NIST Special Publication 500
NIST SP 500-343

Face Recognition Technology Evaluation (FRTE)

Preparation of Compact Face Images for 2D Barcodes

Patrick Grother
Mei Ngan
Austin Hom

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.500-343

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
U.S. DEPARTMENT OF COMMERCE

NIST Special Publication 500
NIST SP 500-343

Face Recognition Technology Evaluation (FRTE)

Preparation of Compact Face Images for 2D Barcodes

Patrick Grother
Mei Ngan
Austin Hom
Information Access Division
Information Technology Laboratory

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.500-343

September 2025

U.S. Department of Commerce
Howard Lutnick, Secretary

National Institute of Standards and Technology
Craig Burkhardt, Acting Under Secretary of Commerce for Standards and Technology and Acting NIST Director

NIST Technical Series Policies
Copyright, Use, and Licensing Statements
NIST Technical Series Publication Identifier Syntax

Publication History
Approved by the NIST Editorial Review Board on 2025-09-24

How to cite this NIST Technical Series Publication:
Grother PJ, Ngan M, Hom A (2025) Face Recognition Technology Evaluation (FRTE). (National Institute of Standards and Technology, Gaithersburg, MD), NIST SP 500-343. https://doi.org/10.6028/NIST.SP.500-343

Author ORCID iDs
Patrick Grother: 0000-0003-1996-4363
Mei Ngan: 0009-0008-3438-7756
Austin Hom: 0009-0000-4890-199X

Contact Information
frvt@nist.gov

Abstract

Compact face photographs, stored and digitally signed in 2D barcodes, form a low-cost tamper-resistant identity credential. The cryptographic signature allows binding of the 2D barcode to the issuer, and biometric authentication binds a live photograph of the user to the photo stored in the barcode. This report documents options for the detection, cropping, resizing, and compression steps needed to instantiate a barcode-ready photograph. The compression step is necessarily lossy, removing fine spatial structure from the face photograph. The report documents the accuracy consequences and provides assurance that fully open standards-based compact faces can be verified with very high accuracy.

Keywords

Biometrics; Face Recognition; Verifiable Credentials; 2D Barcodes.

Table of Contents

List of Tables

List of Figures

Preface

This work has been conducted to support the expanding market for storage of digitally signed face photographs on two dimensional barcodes. In particular it is intended to give quantitative support to a formal consensus standard for preparation of compact face images. Such a document would reside in a layered set of standards that include compression, cryptographic, data serialization, and barcode standards. The authors are grateful for discussions with experts at ID4Africa, ICAO, and the ISO/IEC JTC 1 SC 37 Biometrics subcommittee for encouraging this work.

Acknowledgments

The authors are grateful for the support and collaboration of the Department of Homeland Security's Science & Technology Directorate (S&T) and the Office of Biometric Identity Management (OBIM).

Additionally, the authors are grateful to staff in the NIST Biometrics Research Laboratory for infrastructure supporting rapid evaluation of algorithms.

Institutional Review Board

The National Institute of Standards and Technology's Research Protections 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).

Disclaimer

Certain equipment, instruments, software, or materials, commercial or non-commercial, are identified in this paper in order to specify the experimental procedure adequately. Such identification does not imply recommendation or endorsement of any product or service by NIST, nor does it imply that the materials or equipment identified are necessarily the best available for the purpose.

Author Contributions

Patrick Grother: Conceptualization, Methodology, Drafting.
Mei Ngan: Data curation, Software, Execution.
Austin Hom: Research, Computational Cost, Drafting.

Executive Summary

Role of biometrics on standardized 2D barcodes: QR codes (a type of 2D barcode) carrying digitally signed biometric data form a low cost identity credential suitable for printing or display. The digital signature, once verified, binds the credential to its issuer; the biometric data binds the credential to the person presenting it. With additional application-specific metadata, error correction data, and the signature, the biometric on a 2D barcode needs to have a size of around one kilobyte. Overall capacity is limited by small physical dimensions and limited display resolution.

Compact faces images: This report addresses the viability of a biometric payload that is a typical face portrait but compressed to about 1 kilobyte. The compression is lossy, removing fine detail from the image. The report documents a large empirical trial data showing that such faces can be verified against subsequent authentication photos with minimal accuracy loss relative to typical non-compact passport-style photographs. The report shows *how* this is achieved using open and formally standardized tools.

The report also guides photographers and their subjects to assist compression.

Human role in authentication: In typical operational settings, human review is needed to adjudicate whether a rejection is either a (quality-related) biometric failure or the correct rejection of an impostor. Human review is known to be error prone, and will be degraded further with the use of compact images. The report uses score-distribution metrics to finely quantify damage inherent in the use of lossy compression.

Images of other biometric modalities: The biometric payload considered here is a human-interpretable face photograph. Analogous data elements derived from fingerprint images, irises, or other modalities have not yet been evaluated and standardized at the compact sizes needed for QR codes.

Biometric templates: Recognition algorithms operate by comparing templates extracted from images. While templates are sometimes small enough for use on QR codes, they are generally non-interoperable - a single developer's software must be used for their generation and verification. The one notable exeception to this is the ISO/IEC 39794-2 standardized fingerprint templates whose size is well within the capacity of a QR code, and whose interoperability is well documented.

1. Background

Biometric data has been stored in portable form for many decades originally in printed media and, more recently, electronically. The data is usually a photograph of a face or fingerprint, but it can also be feature data extracted from such samples, such as a fingerprint minutia template.

The virtue of storing photographs is that they are universally interpretable by humans, and consumable by biometric algorithms: they offer interoperability. In contrast template data is non-interoperable; it is produced and consumed by paired algorithms from one (commercial) developer. An exception is the multi-party ecosystem of algorithms producing and consuming instances of the ISO/IEC 39794-2:2023 fingerprint minutia record as tested in NIST MINEX.

In recent years biometric data has been encoded into 2D barcodes. The barcodes are defined in formal standards including those for ISO/IEC 18004:2024 QR code, ISO/IEC 16022:2024 Data Matrix, the ISO/IEC 24778:2024 Aztec code, and the ISO/IEC 15438:2015 PDF417 code. Those standards allow encoding of arbitrary data (i.e. any sequence of bytes). They are most useful when that data is itself defined by a standard, for example a URL conformant to RFC 3986, or a structured verifiable credential such as ICAO's digital seals or MOSIP's claim 169. The key feature of the 2D barcodes is that they can include cryptographic signatures of the data. These ensure data integrity such that an attacker cannot modify the payload. This is useful for an identity credential - the signature strongly binds the QR code to the issuer, and the biometric element binds the QR code to the holder.

However, the 2D barcodes have capacity limited by two related factors. One is resolution: all media whether ink-on-substrate or mobile phone electronic display have finite resolution such that 2D barcodes have high, but limited, density. This usually cannot be avoided by using larger media due to the second factor: limited space media, for example the dimensions imposed on credit or ID cards by ISO/IEC 7810 ID-1 requirements.

2. Preparation of compact images

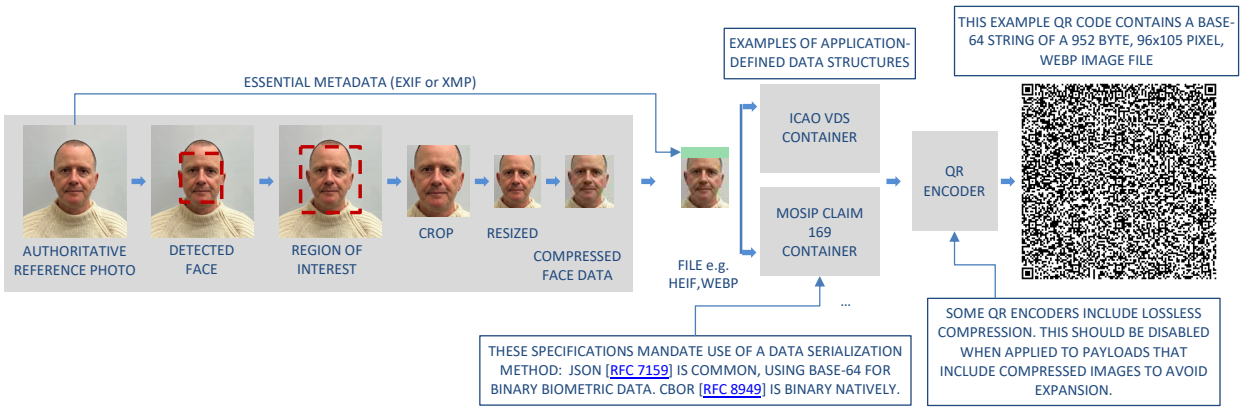

Fig. 1. Face recognition using QR code-based credentials can be built on a set of formal standards. These include standards for compression, credentials, data serialization, integrity protection, and 2D barcode representation. Note most mobile phones are not (yet) equipped to read this example.

This report quantifies the biometric accuracy available when preparing face photographs to sizes that can be stored in a 2D barcode. The preparation requires face detection, cropping, scaling to reduce width and height, and lossy compression as depicted in Figure 1. Thereafter, the compressed face is paired with application-specific metadata (e.g. travel document [ICAO]) and encoded (e.g. with JSON text or more compactly, CBOR binary) then finally encoded as a 2D barcode (a QR code, say, using open source for example).

Detection: Face detection is necessary to support subsequent cropping that removes extraneous background information. In our trials we used the dlib face detector; in practice any tested and capable detector could be used. In many operations a trained human could perform this step also, given a suitable user interface.

Cropping: Cropping is simply production of a rectangular subimage from the parent image; many image handling tools support this. The process begins with a definition of the region. In our case we cropped the region defined by a symmetric 25% expansion of the bounding box returned by dlib. We did not survey over tighter or looser crops.

Scaling: Reduction of the size of the input photograph is achieved using an implementation, Pillow, that preserves aspect ratio and uses bicubic interpolation. Critically it also blurs the image before resizing to mitigate aliasing effects. Without this, the resized image can include aliasing artifacts that will burden the compression step that follows.

Table 1. Table of compression codecs available for compact image preparation. JPEG-XL was not considered in this report.

