NISTIR 8232

Tattoo Recognition Technology - Evaluation (Tatt-E) Performance of Tattoo Identification Algorithms

Mei Ngan
Patrick Grother
Kayee Hanaoka
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8232

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

NISTIR 8232

Tattoo Recognition Technology - Evaluation (Tatt-E) Performance of Tattoo Identification Algorithms

Mei Ngan
Patrick Grother
Kayee Hanaoka
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.IR.8232

Last Updated: October 30, 2018

U.S. Department of Commerce
Wilbur L. Ross, Jr., Secretary

National Institute of Standards and Technology
Walter Copan, NIST Director and Under Secretary of Commerce for Standards and Technology

Executive Summary

Background

The Tattoo Recognition Technology - Evaluation (Tatt-E) was organized by the National Institute of Standards and Technology (NIST) in collaboration with law enforcement to assess and measure the capability of algorithms to perform automated image-based tattoo recognition. Tatt-E is the follow-on to the Tattoo Recognition Technology - Challenge (Tatt-C) 2015 activity [1], which engaged researchers with an open dataset of operationally collected tattoo images and challenge problems to advance image-based tattoo recognition research and development. The Tatt-E activity assesses comparative and absolute accuracy, along with run-time measures on larger, operationally realistic datasets than seen in Tatt-C.

Tatt-E Test Activity

The Tatt-E program was open to participation worldwide. The participation window opened on December 1, 2016, and the submission deadine to the final phase was September 29, 2017. There was no charge to participate. Tatt-E was performed as a large scale empirical evaluation of a total of 12 tattoo recognition algorithms, with participation from two providers - one commercial, MorphoTrak, and one university, the Chinese Academy of Sciences (CAS). The test leveraged large operational datasets comprising tattoo images from law enforcement databases, enabling evaluation with enrollment database sizes of up to 100 000. NIST employed a "lights-out", black-box testing methodology designed to model operational reality where software is shipped and used as-is, without algorithmic training. Core tattoo identification accuracy was baselined over tattoo images used as-is, then traded off against gallery size and search speed. The effects of cropping around the primary tattoo content, skintone, contrast, and tattoo-to-image ratio were assessed, and matching accuracy on sketch images and tattoos collected in the short-wave infrared (SWIR) spectrum are also reported. In addition, performance on algorithmic capability to do tattoo detection and tattoo localization as separate tasks are also documented.

Key Results

Key results for the use cases studied are summarized below. A legend mapping the algorithm letter code to the participating organization is provided in the footer of every page of this report and in Table 1.

- **Tattoo Identification:** For this use case, matching different photos of the same tattoo image from the same subject over time is investigated. When searching 11 921 ***uncropped*** probe tattoo images against a gallery of 100 000 tattoos enrolled as-is (i.e., uncropped), the top performing algorithm (B31I) achieved a rank 10 hit rate* of 72.1 % (miss rate of 27.9 %). *Section 4.1.1*

  When the same set of probe imagery was manually ***cropped*** around the primary tattoo content prior to algorithm processing (gallery images remained uncropped), notable accuracy gains were observed. The top performing algorithm (B11I) achieved a rank 10 hit rate of 84.8 % (miss rate of 15.2 %). This quantifies that an additional ˜13 % of searches can be matched when using cropped probes images over uncropped images, which constitutes nearly a 50 % decrease in miss rate. *Section 4.1.2*

  Factors that influenced accuracy included:

  - **Algorithms:** Tattoo recognition accuracy depends on the implementation of the core technology. While algorithm performance varied between the two participants, there were notable accuracy gains in algorithms submitted by both providers in successive phases of the test.
  - **Tattoo-to-Image Ratio:** The size of the tattoo relative to the entire image has an effect on tattoo retrieval rates. For all algorithms, over 60 % of tattoos that are captured as a very small percentage (less than or equal to 1 %) of the entire image were not matched within the top 300 images retrieved. Miss rates decrease as the tattoo-to-image ratio increases, which indicates that tattoos that occupy a larger percentage of the image have a higher chance of being matched. *Section 4.1.3*
  - **Skintone Brightness and Tattoo Contrast:** The impact of skintone brightness and tattoo contrast on matching accuracy depends on the algorithm. All algorithms appear to be sensitive to very low contrast. Some algorithms are not greatly impacted by skintone brightness while other algorithms have difficulty matching images with very low skintone brightness. *Section 4.1.4*

*For the definition of hit rate, true positive detection rate, and localization accuracy, see Section 3. Generally speaking, the higher the hit rate, true positive detection rate, and localization accuracy value, the more accurate the algorithm.

  - **Gallery Size:** Gallery size has an impact on accuracy. While decreases in accuracy is observed across all algorithms when the gallery size is increased, accuracy declines at fairly benign rates. For the more accurate algorithms, the decrease in hit rate ranges between 2 % to 3 % when the gallery is increased by a factor of 5. Such nonlinear trends demonstrate the viability of tattoo recognition on operationally-sized databases. *Section 4.1.5*

- **Sketches:** For this use case, matching sketches against tattoo images enrolled in a database is investigated. On a gallery of 100 000 enrolled tattoos, the best performing algorithm (A30I) achieves a rank 10 hit rate of 40 %. Increasing the number of candidates reviewed yields significant accuracy benefits, with a rank 300 hit rate reaching as high as 71 %. *Section 4.2*

- **Multispectral:** For this use case, matching tattoo images collected in the SWIR spectrum against visible tattoo images enrolled in a database is investigated. In SWIR, all algorithms showed the best match performance on search images collected between the 1100nm to 1300nm wavelengths. The best performing algorithm (B11I) achieved a rank 10 hit rate of 95 % on images collected in the 1100nm and 1200nm wavelengths. On the same set of probe tattoos, the best match performance remains on images collected in the visible spectrum, which is observed across all algorithms. *Section 4.3*

- **Tattoo Detection:** For this use case, detecting whether an image contains a tattoo or not is investigated. On a mixed dataset of 131 662 tattoo images and 125 253 non-tattoo images, the top performing algorithms (B21D, B30D, and B31D) achieve a true positive detection rate of 99.5 % when false positive detection rate is set to 1 %. *Section 4.4*

- **Tattoo Localization:** For this use case, the ability to locate and segment one or more tattoos contained in an image is investigated. On a set of 12 613 tattoos contained in 10 926 images, the top performing algorithms (A20D, A21D, A30D, and A31D) achieve spatial localization accuracy of 89.5 %. *Section 4.5*

- **Speed-accuracy Tradeoff:** There is an observable tradeoff between search speed and accuracy for the two participating organizations. The faster algorithms tend to be less accurate, and the more accurate algorithms tend to be slower. Rank 10 accuracy ranges from 10.3 % (A10I) to 72.1 % (B31I) while search speeds range from 2.0 seconds (A10I, A11I) to 255.4 seconds (B11I) to search a single tattoo image against a gallery size of 100 000. Both participants leveraged the common open-source deep learning framework, Caffe [2]. For timing, some algorithms ran on a single NVIDIA Tesla K40 GPU, and some submissions ran on a single Intel Xeon E5-2695 v3 CPU. Note that differences in hardware make direct timing comparisons difficult, but measuring absolute timing remains useful as different applications will have different speed requirements. *Section 4.1.6*

Caveats

False positive identification rates are not documented in this report. Tattoos cannot be used as a primary biometric as an arbitrary number of people can have nearly identical tattoos. Through analysis and trial runs on the test data, searching tattoos of subjects that are known not to exist in the gallery (non-mates) sometimes returns the same tattoo on different subjects. This is akin to using face recognition to match twins that exist in the same database. While in theory such near-identical tattoos could be considered false positives, algorithmic capability to retrieve the same tattoo from different subjects warrants consideration due to investigative utility in operations. Removal of such near-identical tattoo images from different people from the test dataset for all possible non-mated scenarios would require substantial human labor, which was not available for the scope of this evaluation.

Topics Not Covered

Based on the outcomes of this test and discussions/suggestions from the tattoo recognition developer and user community (including law enforcement, forensic examiners, and others), studies not covered in this report which warrant consideration by the research community include:

- Multiple image enrollment: assess the impact of multiple image enrollment and whether algorithms can take advantage of fusion of the information across images to enhance accuracy;

- Tattoo ageing: longitudinal study of matching tattoos over time and whether there is an impact on matchability as a tattoo ”ages”;

- Occlusion study: given tattoos are often occluded by clothing or other items when collected in uncontrolled settings, assess algorithm performance on tattoos that are occluded in different ways;
- Tattoo recognition ”in the wild”: study performance of tattoo recognition on images collected under unconstrained settings;
- Automated tattoo image quality assessment: establish criteria to measure and assess tattoo image quality with goals of an automated capability to accept or reject tattoo images during the capture process based on a quality measure;
- Tattoo similarity: matching tattoos based on visually similar content;
- Multispectral: matching tattoos collected in other parts of the infrared spectrum (e.g. near infrared);
- Algorithm fusion;
- Tattoo recognition in video.

Acknowledgements

The authors would like to thank the sponsor of this activity, the Federal Bureau of Investigation, for initiating and progressing this work. In addition, we would like to thank the Michigan State Police and Pinellas County Sheriff's Office, Florida for their contributions, and many thanks to Patricia Flanagan and Karen Marshall at NIST for their numerous hours of data review. The authors are also grateful to Nick Orlans at MITRE and Mike Garris at NIST for their thorough and constructive review of this document.

The authors are grateful to Wayne Salamon and Greg Fiumara at NIST for robust software infrastructure, particularly that which leverages Berkeley DB Btrees [3] for storage of images and templates, and Open MPI [4] for parallel execution of algorithms across our computers. Thanks also to Brian Cochran at NIST for providing highly available computers and network-attached storage.

Disclaimer

Specific hardware and software products identified in this report were used in order to perform the evaluations described in this document. In no case does identification of any commercial product, trade name, or vendor, imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products and equipment identified are necessarily the best available for the purpose.

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

Institutional Review Board

The National Institute of Standards and Technology Human Subjects Protection Office reviewed the protocol for this project and determined it is not human subjects research as defined in Department of Commerce Regulations, 15 CFR 27, also known as the Common Rule for the Protection of Human Subjects (45 CFR 46, Subpart A).

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 purposes. Historically, law enforcement agencies have followed the ANSI/NIST-ITL 1-2011 [5] standard to collect and assign keyword labels to tattoos. This keyword labeling approach comes with drawbacks, which include the limited number of class labels to describe the wide variety of new tattoo designs, the need for multiple keywords to sufficiently describe some tattoos, and subjectivity in human annotation as the same tattoo can be labeled differently by officers. As such, the shortcomings of keyword-based tattoo image retrieval have driven the development of other approaches to improve the ability to find and match tattoos.

In the last decade, a number of novel computer vision and machine learning approaches have pushed unprecedented progress in areas such as image-based object recognition and face recognition. This, along with the availability of curated tattoo datasets [6] [7] [8] for the purposes of research and testing, presents an opportunity to investigate automated image-based tattoo recognition as a means to augment or supersede the traditional keyword-based method.

1.2 The Tattoo Recognition Program

The Tattoo Recognition Technology Program was initiated by the National Institute of Standards and Technology (NIST) to evaluate an operational law enforcement need for image-based tattoo recognition to support law enforcement applications. The program provides quantitative results for tattoo recognition development and best practice capture guidelines. Program activities to date are summarized in Figure 1.

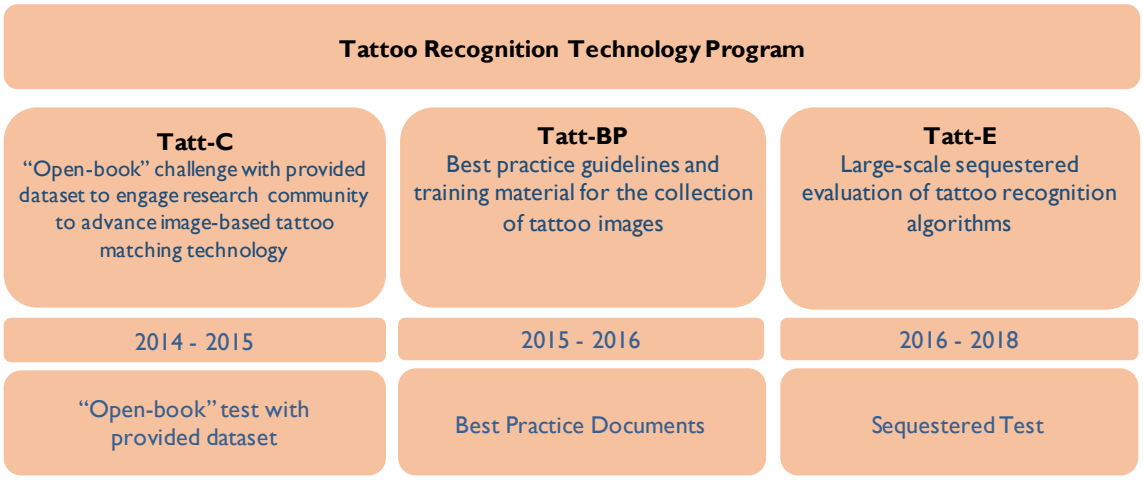


Figure 1: Activities under the Tattoo Recognition Technology Program

- **Tatt-C** was an open research challenge that provided operational data and use cases to the research community to motivate and advance research and development into automated image-based tattoo technologies and to assess the state-of-the-art. Tatt-C culminated with an industry workshop and a published public report on the outcomes and recommendations from the research activity. Please visit the Tatt-C website [6, 9] for more information.
- **Tatt-BP** provides best practice guidance material for the proper collection of tattoo images to support image-based tattoo recognition. Recognition failure in Tatt-C was often related to the consistency and quality of image capture, so Tatt-BP aimed to provide guidelines on improving the quality of tattoo images collected operationally. Please visit the Tatt-BP website [10] for more information.

- **Tatt-E** was a sequestered evaluation intended to assess tattoo recognition algorithm accuracy and run-time performance over a large-scale of operational data. In this test, algorithm software was sent to NIST for evaluation on data that had never been shared with developers. The outcomes of Tatt-E are documented in this report and are also available for download from the Tatt-E website [11].

1.3 Scope

Based on the outcomes of Tatt-C 2015, a follow-on "closed-book" test was recommended by the developer and user community to further evaluate the use cases where researchers reported high accuracy results. Running a sequestered evaluation enabled testing of much larger, operationally-realistic gallery sizes of data that cannot be freely distributed. So in 2017, NIST conducted a follow-on activity called the Tattoo Recognition Technology - Evaluation (Tatt-E). Tatt-E was run as a sequestered evaluation, where algorithm software was sent to NIST and tested on data that developers did not have access to. No test data was provided to developers by NIST as a part of this evaluation. Tatt-E reports measurement of algorithmic capability to perform the following use cases:

- **Tattoo Identification:** matching different instances of the same tattoo image from the same subject over time
- **Sketches:** matching sketch images to tattoo images
- **Multispectral:** matching tattoo images collected in the short-wave infrared (SWIR) spectrum against tattoos collected in the visible spectrum
- **Tattoo Detection and Localization:** determining whether an image contains a tattoo and if so, segmentation of the tattoo

2 Methodology

2.1 Test Environment

The evaluation was conducted offline at a NIST facility. Offline evaluations are attractive because they allow uniform, fair, repeatable, and large-scale statistically robust testing. However, they do not capture all aspects of an operational system. While this evaluation is designed to mimic operational reality as much as possible, it does not include a live image acquisition component, workflow management, or any interaction with real users. Testing was performed on high-end server-class blades running the CentOS Linux [12] operating system, some equipped with graphics processing units (GPUs). The test harness used concurrent processing to distribute workload across dozens of computers.

