NISTIR 8078

Tattoo Recognition Technology – Challenge (Tatt-C) Outcomes and Recommendations *Revision 1.0*

Mei Ngan
George W. Quinn
Patrick Grother

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.IR.8078

National Institute of Standards and Technology
U.S. Department of Commerce

NISTIR 8078

Tattoo Recognition Technology – Challenge (Tatt-C) Outcomes and Recommendations *Revision 1.0*

Mei Ngan
George W. Quinn
Patrick Grother
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
http://dx.doi.org/10.6028/NIST.IR.8078

September 2016

U.S. Department of Commerce
Penny Pritzker, Secretary

National Institute of Standards and Technology
Willie May, Acting Under Secretary of Commerce for Standards and Technology and Acting Director

Tattoo Recognition Technology - Challenge (Tatt-C)

Outcomes and Recommendations

NIST Interagency Report 8078

(Revision 1.0)

Mei Ngan, George W. Quinn, and Patrick Grother

Image Group
Information Access Division
National Institute of Standards and Technology

Last Updated: September 6, 2016

Executive Summary

Background

Tattoos have been used for many years to assist law enforcement in investigations leading to the identification of both criminals and victims. A tattoo is an elective biometric trait that contains discriminative information to support person identification and investigation in addition to traditional soft biometrics such as age, gender and race. While some research has been done in the area of image-based tattoo detection and retrieval [9, 10, 13, 16, 19, 20], it is not a mature domain. Prior to this study, there were no common datasets to evaluate and develop operationally-relevant tattoo recognition applications. To address this, NIST conducted the Tattoo Recognition Technology - Challenge (Tatt-C) as an initial research challenge that provided operational data (provided by the FBI) and use cases to engage the research community into advancing research and development into automated image-based tattoo technologies and to assess the state-of-the-art to determine what methods are effective and viable for pertinent operational scenarios. Tatt-C builds upon earlier NIST efforts in biometric challenge problems [15] [14] which have helped promote development, advance the state-of-the-art, and benchmark progress in different areas of biometrics.

Tatt-C Test Activity

The Tatt-C activity ran from September 23, 2014 to May 4, 2015, with participation from six organizations, namely Compass Technical Consulting (US), Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (DE), French Alternative Energies and Atomic Energy Commission (FR), MITRE (US), MorphoTrak (US), and Purdue University (US). Tatt-C was conducted as an "open-book" test, where participants were provided with the dataset and ground-truth data, ran their algorithm(s) on the data following a specified protocol on their own hardware, and provided their system output to NIST for uniform scoring and analysis. Accuracy was measured for the five Tatt-C use cases, including the impact of gallery size for certain scenarios. Detailed descriptions and image examples of the use cases can be found in *Section 2.2* of this report.

Key Results

Key results for the five use cases studied are:

- **Tattoo Identification** evaluated matching different instances of the same tattoo image from the same subject over time. On a gallery size of 4 375, the top performing algorithm (MorphoTrak) reported a rank 10 hit rate* of 99.4% and mean average precision (MAP)* of 99.4%. *Section 3.1*
- **Region of Interest** evaluated matching a subregion of interest that is contained in a larger image canvas. On a gallery size of 4 363, the top performing algorithm (MorphoTrak) reported a rank 10 hit rate of 97% and MAP of 95.4%. *Section 3.2*
- **Mixed Media** evaluated matching visually similar or related tattoos using different types of non-tattoo imagery (i.e. sketches, scanned print, computer graphics, and graffiti). On a gallery size of 55, the top performing algorithm (MITRE) reported a rank 10 hit rate of 36.5% and MAP of 15.1%. *Section 3.3*
- **Tattoo Similarity** evaluated matching visually similar or related tattoos from different subjects. On a gallery size of 272, the top performing algorithm (MITRE) achieves a rank 10 accuracy of 14.9% and MAP of 5.2%. *Section 3.4*
- **Tattoo Detection** evaluated detecting whether an image contains a tattoo or not. On a mixed dataset of 1 349 tattoo images and 1 000 face images, the top performing algorithm (MorphoTrak) reported an overall detection accuracy* of 96.3%. *Section 3.5*

Factors that influenced accuracy included:

- **Algorithms**: Tattoo detection and matching accuracy depends strongly on the implementation of the core technology as algorithm performance varied substantially. *Sections 3.1, 3.2, 3.3, 3.4, and 3.5*

*For the definition of hit rate, MAP, and overall detection accuracy, see Section 2.3. Generally speaking, the higher the hit rate, MAP, and detection accuracy value, the more accurate the algorithm.

- **Gallery size**: Gallery size has an impact on accuracy. A decrease in accuracy is observed across all algorithms when the gallery size is increased. *Sections 3.1.2 and 3.2.2*
- **Image quality**: Algorithm failure to find the correct match is often related to quality of the images. Inconsistencies in tattoo image angle, orientation, size of tattoo relative to the entire image, and poor collection characteristics including low illumination, low contrast, blur/out of focus, and the existence of clothing and background clutter contributed to algorithm failure for tattoo detection and matching. *Sections 3.1.3, 3.2.3, 3.3.1, 3.4.1, and 3.5.1*
- **Definition of "tattoo similarity"**: Definitional inconsistencies in "visual similarity" observed in the Tattoo Similarity ground-truth data could be a cause for deflated results for that use case. *Section 3.4*

Recommendations

Based on the outcomes of the Tatt-C activity and the discussions and suggestions from the tattoo recognition developer and user community at the Tatt-C Workshop [8], we propose the following recommendations for future work.

Improve Tattoo Image Capture: Algorithm failure to find the correct match is often related to the consistency and quality of image capture. Notably, inconsistencies in image angle, orientation, size of tattoo relative to the entire image and poor collection characteristics such as low illumination, low contrast, blur/out of focus, and existence of clothing and background clutter caused failures for tattoo detection and matching algorithms. As such, recommendations for improving the quality of tattoo images to support image-based tattoo recognition include:

- **Best Practices**: Develop best practice guidelines for the capture of a tattoo image including photography guidelines and image quality definitions and standards for tattoos. Similar guidelines for the capture of face [1], iris [25], and fingerprints [24] have been developed and successfully employed by law enforcement.
- **Training Material**: Develop and distribute simple reference/training material such as a poster with examples of good quality, well-captured tattoos versus bad quality, poorly-captured tattoos, similar to those that have been developed and successfully utilized for other modalities [3].
- **Software**: During photo capture, use software that makes image quality assessments (e.g., illumination, contrast, focus, existence of distractors around tattoo image) and determines whether to accept or reject the image. Such a software concept is similar to NIST's Fingerprint Image Quality (NFIQ) [26] software, but applicable to tattoos.

Build Larger Tattoo Datasets for Researchers: Nineteen organizations requested and received the Tatt-C dataset (eight commercial, six academic, and five research entities). This is a good indicator of industry and research interest in this area. Following the closing of the Tatt-C submission deadline, NIST continues to receive participation interest and requests for the Tatt-C dataset. As an outcome from the Tatt-C workshop, the community identified a need to provide researchers with larger, ground-truthed tattoo datasets with larger gallery sizes (>100K) closer to operations, and more search images per use case, ideally covering the categories of "challenging images" identified in this report. Larger image sets will impose a level of software and computational robustness on algorithm developers and can support advancing the technology further.

Conduct Sequestered Evaluation: Tatt-C was conducted as an "open-book" test based on the honor system where participants were only required to submit their system output to NIST for uniform scoring and analysis. Such an evaluation protocol does not preclude gaming such as algorithm training on test data, candidate list manipulation, etc. For the Tatt-C use cases that resulted in high accuracy - Tattoo Identification, Region of Interest, and Tattoo Detection, we recommend running a "closed-book" evaluation of tattoo recognition algorithms, where participants send software to NIST to be tested with sequestered data. This will allow consistent assessment of run-time performance (enrollment/search time and memory usage) on a uniform hardware infrastructure, which is important as specific operational applications will have specific run-time requirements. Sequestered evaluations can leverage much larger, operationally-realistic gallery sizes of data that cannot be freely distributed.

Evaluate Tattoo Localization: The presence of clutter and occlusions such as clothing, background, furniture, hair, and other body parts in the operationally-collected data makes tattoo localization an important aspect of the recognition process. Localization and matching are generally two distinct tasks, and matching is critically dependent on correct localization. We recommend evaluation of tattoo localization as a separate use case in future evaluations.

Refine Definition of Tattoo Similarity Specific to Law Enforcement Applications: A common trend in the ground-truth for Tattoo Similarity included images that were semantically relevant, but lacked commonality in visual features, which is likely a result related to the "semantic gap". Human beings are able to interpret images at different levels, both in low-level features (color, shape, texture) and high-level semantics (abstract object, concept, event). Computers are only able to interpret images based on low level-image features. This introduces an interpretation inconsistency between image descriptors and high-level semantics that is known as the semantic gap [28]. While some research has been done in bridging the semantic gap for image retrieval, it remains a challenging, unsolved problem [18]. A set of well-defined criteria for tattoo similarity or relevance specific to law enforcement applications needs to be developed for consistent ground-truthing of data that enables research and development. The feasibility of such an endeavor has yet to be determined. An initial step could be to focus on specific law enforcement application, which will likely have observable commonalities in the dominant tattoo content.