Codec	Description	.Ext	Licensing	Standard	Source Code
JPEG	The ubiquitous JPEG implementation	.jpg	GNU General Public License	ISO/IEC 10918	link
JPEG 2000	Available open source through openjpg but typically IP encumbered	.j2k	2-clause BSD license	ISO/IEC 15444	link
JPEG-LI	Built off of JPEG-XL by Google	.jxl .jpg	3-clause BSD license	None	link
JPEG-XL	Focused on better quality and compression ratios than jpeg	.jxl	3-clause BSD license	ISO/IEC 18181	link
HEIC	Built on High Efficiency Image File Format (HEIF) with HEVC encoding	.heic	GNU Lesser General Public License	ISO/IEC 23008-12:2017	link
AVIF	Key-frame encoding from AV1 video compressor, also uses HEIF file format	.avif	3-clause BSD License, royalty free	Alliance for Open Media	link
WEBP	Raster graphics file format developed by Google built on VP8 video codec	.webp	3-clause BSD License, royalty free	IETF RFC 9649	Google

Compression:

The key tool in producing small biometric samples is **standardized compression**. Table 1 shows the image compression algorithms that are viable in this study. We did not consider lossless compression, because such schemes do not produce the output encoding sizes needed for 2D barcodes. Instead we applied lossy

compressors. These take an image quality parameter as input, with lower values producing encoded images with smaller size and more information loss.

We applied the compressor to the cropped and resized images described later in section 3. With the aim of producing an encoding at or below one of the test target sizes (600, 800, 960, 1040, 1200 bytes), we called the compressor iteratively specifying successively lower quality values until the target size was reached. Some compressors - for example, JPEG 2000 - can be invoked with a target size as an input parameter avoiding this iterative descent method.

Thus far, due to labor availability limitations, we have not considered JPEG-XL which has previously been shown to outperform JPEG 2000[1] at encoding sizes down to 2.2 kilobytes.

Similarly we have not investigated detailed sets of parameters - compression profiles - tailored to compression of face images. This work may be worthwhile, although any departure from the open-source code defaults would need to be implemented in operational use.

3. Datasets

This report uses a set of mugshot photographs as follows.

- The number of reference images is 465352; these are the input to the compact-image preparation methods.
- The number of verification images is 1251122; these are not used in compact-image preparation.
- The number of subjects is 465352; the number with two or more images is 465352.
- The images have geometry in reasonable conformance with the ISO/IEC 19794-5 Full Frontal image type. Small minorities of images are underexposed, or have downward head pitch angles.
- The images are of variable sizes. The median inter-eye distance is 105 pixels. The mean IOD is 113 pixels. The 1-st, 5-th, 10-th, 25-th, 75-th, 90-th and 99-th percentiles are 34, 58, 70, 87, 121, 161 and 297 pixels.
- The images are JPEG compressed. The 1-st, 5-th, 50-0th, 95-th and 99-th percentiles are 13510, 17673, 33774, 114602, and 227471 bytes.
- The images are of adults collected in the United States.
- The images are all live capture.
- When these images are input to the algorithm, they are labelled as being of type “FaceMugshot” - see the image descriptions in the Face Recognition Technology Evaluation (FRTE) API interface header.

4. Experiments and metrics

We evaluate compact image preparation methods in terms of face recognition accuracy. Any procedure that reduces spatial resolution will generally reduce facial similarity and therefore yield more false negative outcomes. It may also remove discriminative detail thereby producing false positives. When comparing images of different people, a reduction in resolution can increase or decrease false positive outcomes,

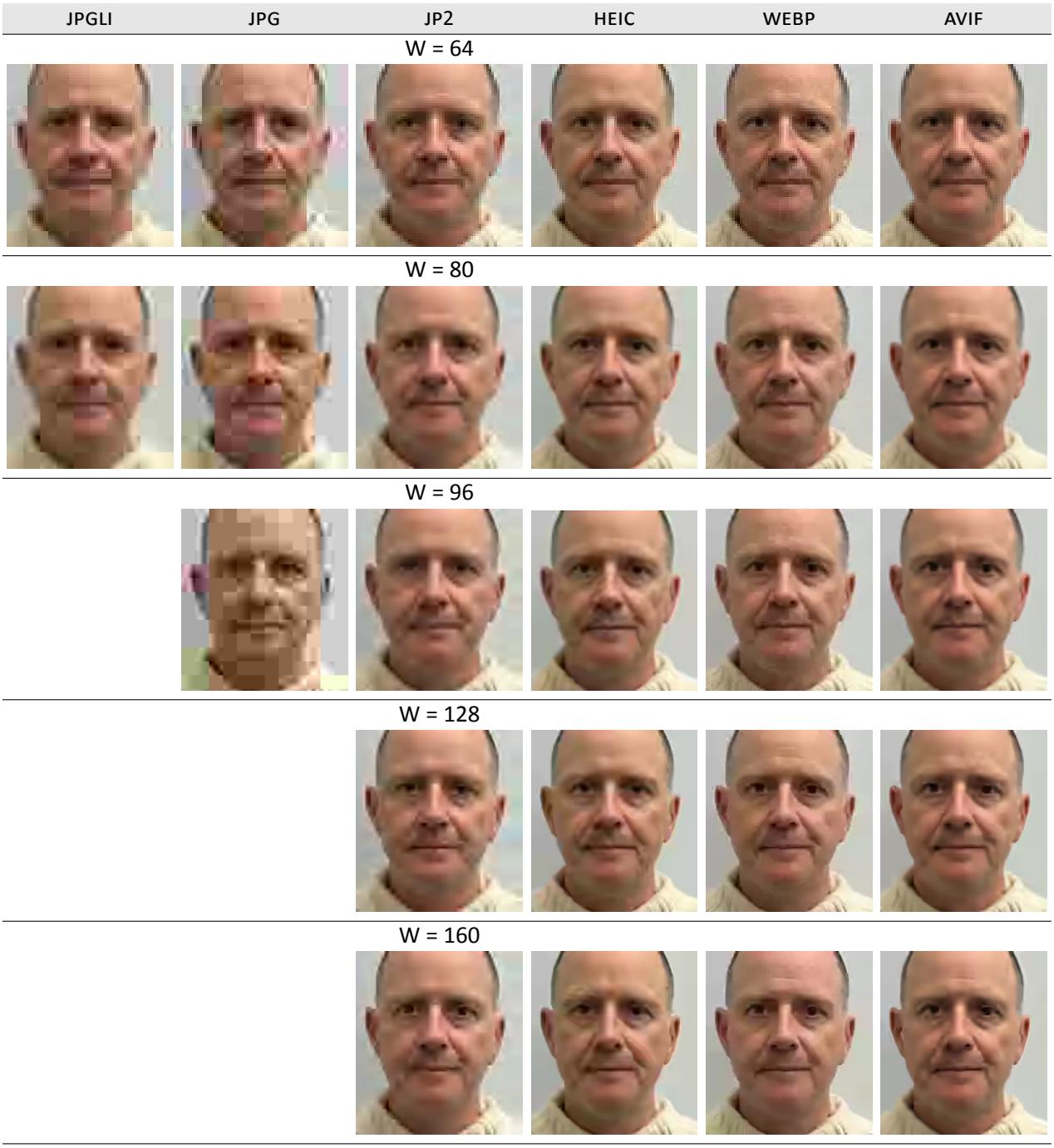

Fig. 2. For a fixed encoded size of 960 bytes, the figure compares the compression algorithms given in the column headers. The images are displayed here at the same width, but the encoded image has the width given in the row header text. Missing entries occur when the given compression algorithm was unable to produce an encoding with size below 960 bytes.

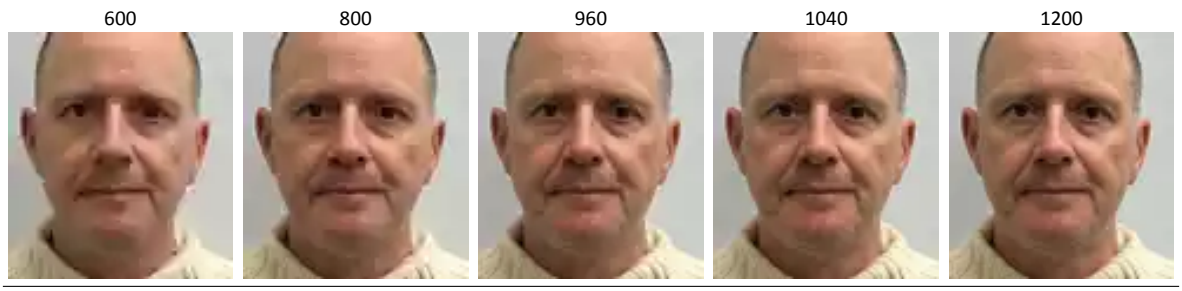

Fig. 3. The figure shows the use of the webp compressor applied to achieve encoded sizes below, from left to right, 600, 800, 960, 1040, and 1200 bytes. The encoded image has width 96 pixels in all cases.

depending on the whether the particular recognition algorithm regards compressed faces as more similar, or instead discounts the loss in facial information.

For each compact image preparation method we repeat the mugshot-mugshot trial that has been used in the FRTE 1:1 evaluation since 2018. It involves executing $M = 1\,251\,122$ mated comparisons and $N = 93\,070\,400$ non-mated comparisons. Each comparison is between one compact image prepared from a reference mughot, and one unprocessed authentication mugshot - see section 3. There are $N_{\mathsf{REF}} = 465\,352$ reference images. The authentication images are used in a variable number of mated comparisons and a fixed number (200) of non-mated comparisons.

For each compact-image preparation method, we process the N_{REF} images independently. This can fail at two stages. First, errors could occur due to the failure of our face detector (dlib) in producing a bounding box. This occurs with $N_{\mathsf{DET}} = 1114$ (0.24%) of the reference images. This failure rate is common to all compact image preparation methods.

The second stage of failure is when the lossy compression algorithm is unable to produce an encoded output sized at or below the target size. When a compressor is applied to an image of width, W, to encode it below a target size, T, there is a non-zero failure rate that depends on T, W, and the image itself, particularly its spatial frequency properties. The number of failures is N_{WT}.

