D), a Planar Image (2D), or a Heightmap (3D).

4.1.2
Comparison Software should assist the forensic examiner in a number of ways. The comparison algorithm should be able to compare one object to another (e.g., individual comparison), one object to a set of many objects (e.g., database search), or a set of objects to another set of objects (e.g., expanded database search). The resulting similarity scores shall be interpreted in the manner described for the corresponding software Category described in Section 4.2.

4.1.3
1D Data

4.1.3.1
Data Format: For the representation and exchange of 1D Topographic Data, comparison software shall support either X3P or CSV.

4.1.3.2
Data Source: Linear profiles are typically obtained from a 1D profilometer or by processing (e.g., cross-section) 3D topographic data.

4.1.3.3
Data Representation: 1D Topographic Data shall be measured in standard units of length (e.g., micrometers).

4.1.4
2D Data

4.1.4.1
Data Format: For the representation and exchange of 2D Topographic Data, comparison software shall support any commonly established lossless image file format (e.g., TIFF, PNG).

4.1.4.2
Data Source: Planar Images are typically acquired using a traditional digital camera attached to a comparison microscope or similar imaging optics.

4.1.4.3
Data Representation: 2D Topographic Data shall be measured on a grid of fixed dimension sample points. The interpoint spacing may or may not be measured in standard units of length (e.g., micrometers).

4.1.5
3D Data

4.1.5.1
Data Format: For the representation and exchange of 3D Topographic Data, comparison software shall support the X3P file format. The X3P data shall follow the specifications of FDIS ISO 25178-72 with the following addition. Record 2 (Metadata), which is an optional data record in the ISO document, shall be a required record for toolmark analysis. Three fields (Version, ProbingType:Identification, and Comment) may be left blank, but all Record 2 fields specified in the ISO document shall be included in the X3P file. It is recommended that data from cartridge cases and bullets also include an optional Record X which specifies toolmark metadata. Record X shall have a <VendorSpecificID> field equal to http://www.openfmc.org/firearm and shall include the fields specified by the OpenFMC group. The X3P may also include vendor specific records.

4.1.5.2
Data Source: Heightmaps shall be measured on imaging hardware compliant with the "Standard for 3D Measurement Systems and Measurement Quality Control for Firearm and Toolmark Analysis" document and with procedures compliant with the "Standard for Implementation of 3D Technologies in Forensic Laboratories for Firearm and Toolmark Analysis" document.

Measurements obtained using hardware that does not comply with the "Standard for 3D Measurement Systems and Measurement Quality Control for Firearm and Toolmark Analysis" document shall be considered Noncompliant 3D Measurements and the value of their interpretation may be limited. Noncompliant 3D measurements shall only be used with Category 0 software. Noncompliant 3D measurements shall not be used with Category 1 or Category 2 software. Noncompliant 3D measurements shall not be exchanged between labs.

4.1.5.3
Data Representation: 3D Topography Data shall be measured on a grid of fixed dimension sample points. The interpoint spacing and heights (x, y, and z) shall all be measured in standard units.

4.1.5.4
3D data shared between labs shall be captured on hardware compliant with the "Standard for 3D Measurement Systems and Measurement Quality Control for Firearm and Toolmark Analysis" document and shall be stored in X3P format. 3D data not collected on compliant hardware shall not be exchanged between labs.

4.2 Software Categories (Mandatory)

4.2.1
Comparison Software seeks to provide, in an explainable manner, a numeric measure (e.g., degree of certainty) quantifying the geometric support for common origin.

4.2.2
Comparison Software (and its associated Scoring Function) shall be evaluated before its use by forensic science service providers. Software evaluation shall advance through a series of Categories. When a software satisfies the requirements of a category it can be used for the purposes and in the manner described for the corresponding category. Software shall not be used for the purposes or in the manner of categories whose requirements have not been satisfied.

4.2.3
In accordance with the "Standard for Implementation of 3D Technologies in Forensic Laboratories for Firearm and Toolmark Analysis" document, all software, regardless of Category, shall be validated prior to use. Therefore, developmental and deployment validation studies must be completed.

4.2.4
Category 0: Rank-Scores Only Scoring Function

4.2.4.1
Criteria for Category 0: Any comparison software may attain a Category 0 designation by completing developmental and deployment validation studies (Section 4.3). Any measurement hardware may be used with Category 0 software. That is, measurement hardware does not need to comply with the "Standard for 3D Measurement Systems and Measurement Quality Control for Firearm and Toolmark Analysis" document to be used with Category 0 software. Category 0 software is considered Rank-Scores Only.

4.2.4.2
Use of Category 0 Software: Rank-Scores Only software shall only be used as part of a hit finding or

sorting process. The numeric score (e.g., similarity score or any other numeric measure of match quality) shall not be mentioned in reports or discussion. That is, the scores shall not be used in any way secondary to ranking. For example, one shall not state that the pair of cartridge cases has a match score of X, which is very high. One may only state that the pair of cartridge cases ranked in the top Y of the database search. (where Y may be a number, e.g., 10, 20). If asked about the significance of this statement the only acceptable response shall be that there is no statistical confidence established for any match results for a rank-scores only, non-statistically validated scoring function.