2.2 Algorithms

The Tatt-E program was open to participation worldwide. The participation window opened on December 1, 2016, and the submission deadine to the final phase was September 29, 2017. There was no charge to participate. The process and format of algorithm submissions to NIST were described in the Tatt-E Concept, Evaluation Plan, and Application Programming Interface (API) document [11]. Participants provided their submissions in the form of libraries compiled on a specified Linux kernel, which were linked against NIST's test harness to produce executables. NIST provided a validation package to participants to ensure that NIST's execution of submitted libraries produced the expected output on NIST's test machines.

Tatt-E had three submission phases where participants could submit up to two algorithms per phase to NIST. All participants submitted the maximum allowed number of algorithms in each phase of the test. This report documents the results

of all algorithms submitted to all phases of the test. Table 1 lists the Tatt-E participants who submitted algorithms for tattoo detection and localization (Class D) and tattoo identification (Class I), and the alphanumeric code associated with each of their submissions. For each participant, the algorithms are labeled alphanumerically, for example, A10I, with A = letter code assigned to the participant, 1 = algorithm submission phase, 0 = sequence number of submission, I = class of algorithm (I for identification, D for detection and localization). The letter codes assigned to the participants are also located at the bottom of each page for reference. The GPU column indicates whether the submitted algorithms ran on the GPU or the CPU.

Participant	Letter Code	GPU	Submissions		
			Phase 1 (Mar. 2017)	Phase 2 (Jun. 2017)	Phase 3 (Sept. 2017)
Chinese Academy of Sciences	A	Yes	A10D, A11D A10I, A11I	A20D, A21D A20I, A21I	A30D, A31D A30I, A31I
MorphoTrak	B	No	B10D, B11D B10I, B11I	B20D, B21D B20I, B21I	B30D, B31D B30I, B31I

Table 1: Tatt-E Participants

2.3 Image Dataset

Testing was performed over sets of operationally-collected and research-collected tattoo images. Details are summarized in Table 2. For the test datasets used, human examiners manually verified that tattoo images contained at least one tattoo; verified subsets of tattoos linked with an anonymized ground truth identifier were indeed the same tattoo; drew bounding boxes around the primary tattoo content; and verified that non-tattoo images did not contain any tattoos.

Use Case	Tattoo Identification	Sketches	Multispectral	Tattoo Detection	Tattoo Localization
Description	Match different instances of the same tattoo from the same subject over time	Match sketches to corresponding tattoo image	Match tattoo image collected in SWIR to visible tattoo images	Detect whether an image contains a tattoo	Segment around the tattoo content
Task	One-to-many search	One-to-many search	One-to-many search	Classification	Localization
Types of images	Tattoos	Sketches and tattoos	Tattoos	Tattoos, non-tattoos	Tattoos
# probes	11 921	1036	150	131 662 tattoos, 125 253 non-tattoos	10 926 images containing 12 613 tattoos
Gallery size	100 000	100 000	100 000	NA	NA
Median Image Size	Probe: 480x600, Gallery: 640x480	Sketches: 510x594, Gallery: 640x480	SWIR: 1280x1024, Gallery: 640x480	Tattoos: 768x960, Non-tattoos: 500x375	480x600

Table 2: Image Datasets

3 Metrics

The following performance measures are reported in the assessment of tattoo matching, detection, and localization accuracy.

3.1 Identification Metrics

Tattoo identification requires submitted algorithms to identify whether a particular tattoo image is present in a database and to retrieve matching entries. An image is searched against a gallery and a list of candidates is returned. Candidates at the top of the list are expected to have a higher probability of being the same tattoo to the searched tattoo. In contrast to most biometric modalities, tattoo recognition is not a "lights-out" operation, meaning human examination of the candidate list is assumed and required. Given this scenario, rank-based metrics are relevant and can be leveraged to support workload assessment.

3.1.1 Cumulative Match Characteristic

The Cumulative Match Characteristic (CMC) is a rank-based metric used to show core identification accuracy, which is defined as the proportion of searches that yield correct retrievals (i.e. mates) within the top K ranks. This shows the fraction of searches that return the correct image as a function of the candidate list length. The longer the candidate list, the greater the probability that mated images are on the list. In this report, the primary identification metric is reported as rank 10 accuracy or hit rate.

3.1.2 False Positive Identification Rate

While rank-based metrics such as the CMC measure algorithm accuracy when the tattoo *exists* in the gallery (e.g., closed-universe), most applications are naturally open-universe, where some proportion of query tattoos will not have a corresponding mate entry among the previously enrolled photos. From a testing perspective, open-universe is accomplished by running both mated searches and non-mated searches. The non-mated searches afford the ability to generate a false positive identification rate against a threshold. Ideally, a tattoo recognition algorithm must minimize both false positive errors where an enrolled non-mated tattoo is mistakenly returned, and false negative errors where an enrolled mated tattoo is missed.

False positive identification rates are not documented in this report. Tattoos cannot be used as a primary biometric as an arbitrary number of people can have nearly identical tattoos. Through analysis and trial runs on the test data, searching tattoos of subjects that are known not to exist in the gallery (non-mates) sometimes returns the same tattoo on different subjects. This is akin to using face recognition to match twins that exist in the same database. While in theory such near-identical tattoos could be considered false positives, algorithmic capability to retrieve the same tattoo from different subjects warrants consideration due to investigative utility in operations. Removal of such near-identical tattoo images from different people from the test dataset for all possible non-mated scenarios would require substantial human labor, which was not available for the scope of this evaluation.

Similarly, searching tattoos of subjects that are known to be in the gallery occasionally yields the same tattoo on different subjects. Tattoo recognition is not a fully-automated, lights-out application where decisions are made without human intervention. There needs to be human adjudicators reviewing tattoos on candidates lists, so therefore, stating accuracy at rank 10 is appropriate, accounting for some cases where the same exact tattoo is retrieved but from different people. Statement of accuracy for ranks up to 300 are reported.

3.2 Detection Metrics

Tattoo detection requires submitted algorithms to determine whether a particular image contains a tattoo or not. Given an image, algorithms reported 1) a binary decision on whether the image contains a tattoo or not and 2) a confidence score on $[0, 1]$ representing the algorithm's certainty about whether the image contains a tattoo. A true positive detection occurs when the algorithm correctly detects a tattoo in a tattoo image. To calculate false positives, non-tattoo images (images that do not contain a tattoo) are provided to the algorithm. A false positive detection occurs when the algorithm incorrectly detects a tattoo in the image. We analyze detection accuracy by analyzing the confidence score returned by the algorithm. 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.

For the detection problem, receiver operating characteristic (ROC) curves show the tradeoff between the true positive detection rate and the false positive detection rate as the decision threshold is adjusted. A true positive occurs when a tattoo is correctly classified as a tattoo, and a false positive occurs when a non-tattoo is misclassified as a tattoo. Let x_i represent the confidence score for the i-th tattoo image ($i = 1, ..., M$), and y_j the confidence score for the j-th non-tattoo image ($j = 1, ..., N$). Then the error rates at any particular decision threshold, t, are

$$\text{True Positive Detection Rate}(t) = \frac{1}{M} \sum_{i=1}^{M} H(x_i - t), \tag{1}$$

$$\text{False Positive Detection Rate}(t) = \frac{1}{N} \sum_{j=1}^{N} H(y_j - t). \tag{2}$$

where $H(x)$ is the step function [13]:

$$H(x) = \begin{cases} 1, & x \geq 0, \\ 0, & x < 0. \end{cases} \tag{3}$$

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

3.3 Localization Metrics

Previously, we discussed metrics for detecting the presence of a tattoo(s) in an image. In this section, we present metrics for finding the location of a tattoo(s) in an image. For tattoo localization or segmentation, given an image, the algorithm is asked to produce bounding boxes around regions where a tattoo is detected. The algorithm is evaluated based on the area of the bounding box that overlaps with the ground truth. Figure 2 illustrates common localization scenarios as observed from the data. Note that masks could be an alternative/refined localization representation and would be more shape-sensitive than bounding boxes. Bounding box representation was chosen to maximize the amount of data generated for the test.

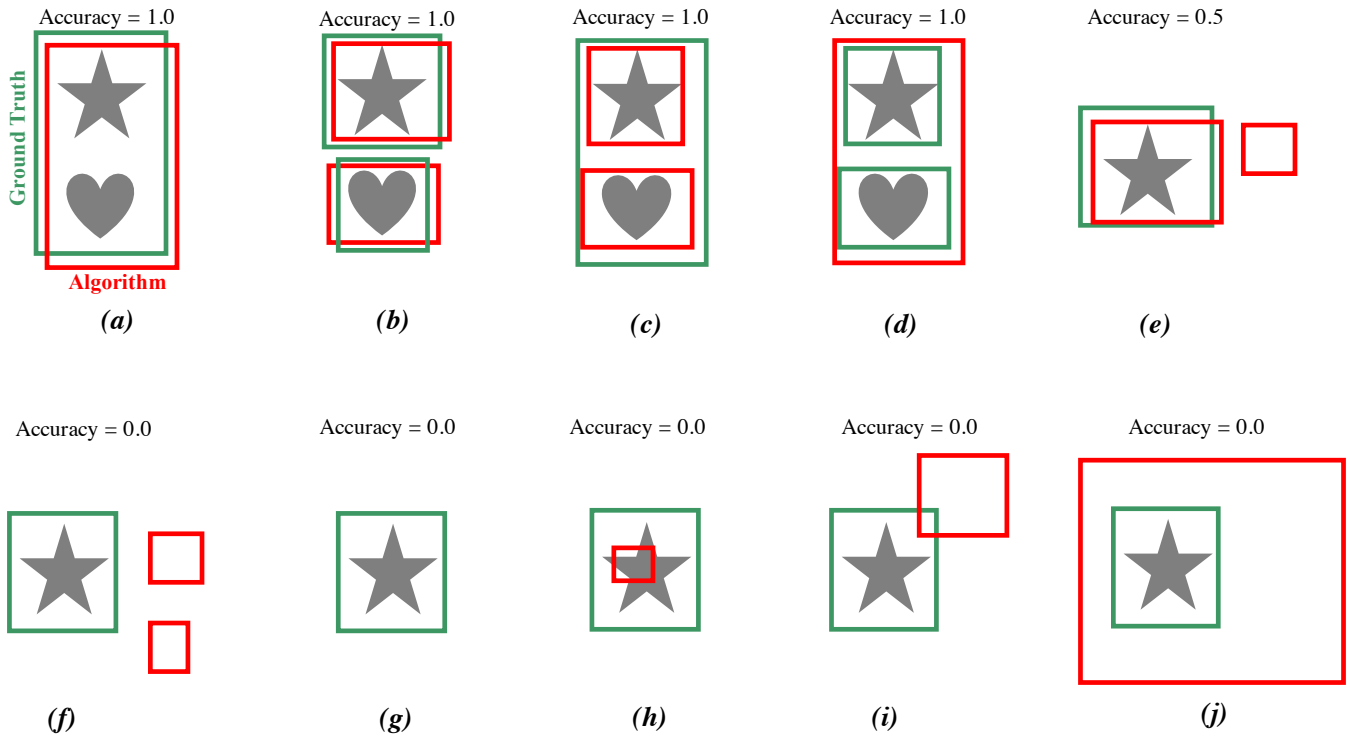


Figure 2: Examples of observed localization scenarios. (a) and (b) one-to-one correspondence between ground truth region(s) and detected region(s); (c) split region (ground truth region is associated with multiple detected regions); (d) merge region (detected region is associated with multiple ground truth regions); (e) one correct detection region and one false detection; (f) one detection failure and two false detections; (g) one detection failure (no detected region returned); (h) detected region too small; (i) insufficient overlap between detected and ground truth region; (i) detected region too large.

Figure 2(a) and 2(b) represent a majority of the localization scenarios where there is a one-to-one correspondence between the algorithm detected region and the ground truth region. Sometimes the localization procedure is subjective, since tattoos may contain several parts, and it is not always easy to determine if it is a single tattoo or several different tattoos. Figure 2(c) and 2(d) illustrate situations where the algorithm may split or merge the detected tattoo regions when compared to the ground truth, both of which should not penalize the performance of the algorithm. Figure 2(e) and 2(f) show examples of false region detections and region detection failures.

Considering one image, we compute a matrix of Intersection over Union (IoU) values between all ground truth bounding boxes and all algorithm detected bounding boxes. So given $N \geq 1$ ground truth bounding boxes, G_i, and $M \geq 0$ detected bounding boxes, A_j, the resulting matrix will contain NxM IoU scores [14].

$$C_{ij} = \frac{G_i \cap A_j}{G_i \cup A_j}, \tag{4}$$

where $\cap$ denotes the area of overlap between the ground truth bounding box and the detected bounding box, and $\cup$ denotes the area encompassed by both the ground truth bounding box and the detected bounding box.

We add the IoU values in each row or column, defining two auxiliary vectors:

$$L_i = \sum_{j=1}^{M} C_{ij} \;\; i \in 1, ..., N \tag{5}$$

$$D_j = \sum_{i=1}^{N} C_{ij} \;\; j \in 1, ..., M \tag{6}$$

A **False Detection** occurs when $D_j = 0$, that is, an algorithm detected bounding box does not overlap with any ground truth bounding box. A **Detection Failure** occurs for G_i if $L_i = 0$, that is, where a ground truth bounding box is not overlapped by any algorithm detected bounding box.