From the $N_{\mathsf{REF}} - N_{\mathsf{DET}} - N_{\mathsf{WT}}$ images that are ultimately produced, we run face recognition against the fixed set of unprocessed authentication images to produce K mated similarity scores. We then compute accuracy as the number of scores, s, at or above threshold, τ:

$$\mathsf{TAR}(\tau) = \frac{1}{M}\sum_{i}^{K} H(s_i - \tau) \tag{1}$$

where $H(x)$ is the unit step function whose value is 1 for $x \geq 0$ and 0 otherwise. The denominator, M, is the number of comparisons we would have made if all images had been produced. Generally $K \leq M$. We then state a false reject rate (FRR) to be consistent with other NIST reports that plot error rates.

$$\mathsf{FRR}(\tau) = 1 - \mathsf{TAR}(\tau) \tag{2}$$

Note that this is not a false non-match rate (FNMR), because it includes errors from detection and compression beyond one-to-one biometric matching. This allows comparison of compact image preparation

schemes. FNMR is computed also as the proportion of comparisons yielding scores below threshold

$$\text{FNMR}(\tau) = 1 - \frac{1}{K}\sum_{i}^{K} H(s_i - \tau) \tag{3}$$

Necessarily FNMR $\leq$ FRR.

4.1. Comparing algorithms on false match response

To determine whether recognition algorithms produce false positives from compressed images we compute a pure false match rate (FMR). This is the proportion of non-mated comparisons that incorrectly associate two individuals with score, v, above threshold.

$$\text{FMR}(\tau) = \frac{1}{L}\sum_{i=1}^{L} H(v_i - \tau) \tag{4}$$

where the denominator L is the number of non-mated comparisons executed using the available images; generally $L \leq N$ as some comparisons are not made due to detection and compression failures. FMR is useful for comparing recognition algorithms; it is not used to compare compact-image preparation schemes.

4.2. Thresholds

The threshold, τ, can take on any value, and FRR, FNMR, and FMR can be estimated and plotted at many thresholds. Here the thresholds are quantiles of the observed non-mate scores produced in the main FRTE 1:1 trial, running on non-compact images. Given an interesting false match rate range, $[\text{FMR}_L, \text{FMR}_U]$, we produce K thresholds correspond to FMR measurements evenly spaced on a logarithmic scale. This is done because typical operations target false match rates spanning several decades e.g., from 10^{-6} to as high as 10^{-2}.

$$\tau_k = Q_v(1 - \text{FMR}_k) \tag{5}$$

where Q_v is the quantile function, and FMR_k comes from

$$\log \text{FMR}_k = \log \text{FMR}_L + \frac{k}{K}\left[\log \text{FMR}_U - \log \text{FMR}_L\right] \tag{6}$$

using base-10 logarithms.

4.3. Metrics summarizing score distribution shifts

Compression changes facial appearance. This reduces mated similarity scores. As an alternative to biometric error rates, which summarize the left tail of the mated distribution, we can compute statements of distributional shifts. The most common is the Kullback-Leibler Divergence (KLD). A simpler one is the Earth Mover's Distance (EMD), also known as the Wasserstein-1 distance. Both measures summarize the shift of the entire mated (or nonmated) score distribution.

KLD: This measure is probability-weighted sum of log likelihood ratio between a reference distribution $p(s)$ - here the score distribution from non-compact images - and a second distribution $q(s)$ - here the score

distribution from compact images. Small values arise when the two distributions overlap. Larger values indicate the surprise in observing samples from q instead of p.

$$D_{KL} = \int_{-\infty}^{+\infty} p(s) \log \frac{p(s)}{q(s)} ds \tag{7}$$

Given scores from the two trials (non-compact and compact), we compute the empirical range, (s_l, s_h), divide it into 512 bins, and compute densities for each distribution whence

$$D_{KL} = \sum_{i=1}^{512} p_i \log \frac{p_i}{q_i} \tag{8}$$

where if $p_i = 0$ the summand is 0, and divide by zero is avoided by using $max(q_i, \varepsilon)$ with $\varepsilon = 10^{-6}$ in the denominator. The choice of ε matters if p_i is non-zero.

EMD: If the cumulative distribution functions (CDFs) of the compact and non-compact mated scores distributions are $F(s)$ and $G(s)$ respectively, the EMD is the area between the two CDFs

$$D_{EM} = \int_{-\infty}^{+\infty} |F(s) - G(s)| ds \tag{9}$$

As the integral is taken across the entire range of s, which varies by recognition algorithm, the values of EMD are algorithm-specific with the result that EMD cannot be used to compare algorithms. However, it is useful here for comparing compact image preparations. EMD is 0 when the CDFs are identical. Compression usually reduces mated scores so typically $F < G$.

5. Results

5.1. Failure to process

All of the steps involved in a preparation of a compact image can fail, meaning that they do not produce an output. We noted failure of the face detector in section 4. The fraction of images was small (0.002) and arises from the presence of low quality images in our test set.

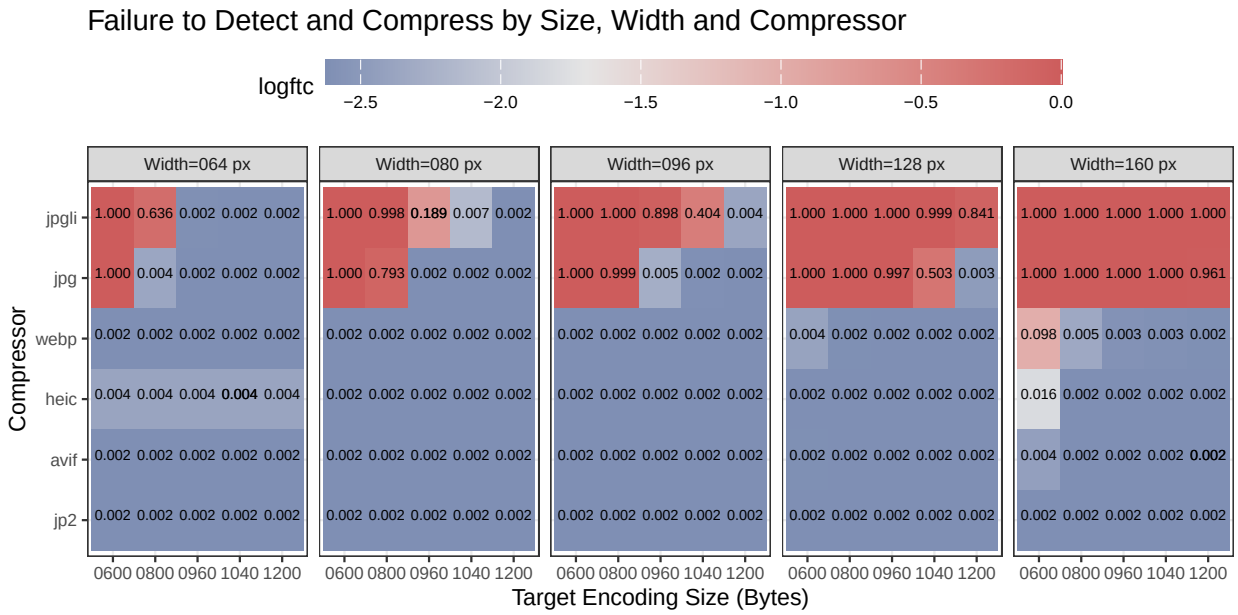

Fig. 4. The heatmap shows the proportion of enrollment images that do not yield a compact image either because our use of the dlib face detector failed or because the compression algorithm could not achieve an encoded size below the target value given on the horizontal axis. These failures precede application of recognition algorithms. The color encodes base-10 logarithm of the text given in each cell. Note that the two JPEG variants are only successful for larger encodings of smaller photographs.

Figure 4 shows overall failure rates - from detection and compression - for all compressors, cropped image widths, and target file sizes. Note particularly that JPEG and JPEG-LI fail except on smaller images ($W \in \{64, 80, 96\}$) and larger target sizes. For the more modern codecs, failure rates are usually very low, except some slight elevation is evident for WEBP at large widths (W = 160) and small target sizes ($T \in \{600, 800\}$), and for HEIC at W = 64 or 160 pixels.

Compressors fail more frequently. Any particular implementation will fail if it is tasked with producing an image of very low size; the quality parameter, particularly, cannot be reduced to a value where the target file size is reached. This occurs early in images with “clutter” i.e. high frequency and high contrast content the encoding of which requires many bits. Figure 5 shows some examples of images for which the WEBP compressor failed to produce encoding of size at or below 960 bytes from an input crop of width 160 pixels.

Fig. 5. From NIST Special Database 32, redacted examples of good quality photographs that stress lossy compression algorithms. The images once cropped will still contain non-facial detail - hair, beards, high-contrast edges, and textured clothing - on and near the face that requires the compressor to dedicate valuable bits from its size budget to encode.

This observation implies guidance to photographers to ask subjects, when possible, to remove clothing around the neck and also hats and glasses.

5.2. Failure to match

Figure 6 shows overall failure rates when preparing compressing and recognizing compact mugshots against non-compact mugshot verification images. Figure 7 shows false non-match error rates for recognition of those images that survive the compact-image preparation step.

