Document Title:	NIJ Face Algorithm Assessment – Phases I, II, and III, Version 1.0
Author(s):	The National Criminal Justice Technology Research, Test, and Evaluation Center
Document Number:	252825
Date Received:	April 2019
Award Number:	2013-MU-CX-K111

JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

11100 Johns Hopkins Road • Laurel, Maryland 20723-6099

AOS-19-0199

NIJ RT&E Center Project 15-FR

March 2019

NIJ FACE ALGORITHM ASSESSMENT – PHASES I, II, AND III

Version 1.0

Prepared for:

NIJ | National Institute of Justice

STRENGTHEN SCIENCE. ADVANCE JUSTICE.

Prepared by:

The National Criminal Justice Technology Research, Test, and Evaluation Center
The Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Rd.
Laurel, MD 20723-6099

Task No.: FGSGJ

Contract No.: 2013-MU-CX-K111/115912

This project was supported by Award No. 2013-MU-CX-K111, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication/program/exhibition are those of the author(s) and do not necessarily reflect those of the Department of Justice.

Products included herein do not constitute an endorsement by NIJ, DOJ, NLECTC, NIJ RT&E Center, or The Johns Hopkins University Applied Physics Laboratory. NIJ, DOJ, NLECTC, NIJ RT&E Center, and The Johns Hopkins University Applied Physics Laboratory assume no liability for any use of publication content. This publication is a reference for educational purposes only. Please carefully consider the particular needs/requirements of your agency and any applicable laws before developing policies or procedures governing the use of any technology.

NIJ FACE ALGORITHM ASSESSMENT - PHASES I, II, AND III

EXECUTIVE SUMMARY

ES.1 Overview of Effort

The National Criminal Justice Technology Research, Test & Evaluation Center (RT&E Center), operated by the Johns Hopkins University Applied Physics Laboratory, conducted an assessment of the facial image processing software developed by Carnegie Mellon University. This effort was divided into three phases – one for each of the three algorithms (Face Detection, Face Recognition, and Periocular Face Reconstruction); the scope of the evaluations is summarized below and each phase is described in detail in their respective section of this document. Results from each algorithm assessment were compared with the corresponding results from benchmark algorithms. All results were computed with a software framework to maximize the consistency of the assessments; in particular, the process was designed to provide a one-to-one comparison between algorithms by processing both the benchmarks and the algorithm being assessed using the same datasets and metrics.

1. Phase I – Face Detection
    a. Algorithm: Ultron (developed by Carnegie Mellon University);
    b. Benchmark algorithms: TinyFace, YOLO, PittPatt;
    c. Datasets: WIDER FACE, CMU-dataset, Challenge Set 3 (CS3); and
    d. Metrics: Precision-Recall (PR) curve [Area Under the PR (AUC-PR) curve].
2. Phase II – Face Recognition
    a. Algorithm: CMU (Dlib CMU, Native CMU);
    b. Benchmark algorithms: OpenFace, PittPatt;
    c. Datasets: Labeled Faces in the Wild (LFW); the Good, Bad, and the Ugly (GBU); CS3; and
    d. Metrics: Receiver Operating Characteristic (ROC) curves [Area Under the ROC (AUC-ROC) curve], Cumulative Matching Characteristic (CMC) curve, and CMC Rank-Percent Table.
3. Phase III – Periocular Face Reconstruction
    a. Algorithm: Dimensionally Weighted K-SVD (DWK-SVD);
    b. Benchmark Algorithm: Principal Component Analysis (PCA);
    c. Face Recognizers: Kernel Class-dependence Feature Analysis (KCFA), PittPatt;
    d. Datasets: NIST Face Recognition Grand Challenge (FRGC), Cropped Yale; and

e. Metrics: ROC curves [AUC-ROC, Equal Error Rate (EER) point for ROC curve], and Peak Signal-to-Noise Ratio (PSNR).

The datasets include several that are commonly used to assess performance, e.g., WIDER FACE, LFW, FRGC, a larger and newer one (CS3), and some that are less well known (CMU-dataset, Cropped Yale). In all cases, the datasets were curated to ensure all processing had the same input. Dataset preprocessing consisted of image resizing and file formatting to satisfy the algorithm input requirements. The assessment did not modify either ground truth or detector bounding boxes in order to make the dataset more consistent with the detector's capabilities; this approach is used in other assessments. The matching of datasets with algorithms is discussed further in Section ES.3.1.

The interpretation and computation of the metrics are discussed in the various metrics sections for the individual phases. In particular, PR curves and associated metrics are discussed in Section I-4.2.1, CMC curves in Section II-4.1.3, ROC curves in Sections II-4.1.2 and III-4.3.2, and finally PSNR in Section III-4.3.1.

ES.2 Detailed Results

Summary results for the assessment are provided for each phase. In each case, the metric values are broken out by algorithm and dataset. Note that for single value metrics (e.g., AUC-PR) the values range between 0.000 and 1.000; and differences in values are often in the second decimal place or smaller. Differences in this range are found in the algorithm comparison literature.

ES.2.1.1 Phase I Results

Representative PR curves for detection algorithms are presented in Figures I-4–5 to I-4–7; and the assessment results are summarized in Table ES-1. The AUC-PR and F1 (harmonic mean of precision and recall) values for both Ultron and TinyFace are larger (with 1 exception) than the values for the other two algorithms. YOLO and PittPatt generally have the lowest AUC-PR and F1 (harmonic mean of precision and recall) scores. The major exception is PittPatt on the CS3 dataset. When the bounding boxes (BBs) are normalized, as discussed in Appendix I-A, the CS3 (esp. PittPatt) results become consistent with the other two datasets.

The detector and ground truth BBs, and their spatial relationship, are the basis for the metrics used to characterize the Phase I assessment. In addition to IOU-based precision and recall curves (and the AUC-PR value), the distributions of BB sizes, the ratio of detector and ground truth BB sizes and the relations between the number of ground truth BBs, detector BBs and the number of associated BBs were compiled to assess Ultron performance. For each detector-dataset combination, Table ES-1 provides values for the mean and std. dev. of the distribution of IOU values and BB size ratios as well as the association rate between detector and ground truth BBs. In general, these results showed the variability of detector generated BBs vs the annotated ground truth BBs. The CS3 dataset has anomalous values for the distribution of BB size ratios. In Phase I, the RT&E Center conducted a follow-on evaluation using the Ultron algorithm and CS3 dataset. The predicted and ground truth BBs were "normalized" according to "recommended" protocols and the performance metrics recalculated. The results were significantly different from the non-normalized case (Phase I, Appendix I-D).

Table ES-1: Phase I Assessment Summary

Detector/Dataset	CMU	CS3	WIDER FACE
Ultron			
AUC-PR	0.7687	0.6092	0.4048
Max. F1 score	0.8346	0.7114	0.5549
IOU Dist. (mean/std. dev.)	0.59 / 0.14	0.56 / 0.14	0.64 / 0.24
Unmatched BBs	519	8815	70
Association rate	85.7%	65.4 %	82.6%
BB size ratios (mean/std. dev.)	1.23 / 0.28	0.65 / 0.11	1.07 / 0.21
TinyFace			
AUC-PR	0.7341	0.6152	0.6283
Max. F1 score	0.7911	0.7002	0.7252
IOU Dist. (mean/std. dev.)	0.55 / 0.14	0.54 / 0.15	0.63 / 0.19
Unmatched BBs	46	8395	55
Association rate	75.2	59.9	82.3
BB size ratios (mean/std. dev.)	1.31 / 0.28	0.68 / 0.12	1.09 / 0.23
PittPatt			
AUC-PR	0.1453	0.6449	0.2064
Max. F1 score	0.3458	0.7313	0.3872
IOU Dist. (mean/std. dev.)	0.43 / 0.14	0.54 / 0.17	0.43 / 0.23
Unmatched BBs	1151	12185	553
Association rate	47.9	68.7	78.3
BB size ratios (mean/std. dev.)	1.43 / 0.27	0.78 / 0.17	1.26 / 0.27
YOLO			
AUC-PR	0.3403	0.2778	0.2245
Max. F1 score	0.5230	0.4848	0.3772
IOU Dist. (mean/std. dev.)	0.49 / 0.16	0.47 / 0.14	0.50 / 0.25
Unmatched BBs	511	14328	479
Association rate	65.2	51.3	83.1
BB size ratios (mean/std. dev.)	1.12 / 0.29	0.67 / 0.15	0.97 / 0.23

ES.2.1.1.1 Phase 1 Results Summary

For detection, the primary performance metric is AUC-PR. The association rate is also investigated since it is a measure of how well an algorithm identifies BBs that "overlap" the ground truth BBs. As shown in Table ES-1, no detector-dataset combination had the best performance for all three datasets based on AUC-PR and association rate values; performance ranking of both metrics, in isolation, varied between detector-dataset combinations. The relative performance of detector-dataset combinations is visualized in Figure ES-1.

- For the CMU dataset, Ultron has an approx. 4.7% higher AUC-PR value than TinyFace; whereas, TinyFace is 115.7% better than YOLO, which is 134.2% greater than PittPatt. Note that the F1 values follow the same trend.
- For the non-normalized CS3 dataset, PittPatt has a 4.9% higher AUC-PR value than TinyFace, which is 1.0% larger than Ultron. YOLO is lowest at 45.6% of the Ultron value. For the normalized CS3 dataset, Ultron is only 0.2% greater than TinyFace and TinyFace is 17.5% greater than YOLO, which is only 2.6% greater than PittPatt.
- For the WIDER FACE dataset, TinyFace has a higher AUC-PR value than Ultron (55.2%); Ultron to YOLO ratio equals 80.3% and YOLO is only 8.7% larger than PittPatt.

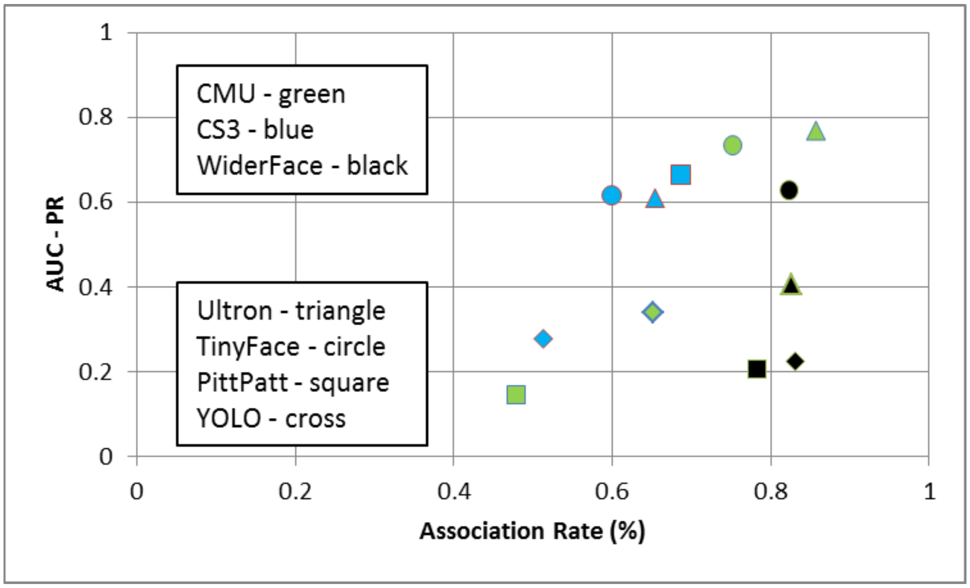


Figure ES-1: AUC-PR vs Association Rate for All Algorithm-Dataset Combinations

In summary, based on the AUC-PR values, this assessment found Ultron performance to be in the same range as TinyFace and that PittPatt has lower values than Ultron/TinyFace (with one exception). YOLO has consistently lower values than Ultron/TinyFace. Based on the average values (across datasets) of the AUC-PR scores, the order of best-to-worst performers is TinyFace, Ultron, PittPatt, YOLO.

ES.2.1.2 Phase II Results

This assessment of face recognition algorithms compared results from four preprocessor-algorithm combinations (comprised of two preprocessors and three algorithms) processing three datasets using two primary metrics: ROC curves and CMC curves. The four preprocessor-algorithm combinations addressed two assessment objectives: compare three algorithms receiving input from the same Dlib preprocessor and compare two algorithms with native (i.e., built-in) preprocessing.

The assessment included two applications of the CMU algorithm – with built-in preprocessing capabilities (Native CMU), with common preprocessing (Dlib CMU). The metrics used in the assessment include the area under the ROC curve, and the rank-percent tabular results of the CMC curve.

The AUC values for the ROC curves (Table ES-2) are all greater than 0.9000 with the one exception (PittPatt). Also, inside the GBU dataset partitions, the AUC order is consistent with the intuitive interpretation. Native CMU consistently yields an AUC value greater than 0.99 with the one exception (GBU-U) which consistently has the lowest scores across all matchers.

These results show that, based on ROC and CMC (Table ES-3) curves, the Native CMU algorithm outperforms the others (higher AUC values, higher percentages at Rank equal 1); the OpenFace algorithm has essentially the same performance as Dlib CMU using ROC curves (based on the average ratio (= 1.0037) of the Dlib CMU to OpenFace AUC-ROC values in Table ES-2) and Dlib outperforms OpenFace using the CMC Curves (Table ES-3, see Figures II-4–10 through II-4–12). Both of the neural network based algorithms modestly outperform the PittPatt algorithm (based on the average ratio (=1.007) of AUC-ROC values for both the OpenFace and Dlib CMU to PittPatt ratios).

Table ES-2: Phase II Assessment Summary AUC-ROC

Dataset	Dlib CMU	OpenFace	PittPatt	Native CMU
LFW	0.9703	0.9913	0.8992	0.9985
Aggregate GBU	0.9778	0.9725	0.9200	0.9933
GBU Partitions	G – 0.9941 B – 0.9798 U – 0.9484	G – 0.9898 B – 0.9749 U – 0.9426	G – 0.9859 B – 0.9273 U – 0.8096	0.9980 0.9930 0.9841
CS3_1870Bal	0.9646	0.9834	0.9364	0.9970

The relative recognition performance of the four detectors on the three datasets (Table ES-3) shows a general pattern where the Native CMU value at Rank = 1 is 20 to 30 percentage points higher than the next highest one and the next highest one is Dlib CMU for two of the three datasets. For the fixed preprocessor case, the maximum percentage to which Rank = 1 extends is zero to 20 points higher for Dlib CMU than OpenFace. OpenFace is 10 to 40 points higher than PittPatt.

Table ES-3: Phase II Assessment Summary CMC Rank-Percent Table*

	LFW				GBU				CS3			
Rank at	OpenFace	PittPatt	Dlib CMU	Native CMU	OpenFace	PittPatt	Dlib CMU	Native CMU	OpenFace	PittPatt	Dlib CMU	Native CMU
~ 10%	1	1	1	1	1	1	1	1	1	1	1	1
~ 20%	1	1	1	1	1	1	1	1	1	1	1	1
~ 30%	1	6	1	1	1	1	1	1	1	1	1	1
~ 40%	1	18	1	1	1	1	1	1	1	7	1	1
~ 50%	1	47	1	1	1	3	1	1	2	23	1	1
~ 60%	1	110	1	1	1	8	1	1	5	64	1	1
~ 70%	2	235	4	1	2	18	1	1	13	144	5	1
~ 80%	7	457	15	1	5	40	2	1	32	294	26	1
~ 90%	28	958	116	1	13	88	8	1	90	603	188	1
~ 99%	604	2529	2251	2	83	257	195	52	485	1387	1389	458
~ 100%	2832	3292	3292	2826	259	383	424	411	1389	1779	1778	1757

*The CMC score at rank N is the percentage of time that the correct match is in the top N matches.

ES.2.1.2.1 Phase II Results Summary

The results found that for both the ROC and CMC curves, the two algorithms based on CNN implementations (CMU, OpenFace) out-performed the algorithm based on manually crafted features (PittPatt). The results for both metrics show the Native CMU algorithm is superior; however, in the case of the ROC analysis, Native CMU is only higher than PittPatt by an AUC average of 9.2%. The relative superiority of Native CMU is larger for the CMC analysis.

ES.2.1.3 Phase III Results

The CMU software [Dimensionally Weighted K-SVD (DWK-SVD)] is based on a linear transformation that maps an image of a periocular region to an image of a full face. Standard metrics used to compare periocular-based face reconstruction include the ROC curves (further characterized by AUC-ROC and EER point) and the PSNR values (Figure III–11). Results (Table ES-4) show that DWK-SVD approach has larger EER values than the PCA benchmark in hallucinating a face based on just its periocular representation.

Table ES-4: Assessment EER Results vs. CMU Results

Methods	Equal Error Rate	Dataset
KCFA Original vs Original	**0.014**	**FRGC**
KCFA Recon. DWK-SVD vs. Original	**0.056**	**FRGC**
KCFA PCA vs Original	**0.279**	**FRGC**
PittPatt Original vs Original	**0.053**	**FRGC**
PittPatt Recon. DWK-SVD vs. Original	**0.163**	**FRGC**
PittPatt Recon. PCA vs Original	**0.347**	**FRGC**
PittPatt Original vs Original	**0.2299**	**Cropped Yale**
PittPatt DWK-SVD vs Original	**0.2853**	**Cropped Yale**
PittPatt PCA vs Original	**0.3815**	**Cropped Yale**

ES.2.1.3.1 Phase III Results Summary

The comparisons found for the FRGC dataset include:

- The PSNR comparison (Figure III–11) found the DWK-SVD distribution superior to the PCA, i.e., the mean PSNR is larger for the DWK-SVD distribution;
- The EER comparison (Table ES-4) found the DWK-SVD representation methodology superior to PCA (i.e., lower EER values);
- The EER comparison also found the KCFA face matcher superior (lower EER values) to the PittPatt matcher; and
- The ROC curve comparison (Figure III–12) found the DWK-SVD reconstruction (AUC = 0.8512) to be much larger than the PCA reconstruction (AUC = 0.5770).

The corresponding results for the Cropped Yale dataset are as follows:

- The EER comparison (Table ES-4) found the DWK-SVD representation methodology superior (i.e., lower EER values) to PCA;
- The ROC curve comparison (Figure III–16) found the DWK-SVD representation (AUC = 0.7784) to be larger than the PCA representation (AUC = 0.6491).

Additionally, since the AUC values for different Pittpatt comparisons within the FRGC dataset are universally higher than the corresponding comparisons within the Cropped Yale dataset, the RT&E Center concludes the DWK-SVD periocular reconstruction approach is more suitable for more “natural” datasets. The illumination variability of the Cropped Yale dataset may or may not be relevant to end users of CMU’s algorithm, but to productize the algorithm itself, it should be retrained on a larger dataset of representative images of the end use case. Furthermore, a firm preprocessing stage to detect the periocular region would be required as input to the reconstruction algorithm.