3.3.1 Mean IoU

For K $\gg 1$ images, we report mean IoU as

$$\overline{IoU} = \frac{1}{K} \sum_{k=1}^{K} \frac{\sum_{j=1}^{M} D_{jk}}{M_k} \tag{7}$$

3.3.2 Localization Accuracy

Here, localization accuracy is calculated by applying a minimum threshold on IoU. In the case where $M < N$, the accuracy score for an image is defined as the number of instances where the area of overlap between a ground truth bounding box and detected bounding box is more than 50 %, (i.e. IoU > 0.5), divided by the total number of detected bounding boxes and detection failures:

$$\text{Accuracy} = \frac{\sum_{j=1}^{M} H(D_j - 0.5)}{M + (N - \sum_{i=1}^{N} H(L_i))} \tag{8}$$

where $H(x)$ is the step function:

$$H(x) = \begin{cases} 1, & x > 0, \\ 0, & x \leq 0. \end{cases} \tag{9}$$

For $M \geq N$, accuracy for an image is defined as the number of instances where the area of overlap is more than 50 %, divided by the total number of ground truth bounding boxes and false detections:

$$\text{Accuracy} = \frac{\sum_{i=1}^{N} H(L_i - 0.5)}{N + (M - \sum_{j=1}^{M} H(D_j))} \tag{10}$$

The localization accuracy across all test images is the mean of the accuracy for each image. For K $\gg$ 1 images,

$$\text{Localization Accuracy} = \frac{1}{K}\sum_{k=1}^{K}\text{Accuracy}_k \tag{11}$$

4 Results

Gallery Composition: Operationally, tattoo enrollment would proceed with the image as-is with no further processing (e.g., cropping around primary tattoo content). For all of the experiments, the gallery contains original, uncropped tattoo images.

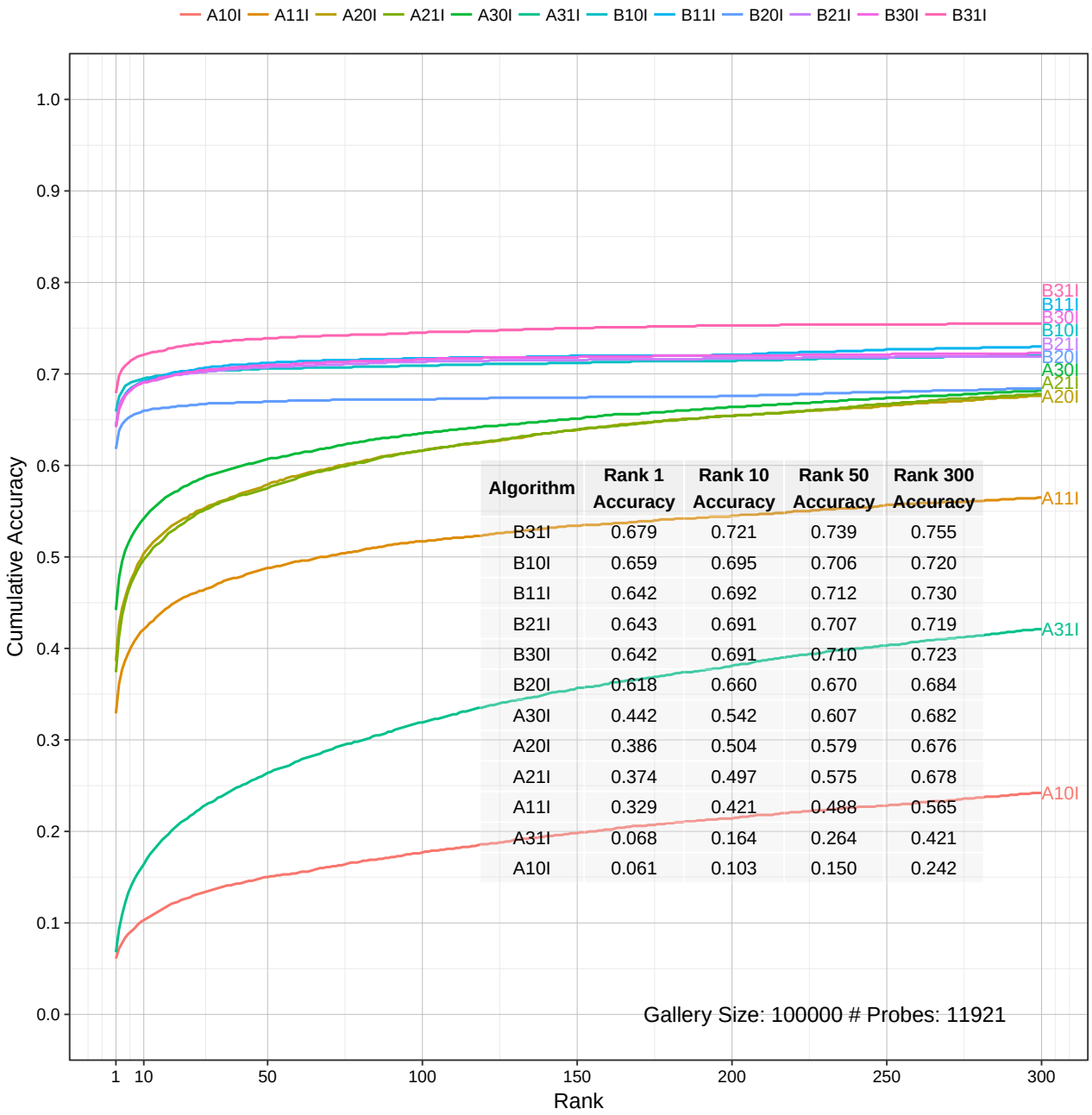
4.1 Identification

This section details the performance of algorithms tasked with matching different photos of the same tattoo from the same subject. The test data for this use case is composed of images of the same tattoo from the same subject collected at different times. A single instance of each mated tattoo is enrolled in the gallery and one or more mated probes are searched.

4.1.1 Accuracy

The results in this section show performance of algorithms when both the enrollment and search image are uncropped, where the entire image is provided to the algorithm and no prior cropping/background removal of the tattoo image(s) was performed. Figure 3 presents representative examples of images used, and Figure 4 shows accuracy results for matching uncropped probe tattoo images to uncropped gallery tattoo images.

Figure 3: Two examples of pairs of uncropped tattoos collected at different times. Note the existence of clothing and objects in the background. These images are representative of what was used for matching uncropped probe images to uncropped gallery images. Image source: NIST.



Algorithm	Rank 1 Accuracy	Rank 10 Accuracy	Rank 50 Accuracy	Rank 300 Accuracy
B31I	0.679	0.721	0.739	0.755
B10I	0.659	0.695	0.706	0.720
B11I	0.642	0.692	0.712	0.730
B21I	0.643	0.691	0.707	0.719
B30I	0.642	0.691	0.710	0.723
B20I	0.618	0.660	0.670	0.684
A30I	0.442	0.542	0.607	0.682
A20I	0.386	0.504	0.579	0.676
A21I	0.374	0.497	0.575	0.678
A11I	0.329	0.421	0.488	0.565
A31I	0.068	0.164	0.264	0.421
A10I	0.061	0.103	0.150	0.242

Figure 4: CMC plot (uncropped). Both enrollment and search imagery were uncropped, meaning no prior cropping/pre-processing of the tattoo image(s) was performed. Number of probes: 11 921, Gallery size: 100 000. The table, sorted by rank 10 cumulative accuracy, shows the algorithm and accuracy at rank 1, 10, 50, 300.

Results and Notable Observations:

- For the best performing algorithm (B31I), rank 10 accuracy is 72.1 % (miss rate of 27.9 %).
- Even for the best performing algorithm, approximately 25 % of searches never return the correct mate within the top 300 candidates. General characteristics of images associated with identification failures can be categorized into detection, localization, and matching failure. Tattoos that did not get detected in an image primarily consisted of low contrast tattoos on very dark skin and images where the tattoo-to-image ratio (the size of the tattoo relative to the entire photo) was small. Images where segmentation of the tattoo from the background was bad or incorrect often contained patterned/graphical clothing in the background or body hair that wasn't part of the tattooed area.

Bad segmentation of the tattoo, on patterned clothing for example, caused false matching on images with similar patterns. False matching of tattoos with similar fonts was observed, and very simple and small tattoos such as teardrops and dots were difficult to match likely due to the lack of unique features. Images with similar body parts, such as fingers and fists sometimes matched with other images with the same body part and/or pose.

- For the algorithms from provider B, the curves flatten out very quickly as a function of rank. This is indicative of diminishing returns when reviewing long candidate lists. Based on the results, if the correct match is not retrieved within the top 50 candidates, extending the review list to 300 would yield minimal return. This has resource implications as there is labor cost associated with looking down long candidate lists in operations.

4.1.2 Effect of Cropping Search Imagery

The impact of tattoo segmentation on tattoo matching performance has been investigated [15], and results show that irrelevant non-tattoo regions in the image can have a negative impact on matching accuracy. Operationally, there is an opportunity to crop the probe tattoo prior to running it through a large database primarily composed of uncropped tattoos enrolled as captured. Here, we assess the performance impact of such a scenario, where cropped probe images are searched against a gallery of uncropped tattoo images. The same 11 921 search images from Section 4.1.1 were cropped around the primary tattoo content, and searched against the same gallery of 100 000 uncropped tattoos.

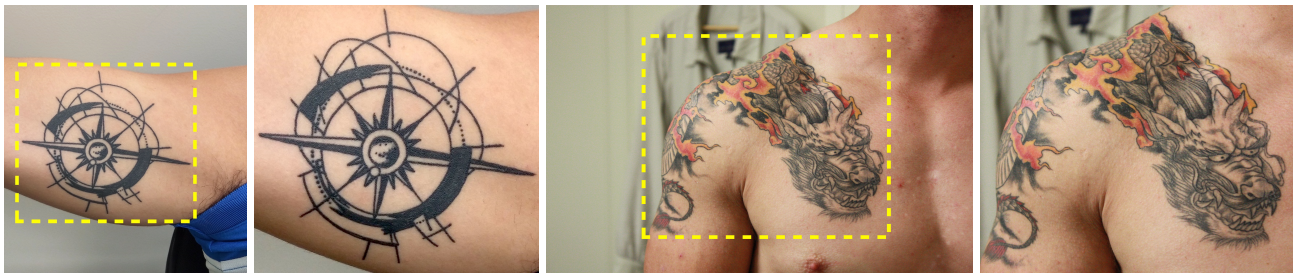


Figure 5: Examples of tattoos cropped from the original image. Image source: NIST.

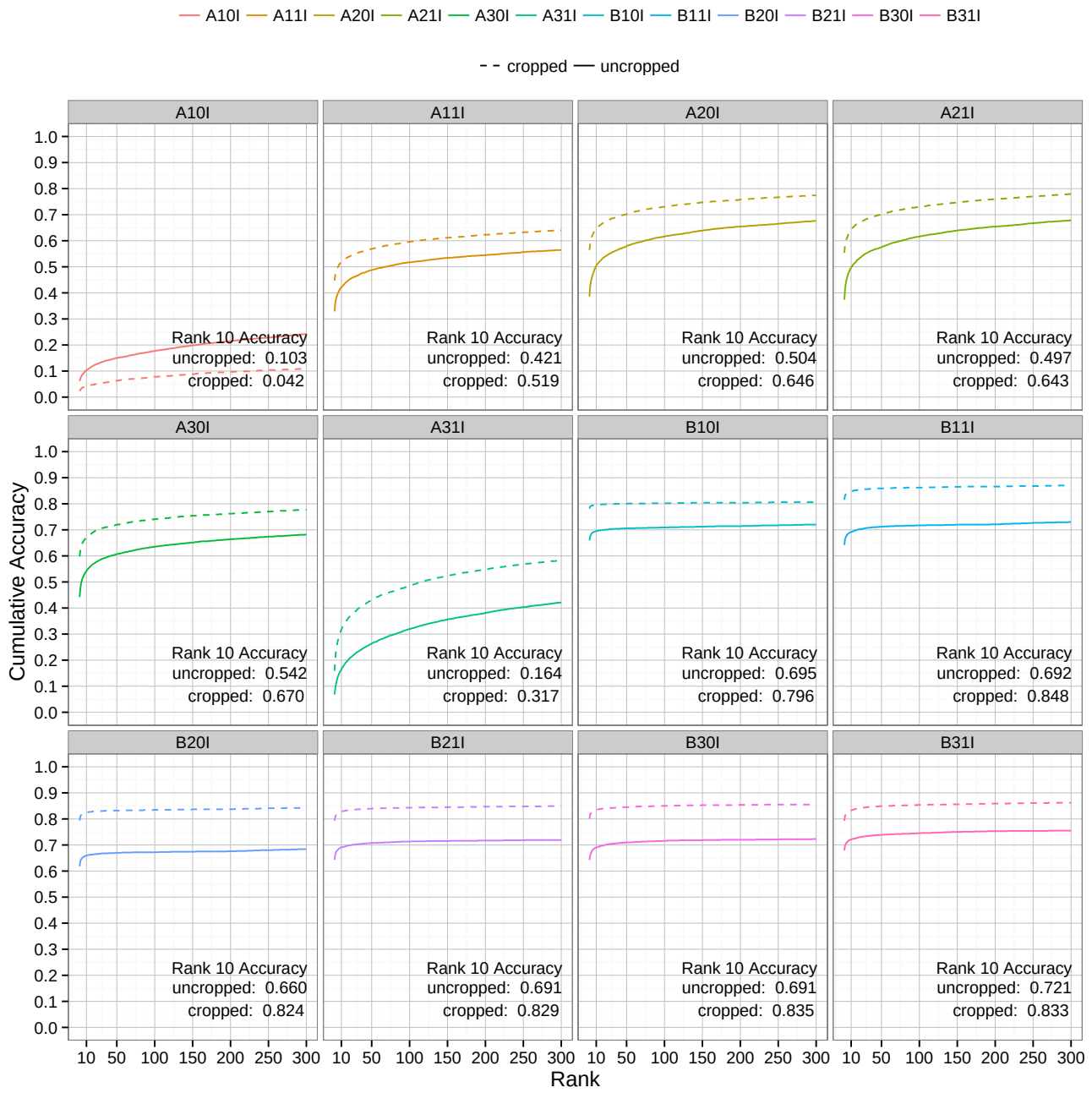


Figure 6: CMC plot (cropped probes). The probes were cropped around the primary tattoo content. Images enrolled in the gallery were "uncropped", meaning no prior cropping/pre-processing of the tattoo image(s) was performed. Number of probes: 11 921, Gallery size: 100 000.

Results and Notable Observations:

- For the best performing algorithm (B11I), rank 10 accuracy is 84.8 % (miss rate of 15.2 %) when the probe image is cropped.
- There are notable accuracy improvements for almost all algorithms when cropping the probe around the primary tattoo content prior to search. From Section 4.1.1, searching with uncropped probe images yields a best-case rank 10 hit rate of 72.1 % (miss rate of 27.9 %). This quantifies that an additional ˜13 % of searches can be matched when using cropped search images over uncropped images, which constitutes nearly a 50 % decrease in miss rate (which means approximately half of the matching failures are fixed by just cropping around the tattoo).
- The improvement from cropping and hit rates going up is due to assistance with localization, which shows that

detection/localization still remains a problem on a percentage of tattoos even in the moderately controlled collection of tattoo images used in this test. The implications to improve tattoo matchability is to maximize the primary tattoo content area in the photo and avoid capturing large area photos with small tattoos, which is covered in the Tatt-BP guidelines for proper collection of tattoo images [10].