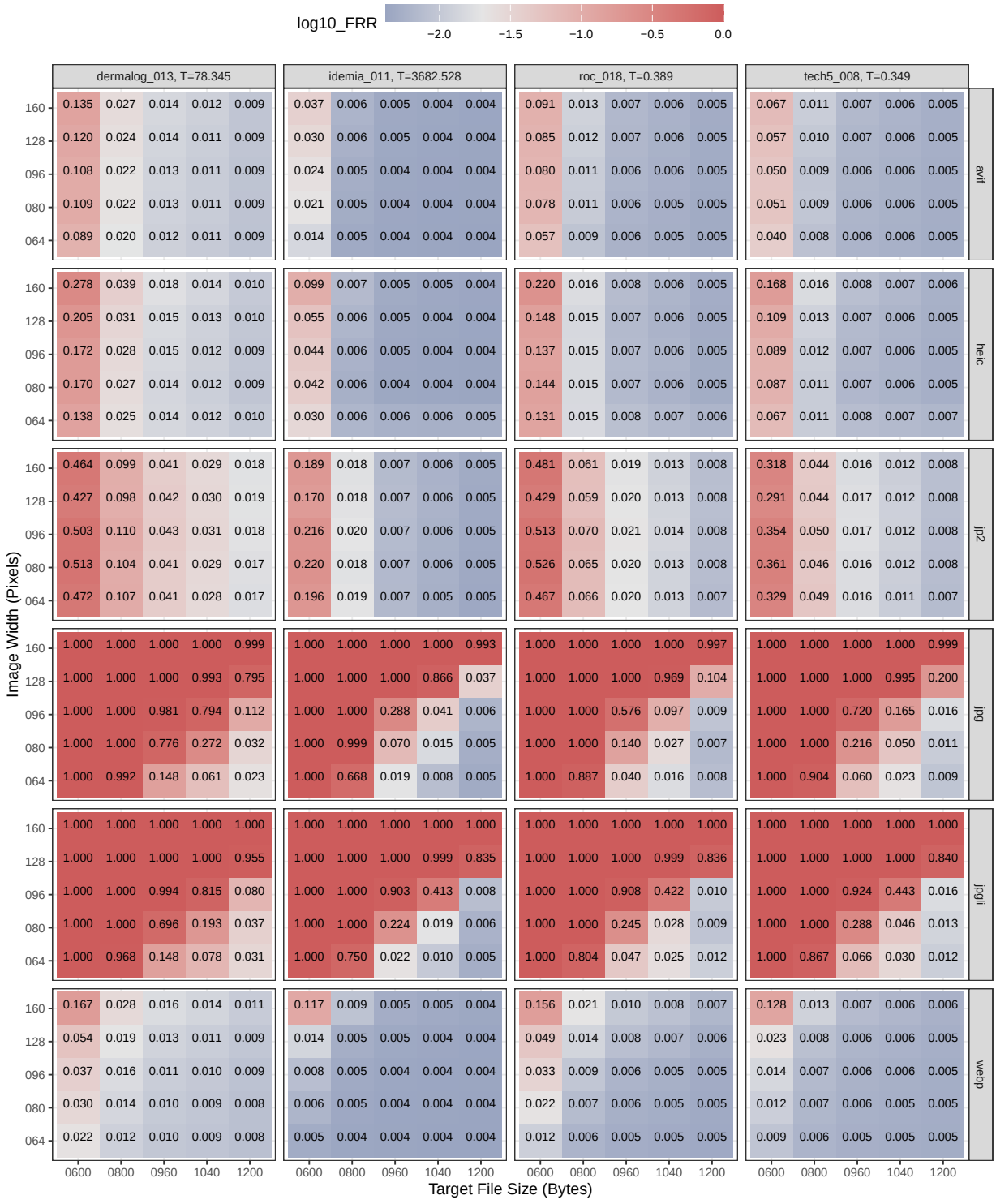

Fig. 6. The heatmap compares compact image preparation methods by showing the combined effects of detection, compression, and recognition failure - see the discussion in section 4. The text values are empirical FRR estimates. The colors encode $\log_{10}$ FRR. At best, FRR exceeds figure 7 by the proportion of the dlib detection failures, 0.0024.

Together the figures show:

- ▷ **JPEG** The ubiquitous JPEG algorithm, and its JPEG-LI derivative, either cannot produce compact images of the required size, or cannot maintain low false negative recognition outcomes. These compressors are deprecated for use in preparation of faces for 2D barcodes.
- ▷ **Other compressors** The lowest recognition error rates are produced by the WEBP compressor, followed by AVIF, HEIC, JPEG 2000 and the two JPEG variants. This ranking applies across all recognition algorithms, all target file sizes, and image widths tested. The only exception to this is that AVIF slightly out-performs WEBP at sizes 800 and 960 bytes and large image widths (128, 160 pixels).
- ▷ **Attainable sizes** At 600 bytes, the smallest encoding size we assessed, we observe elevated false negative errors from all compressors and all compact image widths. At 800 bytes, error rates, while lower than those observed at 600 bytes, are still detectably elevated. At 960 bytes, recognition error rates are close to that achieved when processing non-compact images. The section 5.3 analysis of comparison scores shows that even with low error rates, recognition is being undermined at such sizes.

A different visualization of accuracy loss appears in Figure 8 which plots for all threshold values the FNMR from a compact-image preparation method against FNMR from the unprocessed images. In almost all cases, FNMR is increased, so all traces lie above the no-change diagonal. Moreover the traces rarely cross, so that different compact image preparation methods are preferred at particular thresholds.

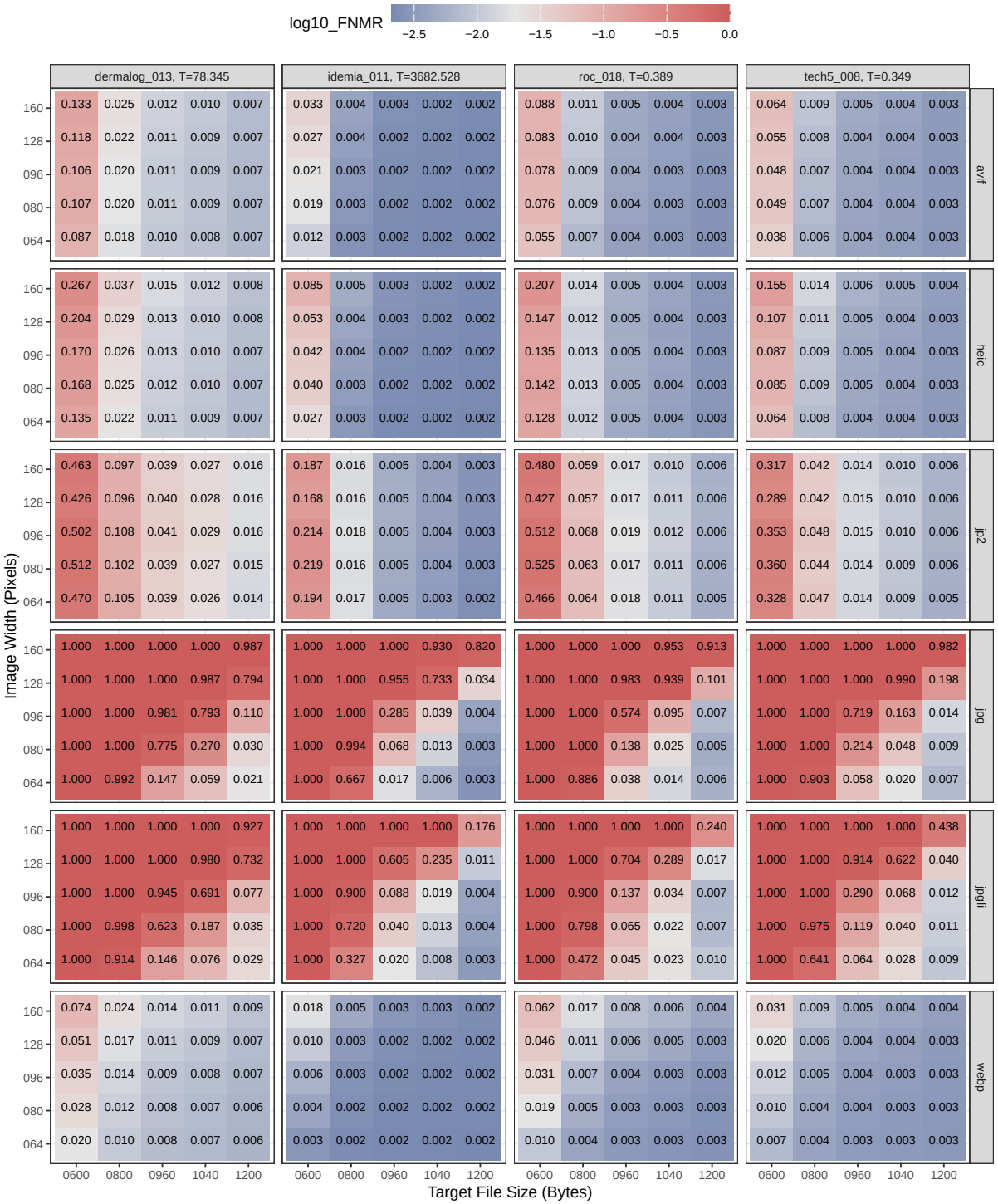

Fig. 7. The heatmap compares compact image preparation methods by showing the false non-match rates measured when comparing successfully compressed compact images with non-compact verification samples. FNMR is stated at the threshold identified in the top row with the respective recognition algorithm. The text values are empirical FNMR estimates. The colors encode $\log_{10}$ FNMR. In cases where no compressed images were produced, we report the value 1.0.