ES.3 Conclusions

ES.3.1 Feasibility of Productization

This assessment of CMU-generated face processing software used multiple datasets (Section ES.1). These datasets reflect the evolution of benchmark datasets for facial processing performance. The dataset descriptions all claim a more "realistic nature" for the data but in all cases, there is some control over the image collection. The detector-ground truth BB analysis in Phase I (Section I-4.3.3) showed performance variability for detector-dataset combinations. The detector that outperforms across all datasets is most likely to be robust to raw image changes. The AUC-PR values for the different datasets (Table ES-1) are a stand-in for the robustness of a detector without tuning between training images datasets and application images.

All trained CNN-based algorithms have preferred characteristics for datasets; e.g., the ratios of ground truth BBs to the target object size or the orientation of faces for periocular reconstruction. For issues such as these, all tested algorithms were treated as application ready in the sense there was no parameter tuning. This included detection algorithms Ultron, TinyFace, PittPatt and to some extent YOLO; recognition algorithms CMU; and, CMU developed periocular reconstruction algorithm DWK-SVD. In general, the Carnegie Mellon University generated algorithms are all tested as they were delivered to the RT&E Center – treated as figurative black boxes.

On the other hand, all data sets have intrinsic parameters and profiles and, for the most part, these are unknown. The parameter choices are defined when the algorithms are broadly trained to optimally process a range of face sizes. This is implemented through the training set used for the CNNs. For example, the three CNN-based algorithms compared in the detection analysis (Phase I) were all trained on the WIDER FACE dataset.

Each assessment approach started with the individual datasets in their original state since the datasets were both intended as "realistic" and a benchmark for algorithm performance. The RT&E Center restricted dataset manipulation to known image attributes required to satisfy the input format for subsequent processing. The primary restriction was resizing the image via cropping, sub-sampling the pixels to make the pixel density (pixels/inch) uniform across the datasets or re-orienting the image. There is no image enhancement such as contrast stretch or edge filtering or image segmentation done as part of the preprocessing.

TinyFace is an example of built-in characteristics that are tuned as part of training. TinyFace specializes in detection of small faces, as small as a few pixels. It contains multiple trained detectors tuned at different scales and aspect ratios in order to "zoom in" on an image at execution and find the TinyFace faces but still be able to "zoom out" and take advantage of detailed features on larger faces. It fine-tunes pre-trained ImageNet models to create 25 templates, where a template is a scale-specific detector, each of which was trained on the WIDER FACE training set.

For the recognition assessment, the preprocessing done with OpenFace (includes Dlib) and CMU (no Dlib) orients faces in order to establish a spatial correspondence between images, and resizes BB to 96 x 96 pixels. The native CMU algorithm (i.e., without Dlib preprocessing) may include image preprocessing steps that may influence some of the dominant control parameters in face recognition, e.g., illumination normalization, zeroing the pixels at the face edges, and removing background clutter.

Similar to above, the periocular reconstruction task requires full frontal (mugshot-like) images that have been cropped and resized. This dataset was provided by Carnegie-Mellon Univ. to the RT&E Center. The Center staff are not familiar with the specific requirements for image cropping that Carnegie-Mellon Univ. employed to create this dataset.

Practical applications of face processing software are likely to include a detection-recognition step or a reconstruction-recognition step. The detection step will have to be flexible and robust enough to be amenable to a variety of system implementations and processing goals. It is unlikely that there will be a perfect “match” between data collected in an operation and data the software is designed for. Regardless, a primary goal of facial image processing is recognition/identification; this step is likely part of the data flow.

Optimum results require knowledge of built-in policies and constraints of face detection software and subsequent “matching” of raw image characteristics to detection software. Users may be unaware of or unable to “optimally” control software performance; e.g., unaware of training set compatibility with operations data. In particular, the parameterizations of the Carnegie Mellon University algorithms are not documented and likely not meant to be controlled by the users. Matching would almost certainly make a “significant” difference in results. In the Phase I assessment, the RT&E Center only “matched” image size to what the algorithm required and used out-of-the-box configurations for preprocessing; the assessment did not match non-native preprocessors to the detection algorithm.

Practical applications of face recognition technology will likely require capabilities with high true positive rate and low false positive rate. This is the region in the vicinity of (0,1) for ROC curves and (rank 1, 100 percent) for CMC curves. Consistently, for assessment of the native preprocessing using the AUC-ROC metric, the Native CMU algorithm outperformed OpenFace by a few percent (see Table ES-2). For assessing the fixed preprocessor (Dlib) combinations on all datasets, PittPatt significantly lagged performance of both OpenFace and Dlib CMU, which were more comparable with each other. In particular, in the AUC-ROC analysis, OpenFace is superior to Dlib CMU in two out of three cases, and in all cases, the differences are about 2% or less.

The above being said, the RT&E Center does not have enough experience with the Carnegie Mellon University algorithms to understand, in general, how much flexibility there is in the input image data. We do know that they gave better performance for detection and recognition (similar to TinyFace performance) across three different datasets often used in similar assessments without any “tuned” preprocessing.

ES.3.2 Algorithm Assessments – Relative Performance

The Carnegie Mellon University generated, CNN-based software generally out-performed the benchmark comparisons; although there are a few exceptions to this statement.

In the **detection domain**, the assessment found that from the perspective of AUC-PR, none of the detectors had superior performance for all three datasets. The Carnegie Mellon University developed Ultron detector (with 50 percent IOU threshold) is superior to the other three detectors for the CMU dataset only; however, Ultron ranks high for the other datasets. TinyFace performance is superior for WIDER FACE and comparable to Ultron for the other datasets. PittPatt has the best AUC-PR only for the CS3 dataset; this performance exception is attributed to PittPatt's stronger detection BB area agreement with ground truth BBs in the CS3 dataset. After BB normalization, the PittPatt performance level was reduced to below that of Ultron and TinyFace (see Appendix I-D). YOLO is one of the lowest performers for all three datasets.

In summary, this assessment found Ultron performance to be comparable to TinyFace. PittPatt performance is inferior to that of Ultron/TinyFace and YOLO has the lowest performance.

In the **recognition domain**, the Carnegie Melon University developed Native CMU outperformed across three datasets (five counting GBU partitions) without any tuning. The results found that for the ROC-AUC metrics, the two algorithms based on CNN implementations (Native CMU, OpenFace) out-performed (by more than 7%) the algorithm based on manually crafted features (PittPatt). Consistently, based on AUC-ROC metric, the Native CMU algorithm outperformed OpenFace by a few percent (see Table II-4–2). PittPatt significantly lagged performance of both OpenFace and Dlib CMU, which were more comparable with each other. In particular, in the ROC-AUC analysis, OpenFace is comparable to Dlib CMU in two out of three cases, and in all cases, the differences are about 2% or less.

The same trend holds in the results using the CMC Curve (see Table II-4–3). For the native preprocessor cases across all datasets, the maximum percentage to which Rank = 1 extends is 30 to 50 points higher for Native CMU than OpenFace. For the fixed preprocessor case, the maximum percentage to which Rank = 1 extends is zero to 20 points higher for Dlib CMU than OpenFace. OpenFace is 10 to 40 points higher than PittPatt.

The results for both metrics show the Native CMU algorithm has better values; however, in the case of the ROC analysis, Native CMU is only higher by an AUC of 0.02 or less. The relative superiority of Native CMU is larger for the CMC analysis.

In the **periocular reconstruction domain**, the goals of the assessment were to compare the performance of the DWK-SVD face representation capability with earlier capabilities as represented by PCA using two primary metrics PSNR and ROC curves and the two ROC curve characteristics (EER, AUC). A by-product of the assessment was to verify the results reported by Carnegie Mellon University.

The DWK-SVD algorithm also did well however, it was only investigated against a comparatively smaller dataset than those used in Phase II. The results for a second dataset (Cropped Yale) were weaker but had the same trend. In addition, the PSNR comparison (Figure III–11) found the DWK-SVD distribution superior to the PCA – the mean PSNR is larger.

The ROC curve comparison of AUC values (Figures III-12 and III-14) and the generated EER results for the FRGC and Cropped Yale datasets (Table ES-4) show the same trend. The DWK-SVD representation has smaller (i.e., better) EER value than the benchmark (i.e., PCA) representation. The EER results for the FRGC dataset are identical (to the CMU results) for four of the six cases. The two cases of different results were investigated; however, the differences could not be resolved.

JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

11100 Johns Hopkins Road • Laurel, Maryland 20723-6099

AOS-18-0706

NIJ RT&E Center Project 15-FR

November 2018

NIJ FACE PROCESSING ALGORITHM ASSESSMENT PHASE I– FACE DETECTION

Version 1.1

Authors: DJ Waddell et al

Prepared for:

NIJ | National Institute of Justice
STRENGTHEN SCIENCE. ADVANCE JUSTICE.

Prepared by:

The National Criminal Justice Technology Research, Test, and Evaluation Center
The Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Rd.
Laurel, MD 20723-6099

Task No.:	FGSGJ
Contract No.:	2013-MU-CX-K111/115912

This project was supported by Award No. 2013-MU-CX-K111, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication/program/exhibition are those of the author(s) and do not necessarily reflect those of the Department of Justice.

PHASE I-CONTENTS

PHASE I-FIGURES

Page

PHASE I-TABLES

Page

PHASE I-EXECUTIVE SUMMARY

The National Criminal Justice Technology Research, Test & Evaluation Center (RT&E Center), operated by the Johns Hopkins University Applied Physics Laboratory, conducted an assessment of the face detection capability of the Carnegie Mellon University - developed face detection software, Ultron. The assessment compared Ultron performance with that of three other face detection software systems, PittPatt, You Only Look Once (YOLO), and TinyFace, using a suite of metrics calculated for three different datasets. Three of these detectors (Ultron, PittPatt, and TinyFace) originated in the Carnegie Mellon University community and reflect the evolution of face detection software; the fourth (YOLO) is a variant of the Convolutional Neural Networks-based approach used in Ultron and TinyFace. The datasets used in the assessment were extracted from the WIDER FACE, CMU-dataset), and Challenge Set 3 [i.e., IARPA (Intelligence Advanced Research Projects Agency) Janus Benchmark-B (IJB-B)] datasets.

A standard graphic used to compare detection algorithm performance is the precision-recall curve. The metrics used in the assessment include the scores generated by the software, both standard precision and recall based on an intersection over union (IOU) threshold, as well as precision and recall values based on an area calculation. These metrics were used to investigate the datasets through the prism of the detectors. In addition, the number, sizes, and size ratios of the face detection software generated bounding boxes (BBs) and the ground truth bounding boxes associated with the dataset were also investigated. Specifically, as part of the assessment, distributions of IOU values were calculated to explore the relationship between detector bounding boxes and ground truth bounding boxes, the effect of IOU threshold on metric scores and the distribution of face sizes vs. bounding box sizes. These results show that the Ultron and TinyFace IOU distributions of the ratio of bounding box sizes are statistically identical and distinct from either the PittPatt or YOLO distributions, which are distinct from each other.

The assessment did not modify either set of bounding boxes in order to make the dataset more consistent with the detector’s capabilities; this approach is used in other assessments.

The results for these metrics show the relative superiority of Ultron and TinyFace for two of the three datasets tested. The results also show that the Ultron detector generated bounding boxes have fewer missed associations than TinyFace, PittPatt and YOLO.

The precision and recall metric values shown in Section I-4.3.2 are comparable to the values reported by CMU; however, the RT&E Center values are slightly lower.

I-1 INTRODUCTION

In September of 2013, the Johns Hopkins University Applied Physics Laboratory (JHU/APL) was selected by the U.S. Department of Justice, National Institute of Justice (NIJ) to establish the National Criminal Justice Research, Test, and Evaluation Center (RT&E Center) within the National Law Enforcement and Corrections Technology Center System. The RT&E Center has been tasked to perform an assessment of the Carnegie Mellon University facial processing algorithms using the technical approach documented in the RT&E Center's Proposed Technical Approach for Project 15-FR [1].

The goal of this effort is a side-by-side comparison of Ultron capabilities with other algorithmic approaches. The effort has been divided into three phases: detection, face recognition, and periocular reconstruction. This report documents the Phase I results – the face detection performance of four algorithms on three different datasets as measured by two primary metrics (precision, recall, defined in Section I-4.2 *Metrics*). Other attributes of the algorithms (e.g., software quality, throughput, memory requirements, operating environment, and usability) are out of scope for this assessment but may be provided in later phases when available.

I-2 BACKGROUND

I-2.1 Face Detection

Face detection algorithms focus on identifying human faces in digital images and videos. The goal of face detection is to find all faces present in an image, and not falsely detect non-faces. This is often the first step in a sequence of processes intended to assign a name to a face. The result of the detection step is a bounding box (BB) that ideally overlaps only the area of a face in an image. The face detectors evaluated within the scope of this study return coordinates of the rectangular frames, or BBs around the faces that each has detected, and a confidence value for the detection. The (x,y) coordinates provided are the top-left corner of the box as well as the width and height. The boxes are displayed on the images containing the faces for comparison with the manually labeled (i.e., ground truth) BBs provided in the datasets. These results, computed on either per image or per face basis, are used to calculate the precision and recall metrics that characterize the performance of the algorithms.

One of the biggest challenges of face detection is being able to identify more naturally occurring faces in images and/or videos. These include painted faces, faces at certain angles, masks that only reveal a certain part of the face, etc. Obstructed and occluded faces are important obstacles to overcome because realistically, there are more "faces in the wild" that need to be detected than faces posed perfectly for the camera. While all the candidate algorithms aim to remedy that, Ultron aims to have superior performance.

I-2.2 Example Applications for Law Enforcement

The expected user community for this assessment is the image processing research and development (R&D) community and law enforcement agencies (LEAs). These two groups are expected to have very different technical capabilities and face detection applications. For the purpose of this assessment, we identified three categories of LEAs and their potential use of face detection and recognition:

- Local community police force (e.g., township police force) to a jurisdictional force (e.g., county level) – this category is assumed to have one or more personnel trained to use an application to collect, manage (transmit, receive, annotate, document, archive) and manipulate (modify appearance, crop, etc.) digital images, and request image processing services from elsewhere. Personnel at this level need to be able to select specific faces, acquire information related to the faces, and manage the data/metadata for the faces.
- State or regional LEAs (e.g., a state data fusion center) – this category supports the local police forces and acts like a broker between local forces and national capabilities. Face detection and recognition, face databases, and general image processing are likely to occur here. This group's need is to receive, process, and respond to face related information requests from local LEAs and manage the data/metadata for these faces.
- National level (e.g., the Federal Bureau of Investigation) – this category will act like a resource for state or regional LEAs as well as an interface with international resources.

As a use case, consider an LEA trying to manage an incident with street demonstrators. Images and videos of the demonstration are collected from LEA cameras, news-network cameras, the local population, local businesses, and transportation centers (bus stations, train stations, airports). The goal is to leverage the images to find things such as particular individuals, groups of individuals, and the behavior over time of these individuals. Once the images have been isolated, the LEA can use their local resources to build a case and make information requests to the state/regional LEAs.

To leverage video data requires collecting, formatting, and storing it for subsequent analysis. Based on a report that "bad guys" were observed in the bus terminal, local LEA personnel run face detection software on the available image data. Possible individuals and groups of interest are selected and their images passed to state or regional levels to get face identification and any relevant information.

A second use case includes images collected at a potentially violent demonstration in a small town. The local LEA assessment is that some of the more violent demonstrators are not local. An officer downloads these images to a PC and adds annotations then submits an information request to state police either as a network service or as a work request. He subsequently receives face identification information and people association information that provides value in planning for ongoing demonstrations.

I-2.3 Approach to Assessment

Because each of the four algorithms tested require different input formats and write different output formats, the RT&E Center developed a software framework that implements a consistent and efficient approach to processing the datasets and analyzing the results (see Figure I-2–1). The framework contains four main parts: algorithm classes, dataset classes, image metadata, and analyses.

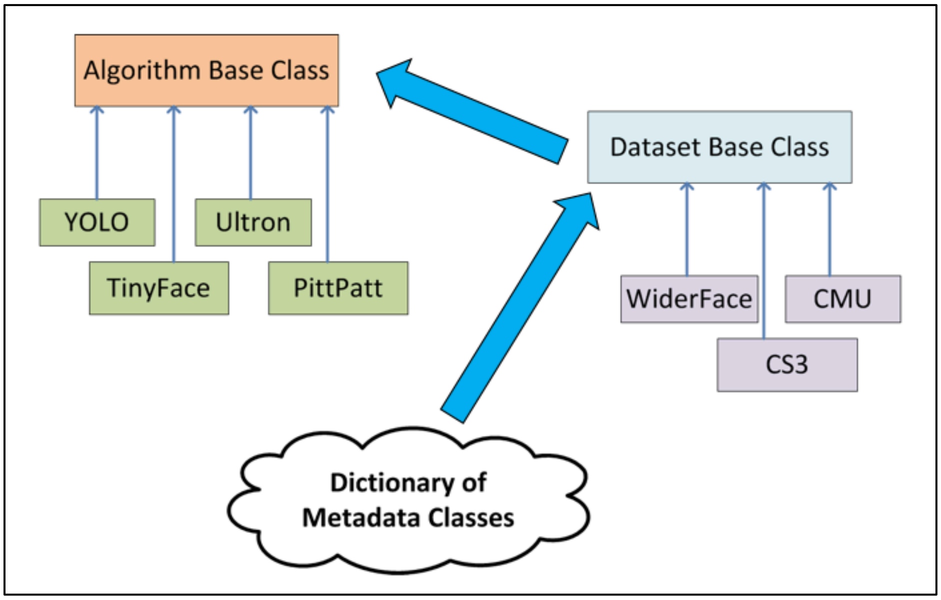


Figure I-2–1: Software Framework Used for Assessment

The algorithm classes are wrappers for each candidate algorithm evaluated. The wrappers implement the execution process of an algorithm in a more user friendly way so that all processes are executed in a uniform and consistent manner. The dataset classes are wrappers for each dataset and configured so that the algorithms may readily access the data. The results are written in a standardized way that can be easily processed for subsequent analysis.

The results from the algorithms such as algorithm-detected BBs, ground truth BBs, and confidence scores compose the image metadata object for every image. That information is then used to plot the BBs and compute metrics to compare performances of all the algorithms on a dataset. Metrics are described in Section I-4.2.

I-3 FACE DETECTION ALGORITHMS

The RT&E Center's assessment evaluated the performance of Carnegie Mellon University's newly developed face detection software, Ultron, vis-à-vis other face detection algorithms using standard metrics, and compared results with other existing face detection algorithms that the NIJ may be interested in using, in particular, PittPatt [1], TinyFace [3], and You Only Look Once (YOLO) v2 [4]. Ultron is the most recent CMU-developed face detection algorithm and its goal is to improve detection of obstructed and occluded faces, or "faces in the wild." These include painted faces, people wearing masks/hats/scarves, or faces at an angle (see Table I-3–1).

Each detection algorithm was acquired and executed as delivered by the source; the RT&E Center did not perform any additional training or parameter adjustments to the algorithms.