4.2.5
Category 1: Interpretable Scoring Function

4.2.5.1
Criteria for Category 1: Category 1 software shall only utilize scan data collected using hardware compliant with the "Standard for 3D Measurement Systems and Measurement Quality Control for Firearm and Toolmark Analysis" document. To satisfy Category 1 and be considered an Interpretable Scoring Function, the scoring function and software shall complete developmental and deployment validation studies (Section 4.3). In addition, for a single sample-to-sample comparison an Interpretable Scoring Function:

4.2.5.1.1
Shall output a single quantified numeric value.

4.2.5.1.2
Shall be reported on a consistent scale. A score of X obtained in two different searches shall mean the same thing and shall carry the same degree of confidence, probability, statistical weight, or likelihood. For example, if cartridge cases C1 and C2 have a similarity score of X and if cartridge cases C3 and C4 have a similarity score of X then the confidence that C1 and C2 have common origin shall be equal to the confidence that C3 and C4 have common origin.

4.2.5.1.3
Shall be explainable, such that a firearms examiner can describe the general principles on which it works.

4.2.5.2
Scoring functions not meeting the Interpretable Scoring Function criteria shall be considered Rank-Score Only Scoring Functions (Category 0). These Rank-Score Only Scoring Functions are either non-quantified, inconsistent, or non-explainable.

4.2.5.3
Use of Category 1 Software: Comparison Software utilizing an Interpretable Scoring function may be used for all uses indicated for Category 0. In addition, the numeric score (e.g., similarity score or any other numeric measure of match quality) may be described in reports or discussion. For example, one may state that a pair of cartridge cases has a match score of X, which is very high. If asked about the significance of this statement for software at Category 1, the only acceptable response shall be that there is no statistical confidence established for any match results for a non-statistically validated scoring function. Interpretable Scoring functions can be used to develop a statistical model of confidence (Category 2).

4.2.6
Category 2: Statistically Validated Scoring Function

4.2.6.1
Criteria for Category 2: To satisfy Category 2 and be considered a Statistically Validated Scoring Function, software shall meet the criteria for an Interpretable Scoring function and have completed extensive developmental and deployment validation studies (Section 4.3) to demonstrate the statistical performance on a large representative test set. In addition Category 2 software shall only utilize scan data collected using hardware compliant with the "Standard for 3D Measurement Systems and Measurement Quality Control for Firearm and Toolmark Analysis" document.

4.2.6.2
Use of Category 2 Software: Comparison Software completing the validation requirements of Category 2 may be used for all uses indicated for all Categories. Software at Category 2 has established statistical confidence for match results. The statistical significance or support of the numeric score (e.g., similarity score or any other numeric measure of match quality) may be described in reports or discussion. One may present statements such as, "the pair of cartridge cases has a similarity score of X. At a threshold of X, the false match probability is Y". If asked about the significance of this statement for software at Category 2, an acceptable answer shall cite relevant validation studies and their relevance to the question at hand. For example, "A 2017 study by W included V test fires from firearms with relevant manufacturing techniques, ammunition types, and materials. The study demonstrated that a match score of X indicates a false match probability of Y."

4.3 Software Validation (Mandatory)

In accordance with the "Standard for Implementation of 3D Technologies in Forensic Laboratories for Firearm and Toolmark Analysis" document, all toolmark analysis software, regardless of Category and regardless of toolmark type being examined, shall be validated prior to use.

The developmental validation shall be conducted by an organization with appropriate knowledge and/or expertise. The deployment validation is a smaller follow-on evaluation and shall be conducted by a lab prior to their implementation of a new technology. A deployment validation shall only be completed after a successful development validation. Validation studies shall be documented as described in the "Standard for Implementation of 3D Technologies in Forensic Laboratories for Firearm and Toolmark Analysis".

Comparison and Analysis software is not subject to traditional Ongoing Performance Checks as a single piece/version of software does not fall out of calibration. Software upgrades that can affect scoring function functionality (such as a major version upgrade) may require additional validation.

4.3.1
Software at Categories 0 and 1: User certification shall be utilized as described in section 4.4. In addition, development and deployment validation must demonstrate:

4.3.1.1
The limitations of the procedure.

4.3.1.2
The conditions under which reliable results can be obtained.

4.3.1.3
Critical aspects of the procedure that shall be controlled and monitored.

4.3.1.4
The ability of the procedure to meet the needs of the given application.

4.3.2
Software at Category 2: User certification shall be utilized as described in section 4.4. The developmental and deployment validation for Category 2 software shall include a large statistical study aimed at demonstrating the statistical significance of the scoring function.