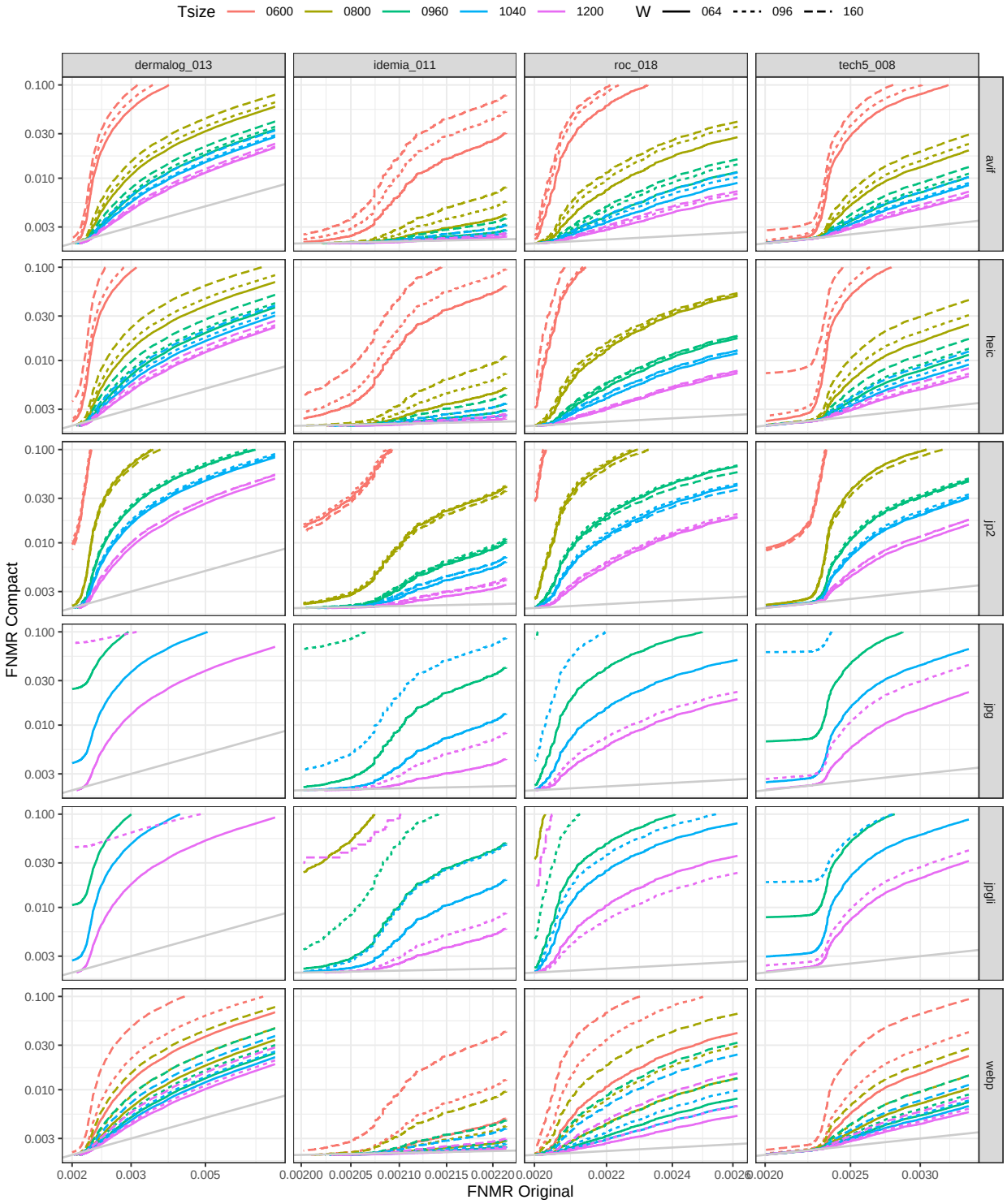

Fig. 8. Each trace plots compact-image FNMR against original, non-compact, FNMR. The plot is parametric on theshold τ, with higher τ giving higher FNMR. The reference line $y = x$ is included to show the no-change outcome.

5.3. Reduced facial similarity

While recognition false non-match rates can remain low, mated similarity scores are reduced relative to those produced when recognizing non-compact parent images. The scores usually remain above threshold τ but are reduced. This occurs because facial similarity is eroded by the combined effects of image size reduction and, especially, lossy compression. We summarize this whole-distribution shift using the Kullback Leibler Divergence (KLD) and the Earth Movers Distance (EMD) of section 4.3 as depicted in Figures 9 and 10. Neither KLD nor EMD can be used to compare algorithms but can be used to compare compact image preparations processed by one algorithm. While FNMR values are more operationally relevant, the KLD and EMD values give more insight into which preparation methods are doing less damage, because they are derived from continuous-valued score distributions whereas FNMR counts binary similarity-score-below-threshold decisions.

The results are that, for almost all combinations of image width, target file size, and recognition algorithm, WEBP yields the lowest KLD and EMD values, followed by AVIF, HEIC, JPEG 2000, and the JPEG variants. The exceptions are that AVIF, and sometimes HEIC, outperform WEBP slightly when W and T are large. Furthermore, at any given target filesize, and for any given recognition algorithm, the lowest KLD and EMD values arise for the smallest width option, $W = 64$. This shows that image size reduction (i.e. downsampling) is doing less damage than is lossy compression.

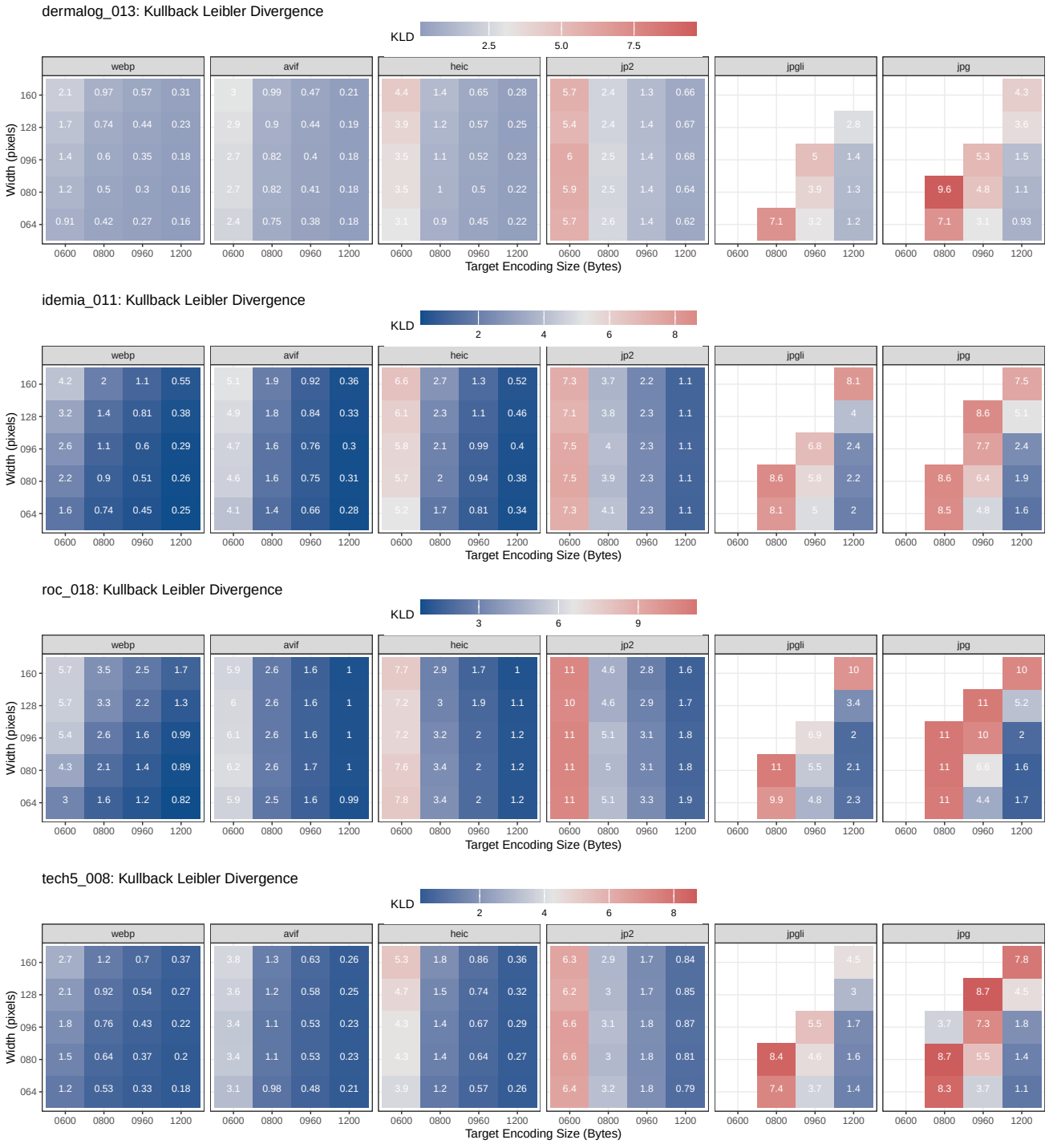

Fig. 9. Kullback Leibler Divergence (KLD) quantifies the distance between the **mated** score distribution for recognition of compact images relative to that for the non-compact parent images. Higher values indicate a larger negative shift in the mated distribution driven by compression-induced change of facial appearance.