Disclaimer

The results in this report are not to be construed, or represented as endorsements of any participant's system or commercial product, or as official findings on the part of NIST or the U.S. Government. Note that the results submitted by developers of commercial products were generally from research prototypes, not commercially available products. Since Tatt-C was an evaluation of research algorithms, the test design required local implementation by each participant. As such, participants were only required to submit their system output to NIST for uniform scoring and analysis. The systems themselves were not independently evaluated by NIST.

The data, protocols, and metrics employed in this evaluation were chosen to support tattoo detection and recognition research and should not be construed as indicating how well these systems would perform in fielded operations. Changes in the data domain, or changes in the amount of data used to train a system, can greatly influence system performance; changing task protocols could result in performance strengths and weaknesses for these same systems.

In summary, this report should not be interpreted as a product testing exercise, and the results should not be used to make conclusions regarding which commercial products are best for a particular application.

Acknowledgements

The authors would like to thank the sponsor of this activity, the Federal Bureau of Investigation (FBI) Biometric Center of Excellence (BCOE), for initiating and progressing this work. In addition, we would like to thank the FBI Cryptanalysis and Racketeering Records Unit (CRRU), Department of Defense (DoD) Defense Forensics and Biometrics Agency (DFBA), Department of Justice (DOJ) National Institute of Justice (NIJ), and the MITRE Corporation for their support and contributions.

Finally, the authors are grateful to Amanda Noxon (Michigan State Police), Dr. Jim Matey (NIST), and Mike Garris (NIST) for their thorough and constructive review of this document.

Release Notes

- **Versioning**: This document is Revision 1.0 of the original report, which was originally published in September 2015.
- **Typesetting**: Virtually all of the tabulated content in this report was produced automatically. This involved the use of scripting tools to generate directly type-settable LaTeX content. This improves timeliness, flexibility, maintainability, and reduces transcription errors.
- **Graphics**: Many of the figures in this report were produced using Hadley Wickham's ggplot2 [29] package running under R, the capabilities of which extend beyond those evident in this document.
- **Contact**: Correspondence regarding this report should be directed to TATTOO at NIST dot GOV.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Background

Tattoos have been used for many years to assist law enforcement in the identification of criminals and victims and for investigative research purposes. Historically, law enforcement agencies have followed the ANSI/NIST [1] standard to collect and assign keyword labels to tattoos. This keyword labeling approach comes with drawbacks: the finiteness of standard class labels to describe the increasing variety of new tattoo designs, the need for multiple keywords to sufficiently describe some tattoos, difficulty assigning any label to abstract tattoos, and subjectivity in human annotation as the same tattoo can be labeled differently by different examiners. As such, the shortcomings of keyword-based tattoo image retrieval have driven the need for automated image-based tattoo recognition capabilities.

1.2 Scope

The Tattoo Recognition Technology - Challenge (Tatt-C) was designed to challenge the commercial and academic community in advancing research and development into automated image-based tattoo matching technology. The activity provided a preliminary assessment of the capability of image-based tattoo recognition algorithms to perform detection and retrieval of tattoos, with the goals to determine which methods are most effective and whether any are viable for the following five operational use cases:

- Tattoo Identification: matching different instances of the same tattoo image from the same subject over time
- Region of Interest: matching a subregion of interest that is contained in a larger tattoo image
- Mixed Media: matching visually similar or related tattoos using different types of images (i.e., sketches, scanned print, computer graphics, and graffiti)
- Tattoo Similarity: matching visually similar or related tattoos from different subjects
- Tattoo Detection: detecting whether an image contains a tattoo or not

1.3 Application Scenarios

With the increase in prevalence of tattoos in the the United States [2], automated image-based tattoo recognition has potential to support a number of applications. In events where primary biometric data such as face or fingerprints are not available, tattoos visible on the body have been used to help narrow the identity of both suspects and victims. In some cases, a tattoo can contain more discriminative characteristics and features for identification of victims of mass disaster, than traditional soft biometrics such as age, gender, race, height, and weight [11]. Tattoos can also support border control operations and investigations for identifying and rescuing victims of child exploitation and missing children [8]. The processes used by law enforcement personnel today to search tattoos is manual, time-consuming, and in many cases, requires ”tribal” knowledge of what has been encountered before. Automated image-based tattoo recognition has potential to provide more effective and consistent mechanisms for searching tattoos.

2 Methodology

Tatt-C was conducted as an ”open-book” test, where participants were provided with the dataset, ran their algorithm(s) on the data with their own hardware, and provided their algorithm output to NIST. There have been a number of datasets with an ”open-book” test protocol similar to Tatt-C: The NIST Multiple Biometric Grand Challenge (MBGC) [15] for face and iris recognition, Labeled Faces in the Wild (LFW) [17] for unconstrained face recognition, and NIST i-vector Machine Learning Challenge [14] for machine learning for use in speaker recognition. These datasets along with Tatt-C help promote development, advance the state-of-the-art, and benchmark progress in computer vision applied to different areas of biometrics.

The format of results submissions to NIST was described in the Tatt-C Dataset, Concept, and Evaluation Plan document [23]. The Tatt-C program was free and open to participation worldwide. The participation window opened on September 23, 2014 and submission to the final phase closed on May 4, 2015.

Protocol: Within each use case, the data was partitioned into five mutually exclusive subsets and participants produced performance results using a 5-fold cross-validation scheme [22]. For each search probe, there are one or more matching images in the gallery, i.e., closed set test.

External Training Data: The use of data outside of the Tatt-C dataset for training purposes was allowed, and participants were asked to disclose the use of external training data upon results submission to Tatt-C.

2.1 Algorithms

Tatt-C had two phases where participants could submit results to NIST. This report documents the results of all submissions across both phases of the Tatt-C activity.

Table 1 lists the Tatt-C participants who submitted algorithm outputs for one or more use cases specified in the Tatt-C Dataset, Concept, and Evaluation Plan [23], and Table 2 summarizes the use cases that each participant submitted to.

Participant	Short Name	Submissions		External Training Data	
		Phase 1	Phase 2	Phase 1	Phase 2
French Alternative Energies and Atomic Energy Commission	CEA	✓		yes	
Compass Technical Consulting	Compass		✓		no
Fraunhofer Institute of Optronics, System Technologies and Image Exploitation	FraunhoferIOSB	✓	✓	yes	no
MITRE Corporation	MITRE	✓	✓	no	no
Morpho/MorphoTrak	MorphoTrak	✓	✓	yes	yes
Purdue University	Purdue	✓	✓	no	no

Table 1: Summary of Tatt-C participation

Participant	Tattoo Identification	Region of Interest	Mixed Media	Tattoo Similarity	Tattoo Detection
French Alternative Energies and Atomic Energy Commission	✓	✓			✓
Compass Technical Consulting	✓	✓	✓	✓	✓
Fraunhofer Institute of Optronics, System Technologies and Image Exploitation	✓	✓			
MITRE Corporation	✓	✓	✓	✓	✓
Morpho/MorphoTrak	✓	✓			✓
Purdue University	✓	✓			

Table 2: Summary of Tatt-C submissions by use case

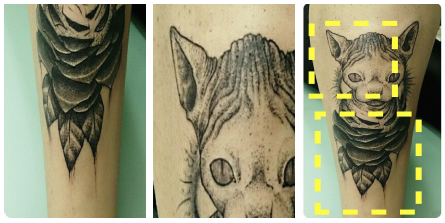
2.2 Dataset

The Tatt-C dataset contains a total of 16 716 tattoo images collected operationally by law enforcement. The dataset is partitioned into five use cases derived from operational scenarios with some partitions containing both original and cropped versions of the images, along with a set of background images used for padding the enrollment gallery for some of the experiments. Table 3 and Figure 1 provides a summary of the dataset and examples of images representative of each Tatt-C use case, respectively.

Use Case	Tattoo Identification	Region of Interest	Mixed Media	Tattoo Similarity	Tattoo Detection
Description	Match different instances of the same tattoo from the same subject over time	Match small region of interest contained in a larger tattoo	Match visually similar or related tattoos across different mediums	Match visually similar or related tattoos from different subjects	Detect whether an image contains a tattoo
Utility Example	Person identification	Person identification	Intelligence gathering	Gang Affiliation [6]	Database construction and maintenance
Task	One-to-many search	One-to-many search	One-to-many search	One-to-many search	Classification
Types of images	Tattoos	Tattoos	Tattoos, sketches, graffiti, computer graphics	Tattoos	Tattoos, faces
# probes	157	297	181	851	1 349 tattoos, 1 000 non-tattoos
# mates	215	157	272	1 361	
Average gallery sizes reported (per fold)	43 and 4 375	31 and 4 363	55	272	
# background images	4 332 (Used in larger gallery experiments)				

Table 3: Summary of Tatt-C dataset. Gallery sizes reported were calculated as an average across all 5 test folds.

(a) Tattoo Identification



(b) Region of Interest

(c) Mixed Media

(d) Tattoo Similarity

(e) Tattoo Detection

Figure 1: Examples representative of images for each Tatt-C use case [5]

2.3 Metrics

The following performance measures are reported in the assessment of tattoo detection and matching accuracy:

2.3.1 Detection Accuracy

Overall Accuracy
Tattoo accuracy is defined as the number of correctly classified tattoo images, TT, divided by the total number of tattoo images, N_{tattoo}. i.e.,

$$\text{Tattoo accuracy} = \frac{TT}{N_{tattoo}} \tag{1}$$

Non-tattoo accuracy is defined as the number of correctly classified non-tattoo images, NT, divided by the total number of non-tattoo images, $N_{non-tattoo}$. i.e.,

$$\text{Non-tattoo accuracy} = \frac{NT}{N_{non-tattoo}} \tag{2}$$

Overall accuracy is defined as the sum of correctly classified tattoo and non-tattoo images divided by the total number of images. i.e.,

$$\text{Overall accuracy} = \frac{TT + NT}{N_{tattoo} + N_{non-tattoo}} \tag{3}$$

Detection Error Trade-off
Participants were required to produce confidence values on $[0, 1]$ for each image describing their algorithm's certainty about whether the image contains a tattoo. Higher values indicate greater confidence that the image contains a tattoo. A reasonable approach to the detection problem is to classify an image as either a tattoo or non-tattoo by thresholding on its confidence value. Accuracy can also be measured using detection error trade-off (DET) plots [21].