Table I-3–1: Assessment Configuration Summary

Algorithm / Source	Paradigm	Training Set	Testing
Ultron	CNN Deep Learning	WIDER FACE Training Set	WIDER FACE validation, CMU-dataset, Challenge Set 3
PittPatt	Feature-centric evaluation using a Bayesian network	Unknown	WIDER FACE validation, CMU-dataset, Challenge Set 3
TinyFace	CNN Deep Learning	WIDER FACE Training Set	WIDER FACE validation, CMU-dataset, Challenge Set 3
YOLO	CNN Deep Learning	WIDER FACE Training Set	WIDER FACE validation, CMU-dataset, Challenge Set 3

All of the face detectors evaluated, excluding PittPatt, are open-sourced, which means that the original source code is free for redistribution and modifications. All of the detectors, again excluding PittPatt, use a machine learning technique called deep learning. PittPatt uses an approach based on a set of prescribed facial feature vectors. Deep learning is typically implemented with Convolutional Neural Networks (CNN), which are a type of artificial neural network comprised of several convolutional layers followed by a classification layer. CNN inputs the pixel value of an image instead of a feature vector or some other numerical representation of an object. All of the CNN-based candidate algorithms were (purportedly) trained or fine-tuned on the WIDER FACE dataset and then published for use.

PittPatt, the state-of-the-art in face detection previous to CNN-based deep learning techniques, also has its origins at Carnegie Mellon University; however, it became the main product of a start-up that was eventually purchased by Google. The RT&E Center could not find access to PittPatt software or documentation through the Internet. We acquired a software copy of PittPatt SDK 5.2.2 through direct contact with Carnegie Mellon University staff. Unlike current deep learning-based solutions, PittPatt does not use a CNN. Instead, it uses feature-centric evaluation, which shares feature values across the image, and a restricted Bayesian network for face detection, a statistical graph used to keep track of the feature values.

TinyFace is a face detector also developed by Carnegie Mellon University. It specializes in detection of small faces, as small as a few pixels. It contains multiple trained detectors tuned at different scales and aspect ratios in order to “zoom in” on an image at execution and find the TinyFace faces but still be able to “zoom out” and take advantage of detailed features on larger faces. It uses pre-trained ImageNet models with 25 templates, where a template is a scale-specific detector, and was trained on the WIDER FACE training set.

You Only Look Once 9000, or commonly known as YOLO v2, is a real-time object detection system. Because it was originally trained as a multi-class object detection system, the RT&E Center re-trained it to detect faces only. YOLO v2 is the second version of YOLO and includes modifications to improve speed, rather than performance, such as utilizing a higher resolution detector to allow image inputs of up to 448x448 pixels, and using batch normalization to lower learning rates and meticulous parameter initialization so the system keeps the same accuracy but with 14 fewer training steps. For this assessment, YOLO v2 was trained on the WIDER FACE training set with a constraint to ignore any faces smaller than 15x15 pixels.

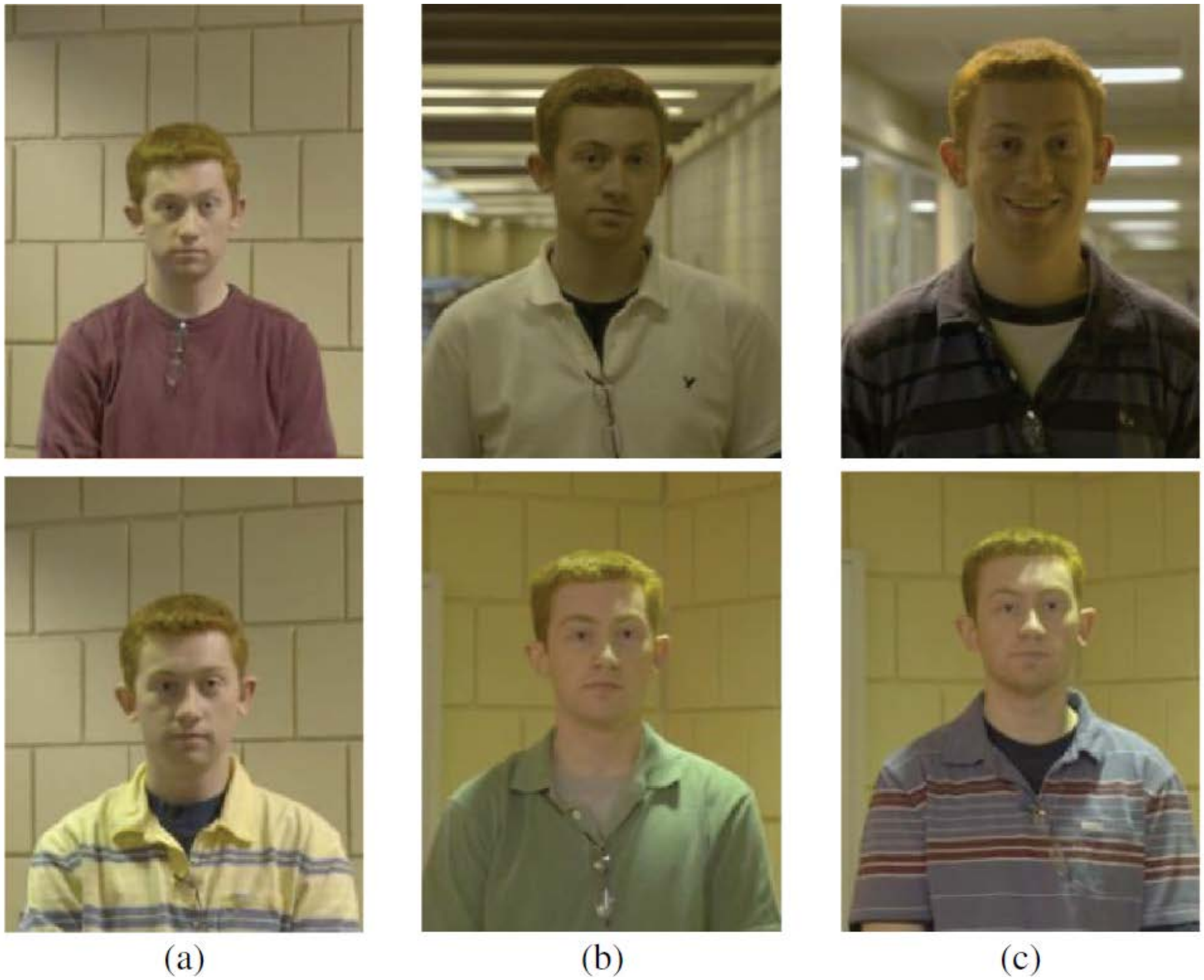
I-4 ASSESSMENT COMPONENTS

This section summarizes the primary components of the assessment task: datasets, metrics, and results. The datasets discussion summarizes the evolution of face datasets and provides examples from some of the datasets used in the assessment. The metrics discussion introduces the “intersection over union” metric (IOU) and other relationships between ground truth and detector BBs and how they determine the two primary metrics (i.e., precision, recall). Finally, Section I-4.3 explores the relationships between detectors, BBs, datasets, and metrics.

The goal of face detection is to determine if one or more faces are in an image and if there are, to determine their location and extent. The face detection component of this assessment consists of processing each of the four algorithms against each of the three datasets. Datasets for image processing development have evolved from “staged” images toward those that are likely to be encountered in typical circumstances of activity. Earlier datasets are characterized as full frontal images with limited pose variability; while subsequent datasets introduced more variability in orientation, lighting, and occlusion.

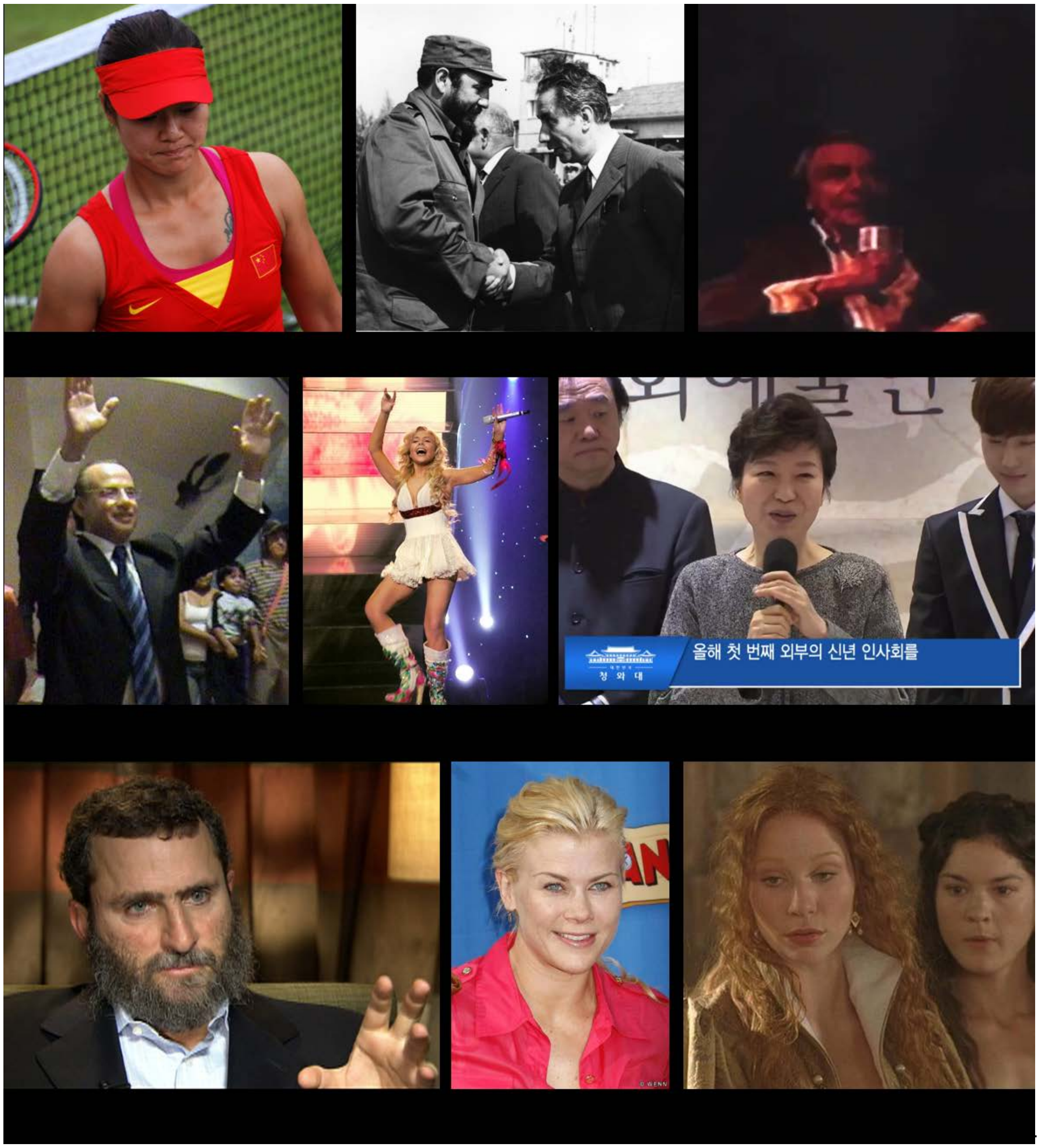
Algorithms for processing faces (detection, recognition) have been an active area of computer vision for the last several decades. Current performance levels of existing algorithms against earlier datasets that feature “staged” images are quite good; however, the performance level against images that are not staged and more representative of people engaged in everyday activities (i.e., the daily activities mentioned in the literature) is weaker. The earlier datasets provided frontal face images and limited variability in pose, occlusion, and scale (number of pixels comprising the image). Examples of good, challenging, and very challenging images from an earlier dataset are shown in Figure I-4–1. A series of more recent datasets (benchmarks) were developed to provide datasets more representative of “naturally occurring” images with larger variations in the parameters describing the images. An illustration of the transition between the earlier and more recent datasets is provided in Figure I-4–2 which shows three successive levels of difficulty: (a) IARPA (Intelligence Advanced Research Projects Agency) Janus Benchmark–B (IJB-B), (b) IJB-A, and (c) LFW (labeled faces in the wild) datasets, where (a) is the top row and most difficult dataset and (c) is the bottom row and least difficult dataset [5]. Example images from a more recent dataset (CMU-dataset) with significant variability in parameters are provided in Figure I-4–3.

Note the challenging imagery that IJB-B contains, including large facial occlusion, poor image quality, and greater illumination variation [5]. In addition to increasing the variability of the images, the number of images in the most recent dataset (i.e., CS3) was increased dramatically as shown in Table I-4–1 to provide better statistical results for the performance metrics. The CMU dataset is recent; however, it has fewer images and faces because it has a very focused topic and was derived from YouTube videos.



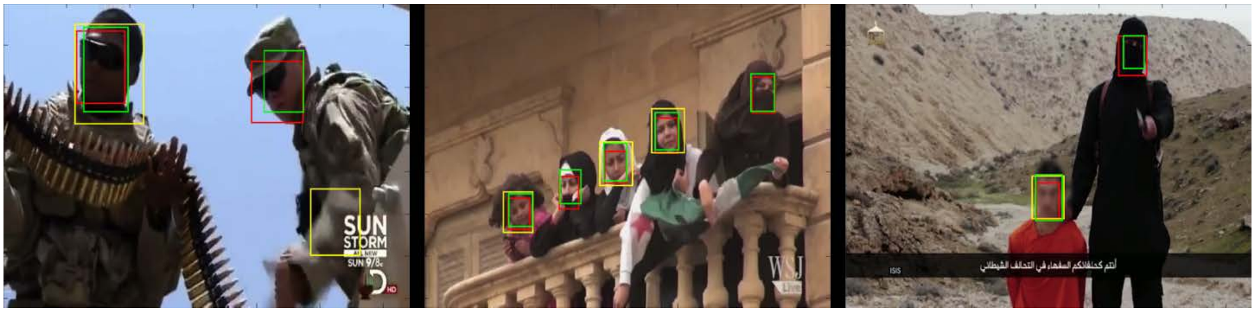
Source: NIST, "An Introduction to the Good, the Bad, & the Ugly Face Recognition Challenge Problem."

Figure I-4–1: Examples of Face Pairs of the Same Person from Each of Three Partitions (a) good, (b) challenging, (c) very challenging



Source: NIST, “IARPA Janus Benchmark-B Face Dataset,” May 2017.

**Figure I-4–2: Sample Imagery of Successive Levels of Difficulty
IJB-B (top), IJB-A (middle), LFW (bottom)**



Source: RT&E Center, "Proposed Technical Approach NIJ RT&E Center," February 2017.

Figure I-4–3: Examples of Images with Significant Variability

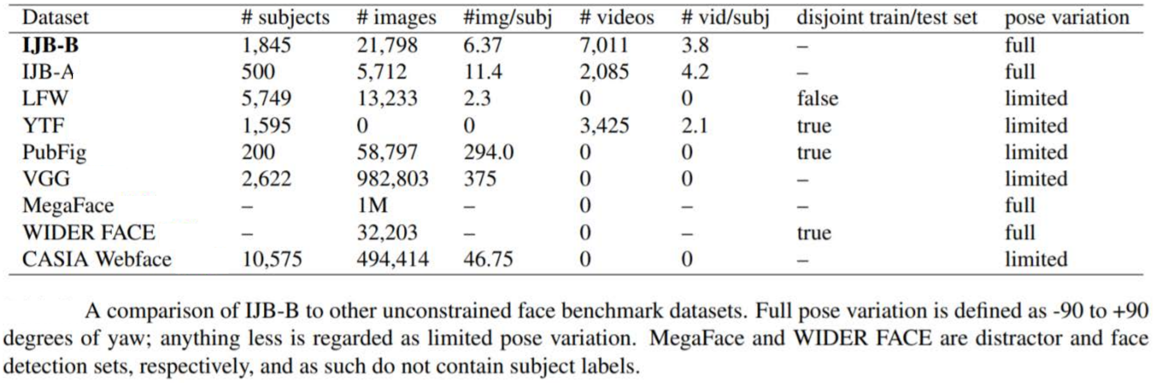
I-4.1 Dataset Descriptions

The three datasets selected for testing are summarized in Table I-4–1. For comparison, Table I-4–2 provides similar information for recent datasets that shows the comparison of face benchmark dataset parameters with the CS3 (i.e., IJB-B) and WIDER FACE datasets. A primary assessment requirement is that each detector sees the same datasets; consequently, each dataset was processed with each of the four algorithms to determine if there were any format or execution problems. About 20 images had to be removed including all ".gif" files and all images YOLO could not process. These latter images were not regular RGB images that the detectors expected and had to be removed or converted to RGB. The next to last two columns of the table provide the numbers of final images and faces. The last column provides some general notes on individual datasets.

Table I-4–1: Summary Comparison of Face Datasets

Dataset	# Subjects	Still images with faces	Images with no face	Video frames	Pose variation	Images–final	Faces–final	Issues
CMU	Unknown	1920	484	0	Yes	2404	3729	Small volume, ground truth based on context
WIDER FACE - Validation Set	Unknown	3226	0	0	Full	3226	39697	WIDER FACE validation set used as is
CS3	1871	11871	8247	55371	Yes	75489	123876	Ground truth BBs created for head – not face. Small number not in a usable format

Table I-4–2: Summary of Recent Face Benchmark Datasets



Dataset	# subjects	# images	#img/subj	# videos	# vid/subj	disjoint train/test set	pose variation
IJB-B	1,845	21,798	6.37	7,011	3.8	–	full
IJB-A	500	5,712	11.4	2,085	4.2	–	full
LFW	5,749	13,233	2.3	0	0	false	limited
YTF	1,595	0	0	3,425	2.1	true	limited
PubFig	200	58,797	294.0	0	0	true	limited
VGG	2,622	982,803	375	0	0	–	limited
MegaFace	–	1M	–	0	–	–	full
WIDER FACE	–	32,203	–	0	–	true	full
CASIA Webface	10,575	494,414	46.75	0	0	–	limited

A comparison of IJB-B to other unconstrained face benchmark datasets. Full pose variation is defined as -90 to +90 degrees of yaw; anything less is regarded as limited pose variation. MegaFace and WIDER FACE are distractor and face detection sets, respectively, and as such do not contain subject labels.

Source: Adapted from IARPA Janus Benchmark-B Face Dataset.

The ground truth BBs for individual datasets are provided by the dataset creators. The BBs are assigned by humans and the process for creating the ground truth BBs was not uniform, either within a dataset or across all datasets.

I-4.1.1 CMU Dataset

The CMU-dataset was created by Carnegie Mellon University personnel and is comprised of images extracted from videos retrieved from a YouTube search using keywords like “Iraq war” and “ISIS.” The videos were then reduced to frames. Faces that met the detection prerequisite were labeled with BBs for each frame and the file names identify the video and frame number. The resulting dataset contains 2,404 images including 484 images with no faces and 1920 images containing 3,729 faces. This dataset contains extreme occlusions that do not show up in many academic datasets that other face detection methods benchmark against, as shown previously in Figure I-4–3. The images are annotated to show ground truth BBs (red), PittPatt detector BBs (yellow), and Ultron detector BBs (green).