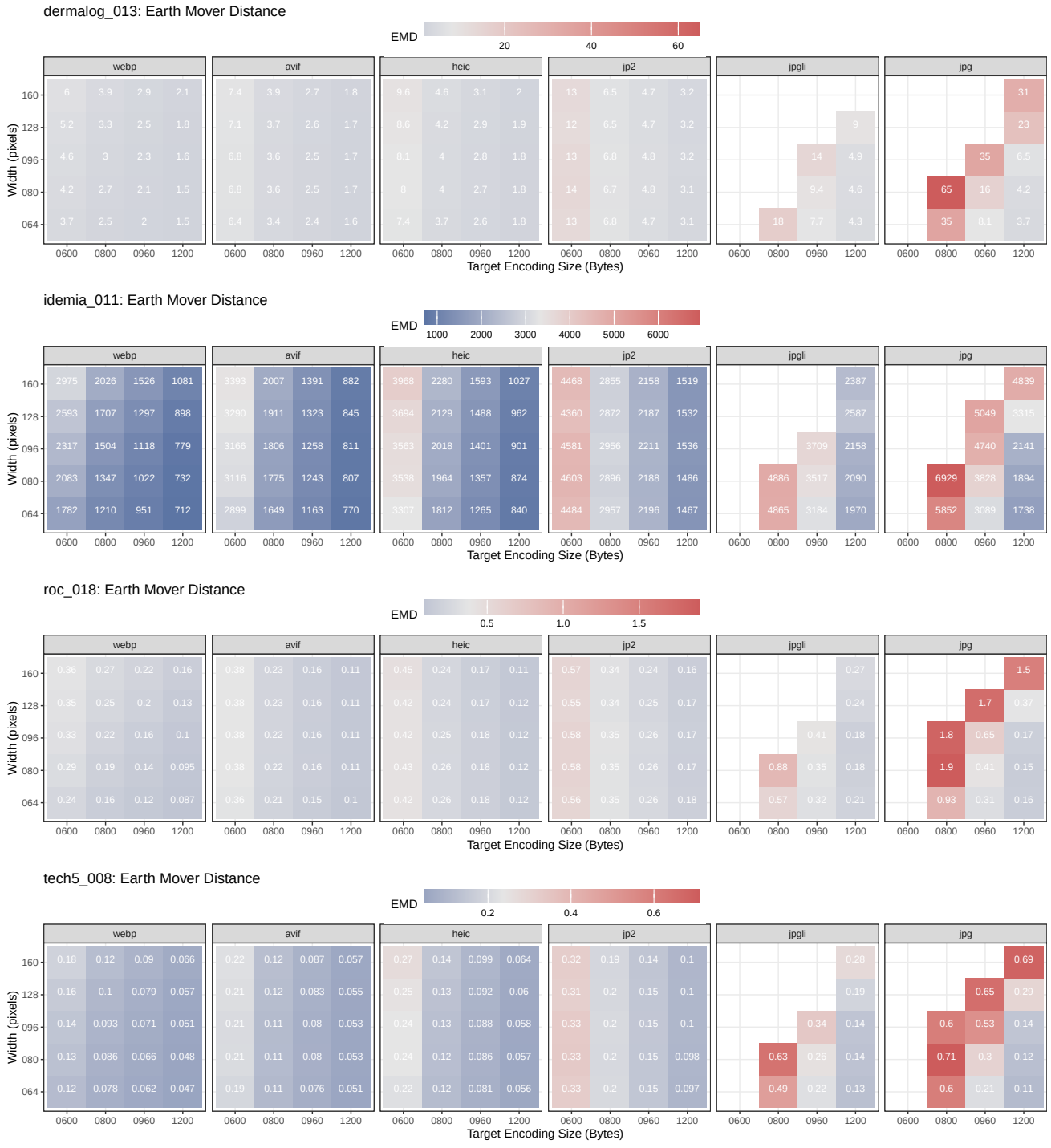

Fig. 10. Earth Movers Distance (EMD), also known as Wasserstein-1 distance, quantifies the distance between the **mated** score distribution for recognition of compact images relative to that for the non-compact parent images. Higher values indicate a larger negative shift in the mated distribution driven by compression-induced change of facial appearance.

5.4. False positives

It is necessary to check that compact images do not elevate false positive outcomes i.e. the incorrect association of the person in a live photo with a different person in the compact reference. Figure 11 shows false match rates for all algorithms, compressors, image widths and target filesizes. The threshold is set to a fixed value for each recogition algorithm. This value gives FMR near 10^{-4}.

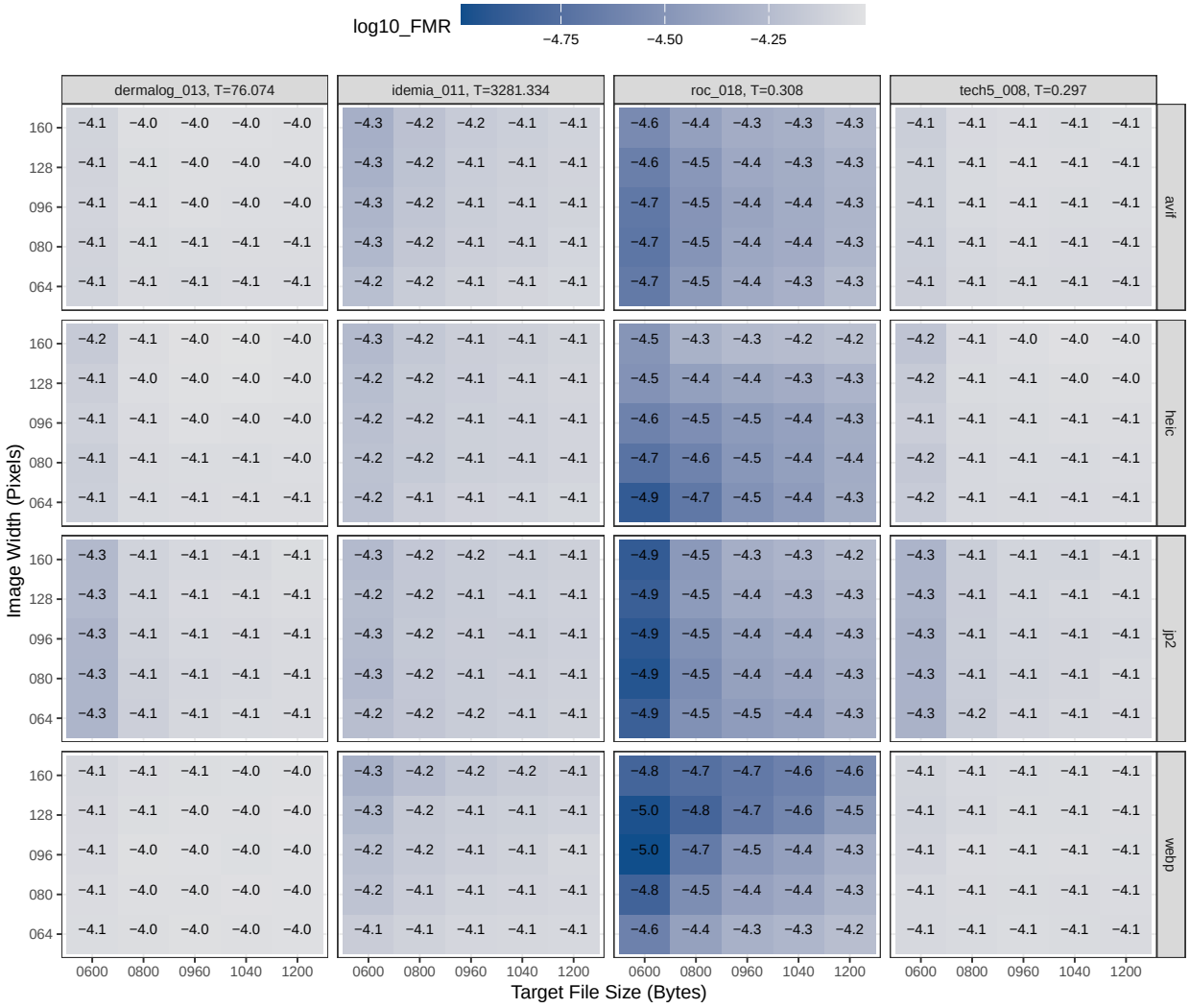

Fig. 11. The heatmap compares compact image preparation methods by showing the false **match** rates measured when comparing successfully compressed compact images with non-compact verification samples. FMR is stated at the threshold identified in the top row with the respective recognition algorithm. The colors and text values are $\log_{10}$ of empirical FMR estimates.

For all recognition algorithms, compressors, target sizes, and image widths, the observed false match rates are at or below that value for non-compact images. While stable FMR would be preferable, the decreases in FMR indicate there is no security downside to the use of compact images.

The distribution shift metrics can be applied to non-mate scores also. Figures 12 and 13 shows much smaller distributional shifts in non-mated scores.

Fig. 12. Kullback Leibler Divergence (KLD) quantifies the distance between the **nonmated** score distribution for recognition of compact images relative to that for the non-compact parent images. Higher values indicate a larger shift in the nonmated distribution driven by the loss of facial detail.

Fig. 13. Earth Movers Distance (EMD), also known as Wasserstein-1 distance, quantifies the distance between the **nonmated** score distribution for recognition of compact images relative to that for the parent images. Higher values indicate a larger shift in the nonmated distribution driven by the loss of facial detail.

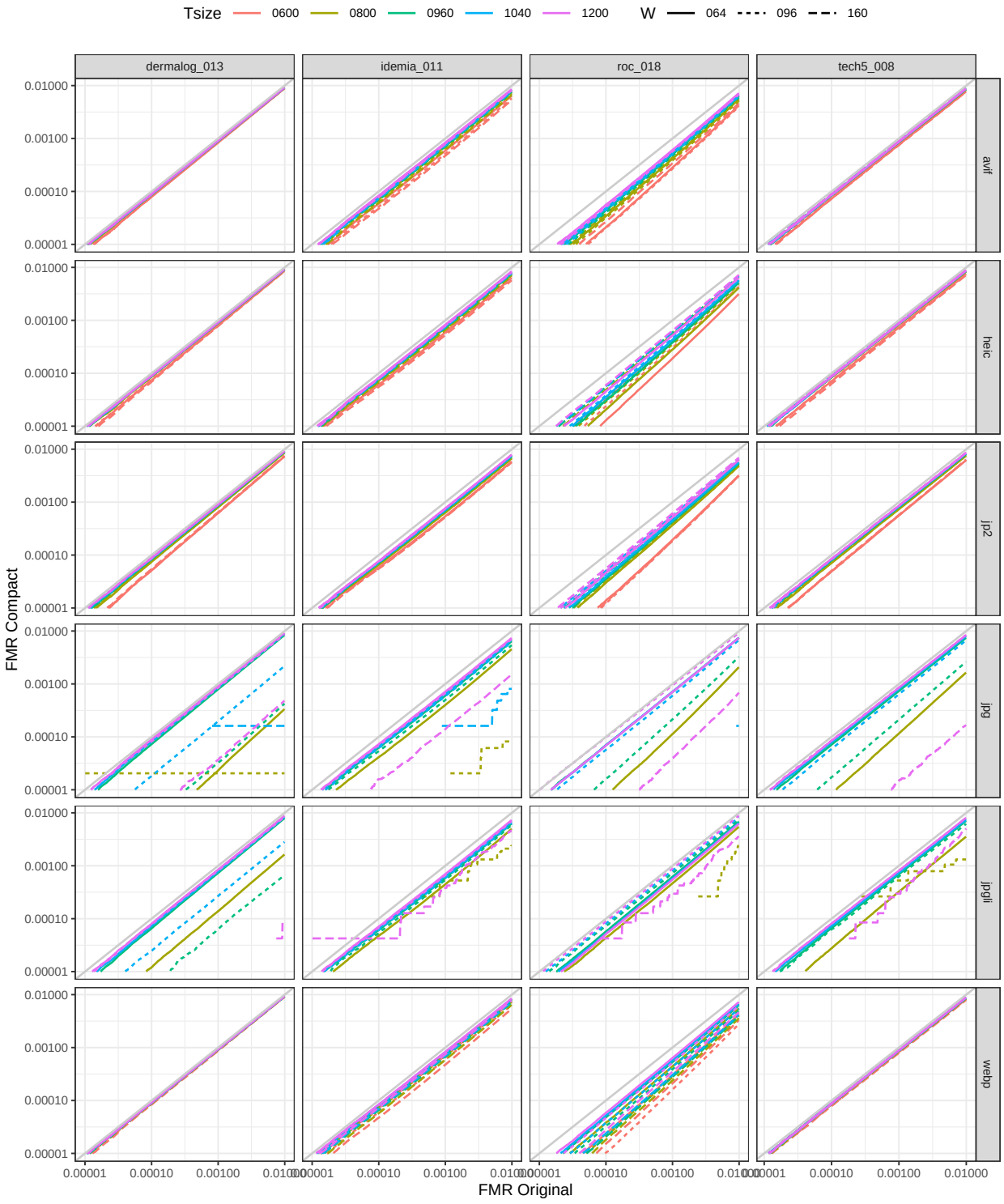

Fig. 14. Each trace plots compact image FMR against original, non-compact, FMR. The plot is parametric on threshold τ, with higher τ giving lower FMR. The reference line $y = x$ is included to show the no-change outcome.