For the detection problem, DET plots show the trade-off between the False Negative Rate (FNR) and the False Positive Rate (FPR) as a decision threshold is adjusted. A false negative occurs when a tattoo is misclassified as a non-tattoo and a false positive occurs when a non-tattoo is misclassified as a tattoo. Let x_i represent the confidence score for the ith tattoo image ($i = 1, ..., X$), and y_j the confidence score for the jth non-tattoo image ($j = 1, ..., Y$). Then the error rates at any

particular decision threshold, τ, are

$$\text{FNR}(\tau) = \frac{1}{X}\sum_{i=1}^{X} H(x_i < \tau), \tag{4}$$

$$\text{FPR}(\tau) = \frac{1}{Y}\sum_{j=1}^{Y} H(y_j \geq \tau). \tag{5}$$

where $H(n)$ is the heaveside step function [12]:

$$H(n) = \begin{cases} 0, & n \geq 0, \\ 1, & n < 0. \end{cases} \tag{6}$$

A high decision threshold should correspond to higher false negative rates but lower false positive rates.

2.3.2 Match Accuracy

In contrast to most biometric modalities, tattoo recognition will never be a "lights-out" operation, meaning there will always be human examination of the candidate list. Given this scenario, rank-based metrics are relevant and can be leveraged to support workload assessment.

Cumulative Match Characteristic
The Cumulative Match Characteristic (CMC) is used to show core accuracy, which is the fraction of searches that return the relevant images as a function of the candidate list length. The longer the candidate list, the greater the probability that relevant images are on the list. For searches that have multiple relevant matches in the gallery, the cumulative accuracy or hit rate at any particular rank is calculated with the best-ranked match and represents a "best-case" scenario.

Precision and Recall
In the case where a search may return multiple relevant tattoos, precision recall plots [27] are a standard way of measuring accuracy. For the tattoo test cases, precision recall plots show the relationship between two performance metrics as the length of the candidate list is adjusted. The first metric relates to *precision*, which, for a given search, is the fraction of candidates on the list that are classified as relevant to the searched tattoo. The second metric, *recall*, is the fraction of relevant candidates in the gallery that are on the candidate list.

$$\text{Precision} = \frac{\text{Num relevant documents on candidate list}}{\text{Length of candidate list}} \tag{7}$$

$$\text{Recall} = \frac{\text{Num relevant documents on candidate list}}{\text{Num relevant documents in gallery}} \tag{8}$$

Both metrics are functions of the candidate list length. Precision is typically interpolated at arbitrary recalls ($r \in [0, 1]$) as the maximum precision attainable when the recall is greater-than-or-equal-to the targeted recall:

$$\text{Precision}_{\text{interp}}(r) = \max_{\acute{r} \geq r} \text{Precision}(\acute{r}) \tag{9}$$

The commonly used 11-point Interpolated Average Precision [7] is created by averaging interpolated precisions across all searches at 11 equally spaced recalls. These curves tend to show a trade-off, where increasing the length of the candidate list increases the recall but decreases average precision.

Mean Average Precision (MAP) is a single-value metric related to precision and recall that describes the discriminating power of an algorithm. It is computationally equivalent to taking the mean of the average area under the precision recall curve across all searches. Generally speaking, the higher the MAP value, the more accurate the algorithm. Average Precision is defined as the average of the (un-interpolated) precision value obtained after each relevant image is retrieved.

The Mean Average Precision across the total number of probes is computed by taking the mean of the average precisions for each probe in the run.

Miss Rate vs. Average Candidate List Length
In addition to CMC and Precision-Recall plots, matching accuracy for the *Tattoo Identification* and *Region of Interest* use cases is presented in the form of Miss Rate vs. Average Candidate List Length plots. Unlike the other plots presented in this report, this type of plot relies on threshold-based, rather than rank-based, metrics. Instead of fixing the length of the candidate list, it allows the length of the candidate list to vary from one search to the next depending on how many candidates return a score greater-than-or-equal-to a decision threshold.

For the identification problem, this type of plot shows the trade-off between two performance metrics: 1) the miss rate, and 2) the average length of the candidate list, as a decision threshold is adjusted. The average candidate list length could be considered a measure of human workload (if we assume each candidate must be manually inspected by an examiner).

Formally, let x_i ($i = 1, ..., X$) be the ith mated score and y_j ($j = 1, ..., Y$) the jth non-mated score, irrespective of which search each comparison score is associated with. Furthermore, let S represent the number of searches performed to populate x_i and y_j. Then

$$\text{Miss Rate}(\tau) = \frac{1}{X} \sum_{i=1}^{X} H(x_i - \tau) \tag{10}$$

$$\text{Average Candidate List Length}(\tau) = \frac{1}{S}\Big(\sum_{i=1}^{X} H(x_i - \tau) + \sum_{j=1}^{Y} H(y_i - \tau)\Big) \tag{11}$$

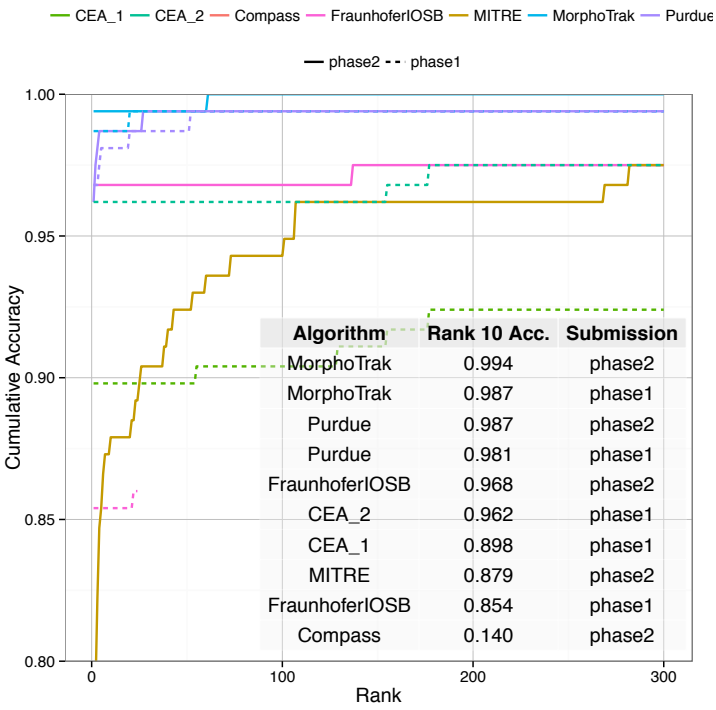
where τ is the decision threshold. When the decision threshold is relaxed (i.e., lowered in value) more candidates will be returned for some searches, increasing the average candidate list length but decreasing the miss rate. Note that the length of the candidate lists are computed by considering all mates and non-mates that return a similarity score at or above the decision threshold. All searches are mated (i.e. a mate is always present in the gallery). In the few cases where more than one mate was enrolled in the gallery, only the highest ranking mate was considered.

3 Results

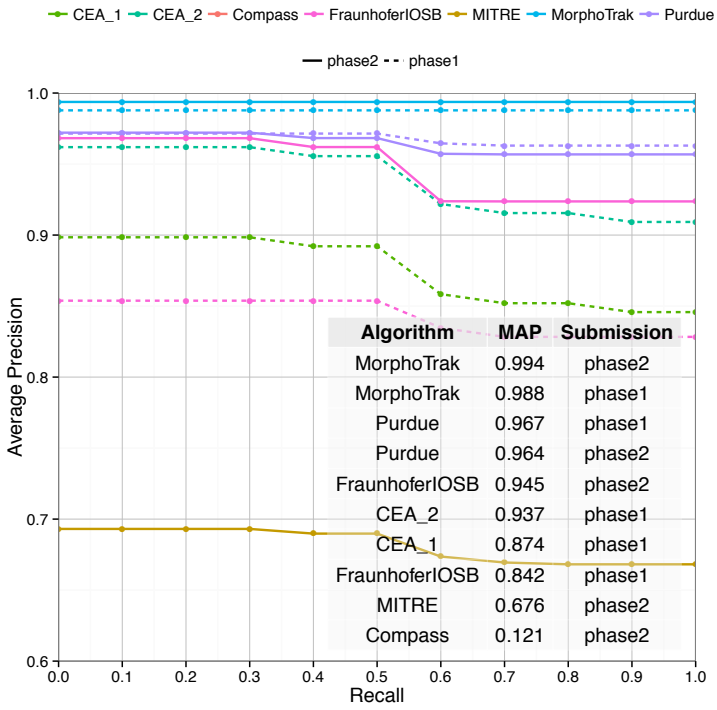
3.1 Tattoo Identification

This section details the performance of algorithms tasked with matching different instances of the same tattoo from the same subject in galleries of different sizes. This use case has application in investigation supporting identification of an individual, for example, in the case of a crime where the suspect is wearing a mask and gloves, video surveillance may be able to record a tattoo exposed on the neck or the arm from the suspect. The test data for this use case is composed of images of the same tattoo from the same subject collected during different encounters. For each probe image, there could be one or more correctly matching tattoo image(s) in the database.