4.1.3 Effect of Tattoo-to-Image Ratio

Tattoos that are very small relative to the total size of the image have anecdotally posed challenges to automated tattoo recognition [9]. We define tattoo-to-image ratio as the area of the bounding box drawn around the primary tattoo content divided by the total size of the entire image. Here, we investigate the accuracy impact of tattoo-to-image ratio by looking at the proportion of searches on cropped probe images where the correct mate was not retrieved within the top 300 ranks (i.e., rank 300 miss rate), broken out by tattoo-to-image ratio.

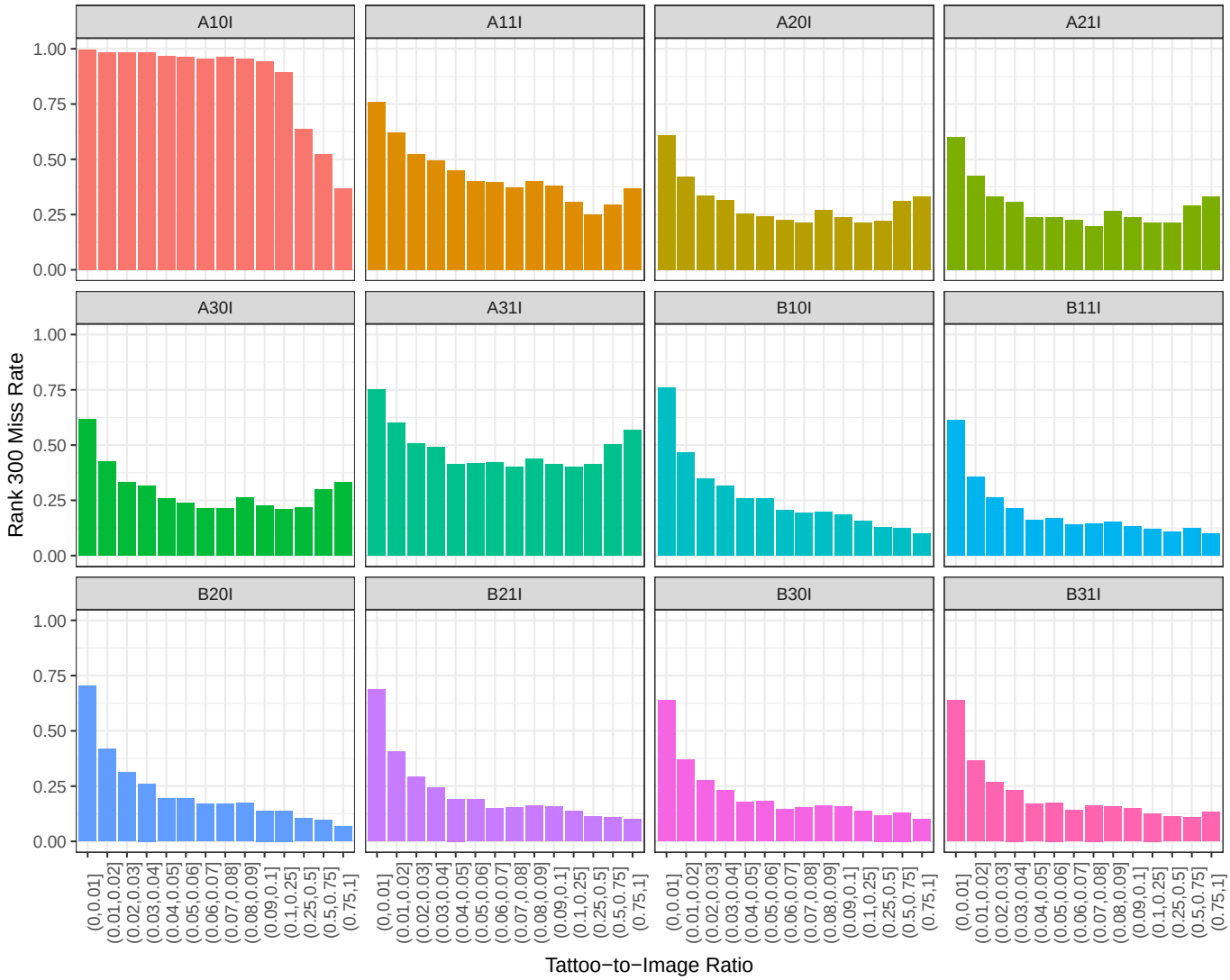


Figure 7: This plot shows rank 300 miss rate on cropped probe images plotted against tattoo-to-image ratio. Rank 300 miss rate is defined as the proportion of searches where the mate was not retrieved within the top 300 ranks. Tattoo-to-image ratio is defined as the size of the tattoo (as a bounding box) divided by the total size of the image. Note non-uniform scale on x-axis.

Results and Notable Observations:

- The size of the tattoo relative to the entire image has an effect on tattoo retrieval rates. For all algorithms, over 60 % of tattoos that are captured as a very small percentage (less than or equal to 1 %) of the entire image were not matched within the top 300 images retrieved.
- Since the probe images have already been cropped around the primary tattoo content, this eliminates the need for the algorithm to perform detection and localization. Through visual inspection, the cropped tattoos with very low tattoo-to-image ratios result in very low resolution, blurry tattoos that could pose challenges to matching.
- Miss rates decrease as the tattoo-to-image ratio increases for most of the algorithms, which indicates that tattoos that occupy a larger percentage of the image have a higher chance of being matched.

4.1.4 Effect of Skintone and Contrast

Images where there is low contrast between the tattoo and skin color are known, qualitatively, to pose challenges to automated tattoo recognition [9]. Here, we quantify the effects of skintone brightness (luminance) and tattoo contrast by defining the following:

Skintone Brightness
Manually created bounding boxes are drawn around each tattoo in an image. The bounding box contains the primary tattoo content as well as a certain percentage of surrounding skin. To calculate skintone brightness, we take the last, bottom row of pixels in the bounding box, P, and for each pixel, p, calculate the relative luminance, Y, in RGB space. Then the median of all luminance values is calculated to generate a single skintone brightness value (see caveat at the end of this Section 4.1.4).

$$\forall p \in P : Y_p = (0.2126 * R_p) + (0.7152 * G_p) + (0.0722 * B_p), \tag{12}$$

$$\text{Skintone brightness} = \tilde{Y} \tag{13}$$

Tattoo Contrast
Given the same bounding box around each tattoo, the contrast of the tattoo, C, is calculated as

$$C = \sqrt{\frac{\sum_{i=0}^{N-1}\sum_{j=0}^{M-1}(I_{ij} - \bar{I})^2}{MN}}, \tag{14}$$

where M and N are the height and width of the bounding box respectively, I_{ij} is the image intensity (normalized in the range $[0, 1]$) at position i and j (within the bounding box), and $\bar{I}$ is the average intensity of all pixel values in the bounding box [16].

Figure 8 plots a heatmap showing the effects of skintone brightness and tattoo contrast on matching accuracy. The results of searching 11 921 cropped search images against a gallery of 100 000 from Section 4.1.2 are binned based on the skintone brightness and tattoo contrast of the search image. The median rank of retrieval of the correct mate is calculated for each bin and expressed as a color on a gradient, with green representing good accuracy with correct retrieval in top ranks, and red representing the correct mate is never retrieved in the top 300 candidates.

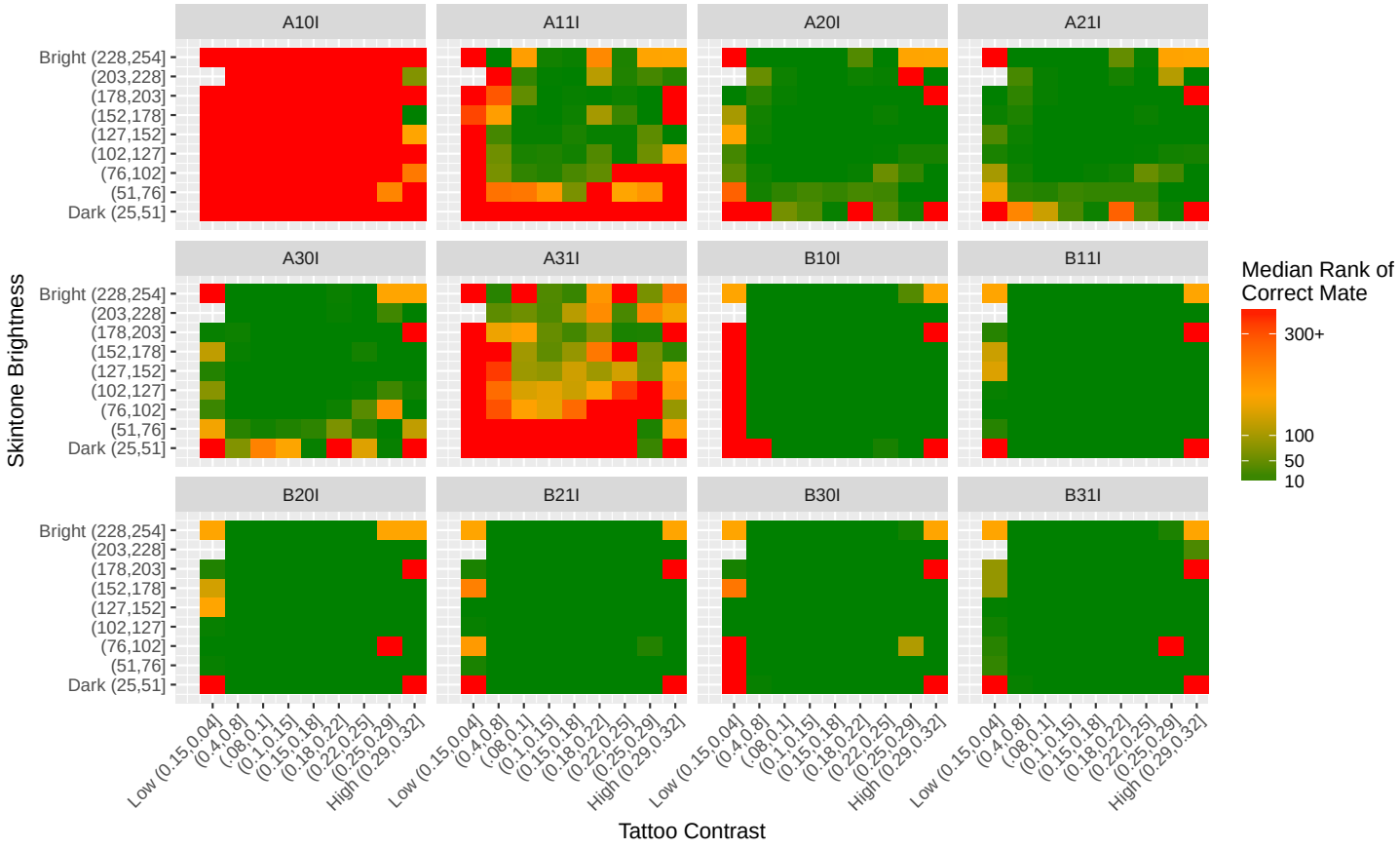


Figure 8: This heatmap plots median rank of the correct mate retrieved, binned by skintone brightness and tattoo contrast of the cropped probe images. The bins on the lower left of the heatmap represent low contrast tattoos on dark skin, and on the upper right of the chart are tattoos on light skin with high contrast. 11 921 probe images were split into 81 bins. Gallery size: 100 000.

Results and Notable Observations:

- The impact of skintone brightness and tattoo contrast on matching accuracy depends on the algorithm. All algorithms appear to be sensitive to very low contrast tattoos. Participant B's algorithms are not majorly impacted by skintone brightness while Participant A's algorithms have difficulty matching images with very low skintone brightness.
- The large area in the middle of the plots where a majority of the bins are green indicate that algorithms perform well and handle a fairly large range of skintone brightness and tattoo contrast conditions, which means performance is degraded primarily in the extreme cases.
- Caveat: Outliers could exist as it is possible for the measurement of tattoo contrast to be unintentionally imprecise from residual background imagery that could be not cropped outside of the primary tattoo content. For example, in cases where the tattoo itself is low contrast against the skin but the cropped image contains high-contrasting clothing or other objects in the field of view. Similarly, the skintone brightness measurement can be impacted by clothing or background at the bottom of the bounding box area.

4.1.5 Effect of Gallery Size

As more tattoo data is collected in operations, scalability related to match performance against larger database sizes becomes important. Figure 9 presents the rank 10 hit rate of searching 11 921 tattoo images against gallery sizes of 20 000, 50 000, and 100 000.

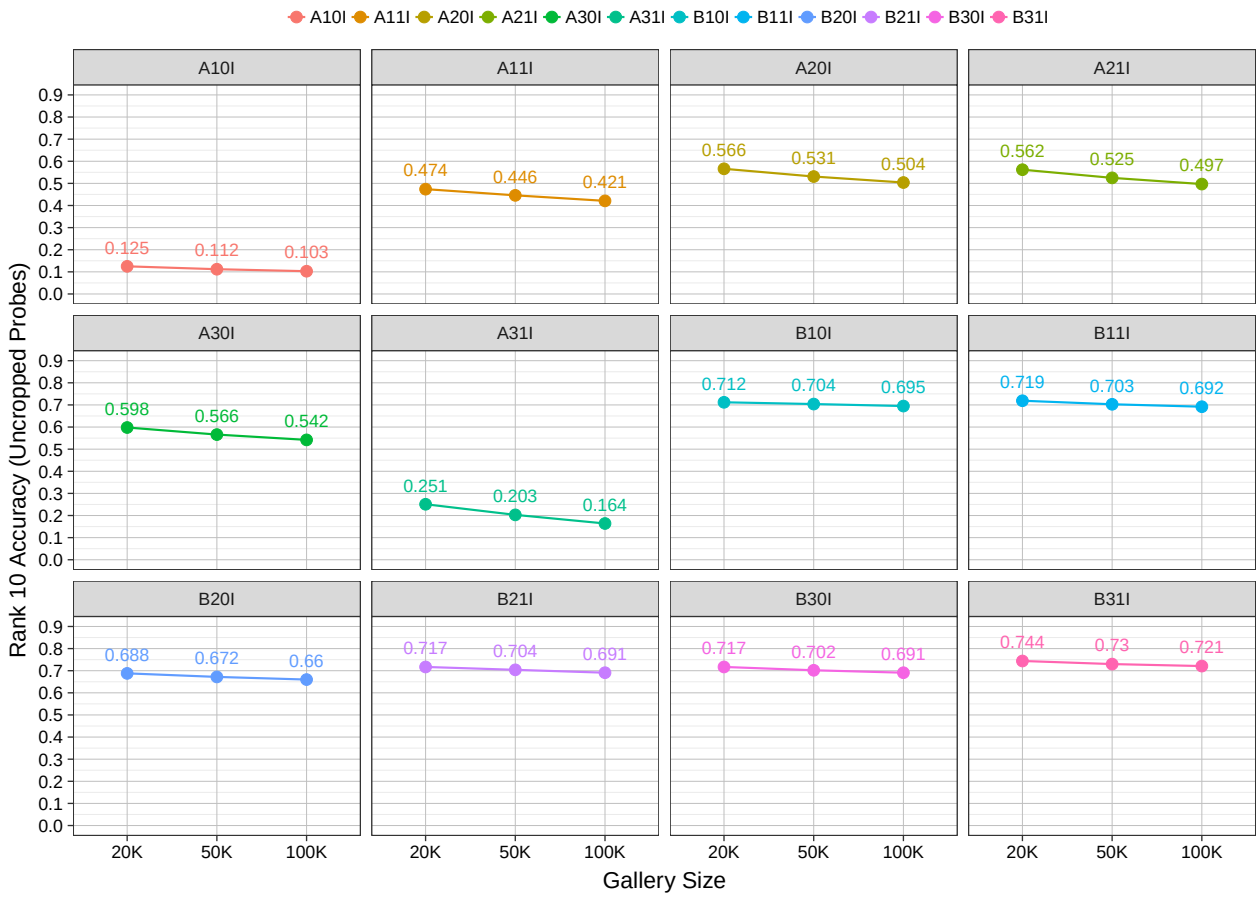


Figure 9: This plot shows the effect of gallery size on the rank 10 accuracy rate for the various algorithms. Both probe and gallery imagery were uncropped, meaning no prior cropping/pre-processing of the tattoo image(s) was performed. Number of probes: 11 921

Results and Notable Observations:

- With an increase in gallery size, a decrease in performance is observed across all algorithms, which is expected behavior and demonstrates that gallery size does have an impact on search accuracy. Accuracy degrades as non-matching gallery images displace the mated gallery image out of the top 10.
- Accuracy decreases at fairly benign rates when the gallery size increases. With a gallery size increase of a factor of 5, the decrease in rank 10 hit rate is modest, ranging from 2 % to 9 % depending on the algorithm. For the more accurate algorithms, the decrease in hit rate ranges between 2 % to 3 %, which demonstrates the viability of tattoo recognition on operationally-sized databases.