I-4.1.2 WIDER FACE

At the time of its creation, the WIDER FACE dataset was considered to be the largest and best annotated face detection image set since, relative to existing datasets, it contained considerably more annotated images that spanned a high degree of variability in key parameters for face detection, i.e., scale, pose, occlusion, expression, makeup, and illumination as shown in examples of [7]. The reference provides a more thorough description of how the dataset was collected and a package evaluation code. Table I-4–2 compares the WIDER FACE dataset with earlier datasets.

WIDER FACE is a face detection benchmark dataset with 32,203 images and 393,703 annotated faces. It only supports protocols, i.e. the set of measures/procedures used to perform an evaluation, designed for face detection, and thus cannot be used to evaluate face verification or identification directly. It is split into three subsets: a training set, a validation set, and a test set. These sets are divided into a 40-10-50 split respectively. The dataset includes the ground truth BBs for the training and validation datasets, which are available on their website. This assessment only used the validation dataset (39,697 faces from 3,226 images, Table I-4–1) for two reasons: because the validation set had ground truth BBs, which the test set does not have, and because this assessment strives to avoid biased results by testing with data used to train the detectors. All of the neural network face detectors leveraged the WIDER FACE training set for detector training. Precision and recall are the primary metrics used by the WIDER FACE site to measure performance of face detection algorithms on the WIDER FACE dataset.

I-4.1.3 Challenge Set 3

IARPA developed datasets for conducting their *Facial Recognition Challenge* program, building on original facial recognition datasets from the National Institute of Standards and Technology (NIST). Information on the NIST Face Projects may be found here: https://www.nist.gov/programs-projects/face-projects. The Challenge Set 3 (CS3) dataset is essentially the same as the IARPA Janus Benchmark-B dataset [5] and was developed by NIST. It is a superset of the IJB-A. IJB-B contains approx. three times as many subjects and four times as many images as IJB-A; it also contains a significant number of images without faces. An "unconstrained" face recognition system should have the ability to perform successful face detection, verification, and identification regardless of subject conditions (pose, expression, occlusion) or acquisition conditions (e.g., illumination, standoff). Both IJB-A and IJB-B were developed in support of this processing goal. The IJB-A dataset was not constrained, however, it was limited in number of subjects in the subsets of the data used for both training and testing which limited the ability to evaluate the algorithms at the lower ends of the receiver operating characteristic (ROC) curve, i.e. there are fewer data points available to define the ROC curve at the curve extremities. A relative comparison of the three datasets is provided in Table I-4–2, which is adapted from [5].

IJB-B consists of 1,871 subjects with human-labeled ground truth face BBs, eye/nose locations, and covariate metadata such as occlusion, facial hair, and skin tone for all images. It also has a more uniform distribution of subjects across the globe. All subjects are ensured to have at least two still images and one video in which their faces appear.

As part of the initial round of processing, the authors found that some of the algorithms (primarily YOLO) could not support some image formats contained within this dataset. Consequently, to ensure a strict one-to-one comparison between the detection algorithms, some images were removed from individual datasets. The results presented in Section I-4.3 span only the subset of images with supporting formats. After removing unsupported images, the remaining dataset contained 21,798 still images (11,871 images with faces, and 8247 images without faces) and 55,371 video frames for a total of 75,489 images (Table I-4–1). A total of nine images were removed; five from the images without faces and four from the images with faces. The authors did not exclude any images from the set of video frames.

The total number of faces from the CS3 dataset used in the assessment is 123,876.

I-4.2 Metrics

The field of face analytics (detection, reconstruction, matching, recognition) has developed to the point where there are well established protocols for designing and executing tests and metrics for performing assessments. These are used by various "challenge" programs including the IARPA Janus Program. The process of face detection includes predicting one or more BBs in an image and assigning corresponding detection confidences. In the performance assessment for an algorithm, the set of predicted BBs are compared with the set of ground truth BBs for that image. The spatial relationship between the two sets of BBs is used to define various face detection metrics, and the metrics are the basis for the analysis in Section I-4.3. Note that for these definitions, metrics range between zero and one.

The analysis in the subsections of Section I-4.3 uses the BB information in the context of IOU-based metrics. For illustration purposes, Figure I-4–4 shows several faces and two (associated) BBs assigned to each face. Throughout this document, the human assigned BB (i.e., ground truth BB) is depicted in green and the detector BB (an output of the face detection algorithm) is depicted in red. Before the relationship between two BBs is explored, the BBs have to be associated. The association algorithm is outlined in Appendix I-A. The output of the association process may be a set of associated BBs, a set of unassociated detector BBs [false positives (FPs)] and a set of unassociated ground truth BBs [false negatives (FNs)].

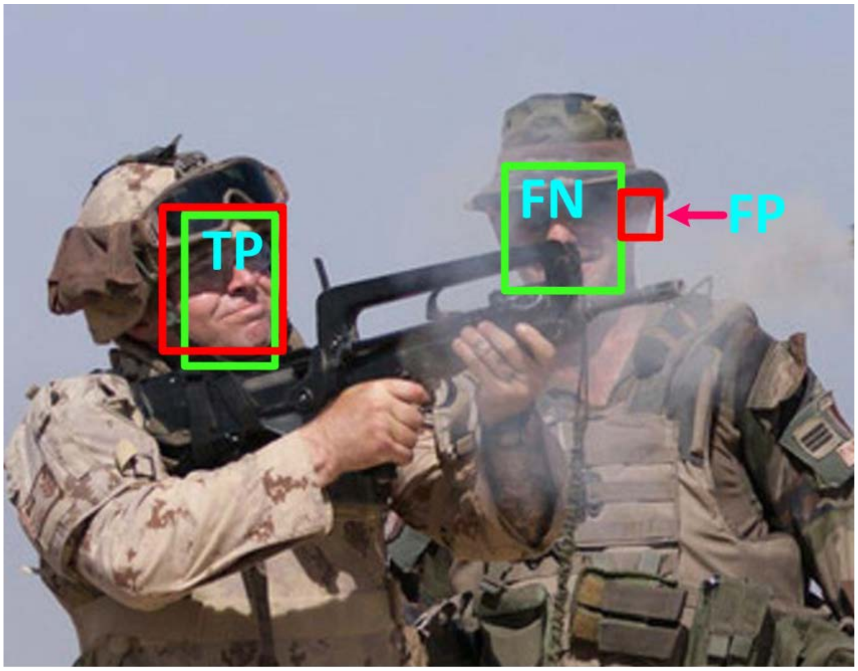


Figure I-4–4: Illustration of Bounding Box Relationships for IOU-based Metrics

I-4.2.1 IOU-based Metrics

For the set of associated BBs and IOU-based metrics, a valid detection occurs when the ratio of the intersection of the two BB areas to the union of the areas exceeds a threshold (the IOU-threshold). The annotation on Figure I-4–4 (an image) shows an example on the left face of the figure where BB regions overlap (and the overlap exceeds the IOU threshold) designated as a valid detection [i.e., true positive (TP)]. The face on the right side of the image is not a valid detection, instead there is one FP since the red box does not overlap any part of a green box and one FN where the green box does not overlap any part of a red box. In summary, this image has one TP, one FP, and one FN. For IOU-based metrics, the integer number of TPs, FPs, and FNs are used to calculate the precision and recall values.

For IOU-based metrics, the standard precision and recall metric definitions are:

- **Precision –** Precision is a measure of the number of “true” detections (i.e., TPs) divided by the total number of detections (= true detections + false detections).
  Precision = TP /(TP + FP)
- **Recall –** Recall is a measure of the number of “true” detections the algorithm generates divided by the number of actual faces in the ground truth data.
  Recall = TP/(TP + FN)

Associated with precision and recall is the Precision-Recall (PR) curve (see example in Figure I-4–5).

These results are computed using the detector confidence score. In particular, starting with a threshold value (= 0) and incrementing the threshold after each calculation, the precision and recall values are calculated for all detections with a confidence value greater than the threshold. This type of curve shows how precision and recall performance can be traded for each detector based on the needs of a particular application.

In lieu of using an eleven point mean average precision (MAP), authors ([8], [9]) recommend using every point to calculate an average precision (AP) to be more discriminative in situations where the value is low. The RT&E Center identified that using every point is proportional to the area under the Precision-Recall curve, and adapted this as a single scalar value metric to describe detector performance. The behavior of the PR curve can be erratic in the vicinity of the left hand axis (see Figure I-4–5). The possible range of PR curves is (0,0) to (1,1); however, many PR curves are only defined for a portion of the range. Consequently, when computing the area under the curve for precision recall (AUC-PR), the computation stops at the largest recall value. When the PR curve has “significant” gaps between the vertical axis and the first recall value of the curve, the PR data has to be extrapolated; the extrapolation method influences the AUC-PR calculation. This is most visible for the TinyFace detector processing all three datasets. In some cases, it is a result of numerical resolution limitations in the TinyFace generated confidence value. APL identified two referenced extrapolation methods:

- Straight line segment between the point (0,1) and the first point of the PR curve [9];
- Straight line segment between the precision value of the first point of the PR data and the same precision value at recall equal “0” [11].

The RT&E Center used the second extrapolation method for all precision recall curves to prevent AUC-PR values from inflating if there was a large gap between the precision axis and the leftmost point. The AUC-PR integral is approximated with a Riemann sum with samples taken at the midpoint.

I-4.2.2 F1 Score

The F1 score is the harmonic average of the precision and recall; an F1 score reaches its best value at 1 (perfect precision and recall) and worst at 0. [12]

F1 = 2 x (precision x recall) / (precision + recall)

I-4.2.3 Segmentation Score

The segmentation score is an indication of how well the detector BB overlays the ground truth BB [9]. The TP, FP, FN values are recorded for all images greater than the selected confidence; possible scores range from 0 to 1. For the scope of this assessment

Segmentation = TP / (TP + FP +FN)

I-4.3 Results

The results are divided into three subsections: overview, discussion of PR and related metrics, and a discussion of the relation between ground truth and detector BBs. Each subsection provides background for the analysis in the subsection, a brief description of ideal results, tables summarizing graphical data, and qualitative observations. Most of the graphical data is provided in Appendix I-B.

I-4.3.1 Overview

The primary purpose of this effort is to compare the face detection capability of Ultron with the three other face detectors discussed in Section I-3 for each of the three datasets enumerated in Section I-4.1. The methodology combines standard metrics used in data science (e.g., IOU-based metrics, PR curves) with a statistical analysis of the relation between the BBs that define ground truth and detector output. The approach is summarized as:

- Adapting both dataset and detector to the software framework (Section I-2.3) for processing;
- Executing the algorithms with a specific dataset;
- Recording the metric values (Section I-4.2) and generating graphical displays; and
- Comparing/contrasting results between detector-dataset combinations.

For comparison, the Proposed Technical Approach NIJ RT&E Center [1], provides values for several IOU-based metrics with a threshold value equal to 50% for the Ultron detector and the CMU-dataset. The IOU threshold of 50% is a standard across object detection literature [8], [9].

Many of the data in the figures in this section are distributions of a given metric value. The distributions have a bin size of 0.02 and a range from 0 to 1.0. In most cases, the contents of the lowest bin (i.e., 0–0.02) or the highest bin (0.98–1.00) are extracted to a table to provide the number of identically "0" or "1" values and the number of remaining values in the bin. Since the value in the lowest or highest bins can be quite large, extracting them provides additional resolution of the metric space and a more detailed discussion of the extreme values (i.e., 0, 1).

Section I-4.3.2 presents standard (IOU based) macro precision and recall results seen in academia ([7], [8]) for an IOU threshold value of 50%. To further explore the impact of, and highlight differences between, detector-dataset combinations, Section I-4.3.3 investigates the sizes, size ratios, and association rates of the ground truth and detector BBs for each detector-dataset combination and ground truth.

I-4.3.2 Precision and Recall Metric Values

Precision and recall values are routinely used to characterize an information retrieval system where there are infinitely many true negatives; the ideal capability is both high recall and high precision. The trade-off between these two variables is shown in a PR curve; and, the trade-off is fixed by selecting an operating point on the curve. Figures I-4–5 through I-4–7 present PR curves for each dataset processed by each of the four detectors; and Figures I-B–1 through I-B–4 present the PR curve for each detector processing all three datasets. Table I-4–3 presents AUC-PR values for each detector-dataset combination.

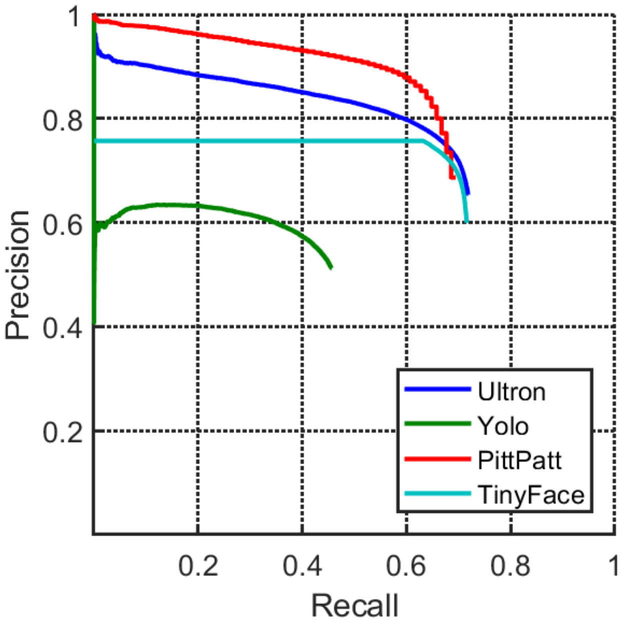


Figure I-4–5: PR Curves for all Detectors on CS3 Dataset

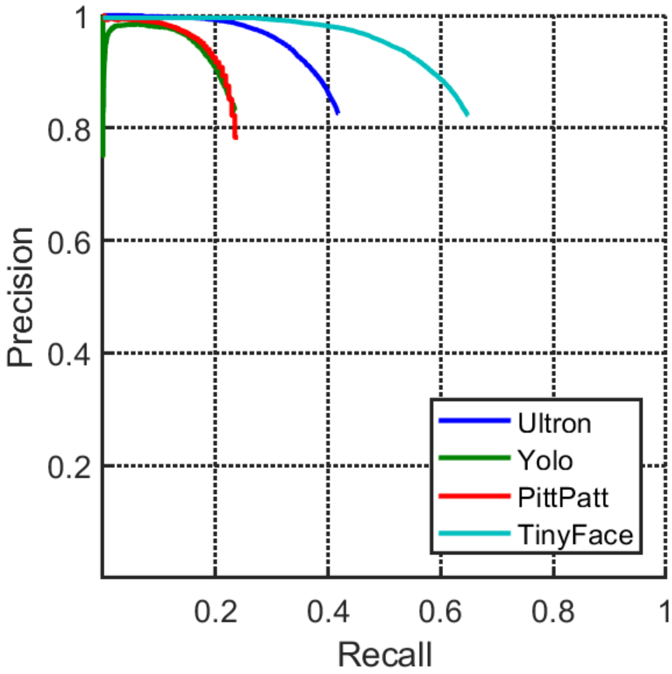


Figure I-4–6: PR Curves for all Detectors on WIDER FACE Dataset

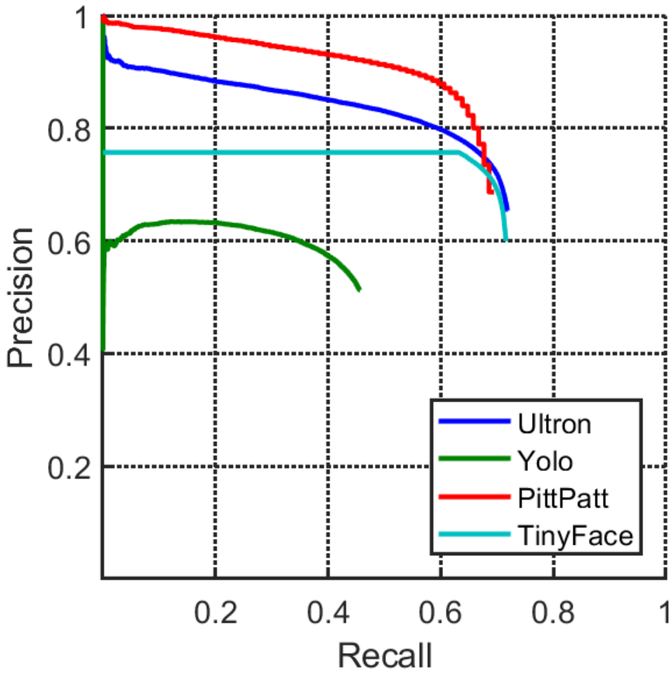


Figure I-4–7: PR Curves for all Detectors on CMU Dataset

Table I-4–3: Area of Precision-Recall Curves (AUC-PR)

Detector\Dataset	CMU	CS3	WIDER FACE
PittPatt	0.1452	0.6434	0.2301
TinyFace	0.7059	0.5384	0.6278
Ultron	0.7687	0.6090	0.4048
YOLO	0.3402	0.2777	0.2244

Note that in Figure I-4–5 the PR curve for the TinyFace detector appears to extend horizontally to a Recall equal to “0. This extrapolation is done for all detectors on all datasets for consistency, but is most apparent in this curve because of the nature in which TinyFace was trained versus the dataset it was presented with. This approach is justified in [11], and is necessary to calculate the AUC-PR value with consistent lower bound of “0” for the integration range over the recall dimension.

The maximum F1 score for each detector-dataset combination is provided in Table I-4–4.

Table I-4–4: Max. F1 Scores for Detector-Dataset Combinations

Detector\Dataset	CMU	CS3	WIDER FACE
Ultron	0.8346	0.7114	0.5549
TinyFace	0.7911	0.7002	0.7252
PittPatt	0.3458	0.7313	0.3672
YOLO	0.5230	0.4848	0.3672

The 50% IOU-based metric values for Ultron processing the CMU dataset that were generated by Carnegie Mellon University [1] and the corresponding values generated by RT&E Center are compared in the conclusions section. The key results are provided in Table I-4–5. The reason for the differences is currently unknown.

Table I-4–5: Comparison of CMU and RT&E Center Generated Precision and Recall Values

Organization	50% IOU-based Precision	50% IOU-based Recall
CMU	0.8856	0.8305
RT&E Center	0.860	0.806

Summary observations based on the PR curves, the AUC-PR values, and the max. F1 scores include:

- No single detector outperforms the others on all three datasets,
- YOLO underperforms Ultron on all three datasets, and
- Ultron and TinyFace perform similarly for the CS3 and CMU datasets.