3.1.1 Accuracy



(a) CMC plot. Accuracy was calculated with the best-ranked match.



(b) Interpolated Average Precision Recall plot.

Figure 2: Tattoo Identification use case results. Number of probes: 157, Average gallery size: 4375. Accuracy was averaged over 5 folds. Note: Some algorithm results may not be visible on the plots given range cut-offs.

Algorithm	Submission	Rank 1 Accuracy	Rank 10 Accuracy	Rank 100 Accuracy	Rank 300 Accuracy	MAP
CEA_1	phase1	0.898	0.898	0.904	0.924	0.874
CEA_2	phase1	0.962	0.962	0.962	0.975	0.937
Compass	phase2	0.089	0.140	0.586	0.650	0.121
FraunhoferIOSB	phase1	0.854	0.854	NA	NA	0.842
FraunhoferIOSB	phase2	0.968	0.968	0.968	0.975	0.954
MITRE	phase2	0.529	0.879	0.943	0.975	0.676
MorphoTrak	phase1	0.987	0.987	0.994	0.994	0.988
MorphoTrak	phase2	0.994	0.994	1.000	1.000	0.994
Purdue	phase1	0.968	0.981	0.994	0.994	0.967
Purdue	phase2	0.962	0.987	0.994	0.994	0.964

Table 4: CMC and MAP statistics for Tattoo Identification use case. Number of probes: 157, Average gallery size: 4375.

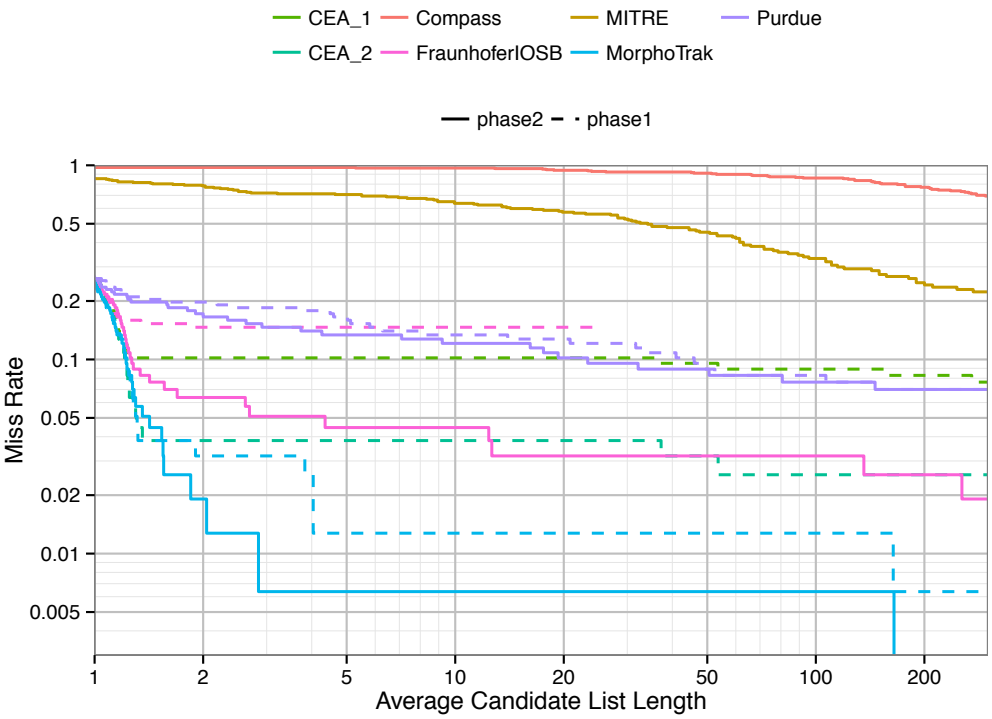


Figure 3: Miss rate vs. Average candidate list length based on score threshold for the Tattoo Identification use case. Number of searches: 157, average gallery size: 4 375. This curve presents a threshold-based trade-off between miss rate and the average number of images returned on the candidate list at or above a certain score threshold. This type of analysis can support workload assessment in cases where the amount of human labor available is constrained to reviewing an average number of candidates per search. The graph evaluates the average miss rate for any given mated search where the threshold is varied to yield an average number of images expected to show up on a candidate list (at or above threshold) for examination.

Results and notable observations:

- Six out of ten algorithms achieve a rank 1 hit rate above 90% and rank 10 hit rate above 96%, with the top performing algorithm (MorphoTrak) achieving both rank 10 hit rate and MAP of 99.4%.
- For many of the algorithms, no substantial accuracy improvements are observed past rank 100, which may be attributed to the law of diminishing returns for this particular use case.
- Per Figure 3, at a threshold set to yield an average of 10 candidates to be returned for any given mated search, the top performing algorithm (MorphoTrak) achieves an expected miss rate of .63%.

3.1.2 Effect of Gallery Size

As more tattoo data is collected in operations, scalability related to match performance against larger database sizes becomes important. Figure 4 presents the rank 10 hit rate across two different gallery sizes.

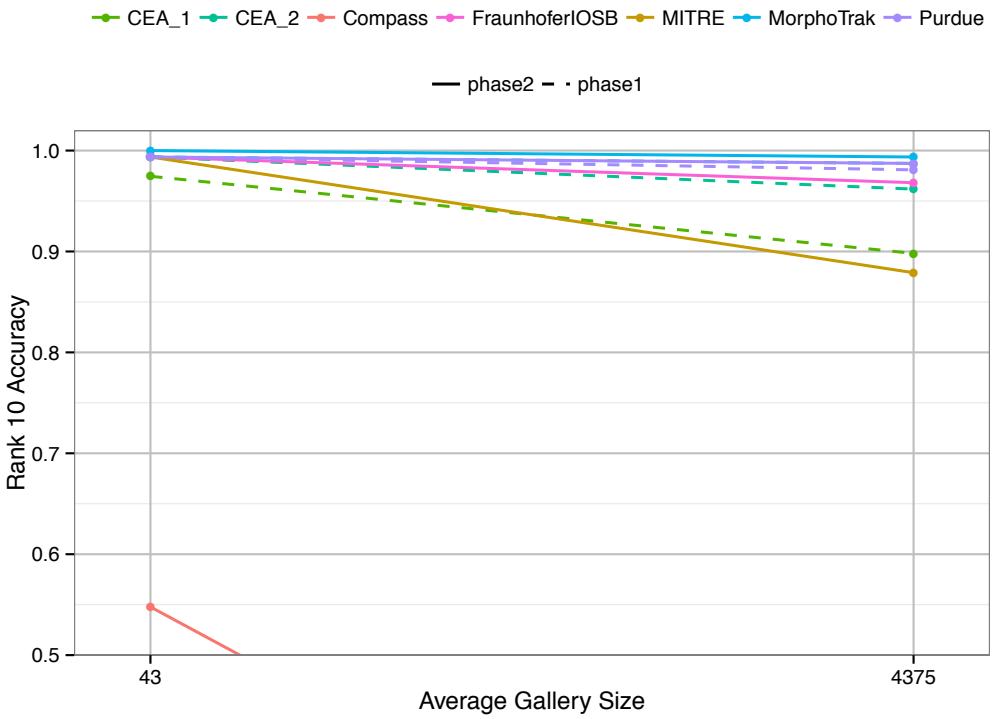


Figure 4: Plot comparing rank 10 hit rate across two different gallery sizes for Tattoo Identification. Number of probes: 157.

Results and notable observations:

- With an increase in gallery size, a decrease in performance is observed across all algorithms, demonstrating that gallery size has an impact on search accuracy. The decrease in rank 10 hit rate between a small gallery (43) versus larger gallery (4 375) ranged from .6% to 41% depending on the algorithm.
- Further studies with much larger gallery sizes (>100K) to reflect operations is warranted.