4.1.6 Speed - Accuracy Tradeoff

This section presents tradeoff analysis between rank 10 hit rate on uncropped tattoo images and median search time (the time it takes to search a single tattoo image against a gallery of 100 000 enrolled tattoos and return a list of candidates). In NIST's test harness, all functions were wrapped by calls to `std::chrono::duration` which enables duration measurements with nanosecond resolution. Timing was measured on a dedicated server-class blade equipped with an Intel Xeon E5-2695 v3 2.3 GHz CPU[1] and a NVIDIA Tesla K40 GPU board[2]. Search was executed in a single process, either on the GPU or CPU depending on the implementation's hardware requirements. Implementations were not allowed to

[1] https://ark.intel.com/products/81057/Intel-Xeon-Processor-E5-2695-v3-35M-Cache-2_30-GHz
[2] https://www.nvidia.com/content/tesla/pdf/NVIDIA-Tesla-Kepler-Family-Datasheet.pdf

multithread. The computer was not running any other processes except those back-grounded as part of the operating system. Timing does not include any pre-processing steps performed by the test software such as loading the image from disk or extracting image data from a compressed JPEG [17] file. Timing measurements do not include disk access unless the implementation under test elected to access data on disk during the core operation.

Both participants leveraged the common open-source deep learning framework, Caffe [2]. The algorithms from participant A all ran on the GPU, and the participant B submissions all ran on the CPU. Note that differences in hardware make direct timing comparisons difficult, but measuring absolute timing remains useful as different applications will have different speed requirements.

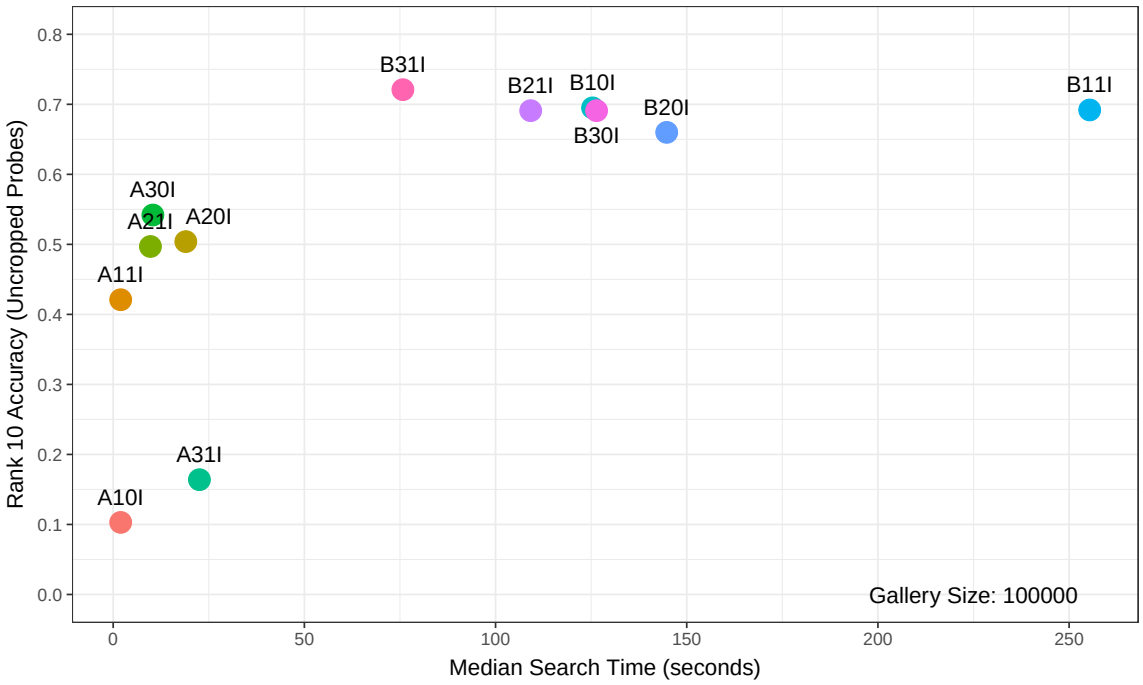


Figure 10: This plot shows a tradeoff between rank 10 accuracy and median search time (in seconds) on a gallery of 100 000. Both enrollment and search imagery were uncropped, meaning no prior cropping/pre-processing of the tattoo image(s) was performed. All A algorithms ran on the GPU, and all B algorithms ran on the CPU.

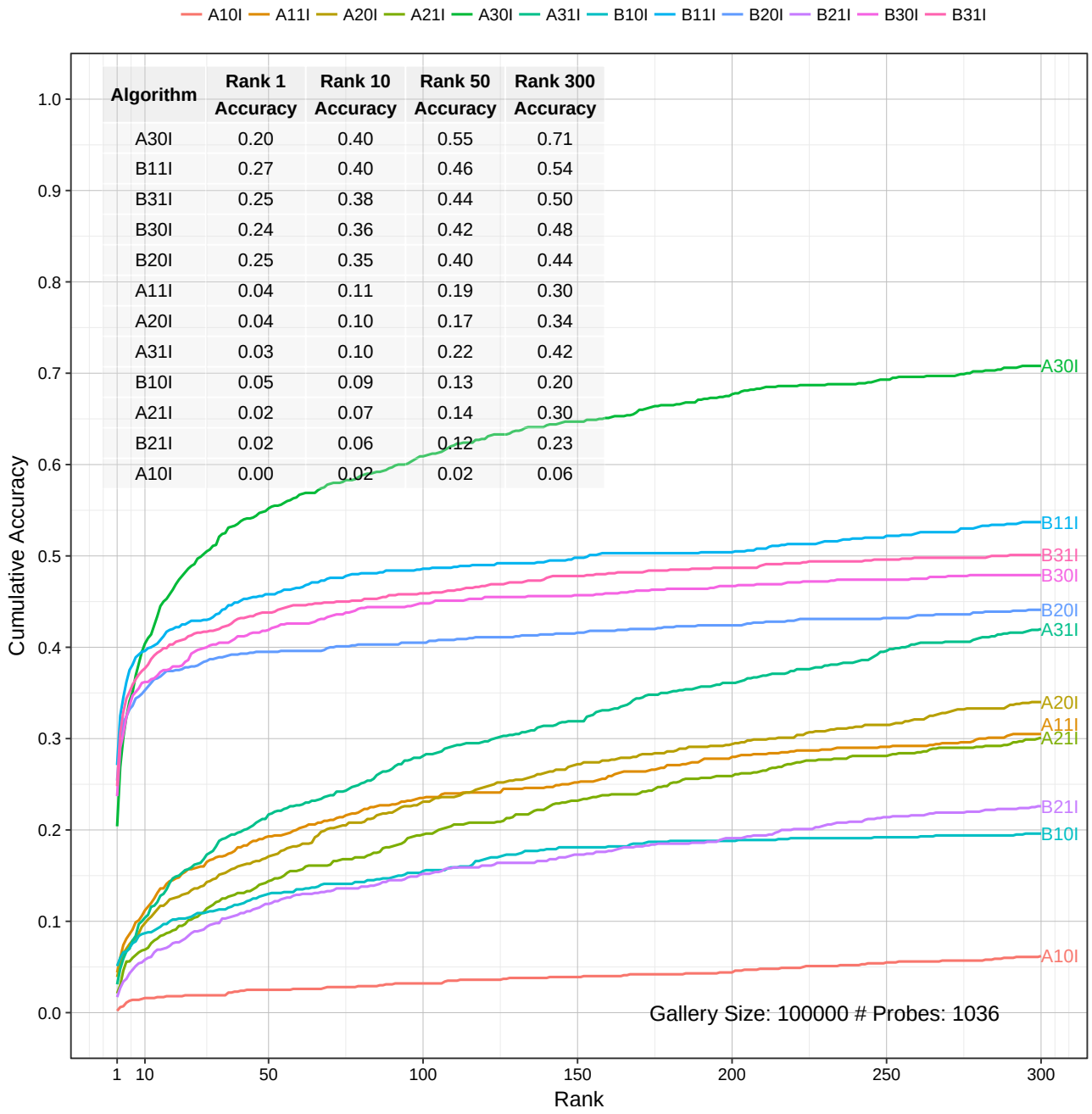
Results and Notable Observations:

- There is observable tradeoff between search speed and accuracy between the two participating organizations. The faster algorithms tend to be less accurate, and the more accurate algorithms tend to be slower. Rank 10 accuracy ranges from 10.3 % (A10I) to 72.1 % (B31I) while search speeds range from 2.0 seconds (A10I, A11I) to 255.4 seconds (B11I) to search a single tattoo image against a gallery size of 100 000. The most accurate algorithm (B31I) takes 75.8 seconds to search a single tattoo image.
- Speed requirements are driven by operational needs and may differ between applications, so users are cautioned to take consideration when timing is critical to their operations. In low volume investigative applications, speed may not be as important, while in high volume, real-time applications, speed is generally more critical. The type of hardware used by the algorithm is a developer design decision and will have resource and cost implications. More details on timing can be found in Section 4.7.

4.2 Sketches

This section details the performance of algorithms on matching sketch images to its corresponding photo of the actual tattoo. Querying sketches of tattoos drawn based on provided descriptions has forensic value, and there has been some research done on the ability of computer vision methods on matching sketches to tattoo photo images [18]. The test data for this use case is composed of sketches drawn from 1 036 tattoo images. A single instance of each mated tattoo is enrolled into a gallery size of 100 000 tattoo images, and the sketch image is used to search the gallery. Figure 11 presents representative examples of images used, and Figure 12 shows accuracy results for the sketch to tattoo matching experiment.

Figure 11: Examples representative of images used for sketch to tattoo matching. Image source: NIST.



Algorithm	Rank 1 Accuracy	Rank 10 Accuracy	Rank 50 Accuracy	Rank 300 Accuracy
A30I	0.20	0.40	0.55	0.71
B11I	0.27	0.40	0.46	0.54
B31I	0.25	0.38	0.44	0.50
B30I	0.24	0.36	0.42	0.48
B20I	0.25	0.35	0.40	0.44
A11I	0.04	0.11	0.19	0.30
A20I	0.04	0.10	0.17	0.34
A31I	0.03	0.10	0.22	0.42
B10I	0.05	0.09	0.13	0.20
A21I	0.02	0.07	0.14	0.30
B21I	0.02	0.06	0.12	0.23
A10I	0.00	0.02	0.02	0.06

Figure 12: CMC plot (sketches). The gallery consists of "uncropped" tattoos, meaning no prior cropping/pre-processing of the tattoo image(s) was performed, and the probes were sketch images. Number of probes: 1 036, Gallery size: 100 000. The table, sorted by rank 10 cumulative accuracy, shows the algorithm and accuracy at rank 1, 10, 50, 300.

Results and Notable Observations:

- For the best performing algorithm (A30I), rank 10 accuracy is 40 %.
- Increasing the size of the candidate list from 1 to 50 can yield significant accuracy benefits (between 10 % to 35 % increase in hit rate) when matching sketch images to a gallery of tattoos. For the best performing algorithm, reviewing the first 300 candidates yields a hit rate as high as 71 %.
- Caveat: The accuracy numbers documented represent a best-case scenario, because in practice, the sketch artist will not have a photo to work with. In operational cases, accuracy of this process will depend on the recollection of the

person describing the tattoo to a sketch artist in creating a faithful representation of what they saw. While accuracy figures are much lower than matching tattoo images to tattoo images, this has to be compared with not having an automated sketch search capability at all. In such cases, the hit rate is 0 % (miss rate is 100 %) and investigative leads have to be developed without automated sketch search capability. As such, the 40 % hit rate just by reviewing the first 10 candidates represents a practical resource in otherwise cold cases.

4.3 Multispectral

The use of different infrared bands for image collection is used in some operational processes. The accuracy of algorithms on matching tattoo images collected in various wavelengths within the short-wave infrared (SWIR) spectrum is reported in this section. A single instance of each mated tattoo collected in the visible spectrum is enrolled in the gallery, and the SWIR tattoo image is searched. The dataset used in this experiment and image examples are published in [8].