I-4.3.3 Characterize BBs for Datasets

This subsection explores the distribution of IOU values and BB sizes for the different detectors to provide further insight into their relative performance. There are three metrics that it explores: (1) a global IOU metric defined as segmentation accuracy by [9], (2) distributions on average IOU on a per image basis, and (3) an analysis of the variability in ground truth BB subjectivity across the three datasets this study spans. The BB sizes are given as the number of pixels in the BB; the IOU calculation and related metrics are discussed in Section I-4.2.1. Intuitively, an ideal result is when the distribution of ground truth and associated detector BB sizes are closely aligned; and, the distribution of IOU values (ratio of detector to ground truth BB areas) indicates agreement between the two areas.

I-4.3.3.1 Segmentation Accuracy

Table I-4–6 presents the maximum segmentation score for each detector-dataset combination.

The values indicate that:

- YOLO has the smallest overlay for all three datasets (one exception);
- PittPatt is best on CS3;
- Ultron is best on CMU-dataset; and
- TinyFace is best on WIDER FACE.

Table I-4–6: Maximum Segmentation Scores

Detector\Dataset	CMU-dataset	CS3	WIDER FACE
PittPatt	0.2090	0.5273	0.2249
TinyFace	0.6300	0.4840	0.5686
Ultron	0.7110	0.5204	0.3840
YOLO	0.3541	0.3183	0.2249

I-4.3.3.2 IOU Distribution

A second perspective on the relationship between ground truth and detector BBs is presented in the graphs of IOU distributions, the behavior of the detector-dataset combinations at the extremes of the distributions, and the statistical representation of the distributions. The tabular data for the detector-dataset combinations are presented using the same approach as discussed in Section I-4.3.1.

As an example, Figure I-4–8 presents the distribution of average IOU values for the CMU-dataset; analogous results for all the datasets are provided in Figures I-B–5 through I-B–7. The average IOU is the mean for all detections per image. Each figure shows distribution results for all four detectors – one figure for each dataset. Unassociated ground truth and detection BBs were tallied as an IOU of “0.” A point at exactly “0” indicates that no ground truth or detected

BBs bridged the IOU threshold. To fill in the distributions in each figure, the IOU threshold was lowered to a value less than the bin width. To keep an informative scale, and to identify counts of exactly "0" and exactly "1," the first and last bins are omitted from the figure and analyzed in Table I-4–7. The sample mean and sample variance of all distributions are provided in Table I-4–8.

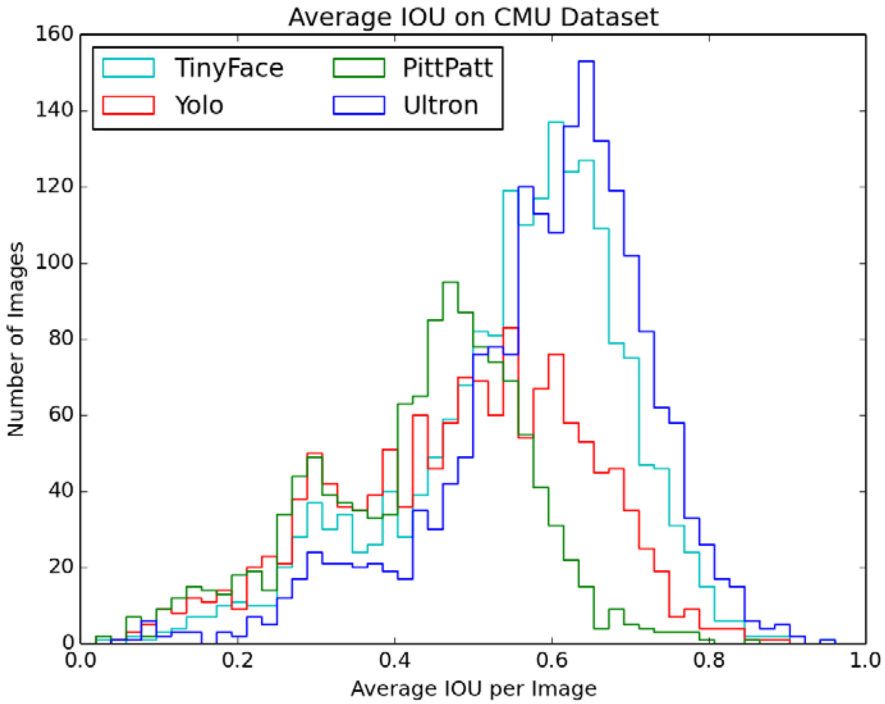


Figure I-4–8: Distribution of Average IOU for CMU-Dataset

In summary, for detections based on a 50% IOU threshold:

- Ultron and TinyFace have fewer images with unassociated detections implying either fewer missed or invalid detections; and
- The mean values for the distributions of Ultron and TinyFace are comparable to each other and superior to PittPatt and YOLO except for CS3 dataset where they are comparable to PittPatt.

Taken together, these observations indicate that, in most cases, Ultron and TinyFace detector BBs have better alignment (i.e., more associations and more overlap per association) with the ground truth BBs than either YOLO or PittPatt.

Table I-4–7: End Conditions for IOU Distributions

(Boundary values for Figure I-4–8, and Figures I-B–5 through I-B–7)

Detector	Dataset	Metric	0	0.0 < x <= 0.02	0.98 <= x < 1.0	1
Ultron	CMU	average IOU	519	0	0	0
PittPatt	CMU	average IOU	1151	9	0	0
YOLO	CMU	average IOU	511	2	0	0
TinyFace	CMU	average IOU	46	0	0	0
Ultron	CS3	average IOU	8815	17	0	0
PittPatt	CS3	average IOU	12185	43	1	0
YOLO	CS3	average IOU	14328	33	0	0
TinyFace	CS3	average IOU	8395	35	0	0
Ultron	WIDER FACE	average IOU	70	3	0	0
PittPatt	WIDER FACE	average IOU	553	43	0	0
YOLO	WIDER FACE	average IOU	479	37	0	0
TinyFace	WIDER FACE	average IOU	55	1	0	0

Table I-4–8: IOU Distribution Statistics for Three Datasets

	Ultron	PittPatt	YOLO	TinyFace
CMU-Dataset				
Statistic				
Mean without 0's	0.59	0.43	0.49	0.55
Standard Deviation without 0's	0.14	0.14	0.16	0.14
CS3 Dataset				
Statistic				
Mean without 0's	0.55	0.54	0.47	0.54
Standard Deviation without 0's	0.14	0.17	0.14	0.15
WIDER FACE Dataset				
Statistic				
Mean without 0's	0.64	0.43	0.50	0.63
Standard Deviation without 0's	0.24	0.23	0.25	0.19

I-4.3.3.3 BB Variability

The precision-recall analysis is based on the relation between the ground truth and detector BBs through IOU values. This section analyzes the number, size, and association of the BBs generated by the face detectors and ground truth. In the ideal case all BBs would be associated and have the same size as the BB they are associated with. The detector to ground truth BB size ratios and IOU values are similar; however, the size ratio speaks directly to the relative sizes whereas the IOU speaks to both size ratio and spatial relationship. The distribution of sizes for each detector processing the CMU-dataset and the ground truth are presented in Figure I-4–9. To investigate BB variability, the distributions of BB sizes, as represented by the number of pixels in the BB (Figure I-4–9 and Figures I-B–8 through I-B–10); the ratio of detector to ground truth BB sizes; (Figures I-B–11 through I-B–13) and the number of ground truth and detector BBs for each dataset (Tables I-B–1 and I-B–2) are tallied in this section.

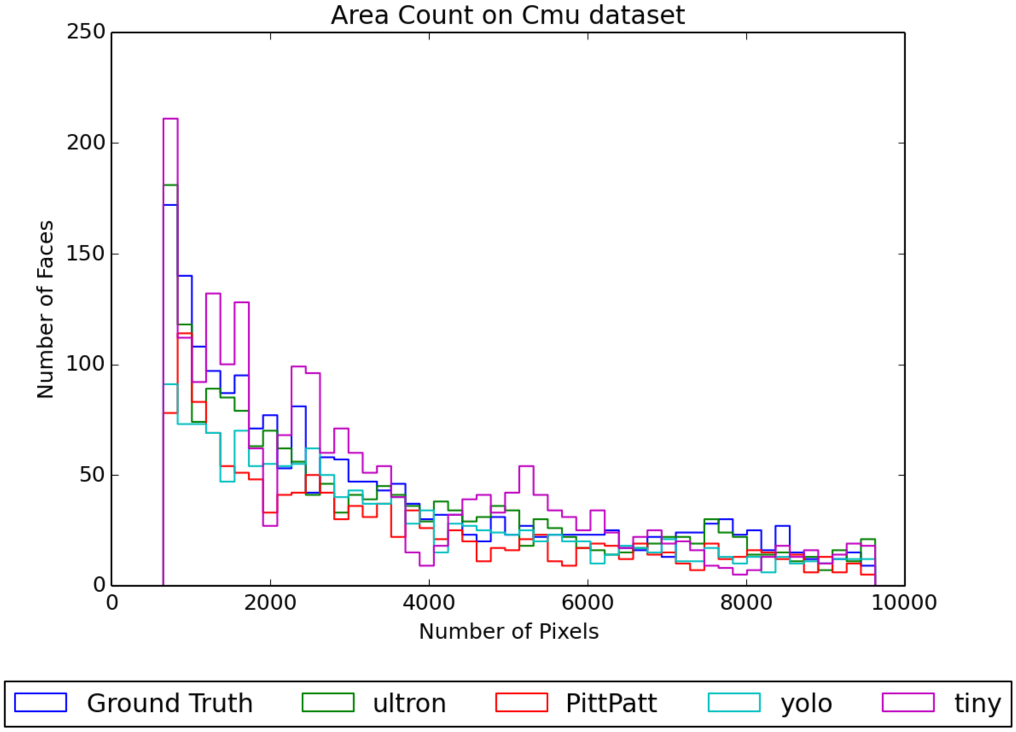


Figure I-4–9: Bounding Box Size in Number of Pixels for CMU-Dataset

To make the BB size distribution (Figure I-4–9 and Figures I-B–7 through I-B–10; and Tables I-B–1 and I-B–2) more readable by eliminating outliers, only BBs with up to 75 percent of the maximum number of pixels in the detector or ground truth BBs for each dataset are presented in the plots. This new sub-region is then divided into five equally spaced bins that span from zero pixels to the maximum number of pixels. More than 99% of BB sizes fall in these five bins. These five bins are presented in Table I-B–1. The number of remaining BBs for each distribution is summarized in Table I-B–2.

Table I-B–1 provides counts for the bins for the following quantities: the range of BB areas (in pixels) for each bin, the total number of detector and ground truth BBs in each bin, the sum of the number of detector BBs for all bins, the total number of associated BBs for the detector and the ratio (as a %) of the number of associated BBs to the total number of detected BBs.

The other parameter directly related to the BB characterization is the ratio of the detector BB size to the ground truth BB size (number of pixels). Distributions of these ratios are provided in Figures I-B–11 through I-B–13. These data include only associated BBs that exceeded the IOU threshold value (0.5). The mean values of the distributions are provided in Table I-B–3.

In addition to the statistical characterization of the distributions of the ratio of detector to ground truth BB sizes in Table I-B–3, the statistical independence of these distributions was compared using the pair-wise hypothesis tests described in Appendix I-C. These results show that the BB size ratio distributions for Ultron and TinyFace are statistically identical and the YOLO and PittPatt distributions are distinct from both Ultron and TinyFace and each other.

Assuming the ground truth BB is correctly assigned, the ideal behavior for a detector BB is to overlay (i.e. IOU = 1.0) the ground truth BB. These results show the CS3 dataset is more generous (i.e., larger ground truth BBs) than the WIDER FACE dataset which, in turn, is more generous than the CMU-dataset. A narrow ratio (i.e., small standard deviation) of the detector BB size to the ground truth BB size distribution implies a consistent relationship between the BBs; therefore, all four detectors have a more consistent, albeit undesirable, relationship with ground truth BBs for the CS3 dataset. The next most consistent is the WIDER FACE dataset and least consistent is the CMU-dataset. The standard deviations for all four detectors give reasonably consistent results across a given dataset. The Ultron and TinyFace detectors show more similar distributions across all three datasets and the analysis provided in Appendix I-C confirms they are statistically similar distributions.

With regard to the distribution of BB size ratios (Table I-B–3 and Figures I-B–11 through I-B–13):

- For the CMU-dataset (Figure I-B–11), the TinyFace and Ultron distributions contain significantly more associations than YOLO and PittPatt; also, all four detectors have size ratios greater than 1.0 indicating the CMU-dataset ground truth BBs are small.
- For the CS3 dataset (Figure I-B–12) the TinyFace, Ultron and PittPatt distributions contain a nearly identical number of associations; all four detectors have mean values less than 1.00; also, PittPatt has the mean value closest to 1.0 which leads to a larger number of detections which, in turn, explains the superior PR curve of PittPatt on this dataset (Figure I-4–5)
- For the WIDER FACE dataset (Figure I-B–13) the TinyFace and Ultron distributions contain significantly more associations than YOLO and PittPatt; and three of the detectors have median value of approx. 1

In summary, the relationship of associated detector BBs and ground truth BBs varies with dataset. For the CMU-dataset, the Ultron and TinyFace detectors generate a number of associated BBs more consistent with ground truth than the other two detectors; however, the results for smaller BB sizes are more erratic than for larger BB sizes. For the CS3 dataset, PittPatt generates a distribution that is statistically different form all other detectors and one that has the highest mean. Finally, for the WIDER FACE dataset, TinyFace and Ultron generate a larger percentage (65% and 42% respectively) than PittPatt/YOLO (approx. 24%) of associated BBs, and the variability is large.

I-5 CONCLUSIONS

The assessment of the Phase 1 effort was conducted to evaluate the performance of the Carnegie Mellon University-developed Ultron face detection software and compare performance results with those of other face detection software capabilities on representative datasets. The selected datasets included the CMU-dataset (extracted from YouTube videos) provided by Carnegie Mellon University, the WIDER FACE dataset (because of its variability and wide-spread use), and the IARPA Challenge Set-B (CS3), which is a large and recent dataset (see Section I-4.1). A summary of the assessment findings are provided in Table I-5–1.

The detector and ground truth BBs, and their spatial relationship, are the basis for the metrics used to characterize this assessment. In addition to IOU-based precision and recall metrics, the distributions of BB sizes, the ratio of detector and ground truth BB sizes and the relations between the number of ground truth BBs, detector BBs and the number of associated BBs was compiled to assess Ultron performance.

The impact of BB annotation on metric calculations is well documented and, approaches to adjusting the annotation to match a detector's output are available. The latter include identifying thresholds for the minimum BB size in a dataset that can be used to compute the metrics ([8], [13]). This assessment adopted the naive approach in which inherent assumptions of all the detectors were ignored. This decision was driven by two factors: (1) the desire for an unconstrained and direct comparison across different (and relevant) datasets, and (2) the lack of information regarding inherent assumptions of the detectors. Other assessments have shown that adjusting the BB size range to the detector BBs result can improve overall results (see Appendix I-D). By comparing the four detectors against the same datasets, the assessment uncovered some differences that were impacted by ignoring assumptions about the detector BBs.

The assessment results are summarized in Table I-5–1. The content categories, defined below, summarize the analysis results; sources for the reported values are provided:

- AUC-PR – the area under the PR curve from Table I-4–3.
- Max. F1 score – product of precision and recall and interpreted as the detector accuracy from Table I-4–4.
- IOU Dist. Mean and standard deviation. – a measure of the desirability of the IOU distribution from Table I-4–8 and Figures I-B–5 through I-B–7.
- Unmatched BBs – a measure of the number of unmatched boxes identified in computing IOU distributions from Table I-4–7.
- Association rate – a measure of the number of detector BBs that were associated with a ground truth BB from Table I-B–1.
- BB size ratios – qualitative assessment of desirability of detector to ground truth BB size distribution in terms of mean and standard deviation from Table I-B–3.

Table I-5–1: Assessment Summary

Detector/Dataset	CMU	CS3	WIDER FACE
Ultron			
AUC-PR	0.7687	0.6092	0.4048
Max. F1 score	0.834632	0.711398	0.554899
IOU Dist. (mean/std. dev.)	0.59/0.14	0.56/0.14	0.64/0.24
Unmatched BBs	519	8815	70
Association rate	85.7%	65.4 %	82.6%
BB size ratios (mean/std. dev.)	1.23 / 0.28	0.65 / 0.11	1.07 / 0.21
TinyFace			
AUC-PR	0.7341	0.6152	0.6283
Max. F1 score			
IOU Dist. (mean/std. dev.)	0.55/ 0.14	0.54/ 0.15	0.63/ 0.19
Unmatched BBs	46	8395	55
Association rate	75.2	59.9	82.3
BB size ratios (mean/std. dev.)	1.31 / 0.28	0.68 / 0.12	1.09 / 0.23
PittPatt			
AUC-PR	0.1453	0.6449	0.2064
Max. F1 score			
IOU Dist. (mean/std. dev.)	0.43/ 0.14	0.54/ 0.17	0.43/ 0.23
Unmatched BBs	1151	12185	553
Association rate	47.9	68.7	78.3
BB size ratios (mean/std. dev.)	1.43 / 0.27	0.78 / 0.17	1.26 / 0.27
YOLO			
AUC-PR	0.3403	0.2778	0.2245
Max. F1 score			
IOU Dist. (mean/std. dev.)	0.49/ 0.16	0.47/ 0.14	0.50/ 0.25
Unmatched BBs	511	14328	479
Association rate	65.2	51.3	83.1
BB size ratios (mean/std. dev.)	1.12 / 0.29	0.67 / 0.15	0.97 / 0.23

I-5.1 Macro Results

Performance of face detection software is typically measured using IOU-based precision and recall with an IOU threshold of 0.5. The RT&E Center results (and dataset parameters) for the Ultron detector processing the CMU-dataset are presented in Table I-5–2; the Carnegie Mellon results are taken from [1]. The reason for the difference is currently unknown.

Table I-5–2: Comparison of Carnegie Mellon and RT&E Center Results

Variable	Carnegie Mellon Result	RT&E Center Result
Number of images	1920	1920
Number of ground truth BBs	3729	3729
Number of detected faces	Unknown	3507
Number of matched faces	Unknown	3007
Accuracy	88.56	85.98
Recall	83.05	80.63
Number of false detections	400	490

The assessment found that from the perspective of AUC-PR, none of the detectors has superior performance for all three datasets. The Ultron detector (with 50 percent IOU threshold) is superior to the other three detectors for the CMU-dataset only (Figure I-4–7); however, Ultron ranks high for the other datasets. TinyFace performance is superior for WIDER FACE and comparable to Ultron for the other datasets. PittPatt has the best AUC-PR only for the CS3 dataset (Table I-4–3); this performance exception is attributed to PittPatt's stronger detection BB area agreement with ground truth BBs in the CS3 dataset. YOLO is one of the lowest performers for all three datasets.