3.1.3 Failure Analysis

Image characteristics associated with match failure included:

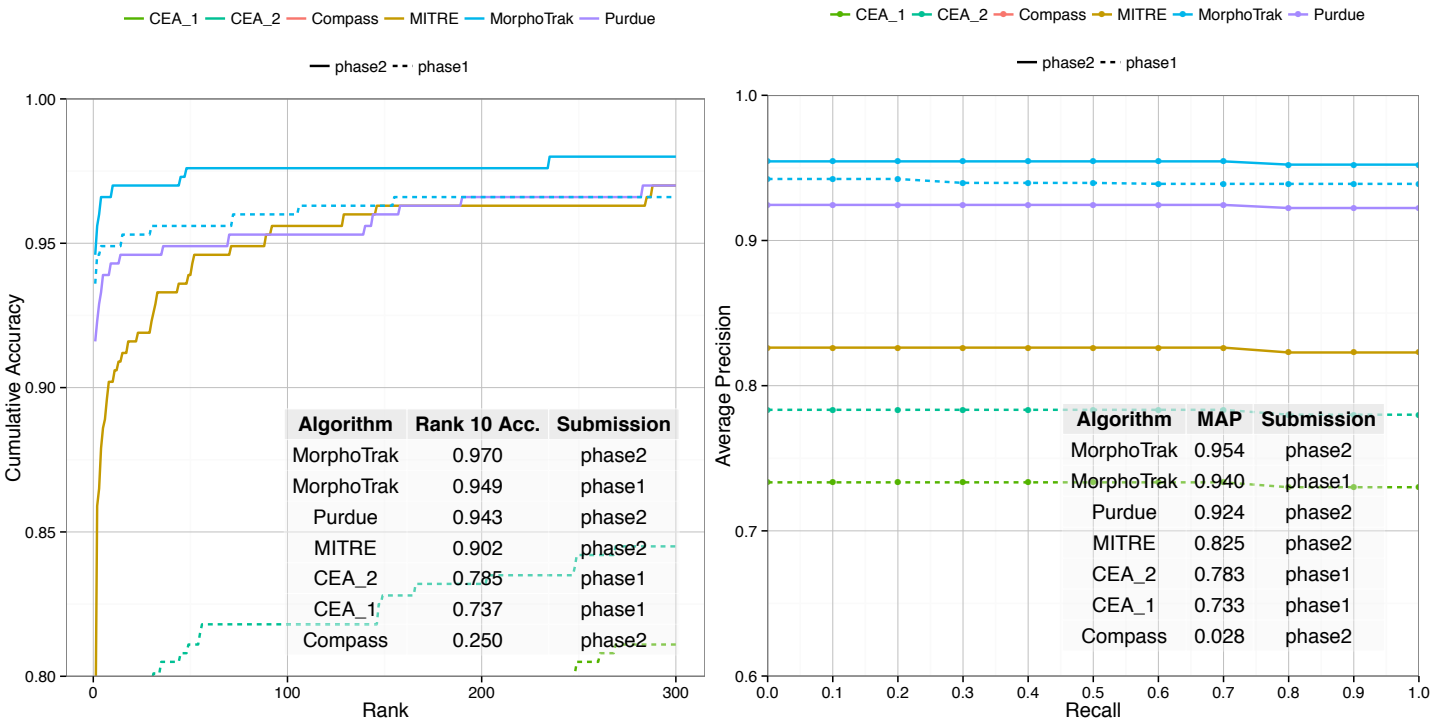
- Major additions made to the original tattoo image over time, altering the overall content of the tattoo; hair in this particular image also causes distractions to the primary tattoo content.
- Very simple and small tattoos which may lack distinctive features for matching.
- Major orientation differences between the probe and gallery tattoo image, along with patterned clothing in the background which causes distractions during feature extraction.
- Low contrast and blurry tattoos.
- Images that contain background clutter (image zoomed in for emphasis) that can cause false matching on the background instead of the actual tattoo content.

3.2 Region of Interest

This section details the performance of algorithms when given a subregion from an image of a tattoo, the ability to match a larger tattoo image (collected at a different time) from which the subregion is contained in galleries of different sizes. This use case has application in investigation supporting identification of an individual, for example, in the case where a

suspect's face is occluded but video surveillance contains a portion of a tattoo that is exposed on the arm. The test data for this use case is derived from images from the Tattoo Identification use case and is composed of multiple tattoo sections (probes) that belong to a larger tattoo canvas (gallery image) from the same subject collected during different encounters. In the experiment with the small gallery, there exists one correctly matching tattoo image in the database for each probe. In the larger gallery experiment, there were a small number of probes which had more than one or more matching gallery image.

3.2.1 Accuracy



(a) CMC plot. Accuracy was calculated with the best-ranked match.

(b) Interpolated Average Precision Recall plot.

Figure 5: Region of Interest use case results. Number of probes: 297, Average gallery size: 4 363. Accuracy was averaged over 5 folds. Note: Some algorithm results may not be visible on the plots given range cut-offs.

Algorithm	Submission	Rank 1 Accuracy	Rank 10 Accuracy	Rank 100 Accuracy	Rank 300 Accuracy	MAP
CEA_1	phase1	0.731	0.737	0.771	0.811	0.733
CEA_2	phase1	0.781	0.785	0.818	0.845	0.783
Compass	phase2	0.175	0.250	0.325	0.475	0.028
MITRE	phase2	0.771	0.902	0.956	0.970	0.825
MorphoTrak	phase1	0.936	0.949	0.960	0.966	0.940
MorphoTrak	phase2	0.946	0.970	0.976	0.980	0.954
Purdue	phase2	0.916	0.943	0.953	0.970	0.924

Table 5: CMC and MAP statistics for Region of Interest use case. Number of probes: 297, Average gallery size: 4 363.

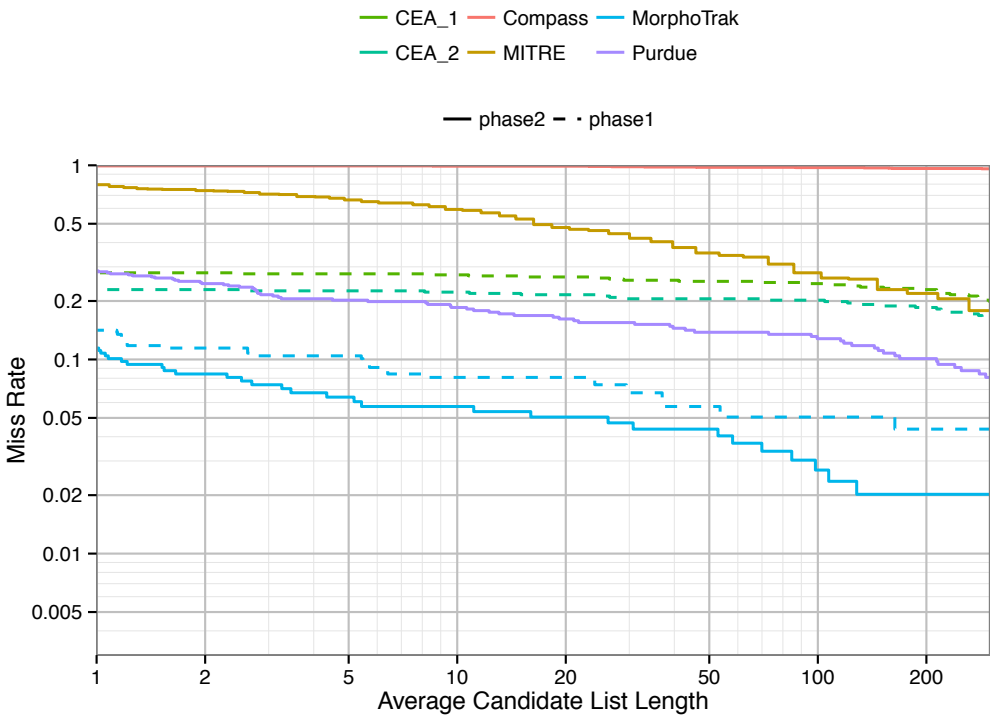


Figure 6: Miss rate vs. Average candidate list length based on score threshold for the Region of Interest use case. Number of searches: 297, average gallery size: 4 363. This curve presents a threshold-based trade-off between miss rate and the average number of images returned on the candidate list at or above a certain score threshold. This type of analysis can support workload assessment in cases where the amount of human labor available is constrained to reviewing an average number of candidates per search. The graph evaluates the average miss rate for any given mated search where the threshold is varied to yield an average number of images expected to show up on a candidate list (at or above threshold) for examination.

Results and notable observations:

- Three out of seven algorithms achieve a CMC rank 1 hit rate above 90%, and one out of seven achieve rank 10 hit rate above 96%, with the top performing algorithm (MorphoTrak) achieving rank 10 hit rate of 97% and MAP of 95.4%.
- Per Figure 6, at a threshold set to yield an average of 10 candidates returned for any given mated search, the top performing algorithm (MorphoTrak) achieves an expected miss rate of 5.7%. For the same algorithm, if the threshold is relaxed to yield an average candidate list length of 100, the miss rate is decreased by approximately a factor of two.
- Accuracy for this use case is lower across all algorithms than that observed from the Tattoo Identification results, which could be due to less information or features available in the probes (tattoo sections) used for matching.

3.2.2 Effect of Gallery Size

As more tattoo data is collected in operations, scalability related to match performance against larger database sizes becomes important. Figure 7 presents the rank 10 hit rate across two different gallery sizes.

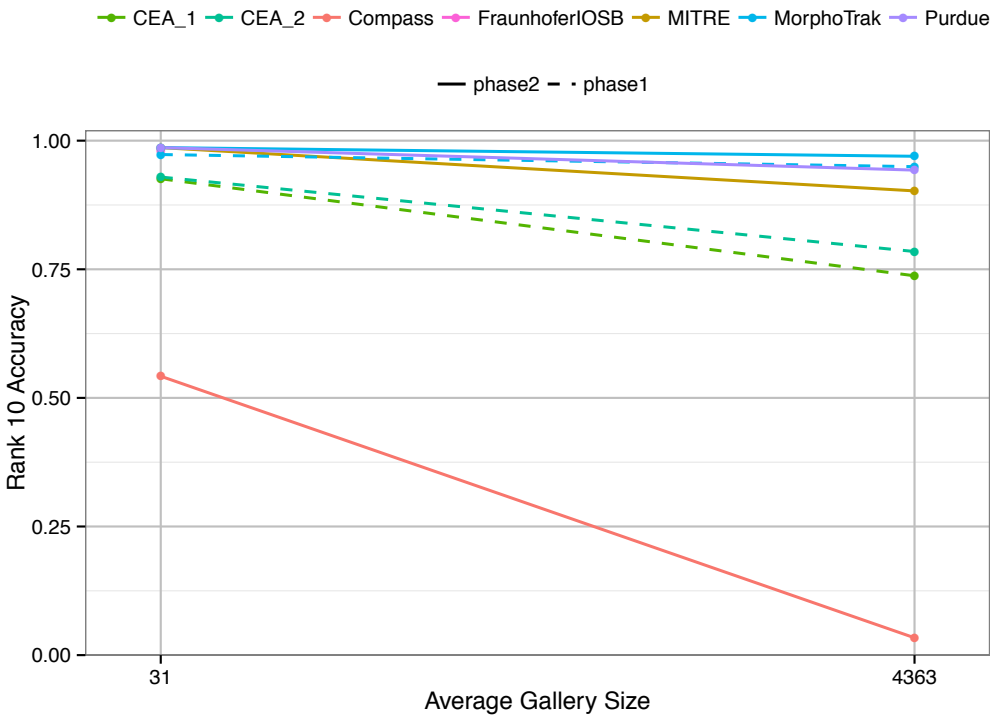


Figure 7: Plot comparing rank 10 hit rate across two different gallery sizes for Region of Interest. Number of probes: 297.