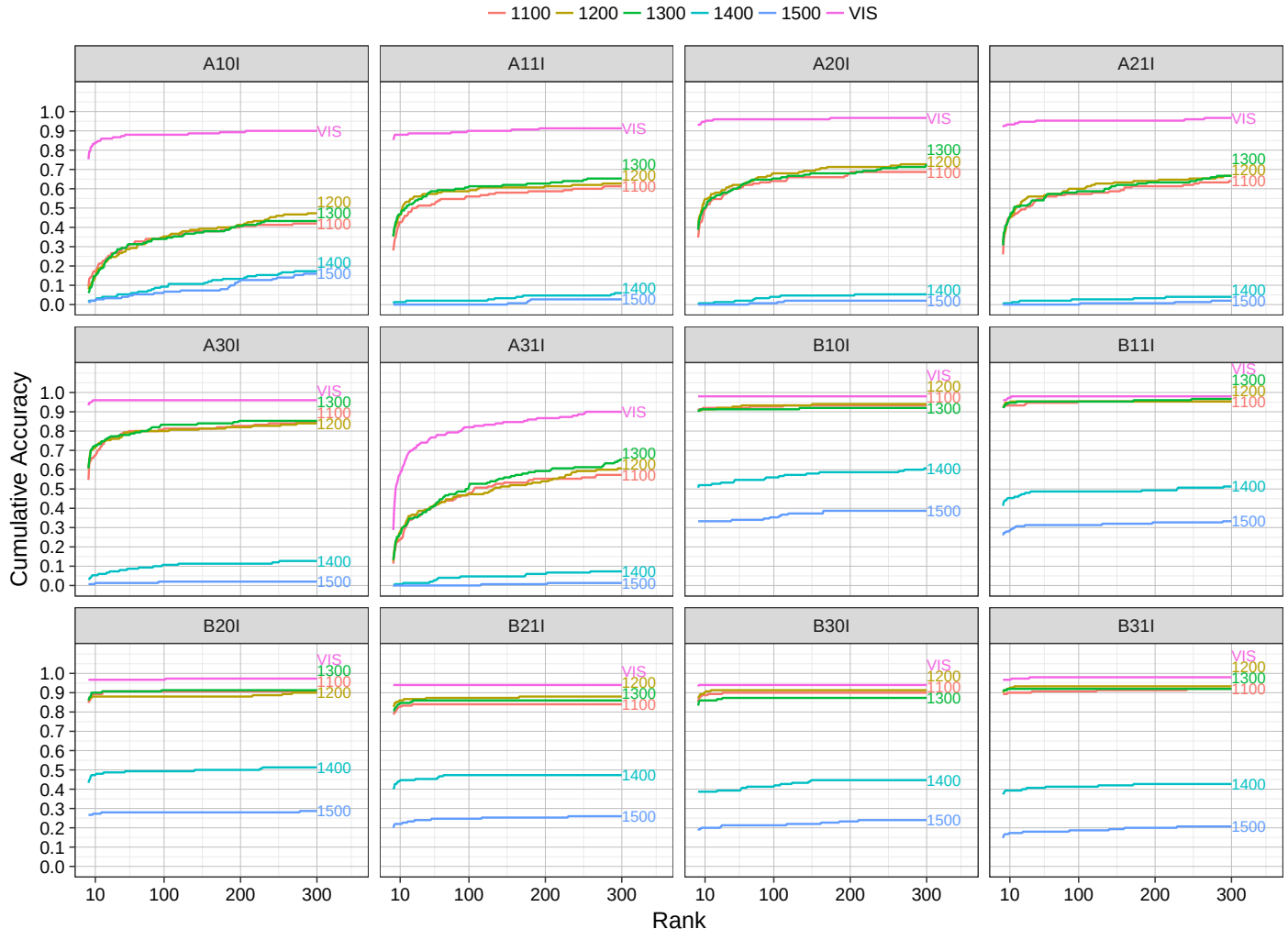


Figure 13: CMC plot (SWIR). The gallery consists of uncropped tattoos, meaning no prior cropping/pre-processing of the tattoo image(s) was performed, collected in the visible spectrum. The probes were tattoo images collected in the SWIR wavelength range of 1100 nm and 1500 nm. For reference, search results for probes collected in the visible spectrum (VIS) are also reported. Number of probes in each wavelength: 150, Gallery size: 100 000. Note: Approximately 50 % of the probe imagery was collected during the same photo session as the enrollment image, which may result in more optimistic performance than what might be expected operationally.

Algorithm	1100nm	1200nm	1300nm	1400nm	1500nm	VISIBLE
A10I	0.17	0.15	0.15	0.02	0.02	0.84
A11I	0.43	0.47	0.47	0.01	0.00	0.88
A20I	0.49	0.55	0.50	0.01	0.00	0.95
A21I	0.45	0.45	0.45	0.01	0.00	0.93
A30I	0.67	0.71	0.73	0.05	0.01	0.96
A31I	0.24	0.28	0.27	0.01	0.00	0.58
B10I	0.92	0.91	0.91	0.52	0.33	0.98
B11I	0.93	0.95	0.95	0.45	0.29	0.97
B20I	0.89	0.88	0.90	0.47	0.27	0.97
B21I	0.83	0.86	0.85	0.45	0.22	0.94
B30I	0.89	0.90	0.86	0.39	0.20	0.94
B31I	0.90	0.93	0.92	0.39	0.17	0.97

Table 3: The table shows the rank 10 cumulative accuracy across each of the SWIR wavelengths and the visible spectrum. Number of probes in each wavelength: 150, Gallery size: 100 000.

Results and Notable Observations:

- In SWIR, all algorithms showed the best match performance on visible images, and then on images collected between the 1100 nm to 1300 nm wavelengths. This is largely due to the decrease in visibility of the tattoo on the skin when collected at 1400 nm and above wavelengths, as shown in [8].
- For all algorithms, the best matching performance is still observed on images collected in the visible spectrum, when searched against a gallery of visible spectrum tattoos.
- Caveat #1: Approximately 50 % of the probe imagery was collected during the same photo session as the enrollment image, which may result in more optimistic performance than what might be expected operationally.
- Caveat #2: As SWIR-to-visible tattoo identification was not declared to be part of the study until the last phase of the test, the algorithms are being used in a manner not expressly intended by the providers. While such cross-domain tattoo matching might be considered "off label" usage of what the algorithms were designed for, the results serve as a baseline of current capabilities for users interested in applying tattoo recognition to images collected in this domain. Domain-specific development of tattoo recognition capabilities may yield better accuracy results.

4.4 Detection

This section details the performance of algorithms tasked with detecting whether an image contains a tattoo or not. The tattoo images were collected under moderately-controlled settings and contained one or more tattoos in them. The non-tattoo images were a subset of the Yahoo Flickr Creative Commons 100 Million Dataset (YFCC100M) [19] and included but were not limited to people, objects, scenery, and other common imagery that did not contain any tattoos.

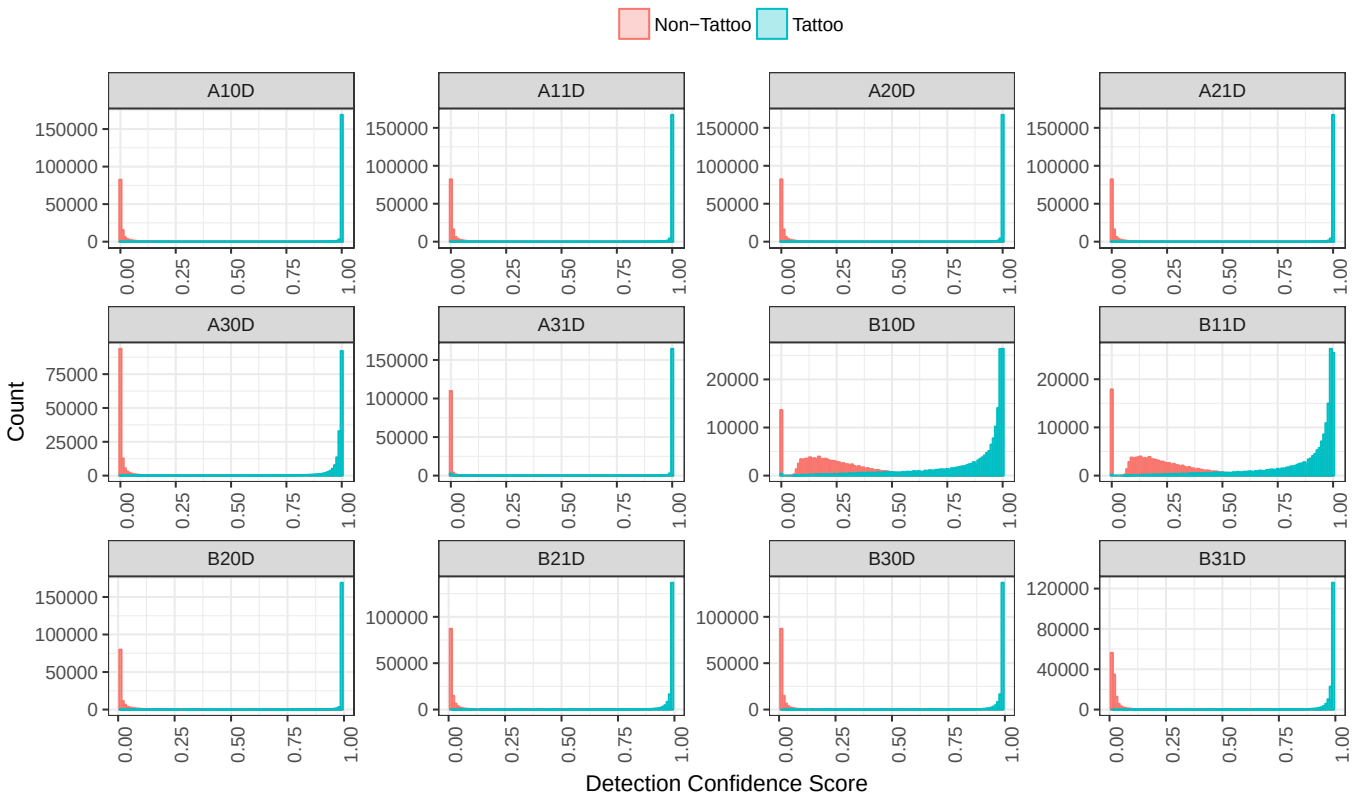


Figure 14: Distribution of tattoo detection confidence scores. The algorithms report confidence values on $[0, 1]$ *for each image describing their algorithm's certainty about whether the image contains a tattoo. Higher values indicate greater confidence that the image contains a tattoo. Number of tattoos: 131 662, Number of non-tattoos: 125 253. Note the different y-axis scales for each algorithm.*

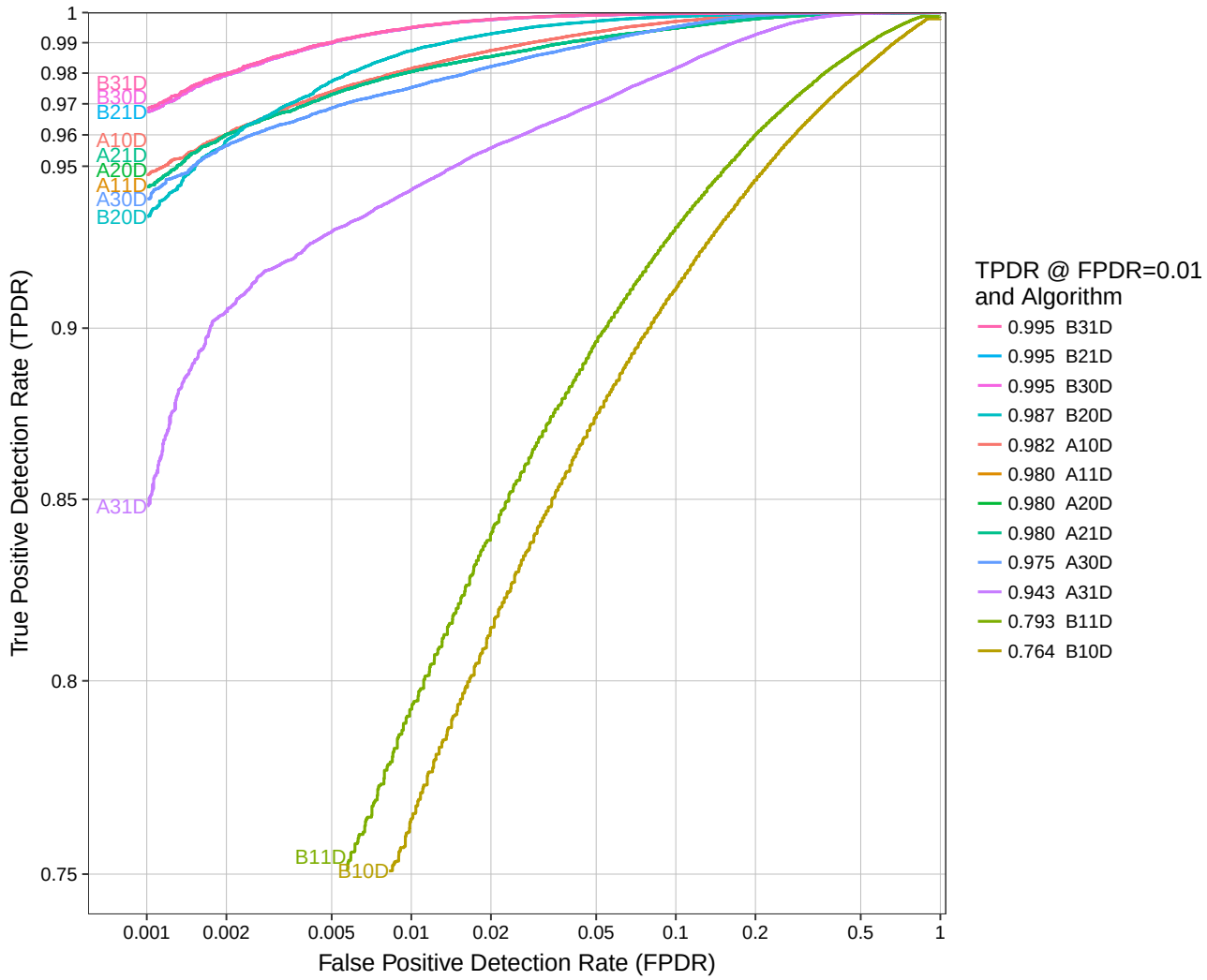


Figure 15: ROC curve based on the tattoo detection confidence score produced by each algorithm over 131 662 tattoo images and 125 253 non-tattoo images.

Results and Notable Observations:

- Figure 15 presents a tradeoff between false positive detection rate (incorrectly classifying a non-tattoo as a tattoo) and true positive detection rate (correctly classifying a tattoo as a tattoo). In a scenario where a system wanted to minimize the number of non-tattoos being classified as a tattoo, (i.e., false positive detection rate), one might reasonably set the false positive detection rate threshold to 1 %, which would yield a true positive detection rate of 99.5 % for the top performing algorithms (B21D, B30D, B31D). Through visual inspection, false positives were often related to images with lots of patterns such as graphics on walls, clothing, artwork, etc.
- Per the confidence score density plots in Figure 14, the more accurate algorithms show a larger separation between tattoo and non-tattoo scores.

4.4.1 NTUTDB

The Nanyang Technological University Tattoo Database (NTUTDB)[3] [20] [21] is a publicly available dataset that is used in academia for benchmarking tattoo detection performance. It contains a total of 10 000 images (5 740 tattoos, 4 260 non-tattoos) collected from Flickr, with diverse image viewpoints and environments with complex backgrounds. The dataset contains different partitions for benchmarking, and the partitions of interest as established in the academic publications contain sets of 2 349 and 10 000 images. Given published results from academia are publicly available for this dataset, we conducted a performance comparison using the Tatt-E algorithms. Table 4 tabulates the performance of the top performing algorithms from each Tatt-E participant against published methods for automatic tattoo detection from the academic literature.