I-5.2 Intersection over Union Variability

In this assessment, a detector's ability to identify a face is characterized by the IOU calculation and IOU threshold. The IOU distributions discussed in Section I-4.3.3 show the variability of IOU across detector-dataset combinations. Figures I-B–5 through I-B–7 show qualitatively, and the Table I-4–8 shows quantitatively that, with one exception, the Ultron and TinyFace detectors have higher mean non-zero IOU metric values for all three datasets than the other two detectors. This indicates better agreement between the detector and ground truth BBs. The exception is that PittPatt and TinyFace have the same mean, to two significant digits, for the CS3 dataset. Also, both Ultron and TinyFace have fewer IOUs with a zero value (Table I-4–7) indicating fewer unassociated BBs compared to PittPatt and YOLO.

The segmentation analysis (Section I-4.3.3) supports these observations since the analysis shows that YOLO has the smallest segmentation scores; PittPatt is best on CS3 dataset; and, Ultron is best on CMU-dataset. TinyFace is the best on WIDER FACE dataset.

I-5.3 Bounding Box Characterization

This section provides information regarding the association and distribution of ground truth and detector BB sizes calculated on a per face basis.

The distribution of detector BB sizes in Figures I-B–8 through I-B–10 shows that for the CMU and CS3 datasets, the total number of TinyFace generated BBs is larger than for ground truth. From Table I-B–1, Ultron, PittPatt, and YOLO have fewer detected faces than ground truth BBs for CMU and WIDER FACE datasets indicating a significant number of missed detections for these detectors. The reverse is true for the CS3 dataset where three of the four detectors have a larger number of detector BBs than ground truth BBs. Finally, the ratio of detector BB area to ground truth BB area for associated BBs varied dramatically across the three datasets (Figures I-B–11 through I-B–13). The statistics are provided in Table I-B–3.

The fact that PittPatt has relatively high association rate on CS3 in Table I-B–1, compared to CMU and WIDER FACE datasets, and relatively lower scores on the other two datasets is attributed to the size ratio distributions in Figure I-B–12. The PittPatt distribution's mean is statistically different from the means of Ultron and TinyFace, as shown in Appendix I-C; it is closer to 1.0 indicating better BB agreement and more detections over the 0.5 IOU threshold for association. The main reason for this is the ground truth subjectivity mismatch between CS3 and the other two datasets.

I-5.4 Overall Conclusions

This assessment ranked four face detection algorithms on three datasets using PR-based metrics, the percentage of detections, BB ratios, and other parameters presented in Table I-5–1. No detector had the best performance for all three datasets; performance ranking varied between detector-dataset combinations. Approximate rankings that weigh the AUC-PR, IOU distribution, and association rate categories of the table are provided:

- For the CMU-dataset, Ultron has marginally better performance than TinyFace, YOLO is a distinct second, and PittPatt is lowest.
- For the CS3 dataset, PittPatt is comparable to TinyFace and Ultron and YOLO is lowest.
- For the WIDER FACE dataset, TinyFace has marginally better performance than Ultron; both of these are superior to PittPatt and YOLO, which are roughly comparable.

In summary, this assessment found Ultron performance to be comparable to TinyFace. PittPatt performance is inferior to that of Ultron/TinyFace and YOLO has the lowest performance.

I-6 ACRONYMS & ABBREVIATIONS

AUC	Area under the curve
BB	Bounding boxes
CMU	Carnegie Mellon University
CNN	Convolutional Neural Networks
CS3	Challenge Set 3
FN	False negative
FP	False positive
IARPA	Intelligence Advanced Research Projects Agency
IJB	IARPA Janus Benchmark
IOU	Intersection over Union
JHU/APL	Johns Hopkins University Applied Physics Laboratory
LEA	Law Enforcement Agency
LFW	Labeled faces in the wild
NIJ	National Institute of Justice
NIST	National Institute of Standards and Technology
PR	Precision/Recall
R&D	Research and Development
ROC	Receiver Operating Characteristic
RT&E Center	Research, Test, and Evaluation Center
TP	True positive

APPENDIX I-A: ASSOCIATION ALGORITHM

The association algorithm is summarized in the following:

- For an image in the dataset, there is a list of ground truth bounding boxes (BBs), detected BBs, and an IOU (intersection over union) threshold value of 0.5.
- Set the following counts to 0. We keep track of separate counts for each image.
  - True Positive (TP)
  - False Positive (FP)
  - False Negative (FN)
- Loop over each ground truth BB such that each is only considered once:
  - Compute the IOU for each BB in the detected BBs with the current ground truth BB. The IOU is the ratio of the number of pixels in the overlap of the two BBs divided by the sum of the number of pixels in the two BBs minus the overlap (see Section I-4.2.1).
  - If the ground truth/detection pair with the highest IOU exceeds the IOU threshold, identify these BB's as associated, increment the TP count by 1. Remove the detection from the list of detections.
  - If the maximum IOU does not exceed the threshold, but is nonzero, increment the FP and FN counts by 1 because the ground truth BB and detected BB are considered independent from one another. The ground truth BB is the FN because no other detection will have a better IOU than the current detected BB. Remove the detected BB from the list of detected BBs because it has been classified as a false positive. This decision opens up the corner case for a FP detection to misclassified as so if there are several clustered ground truth BBs. This corner case is considered a negligible case the will not affect overall results.
  - If there are no nonzero IOUs, increment the to the FN count.
- Add any remaining detection BBs to the FP count.
- Add the TP, FP, and FN counts for the current image to a global count for macro results. Keep track of these counts in three arrays for image-by-image results.

  The TP, FP, and FN values for an image are the final values. Calculate metrics for each image and with the macro totals:

  - Recall = TP/(TP + FN)
  - Precision = TP/(TP + FP)
  - Store these values for this image
- In addition to the IOU based calculations, the assessment also calculated area-based precision and recall metrics. The calculation approach is similar except that instead of summing up the number of precision and recalls across the faces in an image and

calculating an average value, for a detection (IOU exceeds a threshold) the precision and recall values are calculated using TP, FP, and FN areas.

- Repeat the previous steps for each new image until all images in the dataset have been processed.