Results and notable observations:

- With an increase in gallery size, a decrease in performance is observed across all algorithms, demonstrating that gallery size has an impact on search accuracy. The decrease in rank 10 hit rate between a small gallery (31) versus larger gallery (4 363) ranged from 1.7% to 52% depending on the algorithm.
- Further studies with much larger gallery sizes (>100K) to reflect operations is warranted.

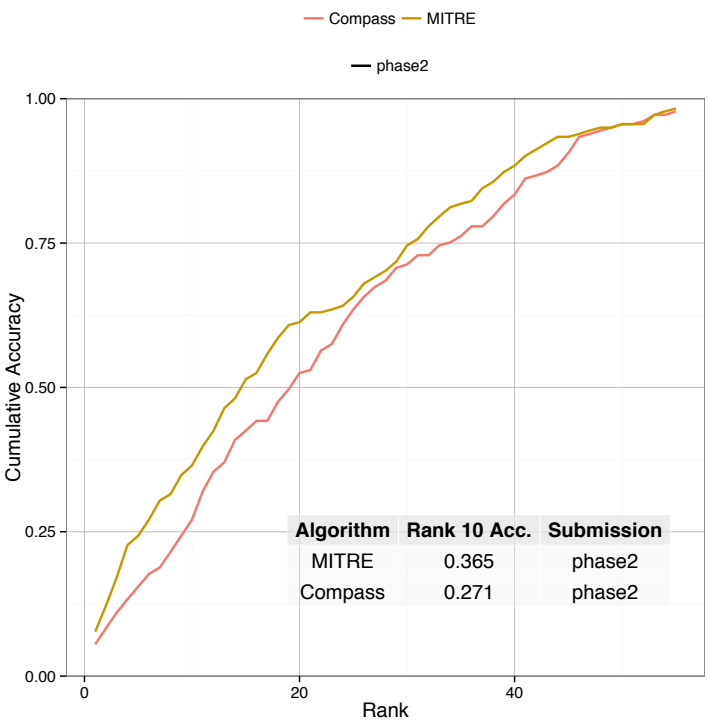
3.2.3 Failure Analysis

Image characteristics associated with match failure included:

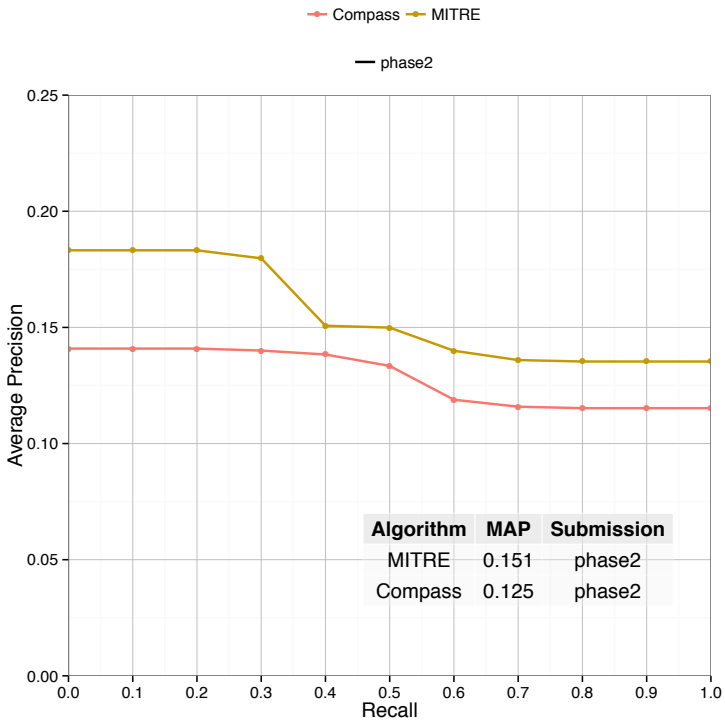
- Large distortion between the probe and gallery images.
- Major viewpoint and perspective differences between the probe and gallery images.
- Very small probe images relative to the larger canvas, along with low contrast and body hair distractions.

3.3 Mixed Media

This section details the performance of algorithms on matching non-tattoo types of images, (e.g., sketches, scanned print, computer graphics, and graffiti) to visually similar or related tattoo images. This use case has application in investigative intelligence gathering. In cases where the image of the tattoo is not captured on a camera, but an individual witnessed someone with a tattoo involved in criminal activity, a description of the tattoo can be provided to a forensic artist, and the sketch can be searched against a database for potential matches. The test data for this use case is composed of groups of visually similar mixed media images and tattoos. For any given mixed media probe image, there are one or more correctly matching tattoo image(s) in the database.



(a) CMC plot. Accuracy was calculated with the best-ranked match.



(b) Interpolated Average Precision Recall plot.

Figure 8: Mixed Media use case results. Number of probes: 181, Average gallery size: 55.

Algorithm	Submission	Rank 1 Accuracy	Rank 10 Accuracy	Rank 20 Accuracy	Rank 30 Accuracy	MAP
Compass	phase2	0.055	0.271	0.525	0.713	0.125
MITRE	phase2	0.077	0.365	0.613	0.746	0.151

Table 6: CMC and MAP statistics for Mixed Media use case. Number of probes: 181, Average gallery size: 55.

Results and notable observations:

- The top performing algorithm (MITRE) achieves a CMC rank 10 accuracy of 36.5% and MAP of 15.1%.

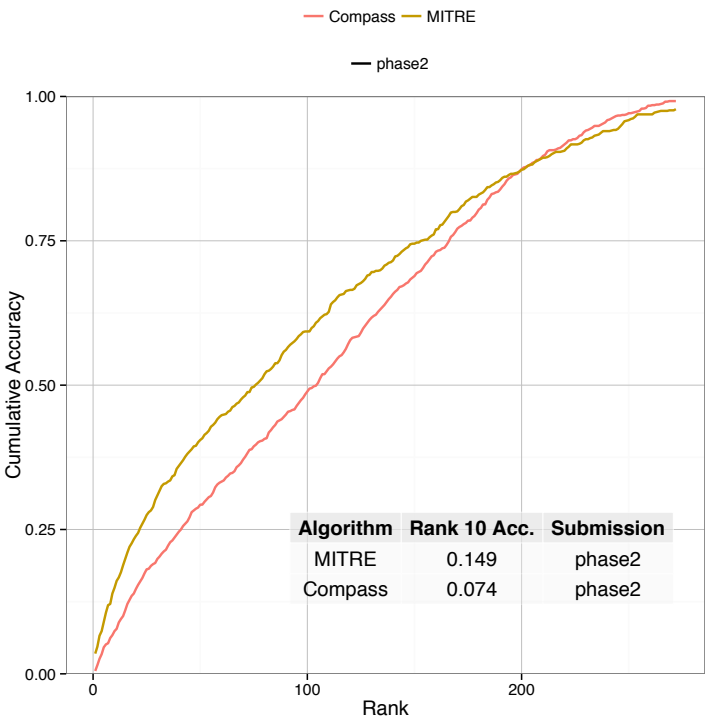
3.3.1 Failure Analysis

Image characteristics associated with match failure included:

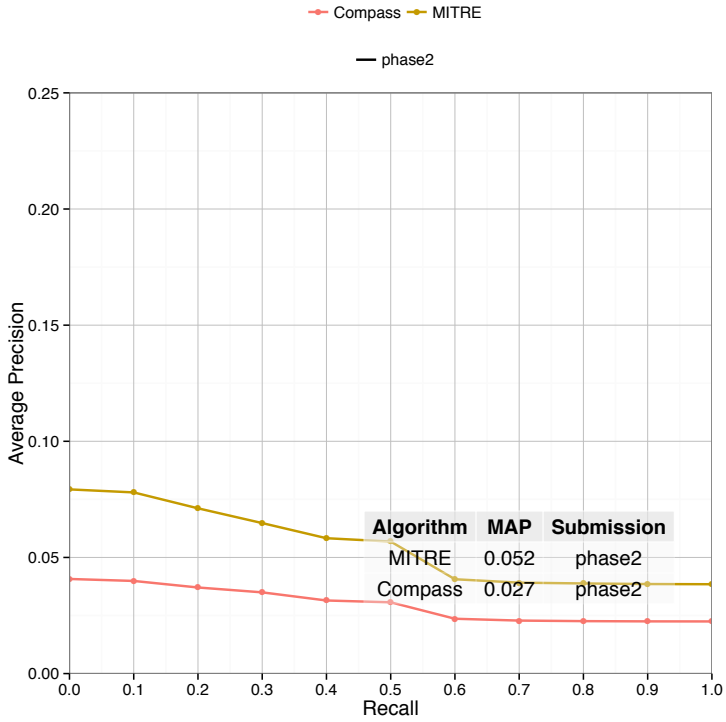
- The existence of clothing and other body parts, and in this case, also the small size of the tattoo relative to the entire image.
- Embellishments such as text and shadows around the primary tattoo content.
- Major variations in the composition of the tattoo. While the dominant content between the probe and gallery images are consistent, there are differences in the overall composition of the images. While such differences are easy to overcome for humans, it remains a challenge to computer vision algorithms.