The NTUTB images are all in the public domain so, in principle, the results could be manipulated either via training, or outright memorization. This is unlikely, however, since the algorithm providers had no prior reason to suspect that this dataset would be part of this evaluation.

Publication	NTU (2349)	NTU (10K)
Xu et al. [20]	84.76 %	-
Sun et al. [22]	-	80.66 %
A31D	98.9 %	99.3 %
A10D	97.4 %	97.5 %
B20D	83.8 %	84.5 %
B31D	77.3 %	77.6 %

Table 4: The table compares detection accuracy performance of published academic methods to the top performing algorithm from each Tatt-E participant on NTUTDB.

4.5 Localization

This section details the performance of algorithms tasked with localizing/segmenting one or more tattoos contained in an image. The algorithm outputs bounding box(es) that correspond to the location(s) of the detected tattoo(s). All images in this experiment contained at least one tattoo.

[3]Portions of the research in this paper use the Nanyang Technological University Tattoo Image Database Version 1. Credit is hereby given to the School of Computer Science and Engineering, Nanyang Technological University, Singapore for providing the database.

Algorithm	Mean IoU	Localization Accuracy (IoU threshold > 0.5)	# False Detections	# Ground Truth Bounding Box Detection Failures (out of 12613 gt bb)	# Algorithm Bounding Boxes
A10D	0.719	0.839	370	1345	11982
A11D	0.719	0.839	370	1345	11982
A20D	0.754	0.895	313	797	12049
A21D	0.754	0.895	313	797	12049
A30D	0.754	0.895	313	797	12049
A31D	0.754	0.895	313	797	12049
B10D	0.620	0.698	1864	1793	12856
B11D	0.492	0.546	2098	1315	13794
B20D	0.563	0.611	431	1108	11258
B21D	0.575	0.644	449	677	11680
B30D	0.691	0.890	519	538	12493
B31D	0.675	0.863	533	848	12128

Table 5: Tattoo Localization Accuracy Statistics. Number of images: 10 926, Number of ground truth bounding boxes: 12 613.

Algorithm	1	2	3+
# images	9568	1128	230
A10D	0.859	0.712	0.602
A11D	0.859	0.712	0.602
A20D	0.910	0.805	0.729
A21D	0.910	0.805	0.729
A30D	0.910	0.805	0.729
A31D	0.910	0.805	0.729
B10D	0.715	0.589	0.525
B11D	0.561	0.456	0.379
B20D	0.629	0.500	0.410
B21D	0.664	0.519	0.421
B30D	0.911	0.771	0.625
B31D	0.885	0.729	0.581

Table 6: Tattoo Localization Accuracy, broken out by the number of tattoos in each image. Number of images: 10 926.

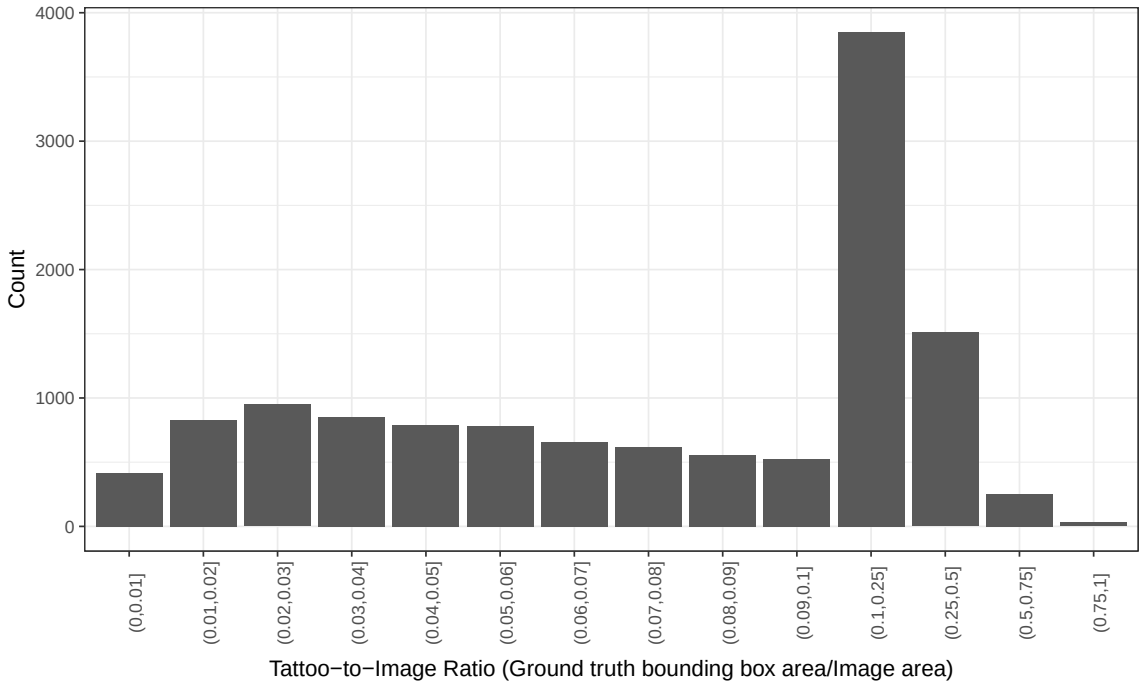


Figure 16: This plot shows the counts of tattoo-to-image ratio for all ground truth bounding boxes used in the localization experiment. Note non-uniform scale on x-axis.

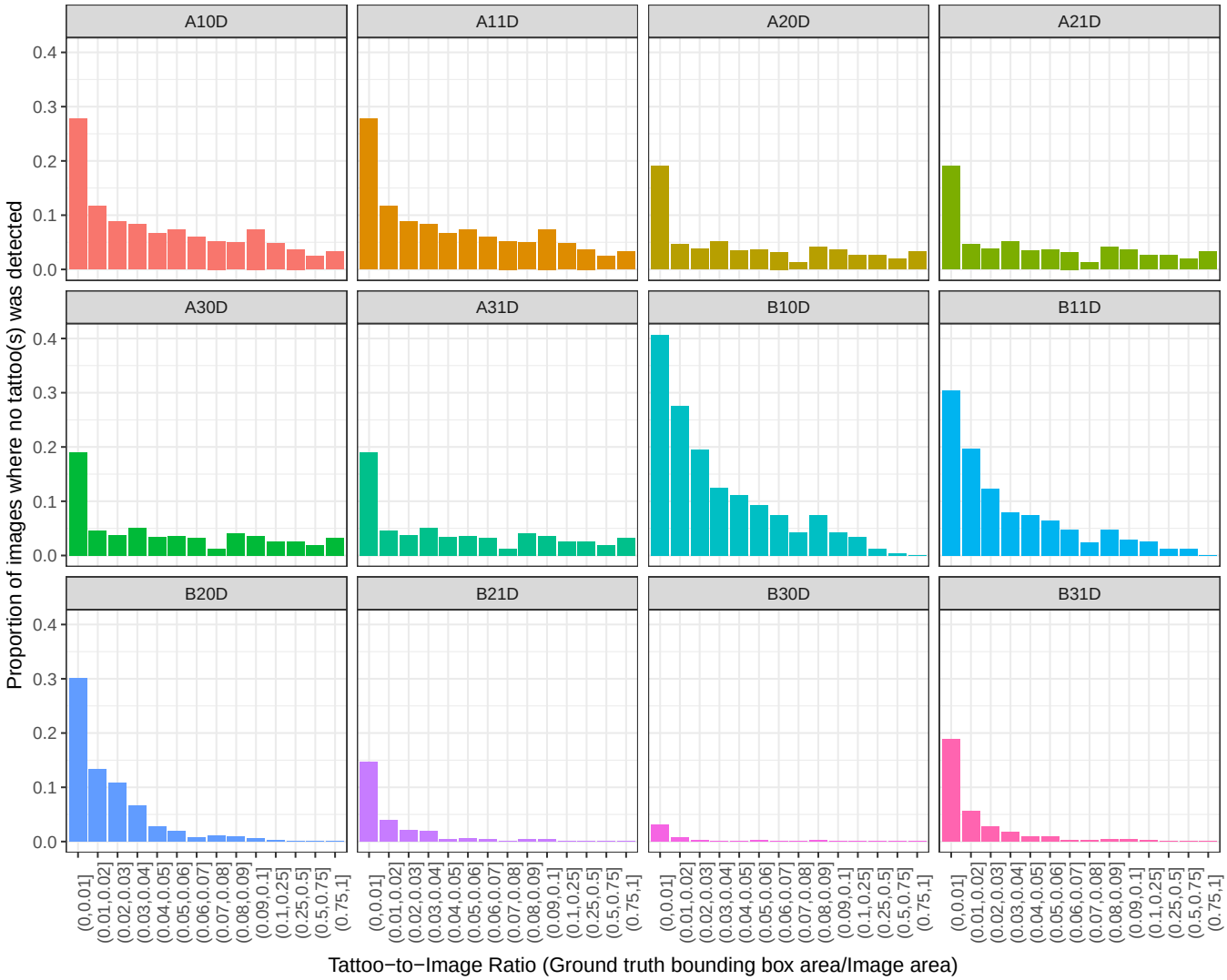


Figure 17: This figure presents the proportion of images where no tattoo(s) was detected by the algorithm (zero bounding boxes were returned by the algorithm) plotted against the tattoo-to-image ratio. Tattoo-to-image ratio is defined as the area of the ground truth bounding box (width X height) divided by the total image area (image width X image height). Note non-uniform scale on x-axis. Number of images processed by each algorithm: 10 926

Results and Notable Observations:

- On a set of 12 613 tattoos contained in 10 926 images, the top performing algorithms (A20D, A21D, A30D, and A31D) achieve localization accuracy of 89.5 %.
- Given an input image, algorithms return zero or more bounding boxes of tattoos detected and segmented. Figure 17 shows the proportion of images where zero bounding boxes were returned (no tattoo was detected/localized in the image), plotted against tattoo-to-image ratio. Tattoos that occupied only a very small percentage (less than or equal to 1 %) of the entire image were the hardest to detect (and therefore, segment).
- Through visual inspection, images that contained a large amount of graphical pattern or lettering on clothing caused algorithms to incorrectly localize as a tattoo, and algorithms often failed to localize on very faded tattoos.

4.6 Template Size

A tattoo template is a proprietary representation of the features extracted from a tattoo image. An enrollment template is what gets stored to disk into a gallery for subsequent searching. A probe template is what gets generated and used to search against what is stored in a gallery. Template size matters as it drives hardware resource needs to operate a tattoo recognition system. Table 7 presents summary statistics on template size, broken out by template and probe type.

Algorithm	
A10I	81920 ± 0
A11I	81920 ± 0
A20I	163840 ± 45486
A21I	163840 ± 45486
A30I	163840 ± 45486
A31I	327680 ± 90972
B10I	176821 ± 234078
B11I	130445 ± 100252
B20I	152069 ± 116917
B21I	117022 ± 95505
B30I	101069 ± 85302
B31I	69381 ± 78771

(a) Enrollment templates

Algorithm	Uncropped	Cropped	Sketches
A10I	81920 ± 0	81920 ± 0	81920 ± 0
A11I	81920 ± 0	81920 ± 0	81920 ± 0
A20I	163840 ± 41658	163840 ± 27291	163840 ± 18536
A21I	163840 ± 41658	163840 ± 27291	163840 ± 18536
A30I	163840 ± 41658	163840 ± 27291	163840 ± 18536
A31I	327680 ± 83315	327680 ± 54582	327680 ± 37073
B10I	133437 ± 79034	24909 ± 44284	19605 ± 21057
B11I	151525 ± 100001	51565 ± 70766	125753 ± 99807
B20I	184709 ± 110841	62309 ± 77729	174713 ± 133943
B21I	143814 ± 90080	69286 ± 75095	178818 ± 133967
B30I	126365 ± 83829	49797 ± 70230	125753 ± 99807
B31I	71965 ± 73729	51565 ± 70290	125753 ± 99807

(b) Search templates, by probe type

Table 7: Median Template Size (in bytes) ± Standard Deviation

Results and Notable Observations:

- Median enrollment template sizes range from 69 381 bytes (B31I) to 327 680 bytes (A30I).
- Most of the algorithms do not have a fixed template size, with the exception of A10I and A11I. Templates sizes appear to be influenced by the amount of data extracted from the image. Template sizes are smaller for cropped images than uncropped images, which means the algorithms may be storing other information extracted from the uncropped images outside of the primary tattoo content.

4.7 Timing

Timing distributions for template generation, search, detection, and localization are presented below. Hardware specifications and details used to measure timing are documented in Section 4.1.6.

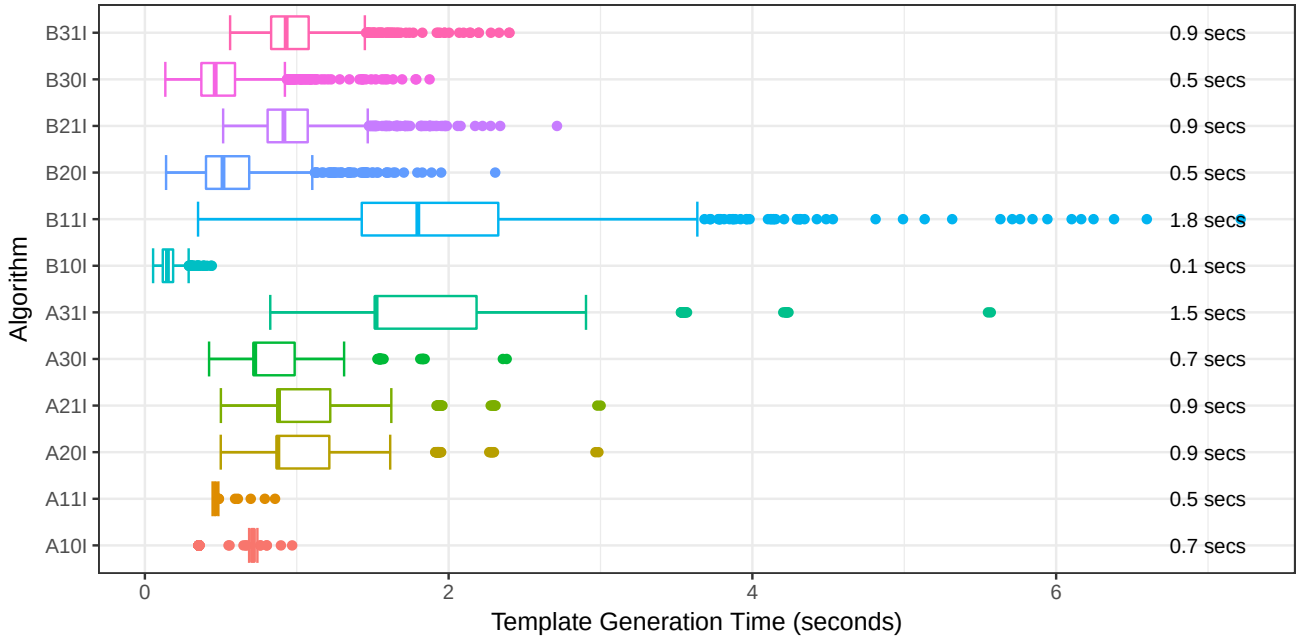


Figure 18: Boxplots of the distribution of template generation times. The values on the right tabulate median template generation time for each algorithm. Plots were generated over 1 000 images. Algorithms had a time limit of 5 seconds for this operation. All A algorithms ran on the GPU, and all B algorithms ran on the CPU.