5.5. Use of already-compressed test data

In our trials, the inputs to the various compact-image preparation methods are JPEG-compressed mugshots. This may be considered non-ideal since it is known that re-compression of images - a second compression applied to an already compressed photograph - can introduce additional artifacts, especially if cropping is also performed.

This could be avoided operationally by configuring the camera and capture subsystem to collect and process raw uncompressed images, and to pass those to a compact image preparation module. However, it is much more typical for the camera to apply (JPEG) compression natively and therefore, for compact image preparation to crop, resize, and re-compress the photo. Thus our experimental design is not optimal, but is operationally representative. Our overall result - that compact images prepared in this way can offer high recognition accuracy - is true despite the recompression step. It may be possible to attain smaller sizes, or better images, or better accuracy if raw uncompressed data had been used; this possibility would need to be tested.

Recall from section 2 that target file sizes are attained by iteratively setting an input "quality" parameter, calling the compressor, and stopping this process only when the target encoded size is reached. The quality values are random variables, because lossy compression is data dependent - i.e. it depends on the visual complexity of the input photograph. Figure 15 shows that lower quality values are needed when preparing the most compact images. It also reveals that, for the WEBP compressor, the needed quality values have some dependence on the JPEG compressed size of the input: Larger input photos tend to need lower quality values.

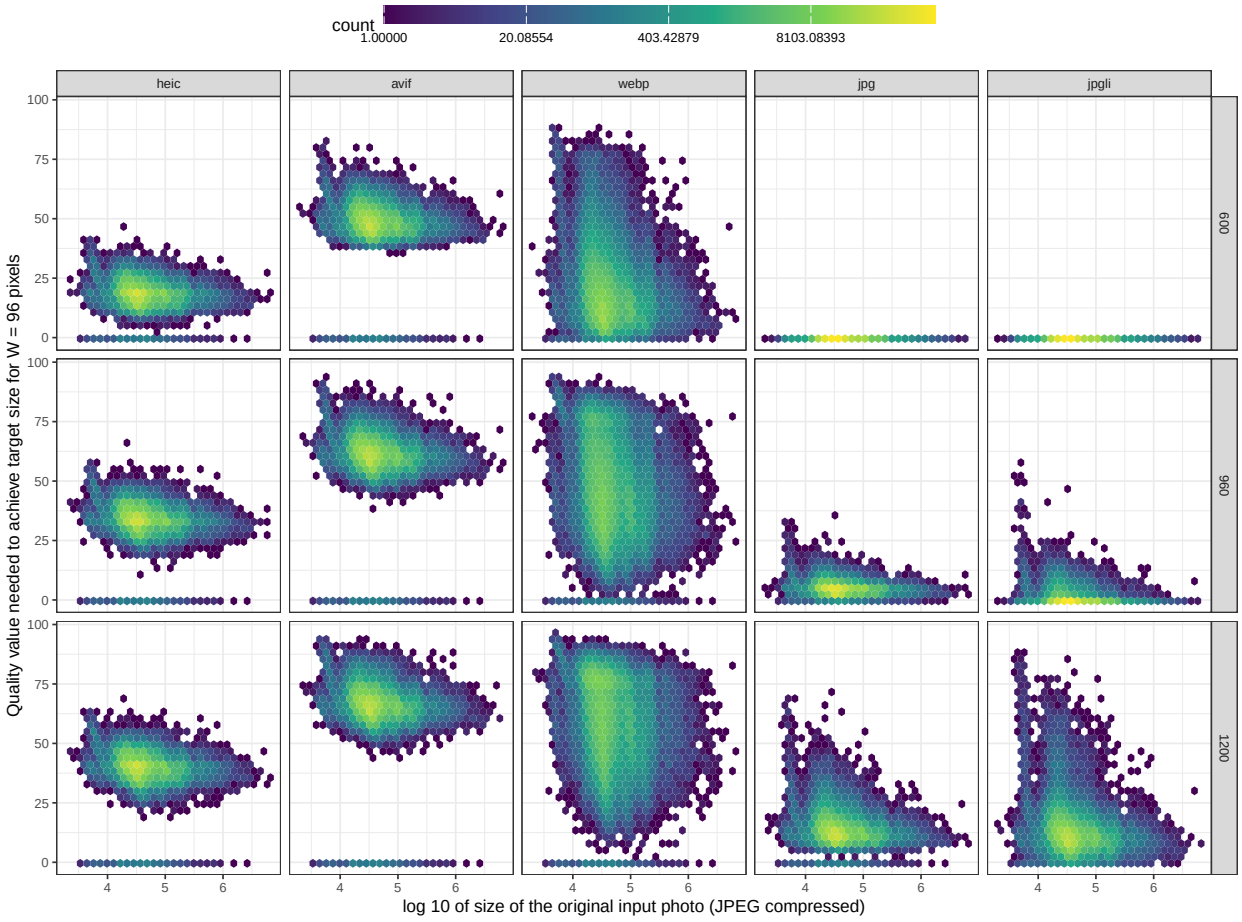

Fig. 15. The figure shows the distribution of quality values input to the given compressor as a function of the filesize of the input. The JPEG 2000 compressor is absent because it produces an encoding of the specified size directly, without iterative reduction of the quality value. The non-compact test set natively uses JPEG compression at, on average, a few tens of kilobytes. The color scale is logarithmic, so yellow indicates many more samples than blue. The semantics of the quality value are specific to the compressor, so compressors cannot be compared on that basis.

5.6. Human review

The results above show that 2D barcodes holding appropriately compressed face images (for example WEBP, size 960 bytes) can give low face verification error rates. However, if we consider an operational context in which some proportion of recognition attempts fail (due to poor illumination in the verification environment, say), the first remedy is to make second and third verification attempts. Thereafter, a common fallback is for a human to compare the reference photo with the face of the live human or a photo thereof. This step is needed if the verification actually includes an impostor - someone presenting the credential illegitimately. There is considerable research showing human accuracy is quite poor[2] and, with some caveats, generally inferior to automated face recognition. Human accuracy varies greatly even among those with

professional training, or natural super-recognizer ability[3]. It also depends on time constraints and is vulnerable to social engineering attacks such as deflection or distraction.

The use of compact credentials makes human review more difficult, because downsampling and compression remove facial detail. However, if compact images must be used, then the results of this report can guide their preparation. The FNMR results of Figure 7 show that smaller image widths are preferred. This is confirmed by the KLD values for mated scores in Figure 9. From the widths studied, this paper suggests use of an image width of 64 pixels corresponding to intereye distances between 19 and 27 pixels. Whether this is ideal for human review or not has not been studied. The alternative, to use a large width, gives slightly elevated FNMR and KLD results, consistent with more compression damage; this would adversely affect human review to some extent. For any of the widths studied here, a human review would almost certainly proceed with a graphical user interface that would allow magnification to a larger display size. So for a compact image of width 64 pixels, the review would be conducted on a larger, software-interpolated image. Figure 16 shows an example: The original crop alongside two 960 byte WEBP-compressed images, input sizes 64 and 160 pixels, displayed at higher resolution using the Chrome web browser. The effects of this interpolation are more evident in the 64 pixel input. A key problem for human review is the presence of specious features - artifacts of the downsampling, compression, and upsampling processes - see the Figure 16 caption. Beyond the results for automated face recognition, we have no data on what processing human reviewers would find optimal, and a human study would be expensive, image sample dependent, and affected by unmeasured or unknown covariates such as age, ethnicity, cosmetics, sun-damage, and photographic underexposure.

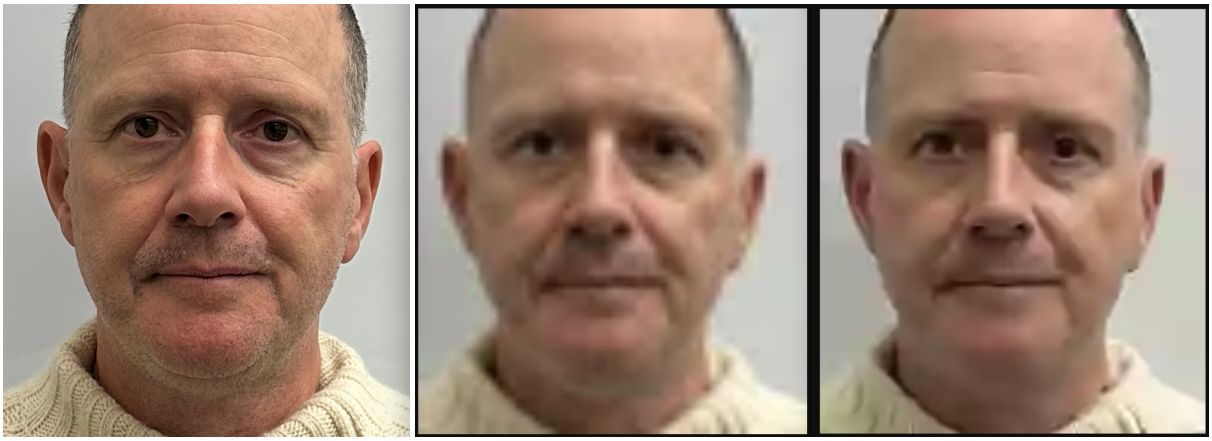

Fig. 16. The figure shows a non-compact image and WEBP-compressed versions of width 64 (middle image) and 160 (right image) pixels. The compact images are resized within a Chrome web browser window then screen-captured to lossless PNG. The browser uses its own in-built interpolation to achieve the same screen size. The images are then included in this document, embedded in PDF, and displayed using software on your computer. The key point is that a human reviewer will see artifacts from all this processing - for example the dark dot on the left lower eyelid in the final image.