3.4 Tattoo Similarity

This section details the performance of algorithms tasked with matching a tattoo image to visually similar or related tattoo images. This use case has application in gang affiliation [6], for example, in cases of identifying individuals in a criminal gang that could lead to other individuals with very similar tattoos that are very likely in the same gang. The test data for this use case is composed of groups of visually similar tattoos from different subjects collected on different occasions. The images for this experiment were cropped around the primary tattoo content. For any given probe image, there are one or more correctly matching gallery image(s) in the database.



(a) CMC plot. Accuracy was calculated with the best-ranked match.



(b) Interpolated Average Precision Recall plot.

Figure 9: Tattoo Similarity use case results. Number of probes: 851, Average gallery size: 272.

Algorithm	Submission	Rank 1 Accuracy	Rank 10 Accuracy	Rank 20 Accuracy	Rank 30 Accuracy	MAP
Compass	phase2	0.005	0.074	0.147	0.199	0.027
MITRE	phase2	0.035	0.149	0.239	0.309	0.052

Table 7: CMC and MAP statistics for Tattoo Similarity use case. Number of probes: 851, Average gallery size: 272.

Results and notable observations:

- The top performing algorithm (MITRE) achieves a rank 10 accuracy of 14.9% and MAP of 5.2%.

3.4.1 Failure Analysis

Image characteristics associated with match failure included:

- The existence of clothing and other body parts is seen in many of the images, which can cause issues during tattoo segmentation. This is a common trend across the use cases.

- Embellishments such as text and shadows around the primary tattoo content.
- Tattoos that are semantically relevant, but lack commonality in visual composition and details. Human beings are able to interpret images at different levels, both in low-level features (color, shape, texture) and high-level semantics (abstract object, concept, event). However, computers are only able to interpret images based on low level-image features. This introduces an interpretation inconsistency between image descriptors and high-level semantics that is known as the semantic gap [28].

3.5 Tattoo Detection

This section details the performance of algorithms tasked with detecting whether an image contains a tattoo or not. This use case has application in database construction and maintenance when large amounts of unlabeled data is comingled, making automatic extraction of different types of images a challenge. An example is the ANSI/NIST Type 10 record where facial mugshot images and scar, mark, tattoo (SMT) images are stored in the same record type, and with a percentage of the data mislabeled or not labeled at all, it presents a challenge to automated extraction of the data based on image content. The test data for this use case is composed of tattoo images and face images extracted from the Multiple Encounter Database 2 (MEDS- II) [4].

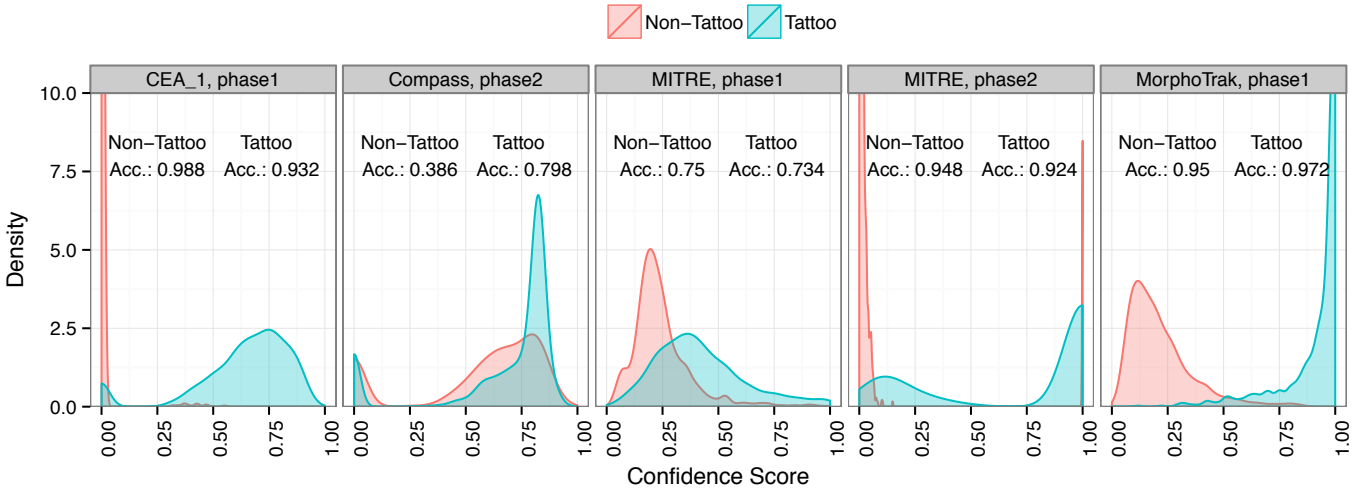
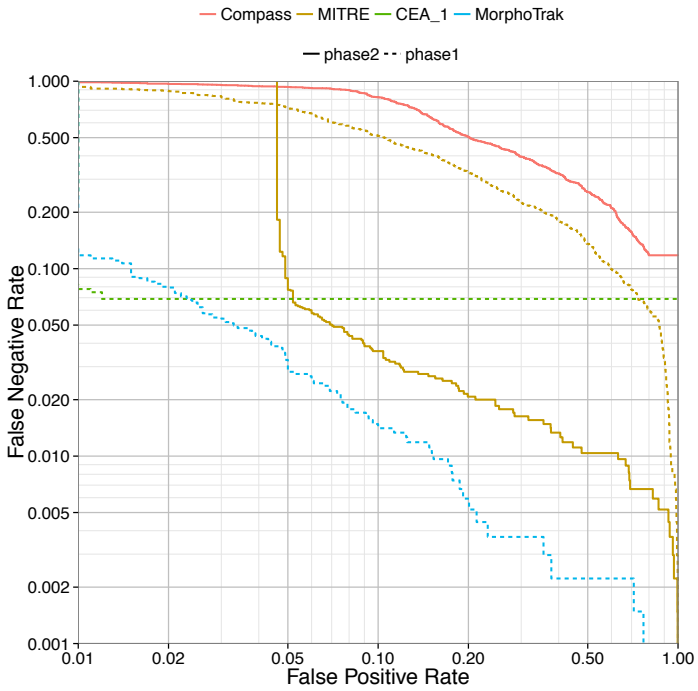


Figure 10: Distribution of tattoo detection confidence scores for the Tattoo Detection use case. Number of tattoos: 1 349, Number of non-tattoos: 1 000. Accuracy was averaged over 5 folds.



(a) DET plot based on the confidence scores.

Algorithm	Submission	Non–tattoo Accuracy	Tattoo Accuracy	Overall Accuracy
CEA_1	phase1	0.988	0.932	0.956
Compass	phase2	0.386	0.798	0.622
MITRE	phase1	0.750	0.734	0.741
MITRE	phase2	0.948	0.924	0.934
MorphoTrak	phase1	0.950	0.972	0.963

(b) Detection accuracy statistics.

Figure 11: Tattoo Detection (DET-1) results. Number of tattoos: 1 349, number of non-tattoos: 1 000. Accuracy was averaged over 5 folds.

Results and notable observations:

- Three out of five algorithms achieve overall tattoo detection accuracy of above 93%, with the top performing algorithm (MorphoTrak) achieving 96.3% in overall accuracy.
- Per the confidence score density plot in Figure 10, the more accurate algorithms show a larger separation between tattoo and non-tattoo scores.
- Figure 11 presents a trade-off between false positive rate (classifying a non-tattoo as a tattoo) and false negative rate (classifying a tattoo as non-tattoo). In a scenario where a system wanted to minimize the number of tattoos rejected as being non-tattoos (i.e., false negative rate), one might reasonably set the false negative rate threshold to 1%, which would yield a false positive rate of 15% for the top performing algorithm (MorphoTrak), meaning 15% of images classified as tattoos are actually non-tattoos. Different thresholds can be set to support specific applications.
- Note the images in this dataset are biased given all non-tattoo images are faces, and some algorithms leveraged face detection algorithms as an approach to classify ”non-tattoos”. Further studies with 1) additional types of non-tattoo imagery and 2) a much larger number of test images to reflect operations is warranted.

3.5.1 Failure Analysis

Image characteristics associated with detection failure included:

- Tattoos of faces: some algorithms used face detection methods where if a face is detected, the image is classified as being non-tattoo. Tattoos of faces are commonly seen in operations.
- Faded tattoos and cases where there is low contrast between the tattoo and skin color.

4 Recommendations for Future Work

Based on the outcomes of the Tatt-C activity and the discussions and suggestions from the tattoo recognition developer and user community at the Tatt-C Workshop [8], we propose the following recommendations for future work.