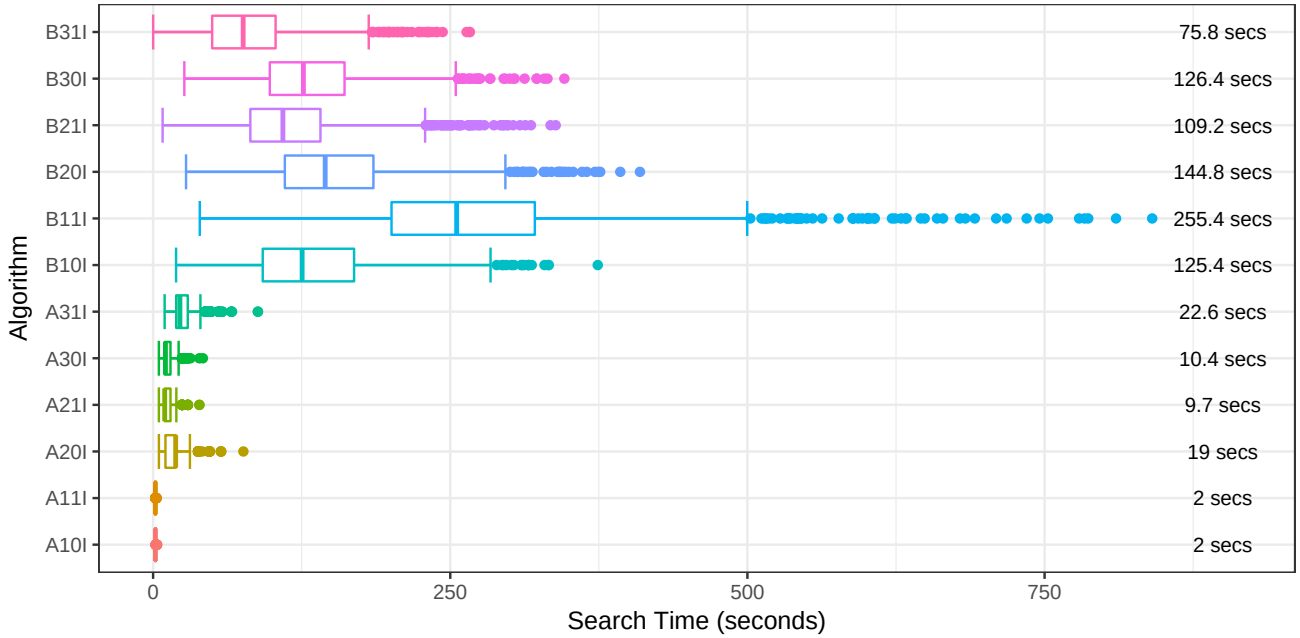


Figure 19: Boxplots of the distribution of search times over an enrollment gallery of size 100 000. The values on the right tabulate median search time for each algorithm. Plots were generated over 1 000 searches with uncropped probes. Algorithms had a time limit of 300 seconds for this operation. All A algorithms ran on the GPU, and all B algorithms ran on the CPU.

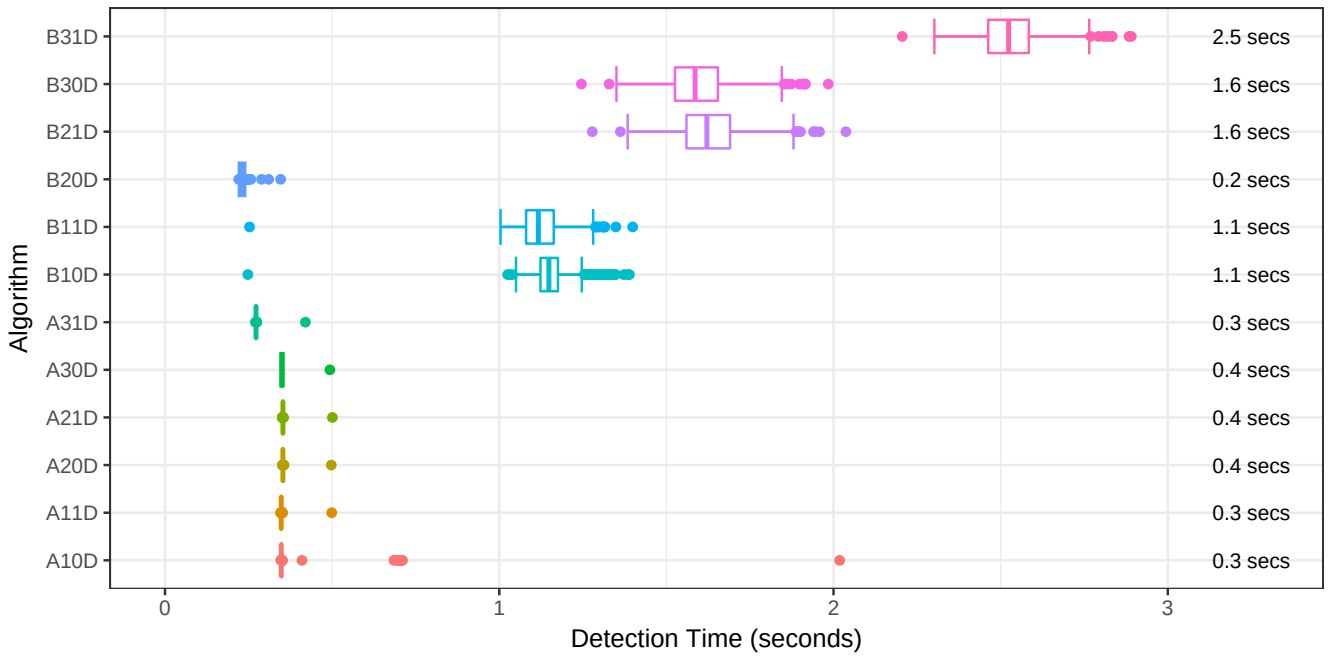


Figure 20: Boxplots of the distribution of detection times. The values on the right tabulate median detection time for each algorithm. Plots were generated over 1 000 images. Algorithms had a time limit of 5 seconds for this operation. All A algorithms ran on the GPU, and all B algorithms ran on the CPU.

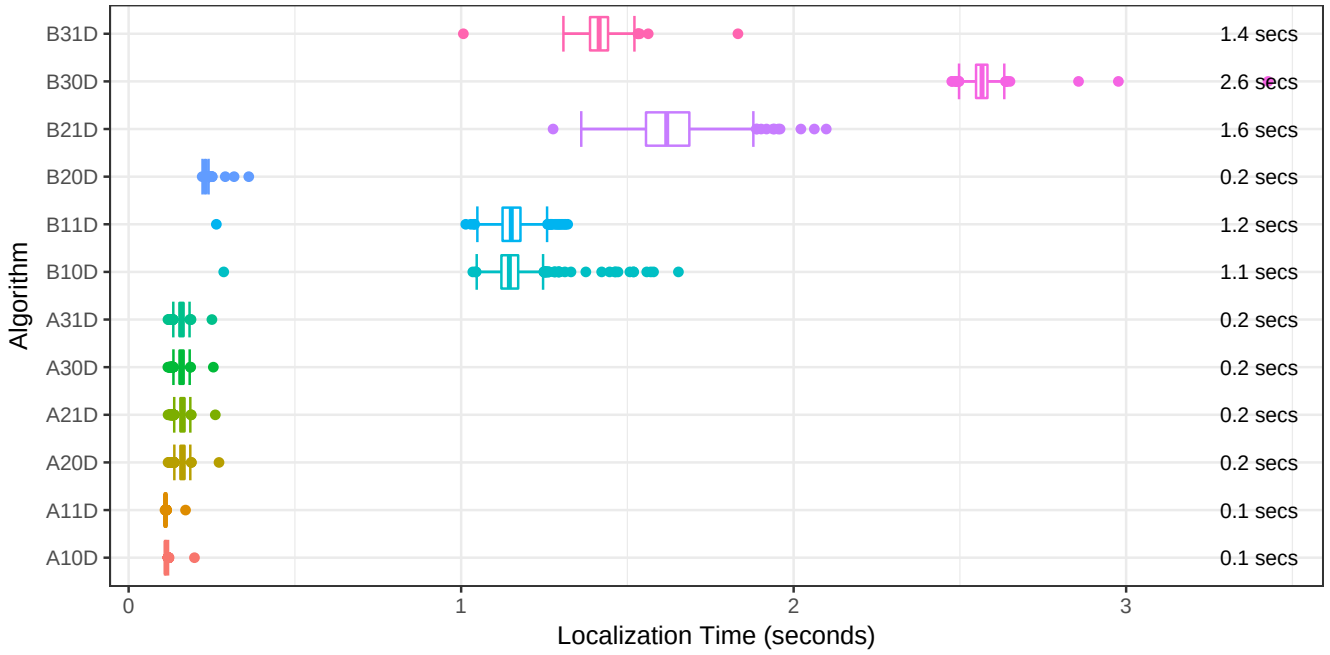


Figure 21: Boxplots of the distribution of localization times. The values on the right tabulate median localization time for each algorithm. Plots were generated over 1 000 images. Algorithms had a time limit of 5 seconds for this operation. All A algorithms ran on the GPU, and all B algorithms ran on the CPU.

5 Topics Not Covered

Based on the outcomes of this test and discussions/suggestions from the tattoo recognition developer and user community (including law enforcement, forensic examiners, and others), studies not covered in this report which warrant consideration by the research community include:

- Multiple image enrollment: assess the impact of multiple image enrollment and whether algorithms can take advantage of fusion of the information across images to enhance accuracy;
- Tattoo ageing: longitudinal study of matching tattoos over time and whether there is an impact on matchability as a tattoo "ages";
- Occlusion study: given tattoos are often occluded by clothing or other items when collected in uncontrolled settings, assess algorithm performance on tattoos that are occluded in different ways;
- Tattoo recognition "in the wild": study performance of tattoo recognition on images collected under unconstrained settings;
- Automated tattoo image quality assessment: establish criteria to measure and assess tattoo image quality with goals of an automated capability to accept or reject tattoo images during the capture process based on a quality measure;
- Tattoo similarity: matching tattoos based on visually similar content;
- Multispectral: explore matching tattoos collected in other parts of the infrared spectrum (e.g. near infrared);
- Algorithm fusion;
- Tattoo recognition in video.

References

[1] Tattoo Recognition Technology - Challenge Website. https://www.nist.gov/programs-projects/tattoo-recognition-technology-challenge-tatt-c.

[2] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv preprint arXiv:1408.5093*, 2014.

[3] Michael A. Olson, Keith Bostic, and Margo Seltzer. Berkeley DB. In *Proceedings of the Annual Conference on USENIX Annual Technical Conference*, ATEC '99, pages 43–43, Berkeley, CA, USA, 1999. USENIX Association.

[4] Open MPI: Open Source High Performance Computing. https://www.open-mpi.org.

[5] ANSI/NIST-ITL 1-2011:Update 2015, 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.

[6] M. Ngan and P. Grother. Tattoo recognition technology - challenge (Tatt-C): an open tattoo database for developing tattoo recognition research. In *IEEE International Conference on Identity, Security and Behavior Analysis (ISBA 2015)*, pages 1–6, March 2015.

[7] M. Martin, J. Dawson, and T. Bourlai. Large Scale Data Collection of Tattoo-Based Biometric Data from Social-Media Websites. In *2017 European Intelligence and Security Informatics Conference (EISIC)*, pages 135–138, Sept 2017.

[8] M. Martin and T. Bourlai. Enhanced Tattoo Image Quality Assessment Through Multispectral Sensing. *IEEE Sensors Letters*, 1(6):1–4, Dec 2017.

[9] Mei Ngan, George W. Quinn, and Patrick Grother. NISTIR 8078 - Tattoo Recognition Technology Challenge (Tatt-C) Outcomes and Recommendations, September 2016. http://nvlpubs.nist.gov/nistpubs/ir/2015/NIST.IR.8078.pdf.

[10] Mei Ngan, George W. Quinn, and Patrick Grother. NISTIR 8109 - Tattoo Recognition Technology Best Practices (Tatt-BP) Guidelines for Tattoo Image Collection, Revision 1.0, September 2016. https://nvlpubs.nist.gov/nistpubs/ir/2016/NIST.IR.8109.pdf.

[11] Mei Ngan and Patrick Grother. Tattoo Recognition Technology - Evaluation (Tatt-E) Concept, Evaluation Plan, and API, December 2016. https://www.nist.gov/programs-projects/tattoo-recognition-technology-evaluation-tatt-e.

[12] The CentOS Project. https://www.centos.org.

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

[14] J. C. Nascimento and J. S. Marques. Performance evaluation of object detection algorithms for video surveillance. *IEEE Transactions on Multimedia*, 8(4):761–774, Aug 2006.

[15] H. Yi, P. Yu, X. Xu, and A. W. K. Kong. The impact of tattoo segmentation on the performance of tattoo matching. In *2015 IEEE International WIE Conference on Electrical and Computer Engineering (WIECON-ECE)*, pages 43–46, Dec 2015.

[16] Eli Peli. Contrast in complex images. *J. Opt. Soc. Am. A*, 7(10):2032–2040, Oct 1990.

[17] JPEG File Interchange Format. https://jpeg.org/jpeg/.

[18] H. Han and A. K. Jain. Tattoo based identification: Sketch to image matching. In *2013 International Conference on Biometrics (ICB)*, pages 1–8, June 2013.

[19] Bart Thomee, David Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. The New Data and New Challenges in Multimedia Research. 03 2015.

[20] Qingyong Xu, S. Ghosh, X. Xu, Yi Huang, and A. W. K. Kong. Tattoo detection based on CNN and remarks on the NIST database. In *2016 International Conference on Biometrics (ICB)*, pages 1–7, June 2016.

[21] N. Q. Huynh, X. Xu, A. W. K. Kong, and S. Subbiah. A preliminary report on a full-body imaging system for effectively collecting and processing biometric traits of prisoners. In *2014 IEEE Symposium on Computational Intelligence in Biometrics and Identity Management (CIBIM)*, pages 167–174, Dec 2014.

[22] Z. H. Sun, J. Baumes, P. Tunison, M. Turek, and A. Hoogs. Tattoo detection and localization using region-based deep learning. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 3055–3060, Dec 2016.