APPENDIX I-B: GRAPHICAL DATA

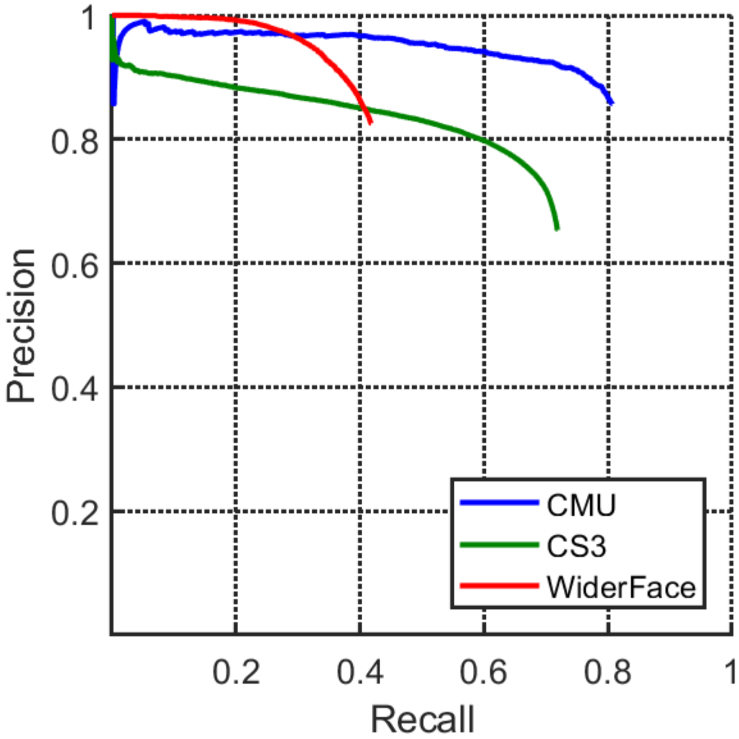


Figure I-B–1: Precision-Recall Curve for Ultron on Three Datasets

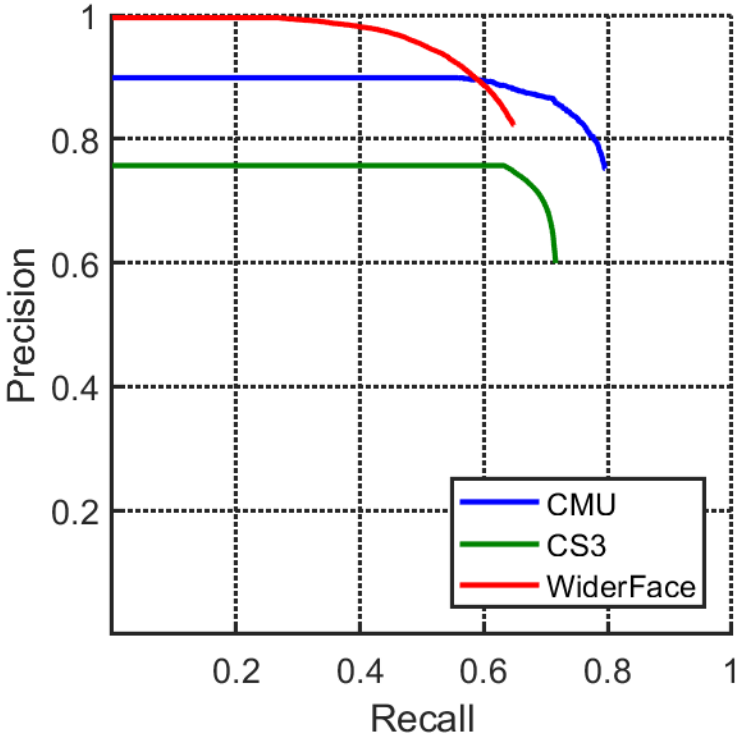


Figure I-B–2: Precision-Recall Curve for TinyFace on Three Datasets

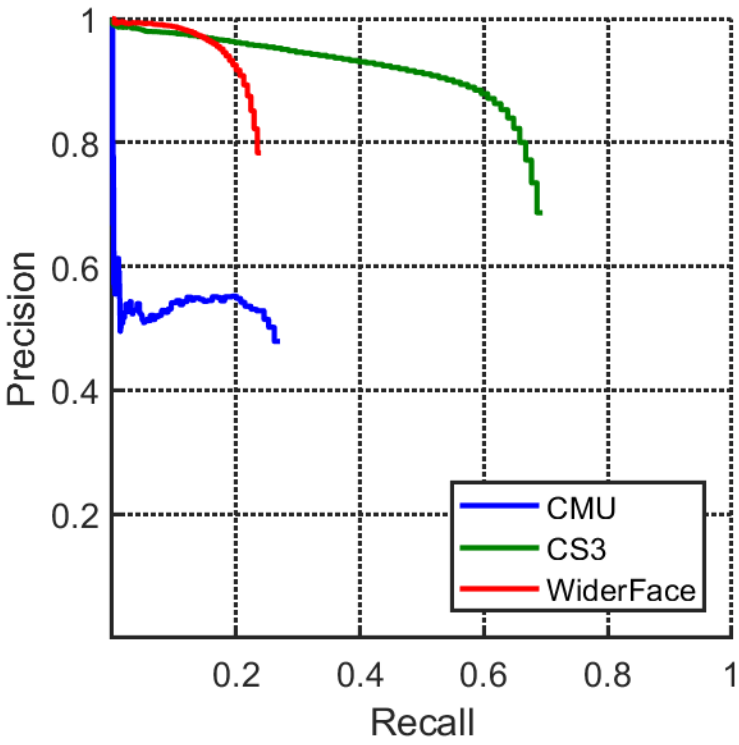


Figure I-B–3: Precision-Recall Curve for PittPatt on Three Datasets

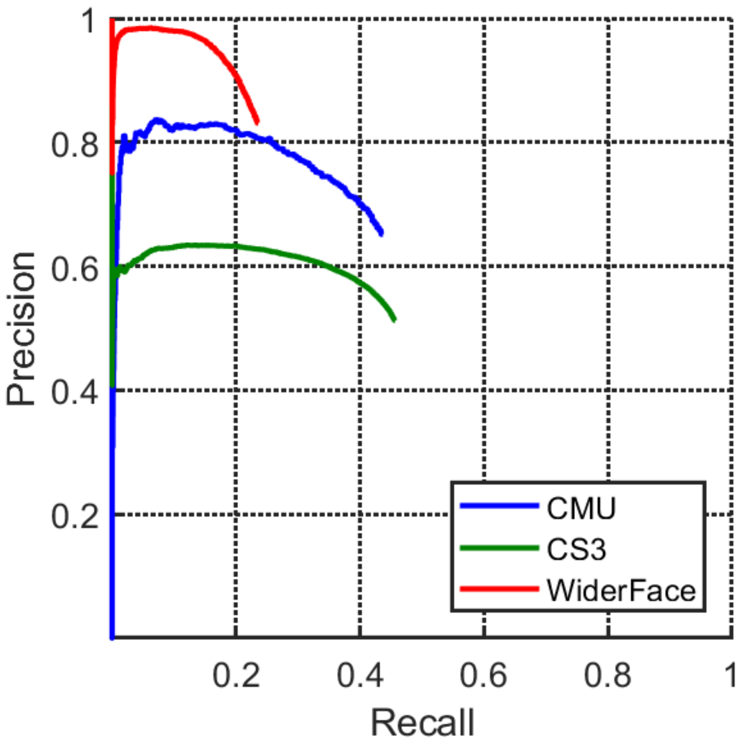


Figure I-B–4: Precision-Recall Curve for YOLO on Three Datasets

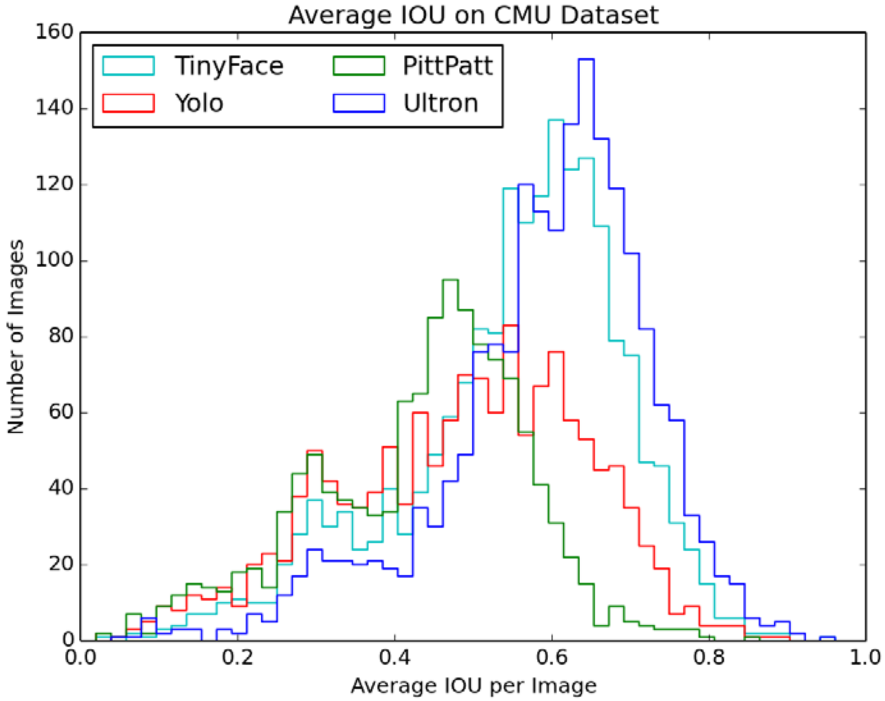


Figure I-B–5: Distribution of Average IOU for CMU Dataset

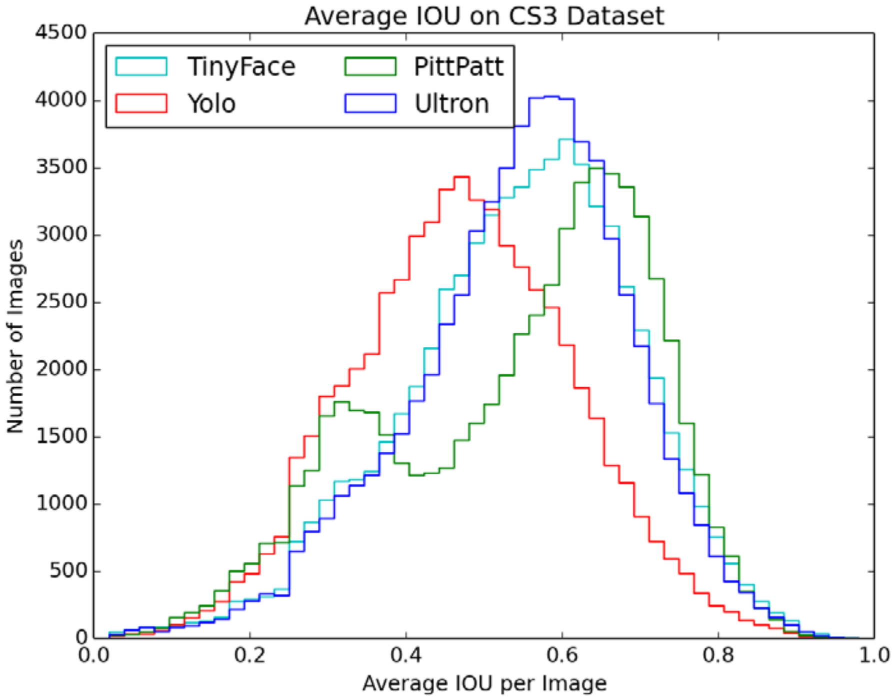


Figure I-B–6: Distribution of Average IOU for CS3 Dataset

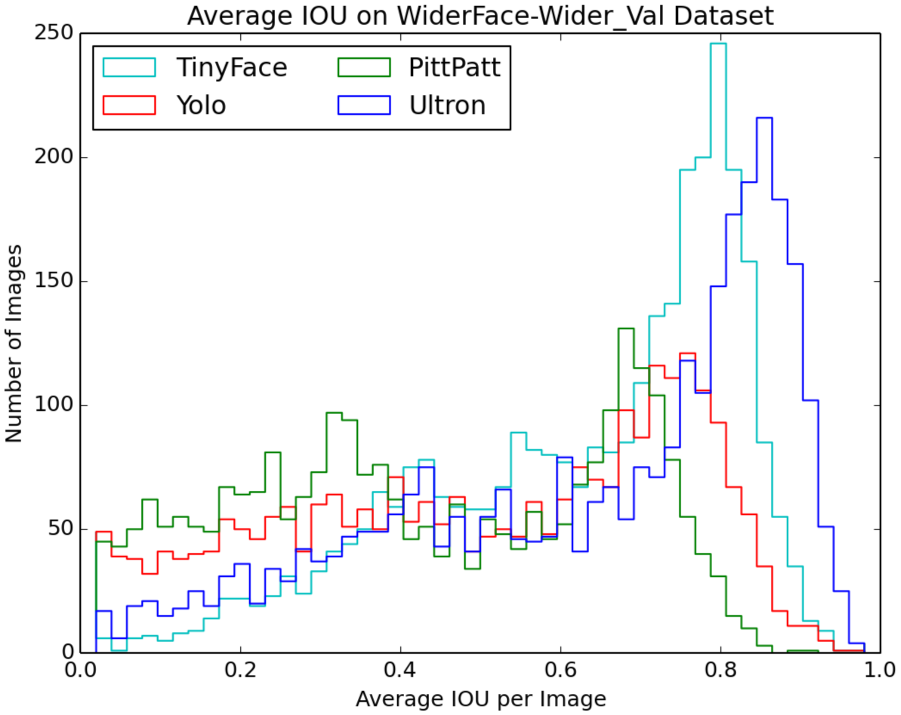


Figure I-B–7: Distribution of Average IOU for WIDER FACE Dataset

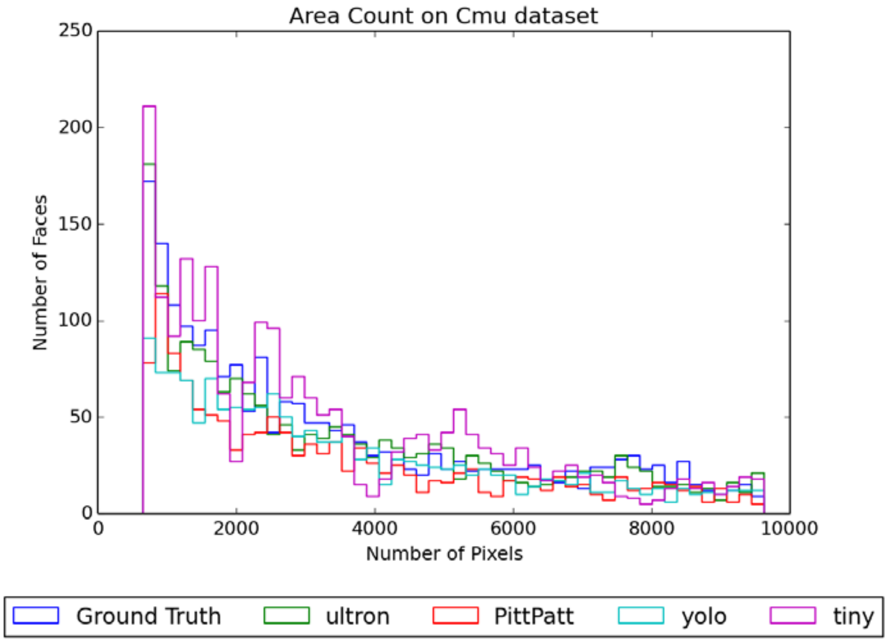


Figure I-B–8: Bounding Box Size in Number of Pixels for CMU Dataset

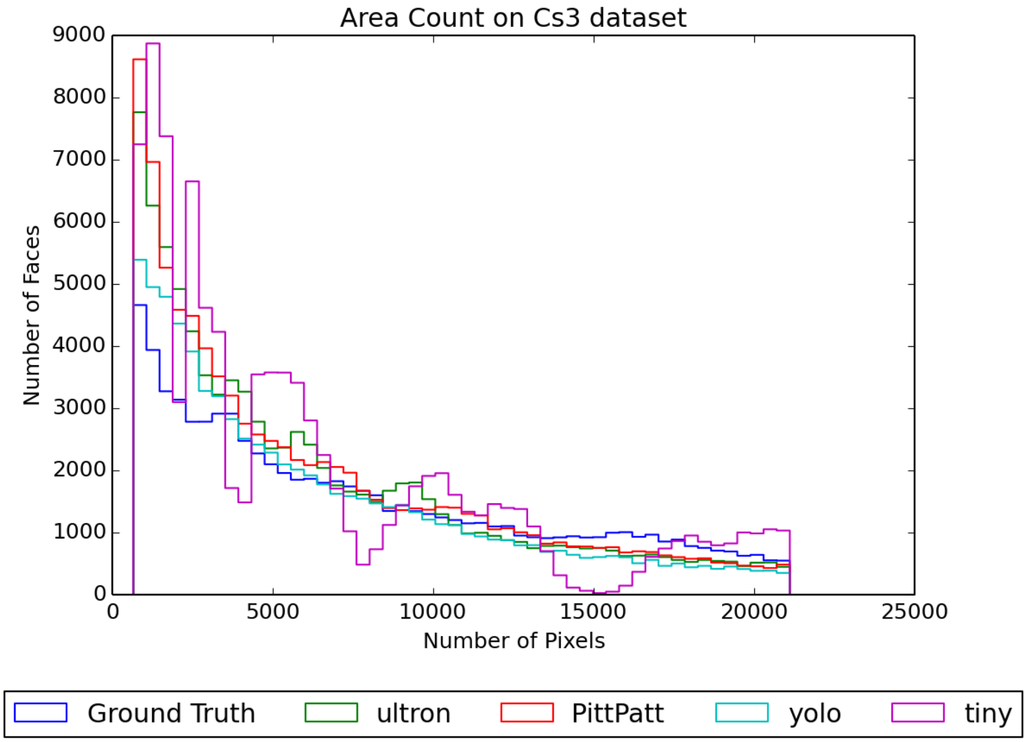


Figure I-B–9: Bounding Box Size in Number of Pixels for CS3 Dataset

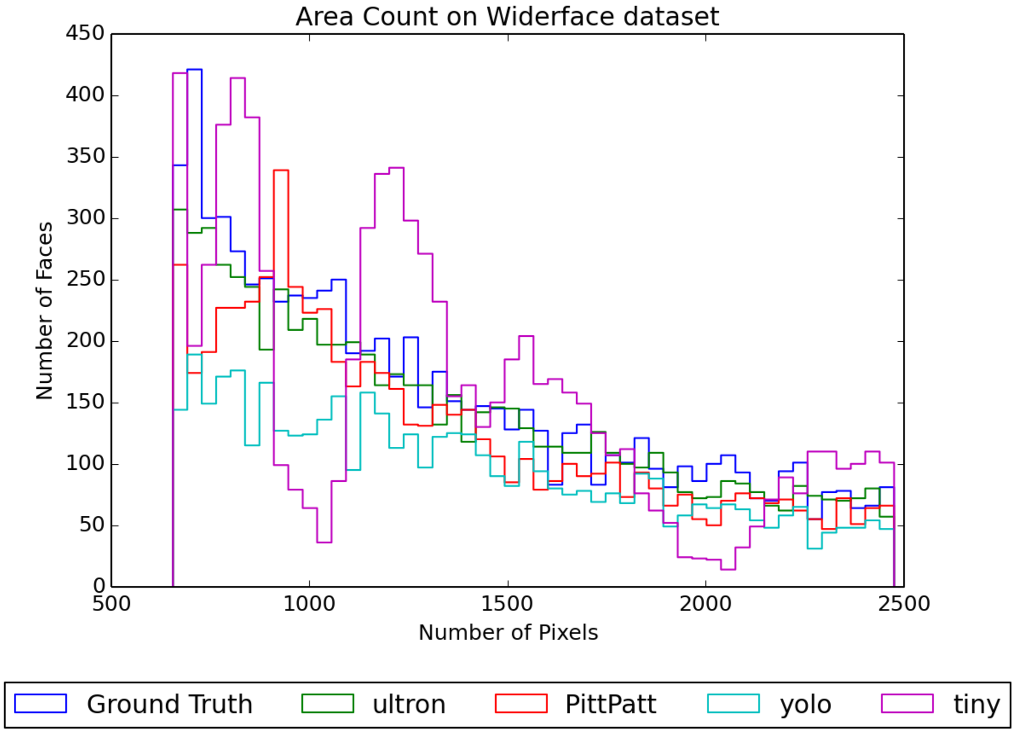


Figure I-B–10: Bounding Box Size in Number of Pixels for WIDER FACE Dataset

Table I-B–1: Bounding Box Sizes and Associations

Right Bin Boundary in Number Pixels	Ultron	TinyFace	PittPatt	YOLO	Ground Truth
CMU Dataset					
18808	3028	3319	1726	2290	3348
37616	332	347	217	152	276
56424	97	208	88	35	72
75232	23	28	42	9	8
94040	2	13	8	2	1
Total	3482	3915	2081	2488	3705
Total Number BBs	3507	3940	2107	2500	3729
Total Number of Associations	3007	2964	1009	1629	N/A
Assoc. Rate (# Assoc. / # BBs)	85.7%	75.2%	47.9%	65.2%	N/A
CS3 Dataset					
347373	135030	146864	123805	109692	121878
694746	629	552	893	397	1192
1042119	202	190	290	102	366
1389492	83	157	128	30	165
1736865	57	44	72	17	90
Total	136001	147807	125188	110238	123691
Total Number BBs	136067	147873	125311	110250	123902
Total Number of Associations	88973	88633	86034	56534	N/A
Assoc. Rate (# Assoc. / # BBs)	65.4%	59.9%	68.7%	51.3%	N/A
WIDER FACE Dataset					
49352	19532	30689	11545	10975	39132
98703	272	286	299	224	276
148055	99	67	116	39	97
197406	62	86	65	11	67
246758	43	26	31	2	45
Total	20008	31154	12056	11251	39617
Total Number BBs	20076	31242	12155	11253	39708
Total Number of Associations	16584	25715	9520	9355	N/A
Assoc. Rate (# Assoc. / # BBs)	82.6%	82.3%	78.3%	83.1%	N/A

Table I-B–2: Uncounted Bounding Boxes

Dataset	Ultron	TinyFace	PittPatt	YOLO	Ground Truth
CMU	25	25	26	12	24
CS3	66	66	123	12	185
WIDER FACE	68	88	99	2	80

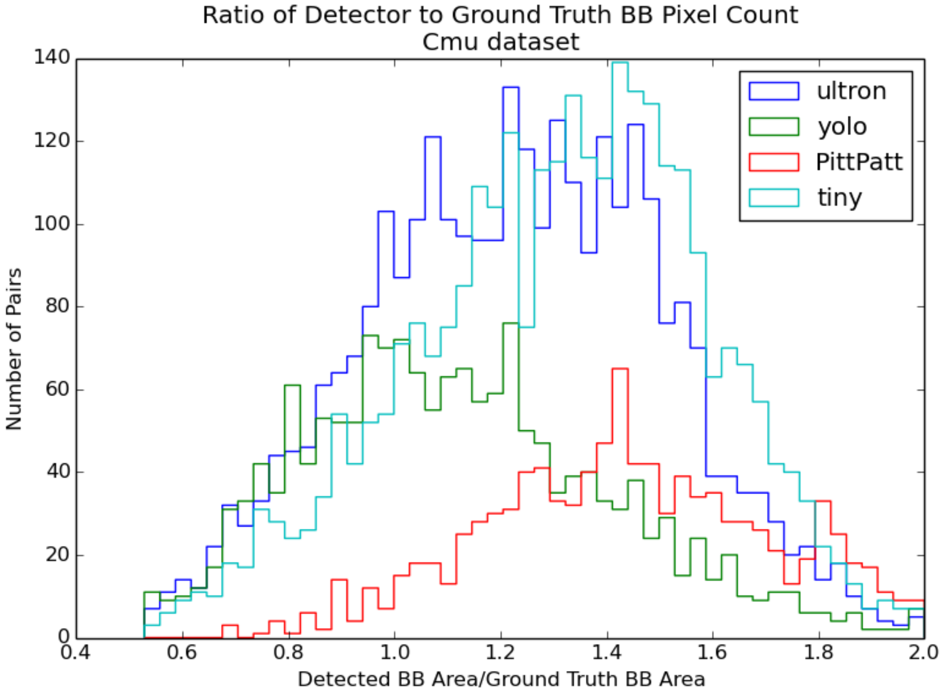


Figure I-B–11: Ratio of Detector to Ground Truth BB Size Distributions for CMU Dataset

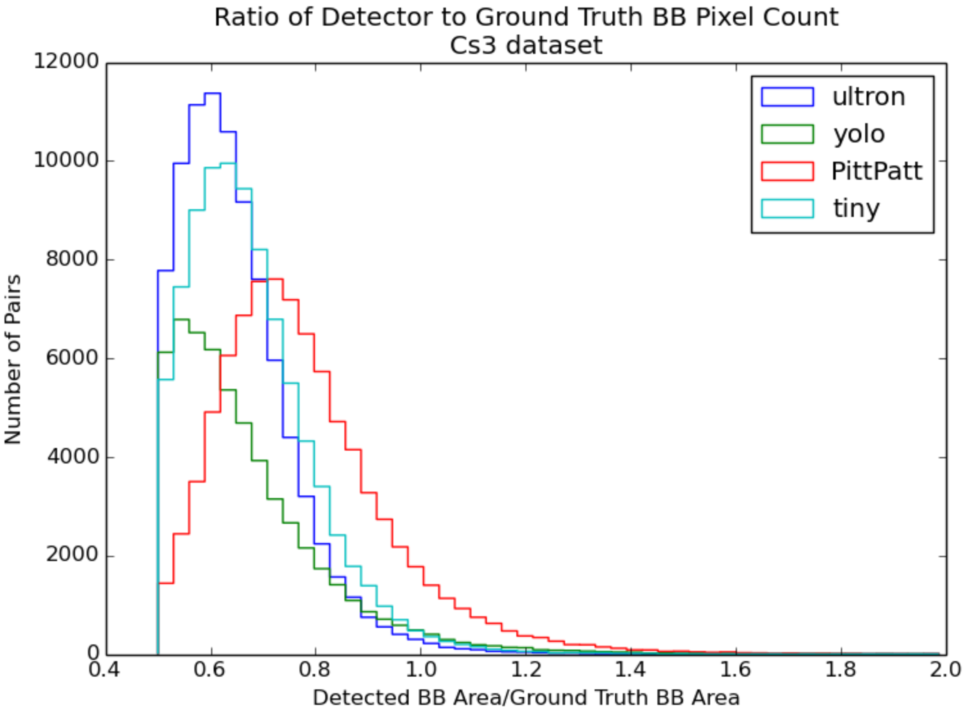


Figure I-B–12: Ratio of Detector to Ground Truth BB Size Distributions for CS3 Dataset

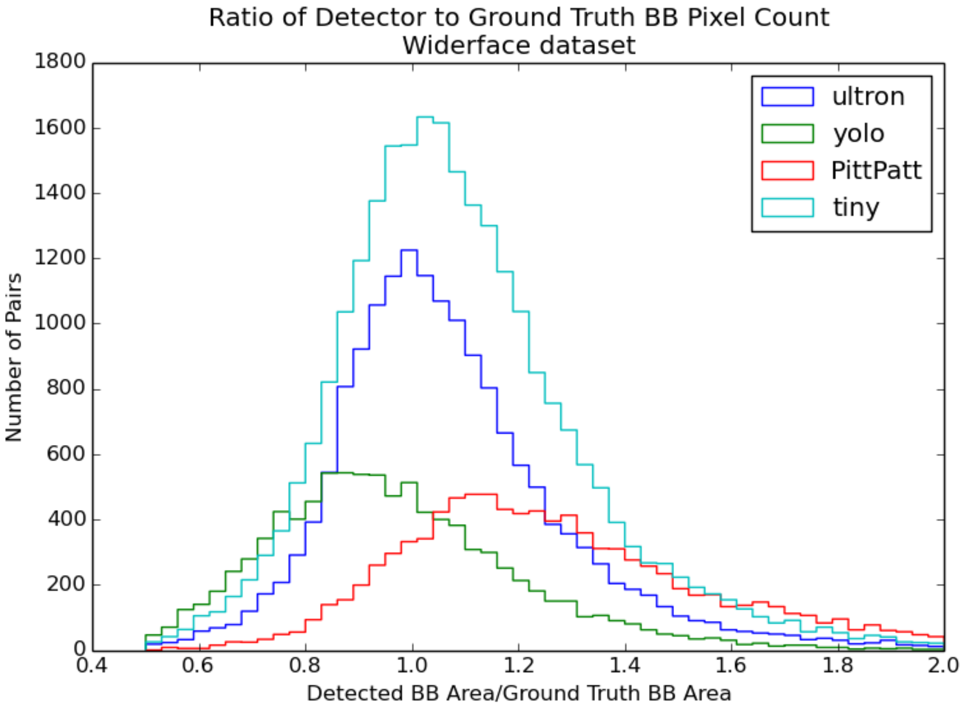


Figure I-B–13: Ratio of Detector to Ground Truth BB Distributions for WIDER FACE Dataset

Table I-B–3: Statistic Values for Ratio of Detector BB to Ground Truth BB Area Distributions

	Ultron	YOLO	PittPatt	TinyFace
CMU Dataset				
Median	1.23	1.10	1.43	1.33
Mean	1.23	1.12	1.43	1.31
Standard Deviation	0.28	0.29	0.27	0.28
Number Associations.	3007	1629	1009	2964
CS3 Dataset				
Median	0.63	0.63	0.75	0.66
Mean	0.65	0.67	0.78	0.68
Standard Deviation	0.11	0.15	0.17	0.12
Number Associations	88973	56534	86034	88633
WIDER FACE Dataset				
Median	1.04	0.94	1.22	1.06
Mean	1.07	0.97	1.26	1.09
Std. Dev.	0.21	0.23	0.27	0.23
Number Associations	16584	9355	9520	25715

APPENDIX I-C: STATISTICAL CHARACTERIZATION OF METRIC DISTRIBUTIONS

Pairwise Comparison of Bounding Box Ratios between Detectors

The following statistical tests were conducted on the ratio of the detector BB size to the ground truth BB size obtained from each detector. These tests sought to validate the statistical differences between the detectors (PittPatt, TinyFace, YOLO, Ultron) when applied to each dataset (CMU, CS3, WIDER FACE). To identify an appropriate pairwise comparison test, a Kolmogorov-Smirnov test was used to identify whether the empirical data was sampled from a normal distribution. The test revealed that all resultant data obtained from all four detectors across all datasets were sampled from non-normal distributions ($\alpha = 0.05$).

An analysis of variance test between non-normal distributions (Kruskal-Wallis One-Way ANOVA) was conducted. It identified that at least one of the detectors is significantly different ($\alpha = 0.05$) from the others reflected across all three datasets. The Mann-Whitney U test was used as a post hoc pairwise comparison between the detectors for each dataset. It was found that, across all three datasets, all four distributions differ from one another except when comparing Ultron versus TinyFace (see Tables I-C–1 through I-C–3 for results). Ultron and TinyFace consistently obtain statistically similar results for the BB ratios across all three datasets, while the remaining detectors obtain statistically dissimilar results.

Table I-C–1: CMU Mann-Whitney U Pairwise Comparison Results where H0 = both detectors have equal means. Green = cannot reject H0, Red = reject H0.

	PittPatt	TinyFace	YOLO	Ultron
PittPatt	NaN	🟥 $1.32e$-215	🟥 $8.12e$-61	🟥 $1.48e$-219
TinyFace	▨	NaN	🟥 $4.00e$-141	🟩 0.33
YOLO	▨	▨	NaN	🟥 $9.35e$-147
Ultron	▨	▨	▨	NaN

Table I-C–2: CS3 Mann-Whitney U Pairwise Comparison Results where H0 = both detectors have equal means. Green = cannot reject H0, Red = reject H0.