4.3.2.1
Developmental Validation: Developmental validation studies shall be robust and report on the tools for which the study was explicitly designed and on which the results are explicitly applicable (e.g., for firearms toolmarks report on the firearm manufacturing techniques, ammunition calibers, ammunition types, ammunition materials, etc...). To guard against overfitting and improve the likelihood that performance generalizes across actual casework data, the validation set shall include both a sufficiently broad selection of makes and models as well as a sufficiently broad selection of substrate types (e.g., for firearms toolmarks substrate types include ammunition materials). The larger and more inclusive the validation set the stronger the validation result. Studies shall be sufficiently large to determine statistical performance to an appropriate degree of certainty. See Annex A1 for an example study set. All developmental validation studies shall estimate error rates and uncertainties on these error rates using established statistical methods.

The developmental validation report shall include overall performance measures indicating at least the Recall Rate, False Positive Rate, and Positive Predictive Value.

4.3.2.1.1
Recall Rate (or true positive rate) shall be defined as the number (or percentage) of cartridge cases containing a Known-Match in the test set for which a Known-Match is correctly determined.

4.3.2.1.2
False Positive Rate shall be defined as the number (or percentage) of Known Non-Matches which are incorrectly determined to be an Identification.

4.3.2.1.3
Positive Predictive Value shall be defined as the prortion of identifications that correspond to true known matches, TP/(TP+FP) where TP (True Positives) is the number of correctly identified Known Matches and FP (False Positives) is the number of Known Non-Matches incorrectly determined to be an Identification.

4.3.2.2
Deployment Validation: The deployment validation shall be more modest in size than the developmental validation. For example, a deployment validation may involve running a small sample of proficiency tests. In completion of the validation study the following aspects shall be documented:

4.3.2.2.1
The limitations of the validation study.

4.3.2.2.2
The conditions under which reliable results can be obtained.

4.3.2.2.3
Critical aspects of the procedure that shall be controlled and monitored.

4.3.2.2.4
The ability of the procedure to meet the needs of the given application.

4.3.2.3
Both development and deployment validation studies shall use test sets that are independent from the initial sets used to build the scoring function or statistical model.

4.3.2.4
Future developments in manufacturing techniques may require additional studies to augment the results of the development and deployment validation studies.

4.4 User Training (Mandatory)

A training process or program shall be established by the authors and/or distributors of all comparison software. The training program shall include an assessment module. The laboratory management shall ensure the initial and continued competence of all users of comparison software within the laboratory. Details of this process are provided in the "Standard for Implementation of 3D Technologies in Forensic Laboratories for Firearm and Toolmark Analysis" document.

4.5 Statistical Models and Interpretation (Mandatory)

4.5.1
Statistical comparison models shall output a statistically grounded metric indicative of whether or not two toolmarks have a common origin or support for common origin.

4.5.2
Statistical Comparison Models shall be constructed only from Interpretable Scoring Functions satisfying Category 1.

4.5.3
Statistical models may employ either a generative or an empirical approach.

4.5.3.1
A generative model estimates the probability of obtaining a similarity score by simulating the toolmark generation and subsequent matching process.

4.5.3.2
An empirical model uses a representative set of relevant population data to estimate the confidence, probability, or frequency of obtaining specific similarity score(s) for a known match or a known non-match. With respect to firearm forensics, the population data shall include a sufficiently broad selection of firearm makes and models as well as a sufficiently broad selection of ammunition types, dependent on the case at hand. An empirical model measures the observed similarity scores over the dataset's known matches and known non-matches. Several statistical methods may be used to compute match or non-match confidence, probability, or frequency.

4.5.4
Statistical Models shall be validated prior to their use.

4.6 Criteria for Interpretation (Mandatory)

4.6.1
Scoring functions and statistical models at Category 2 may be used by an examiner to provide a statistical assessment regarding common origin or weight of evidence.

4.6.2
Conclusions regarding common origin or statements of weight of evidence shall follow the standard operating procedure of the laboratory. Interpretable Scoring Functions and statistical models provide one way for stating the degree of certainty or weight of evidence for common origin and for supporting the finding based on the underlying comparative model.

5 Conformance

Conformance with this *Standard for Topography Comparison Software for Firearm and Toolmark Analysis* document will be accessed utilizing these documents:

ASCLD/LAB-International Supplemental Requirements for the Accreditation of Forensic Science Testing Laboratories, American Society of Crime Laboratory Directors/Laboratory Accreditation Board, Garner, NC, 2011.

ISO/IEC 17025 - General Requirements for the Competence of Testing and Calibration Laboratories, International Organization for Standardization, Geneva, Switzerland, 2005.

Annex A
(informative)

Sample Validation Study Detail

Sample Development Validation Study of Category 2 Software
A 2017 test set for 9mm Luger firearms would include approximately 4800 test fires from at least sixty (60) different firearm manufacturers and two hundred firearm models. The set would include all commonly used manufacturing mark types (e.g. milled, broached, EDM, and filed). The set would include multiple test fires per firearm involving at least six (6) different brands of ammunition. Bullet samples should contain jacket materials that have been forged or chemically bonded that contain lead, copper, steel, and brass metals. Cartridge case sets should contain cartridge cases with both brass and nickel primers. Cartridge case sets should contain cartridge cases with brass, steel, and nickel cases. Different calibers may or may not require different test sets.