6. Summary

In making the decision to use compact face images on 2D barcodes, the following technical factors should be considered.

1. **Trusted capture**: The use of compression with compact images effectively removes the possibility to detect morphing or presentation attacks from that image. For this reason, the parent image should either be collected in trusted setting - typically live capture in the presence of trusted staff - or run through capable attack detection algorithms.
2. **Subject cooperation**: The photographer, if present, should arrange for a uniform background and frontal presentation following the standard ISO/IEC 39794-5 Annex D.1 capture guidance. Per Figure 5, the photographer should also ask the subject, when possible, to remove glasses, hats, scarves and other clothing above the neckline. When a photographer is not present, the subject should be guided to remove such clothing.
3. **Detection and cropping**: Implementers should use detectors of known good accuracy, that either emit landmark points, or eye coordinates. This report did not survey over cropping geometries. Implementers should arrange for an uncluttered background and a tight crop that leaves both ears and chin visible. The crop may exclude some of the hair but should leave forehead and eye brows visible.
4. **Resizing**: The downsampling operation used to convert the cropped image to the small image should filter the input to avoid aliasing in the output. Aliasing artifacts require additional encoding bits thereby causing compression loss elsewhere in the image.
5. **Target file size**: The encoded size should ordinarily be the largest possible size that will fit on a 2D barcode given application-specific constraints imposed by the size requirements of metadata (e.g., age, name, date of birth, date of expiration), digital signatures, and error-correction data. The results in this report suggest implementers should avoid target sizes at or below 800 bytes.
6. **Compression**: Of the compressors considered in this report, AVIF, HEIC, and WEBP all offer high accuracy recognition. Implementers might consider other factors such as software licensing, availability and universal support, processing duration, ease of configuration, and cost. WEBP is strong in most of these aspects.

7. Timing

Timing of the codecs was tested on 250 images of size 960x1440 with various specifications for cropped and resized input image face width (64, 80, 96, 128, 160 pixels) and output compressed image target size (600, 800, 960, 1040, 1200 bytes). Median was chosen as the metric for timing though the difference between median and mean was small.

7.1. Detection

Detection time is the time it took for dlib to detect the face in the image. The median time per image in the timing set is **173 milliseconds**. This step happens before it reaches any codecs and is thus independent of codec, target size, or resize width.

7.2. Cropping

We measured the duration of the cropping operation performed by the dlib library. The median duration measured for an image in timing set is **143 microseconds**, three orders of magnitude faster than detection. This operation occurs prior to resizing and compression.

7.3. Resize

We measured the duration of the image size reduction function of the pillow library. The time taken to resize increases approximately linearly with the requested target width, W. For widths $W = \{64, 80, 96, 128, 160\}$ pixels those durations are **5.7, 6.1, 6.2, 6.6, 7.0 milliseconds** respectively. This operation occurs after cropping and prior to compression.

7.4. Encoding

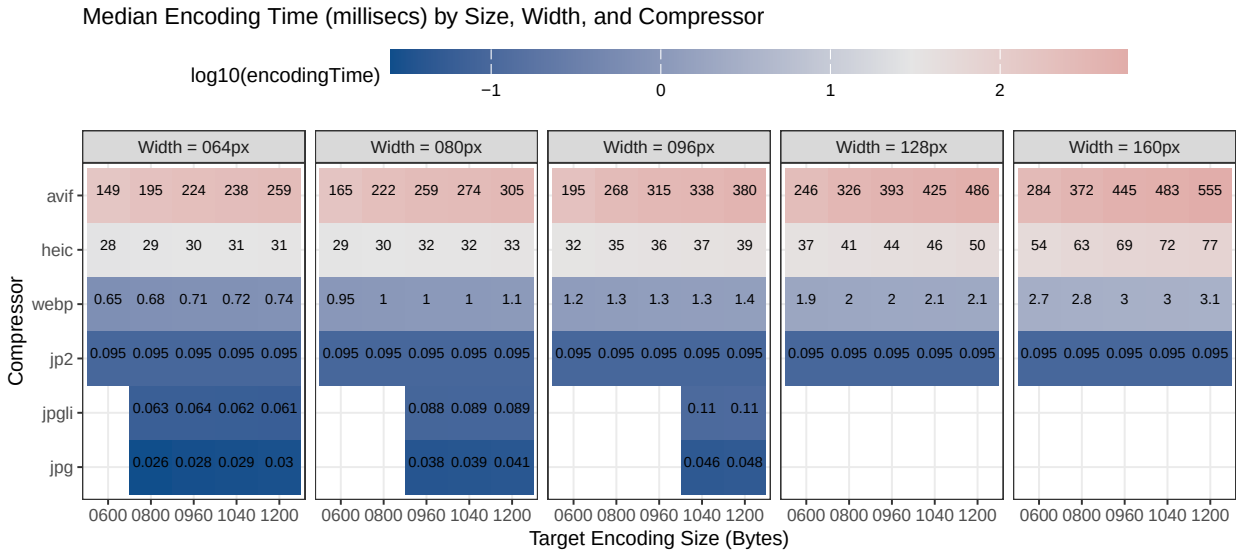

Fig. 17. This plot shows the encoding time across codec, image width, and target file size.

We measured also the duration of the compression operation; this encoding step converts a color image into the binary encoding defined in the respective compression standards. We measure the time taken to produce an encoding at or below the requested number of bytes - the durations therefore include the durations of multiple calls to the underlying compressor. This does not apply to JPEG 2000 because it is able to produce an output of the required size directly, without iteration.

Figure 17 shows the median durations in milliseconds. Note that compression is often a much more expensive operation than cropping and resizing, and this duration varies considerably across compressor.

Note also that while JPEG, JPEG-LI, are extremely fast, they fail on a majority of the images due to not being able to compress to the requested target size. JPEG 2000 is fast because of the lack of iteration. For the remaining three codecs, WEBP is significantly faster than both AVIF and HEIC. AVIF is notably slow with times in excess of 0.5 seconds for $W = 160$ pixel input images.

7.5. Decoding

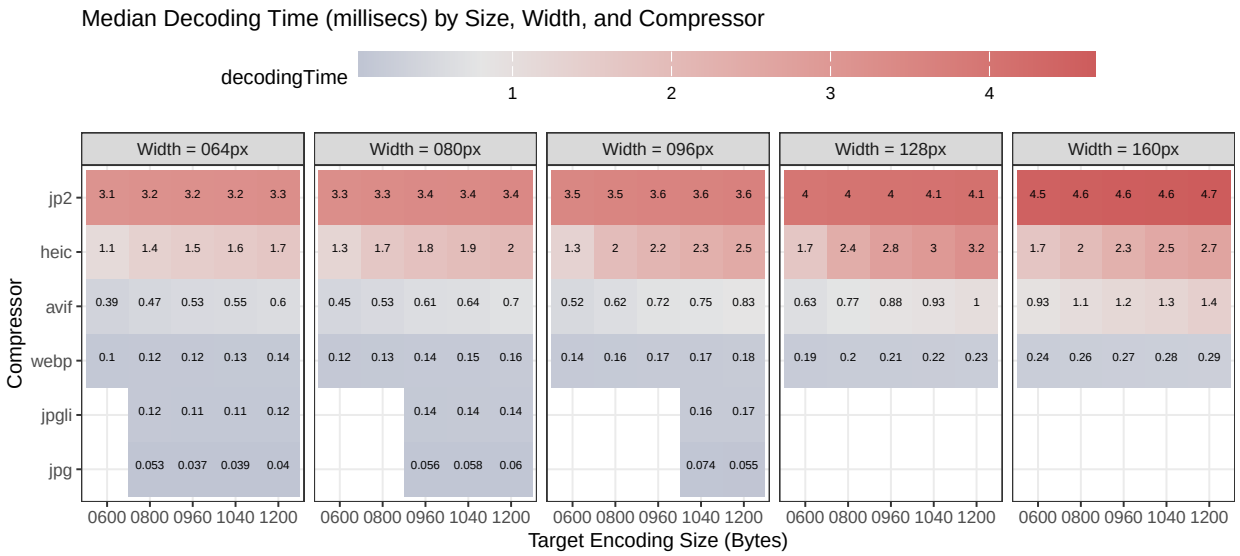

Fig. 18. This plot shows the decoding time across codec, image width, and target file size.

We measured also the duration of the decompression step, representing the time taken to convert an encoded byte stream into a color image usable by a downstream algorithm or human reviewer. Figure 18 shows that decoding an image is orders of magnitude faster than encoding, with the exception of JPEG 2000 which is slower to decode than encode.

References

[1] Schlett T, Schachner S, Rathgeb C, Tapia J, Busch C (2023) Effect of lossy compression algorithms on face image quality and recognition. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp 1–5. https://doi.org/10.1109/ICASSP49357.2023.10095832

[2] White D, Kemp RI, Jenkins R, Matheson M, Burton AM (2014) Passport officers' errors in face matching. *PLoS ONE* 9(8). E103510. doi:10.1371/journal. pone.0103510.

[3] Phillips PJ, Yates AN, Hu Y, Hahn CA, Noyes E, Jackson K, Cavazos JG, Jeckeln G, Ranjan R, Sankaranarayanan S, Chen JC, Castillo CD, Chellappa R, White D, O'Toole AJ (2018) Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences* 115(24):6171–6176. https://doi.org/10.1073/pnas.1721355115. http://www.pnas.org/content/115/24/6171.full.pdf Available at http://www.pnas.org/content/115/24/6171