4.1 Improve Tattoo Image Capture

Algorithm failure to find the correct match is often related to the consistency and quality of image capture. Notably, inconsistencies in image angle, orientation, size of tattoo relative to the entire image and poor collection characteristics such as low illumination, low contrast, blurry/out of focus, and existence of clothing and background clutter caused failures for tattoo detection and matching algorithms. As such, recommendations for improving the quality of tattoo images to support image-based tattoo recognition include:

- **Best Practices**: Develop best practice guidelines for the capture of a tattoo image, including photography guidelines and image quality definitions and standards for tattoos. Similar guidelines for the capture of face [1], iris [25], and fingerprints [24] have been developed and successfully employed by law enforcement.
- **Training Material**: Develop and distribute simple reference/training material such as a poster with examples of good quality, well-captured tattoos versus bad quality, poorly-captured tattoos, similar to those that have been developed and successfully utilized for other modalities [3].
- **Software**: During photo capture, use software that makes image quality assessments (e.g., illumination, contrast, focus, existence of distractors around tattoo image) and determines whether to accept or reject the image. Such a software concept is similar to NIST's Fingerprint Image Quality (NFIQ) [26] software, but applicable to tattoos.

Build Larger Tattoo Datasets for Researchers: Nineteen organizations requested and received the Tatt-C dataset (eight commercial, six academic, and five research entities). This is a good indicator of industry and research interest in this area. Following the closing of the Tatt-C submission deadline, NIST continues to receive participation interest and requests for the Tatt-C dataset. As an outcome from the Tatt-C workshop, the community identified a need to provide researchers with larger, ground-truthed tattoo datasets with larger gallery sizes (>100K) closer to operations, and more search images per use case, ideally covering the categories of "challenging images" identified in this report. Larger image sets will impose a level of software and computational robustness on algorithm developers and can support advancing the technology further.

Conduct Sequestered Evaluation: Tatt-C was conducted as an "open-book" test based on the honor system where participants were only required to submit their system output to NIST for uniform scoring and analysis. Such an evaluation protocol does not preclude gaming such as algorithm training on test data, candidate list manipulation, etc. For the Tatt-C use cases that resulted in high accuracy - Tattoo Identification, Region of Interest, and Tattoo Detection, we recommend running a "closed-book" evaluation of tattoo recognition algorithms, where participants send software to NIST to be tested with sequestered data. This will allow consistent assessment of run-time performance (enrollment/search time and memory usage) on a uniform hardware infrastructure, which is important as specific operational applications will have specific run-time requirements. Sequestered evaluations can leverage much larger, operationally-realistic gallery sizes of data that cannot be freely distributed.

Evaluate Tattoo Localization: The presence of clutter and occlusions such as clothing, background, furniture, hair, and other body parts in the operationally-collected data makes tattoo localization an important aspect of the recognition process. Localization and matching are generally two distinct tasks, and matching is critically dependent on correct localization. We recommend evaluation of tattoo localization as a separate use case in future evaluations.

Refine Definition of Tattoo Similarity Specific to Law Enforcement Applications: A common trend in the ground-truth for Tattoo Similarity included images that were semantically relevant, but lacked commonality in visual features, which

is likely a result related to the "semantic gap". Human beings are able to interpret images at different levels, both in low-level features (color, shape, texture) and high-level semantics (abstract object, concept, event). Computers are only able to interpret images based on low level-image features. This introduces an interpretation inconsistency between image descriptors and high-level semantics that is known as the semantic gap [28]. While some research has been done in bridging the semantic gap for image retrieval, it remains a challenging, unsolved problem [18]. A set of well-defined criteria for tattoo similarity or relevance specific to law enforcement applications needs to be developed for consistent ground-truthing of data that enables research and development. The feasibility of such an endeavor has yet to be determined. An initial step could be to focus on specific law enforcement application, which will likely have observable commonalities in the dominant tattoo content.

References

[1] ANSI/NIST-ITL 1-2011:Update 2013, NIST Special Publication 500-290, Data Format for the Interchange of Fingerprint, Facial and Other Biometric Information. http://www.nist.gov/itl/iad/ig/ansi_standard.cfm.

[2] Fox News Poll: Tattoos arent just for rebels anymore. http://www.foxnews.com/us/2014/03/14/fox-news-poll-tattoos-arent-just-for-rebels-anymore.

[3] Guide to Capturing Iris Images. http://biometrics.nist.gov/cs_links/iris/irexV/IREX_V_Poster_20140612.pdf.

[4] NIST Special Database 32 - Multiple Encounter Dataset 2 (MEDS-II), NISTIR 7807. http://www.nist.gov/itl/iad/ig/sd32.cfm.

[5] Tattoo images are licensed under CC BY 2.0. Face images are from the MEDS-II [4] Dataset.

[6] National Gang Report 2013, National Gang Intelligence Center, 2013, pages 7,12. https://www.fbi.gov/file-repository/stats-services-publications-national-gang-report-2013/view.

[7] NIST Special Publication: SP 500-272, The Fifteenth Text REtrieval Conference (TREC) Proceedings. Appendix - Common Evaluation Measures, 2015. http://trec.nist.gov/pubs/trec15/appendices/CE.MEASURES06.pdf.

[8] Tatt-C Workshop 2015 Online Proceedings, National Institute of Standards and Technology, June 2015. http://www.nist.gov/itl/iad/ig/tatt-c_workshop_proceedings.cfm.

[9] S. T. Acton and A. Rossi. Matching and retrieval of tattoo images: Active contour cbir and glocal image features. In *Proc. of the 2008 IEEE Southwest Symposium on Image Analysis and Interpretation*, pages 21–24, 2008.

[10] J. D. Allen, N. Zhao, J. Yuan, and X. Liu. Unsupervised tattoo segmentation combining bottom-up and top-down cues. In *Proc. of the SPIE 8063, Mobile Multimedia/Image Processing, Security, and Applications 2011*, 2011.

[11] J.-P. Beauthier, P. Lefevre, and E. D. Valck. Autopsy and Identification Techniques, The Tsunami Threat - Research and Technology. In *N.-A. Mörner, ed., InTech*, 2011.

[12] E. J. Berg. *Heaviside's operational calculus as applied to engineering and physics*. Electrical engineering texts. McGraw-Hill book company, inc., 1936.

[13] P. Duangphasuk and W. Kurutach. Tattoo skin detection and segmentation using image negative method. In *Proc. of the 13th International Symposium on Communications and Information Technologies*, pages 354–359, 2013.

[14] C. S. Greenberg et al. The NIST 2014 Speaker Recognition i-Vector Machine Learning Challenge. In *Odyssey 2014: The Speaker and Language Recognition Workshop*, pages 224–230, 2014.

[15] P. J. Phillips et al. Overview of the multiple biometrics grand challenge. In *Advances in Biometrics*, volume 5558 of *Lecture Notes in Computer Science*, pages 705–714. Springer Berlin Heidelberg, 2009.

[16] B. Heflin, W. J. Scheirer, and T. E. Boult. Detecting and classifying scars, marks, and tattoos found in the wild. In *BTAS '12*, pages 31–38, 2012.

[17] G. B. Huang, M. Ramesh, T. B., and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007.

[18] N. Idrissi, J. Martinez, and D. Aboutajdine. Bridging the Semantic Gap for Texture-based Image Retrieval and Navigation. *Journal of Multimedia*, 4(5):277–283, 2009.

[19] A. K. Jain, J.-E. Lee, and R. Jin. Tattoo-id: automatic tattoo image retrieval for suspect and victim identification. In *PCM '07*, 2007.

[20] D. Manger. Large-scale tattoo image retrieval. In *Proc. of the Conference on Computer and Robot Vision*, pages 454–459, 2012.

[21] A. Martin, G. Doddington, T. Kamm, M. Ordowski, and M. Przybocki. The DET curve in assessment of detection task performance. In *Proc. Eurospeech*, pages 1895–1898, 1997.

[22] F. Mosteller and J. W. Tukey. *Data analysis, including statistics*. Handbook of Social Psychology. Addison-Wesley, Reading, MA, 1968.

[23] M. Ngan, P. Grother, M. Garris, and E. Phillips. NIST Tattoo Recognition Technology - Challenge (Tatt-C) Dataset, Concept, and Evaluation Plan, April 2015. http://www.nist.gov/itl/iad/ig/tatt-c.cfm.

[24] Federal Bureau of Investigation. *The Science of Fingerprints: Classification and Uses*. 2006. http://www.gutenberg.org/files/19022/19022-h/19022-h.htm#CHAPTER_IX.

[25] G. W. Quinn, J. Matey, E. Tabassi, and P. Grother. Guidance for iris image collection. NIST Interagency Report 8013, National Institute of Standards and Technology, 2014. http://biometrics.nist.gov/cs_links/iris/irexV/IREX_V_Report.pdf.

[26] E. Tabassi, C. Wilson, and C. Watson. Fingerprint image quality. NIST Interagency Report 7151, National Institute of Standards and Technology, 2004. http://www.nist.gov/customcf/get_pdf.cfm?pub_id=905710.

[27] L. Torgo and R. Ribeiro. Precision and recall for regression. In *Proceedings of the 12th International Conference on Discovery Science*, DS '09, pages 332–346, Berlin, Heidelberg, 2009. Springer-Verlag.

[28] H. H. Wang, D. Mohamad, and N. A. Ismail. Approaches, Challenges and Future Direction of Image Retrieval. *Journal of Computing*, 2(6):193–199, June 2010.

[29] H. Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer, New York, 2009. ISBN 978-0-387-98141-3.