	PittPatt	TinyFace	YOLO	Ultron
PittPatt	NaN	🟥 $2.62e$-26	🟥 0	🟥 $5.38e$-33
TinyFace	▨	NaN	🟥 0	🟩 0.16
YOLO	▨	▨	NaN	🟥 0
Ultron	▨	▨	▨	NaN

Table I-C–3: WIDER FACE Mann-Whitney U Pairwise Comparison Results where H0 = both detectors have equal means. Green = cannot reject H0, Red = reject H0.

	PittPatt	**TinyFace**	**YOLO**	**Ultron**
PittPatt	NaN	0	0.04	0
TinyFace	▨	NaN	0	1
YOLO	▨	▨	NaN	0
Ultron	▨	▨	▨	NaN

APPENDIX I-D: RESULTS WITH NORMALIZED GROUND TRUTH BOUNDING BOXES

I-D.1 Background

The NIJ RT&E Center conducted an assessment of the face detection software developed by Carnegie Mellon University as documented in the National Criminal Justice Technology Research Test & Evaluation Center's Phase I report. The assessment was conducted using well-accepted procedures and metrics; however, there was no modification of either the ground truth or algorithm detector bounding boxes. The results showed significant variability in the relations between detector BBs and the ground truth BBs. For example, the ground truth BBs of the CS3 dataset are large enough to include ears and a full head of hair whereas the detector BBs generated by Ultron are not. Similarly, the WIDER FACE dataset provides smaller bounding boxes, so a detector trained on that dataset is at a disadvantage when processing CS3.

At the NIJ's request, the Center contacted representatives of the CS3 dataset creator (Noblis) and was informed of the BB normalization procedure designed to minimize the impact of weak compatibility between ground truth and detector algorithm generated BBs. Noblis recommended the Center "follow the protocol" which effectively reduced to resizing the ground truth BBs to be more compatible with the BBs generated by the face detector algorithm. A key point is resizing should be part of "the protocol." A Noblis representative stated, "For face detection evaluation, we normalize/resize the detected bounding boxes before associating w/the ground truth bounding boxes." The Noblis staff were under the impression the RT&E Center was doing the assessment using the Janus Evaluation Tools, which included the resizing.

The resizing rationale is discussed in some depth in ([13], [28]), which discusses the impact of several parameters on performance of individual algorithms. The RT&E Center explored the resizing concept in several ways: other instances of re-sizing, approach to resizing, and the impact of resizing on results. This appendix summarizes the findings

I-D.2 BB Normalization

The impact of none or minimal annotation guideline standards for dataset annotation is acknowledged in the face image processing community. This is compounded by natural variability in human results when asked to perform subjective tasks. Dataset protocols designed to reduce mismatches between datasets and detection algorithms will influence the measured detection accuracy. To minimize the impact, dataset creators publish protocols that comprise policies for specific datasets and algorithmic approaches to "normalize" detector BBs for the dataset. For the CS3 dataset, Noblis defined limitations on "acceptable" BB sizes and an algorithm to normalize the BBs.

A possible normalization protocol assumes association has already occurred and normalization is only applied to the N associated sets of BBs that satisfy an Intersection Over Union (IOU) threshold. The normalization correction is then based on the difference between ground truth and detector coordinates for each corner of the BB.

Noblis [13] subsequently altered the process to include an association step after normalization. For each image, the altered process calculates and applies normalization between each detector and ground truth BB combination in an image. Next, the IOU of the normalized pair as well as the percent difference in area between the original detector BB and the normalized BB is calculated. Association is then selecting the lowest value from the list of percent differences sorted in an ascending order. The elements of the selected pair are removed from the list of detector and GT BBs respectively and the process repeated. CMU provided the Center with their python evaluation script for the altered approach, which was the basis for the RT&E Center BB normalization calculations.

Other authors [28] also discussed the impact of mismatches between face image datasets and what the face detection algorithms were "tuned for." Their approach to normalization is a remediation aimed at a global scaling and translation that maximizes overlap. The impact on detector performance in some experiments is demonstrated in Figure I-D–1. The graph are results for multiple algorithms with and without remediation (right and left plots respectively). The general trend is to improve the PR curve – the same trend observed for PR curves in Section I-D.3 of this appendix.

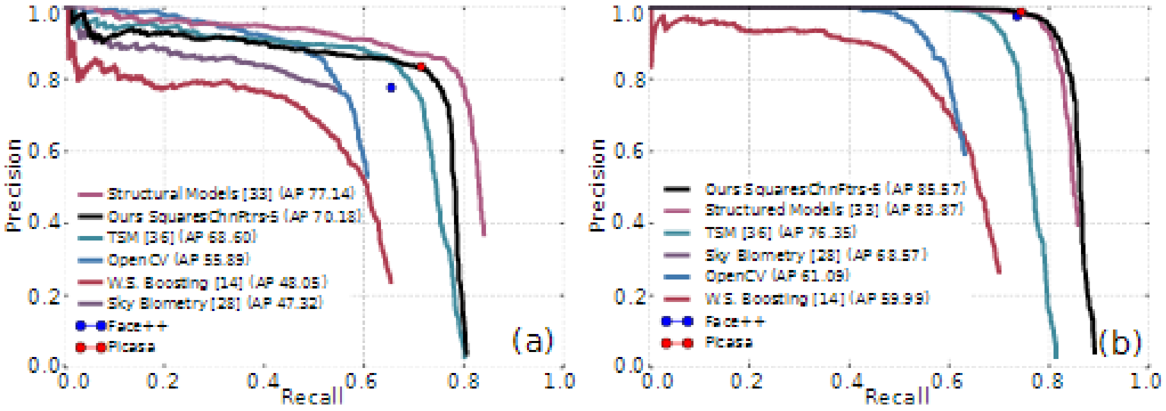


Figure I-D–1: Demonstrating Impact of BB Normalization [28]

I-D.3 BB Normalization Impact on CS3 Precision-Recall Values

The Precision-Recall curve for the four detection algorithms processing the CS3 dataset is provided in Figure I-4–9 and repeated here (Figure I-D–2) for ease of comparison. The PR-AUC values are provided in Table I-D–1. The corresponding results calculated using normalized BBs are provided in Figure I-D–3 for direct comparison with Figure I-D–2.

The order of preference, as indicated by the plots and table entries, after normalizing is more consistent with the analogous results provided in Phase I report for the other datasets. From a PR-AUC perspective in particular, Ultron and TinyFace are both very similar and the top two performers, YOLO is the third lowest and PittPatt is now the worst performer – it was the best for CS3 in the report. This result is an example of the impact of BB normalization on detector performance.

Note that in Figure I-4–5, and Figures I-D–2 and I-D–3 the PR curve for the TinyFace detector appears to extend horizontally to a Recall equal to "0." This extrapolation is done for all detectors on all datasets for consistency (see Section I-4.2.1), but is most apparent in this curve because of the nature in which TinyFace was trained versus the dataset it was presented with.

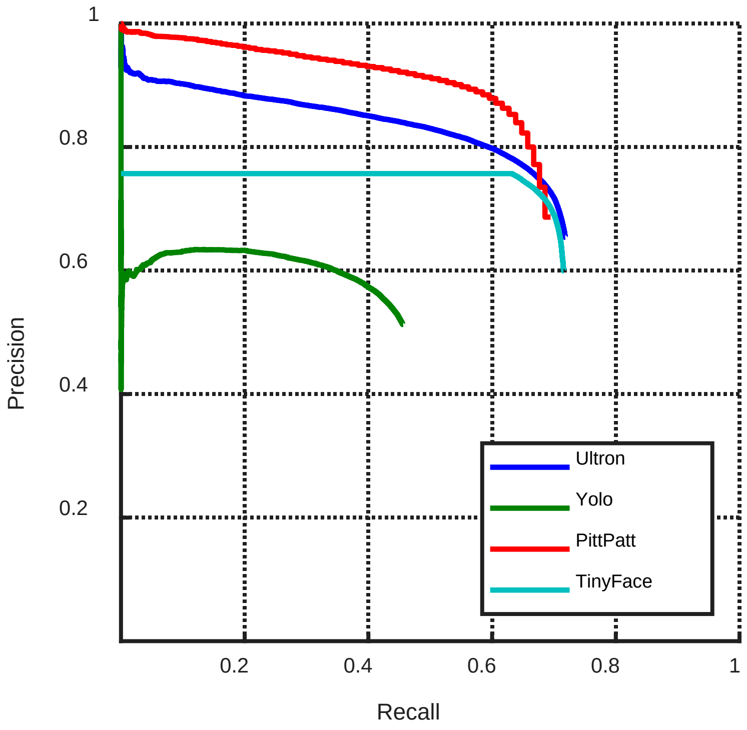


Figure I-D–2: Precision-Recall Plots for Four Detectors with IOU-based Detection on CS3 Dataset (no BB Normalization)

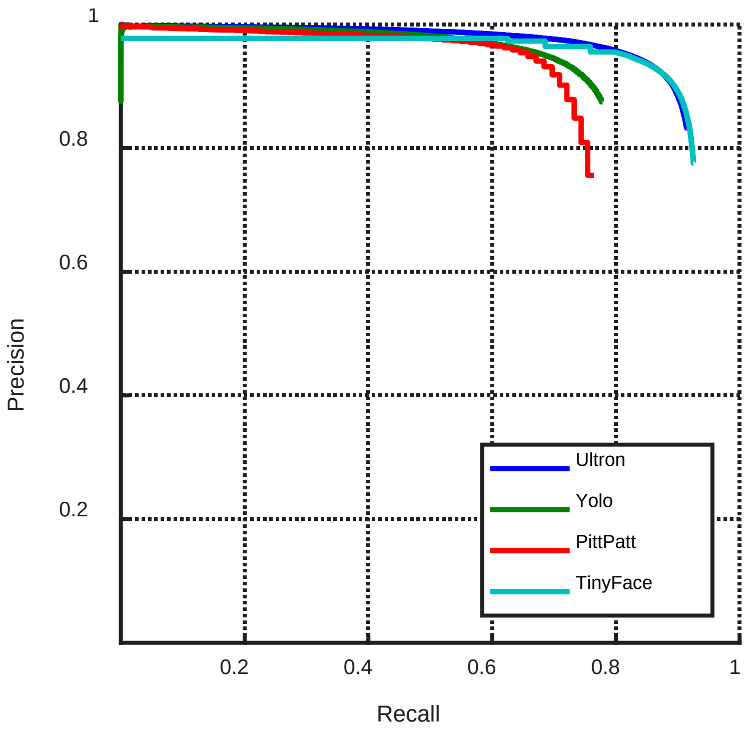


Figure I-D–3: Precision-Recall Plots for Four Detectors with IOU-based Detection on CS3 Dataset (with BB Normalization)

Table I-D–1: Area of Precision-Recall Curves (AUC-PR)

Detector\ Dataset	**CS3 Non-Normalized BB**	**CS3 Normalized BB**
PittPatt	0.6434	0.7428
TinyFace	0.5384	0.8953
Ultron	0.6090	0.8969
YOLO	0.2777	0.7619

I-D.4 Conclusions

As noted in ([13] [28]), the academic community recognizes the issues associated with assessments that compare face recognition algorithms using different datasets. These include items such as size and composition of training sets, the range of face sizes to be detected, environmental factors such as pose and illumination and especially the BB annotation. Some authors ([13] [28]) have introduced global optimization to minimize the impact of non-standard BB annotation.

The problem is that datasets and algorithms have their own sets of "policies" on such parameters as BB size and minimum BB size to try (in pixels), and these policy choices affect the scoring metrics. Noblis suggested it was necessary for the Janus Program to apply this procedure or an alternative to be able to ensure "fair" comparisons.

The creators of the CS3 dataset have included global optimization as part of the CS3 dataset protocol. The RT&E Center performed the Phase I assessment without normalization and repeated the CS3 component of the assessment with normalization at the sponsor's request. The impact of the normalization is evident in the Figure I-D–2 and Figure I-D–3 comparison.

After implementing the suggested normalization, the RT&E Center concluded that it is not appropriate for detection and ground truth association. We note that normalization is but one component that is not accounted for in these assessments. Our primary objection is that an equivalent detection resizing could NOT be done in a real world application since the face images are unlikely to be characterized with respect to what the detection algorithm was tuned to: for example, face size, whole head vs. facial features, pose. Similarly, the face detection algorithm policies may not be known or not amenable to change. For example, TinyFace performance would be impacted by restricting the minimum size BB.

The RT&E Center learned more about the CMU Detection algorithm by using the IOU metric without resizing. It would be better to have more robust capability to get the bounding boxes to match the dataset's bounding box policy (e.g., what portions of the head are within the dataset's bounding boxes). Some other algorithms did just fine with the oversized bounding boxes in CS3. The resizing would promote the CMU algorithm's scores (i.e., Ultron) while leaving others unchanged. The algorithms that don't have a problem would end up with reduced rank relative to the optimized version of the Ultron output. Choice of training sets is an important factor.

Lastly, the approach blurs the lines between the characteristics that separate our detectors currently. It normalizes the correctness of bounding boxes that get part of the face, all of the face, or the face with additional area surrounding the face. Thus, we argue that the raw bounding boxes are more appropriate than the normalized bounding boxes for the Phase I assessment.

JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

11100 Johns Hopkins Road • Laurel, Maryland 20723-6099

AOS-18-1040

NIJ RT&E Center Project 15-FR

October 2018

NIJ FACE PROCESSING ALGORITHM ASSESSMENT PHASE II – FACE RECOGNITION

Version 1.0

Authors: Richard L. (DJ) Waddell, et al.

Prepared for:

NIJ | National Institute of Justice

STRENGTHEN SCIENCE. ADVANCE JUSTICE.

Prepared by:

The National Criminal Justice Technology Research, Test, and Evaluation Center
The Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Rd.
Laurel, MD 20723-6099

Task No.:	FGSGJ
Contract No.:	2013-MU-CX-K111/115912

This project was supported by Award No. 2013-MU-CX-K111, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication/program/exhibition are those of the author(s) and do not necessarily reflect those of the Department of Justice.

PHASE II-CONTENTS

Page

PHASE II-FIGURES

Page

PHASE II-TABLES

Page

PHASE II-EXECUTIVE SUMMARY

The National Criminal Justice Technology Research, Test & Evaluation Center (RT&E Center), operated by the Johns Hopkins University Applied Physics Laboratory, conducted an assessment of the face recognition capability of software developed by Carnegie Mellon University. Analogous to the Phase I assessment, the Phase II assessment compared the Carnegie Mellon University software performance with that of two other face recognition software systems (i.e., PittPatt, OpenFace) using a suite of metrics calculated for three different datasets. Two of these face recognizers (CMU-algorithm, PittPatt) originated in the Carnegie Mellon University community and reflect the evolution of face recognition software; the third (OpenFace) is an implementation of the Convolutional Neural Networks-based technology which is also used in the CMU algorithm. The assessment included two applications of the CMU algorithm – with built-in preprocessing capabilities (denoted Native CMU), and with common preprocessing prior to the Native CMU implementation (denoted Dlib CMU). Face images used in this assessment were extracted from the Labeled Faces in the Wild (LFW); the Good, the Bad, and the Ugly (GBU); and the Challenge Set 3 [i.e., IARPA (Intelligence Advanced Research Projects Agency) Janus Benchmark-B] datasets.

Standard metrics used to compare face recognition algorithm performance include the receiver operating characteristic (ROC) curve and the cumulative matching characteristic (CMC) curve. The metrics used in the assessment include the area under the ROC curve, and the rank-percent tabular results of the CMC curve. These metrics were used to investigate the performance of the recognition algorithms when processing the datasets.

These results show that based on ROC and CMC curves, the Native CMU algorithm outperforms the others, the OpenFace algorithm marginally outperforms Dlib CMU using ROC curves and Dlib CMU marginally outperforms OpenFace using the CMC Curves. Both of the neural network based algorithms significantly outperform the PittPatt algorithm.

II-1 INTRODUCTION

In September of 2013, the Johns Hopkins University Applied Physics Laboratory (JHU/APL) was selected by the U.S. Department of Justice, National Institute of Justice (NIJ) to establish the National Criminal Justice Research, Test, and Evaluation Center (RT&E Center) within the National Law Enforcement and Corrections Technology Center System. The RT&E Center has been tasked to perform an assessment of the Carnegie Mellon University facial processing algorithms using the technical approach documented in the RT&E Center's Proposed Technical Approach for project 15-FR [1].

This effort includes three phases. The first phase assessed the Carnegie Mellon University software capability for face detection and the results are documented in the Phase I section of this report. Similarly, the periocular recognition capabilities of CMU software are discussed in the Phase III section. This section documents the CMU software capabilities for face recognition; in particular, the ability to identify faces in a side-by-side comparison with two other, well-known face recognition capabilities (PittPatt, OpenFace) using three available datasets

- Challenge Set 3 (CS3)
- Labeled Faces in the Wild (LFW)
- The Good, the Bad and the Ugly (GBU)

II-2 BACKGROUND

II-2.1 Face Recognition

Face recognition has two aspects: identification, verification. The identification process compares a query image of unknown identity with a set of target images (a gallery) with known identities. The comparison consists of calculating a pair-wise distance measure between the query image and each gallery image and ranking the comparisons as possible identifications. The verification process, on the other hand, compares a new image with a known identity with an image from a gallery with the corresponding identity, using a pair-wise distance metric calculation. A (usually empirically determined) threshold is used to verify a match [15]. The RT&E Center recognition assessment focuses on the identification component.

In both cases, the process can be defined as a mapping from a pair of face images to a real number representing some measure of the similarity (equivalently, distance) between the images; one image is drawn from a set of query (or unknown) images and the second from a set of target (or known) images. The identification process typically computes a matrix of similarity scores where the matrix elements are the distances (i.e., scores) between query and target images, the verification process just uses the scores for the one target image [15].

II-2.2 Approach to Assessment

The processing framework, shown conceptually in Figure II-2–1 and described in this section, is designed with two objectives:

- Ensure that all recognition algorithms are processed the same way to provide consistent inputs and thus a more direct comparison between the algorithms; and
- Ensure that the CMU algorithm, referred to as the Native CMU algorithm in the rest of this document, is processed in its unadulterated form to get a baseline measure of its performance.

Neither objective was completely obtained since the built-in preprocessing of one algorithm was applied to the other two (for a consistent input) even though it was only required for one of the two. Similarly, for the other objective, a direct comparison can only be made between the two algorithms with built-in preprocessing. The combinations of algorithms and preprocessing are summarized in Figure II-3–1.

The framework was applied to the face recognition algorithms discussed in Section I-3 to create the metrics discussed in Sections II-4.1.2 and II-4.1.3. The essential elements of the process flow are summarized in Figure II-2–1.

Pairs of images are central to the analysis of face recognition capabilities and the pairs information facilitates construction of receiver operating characteristic (ROC) and cumulative matching characteristic (CMC) curves discussed in the metrics section. The facial recognition software requires a pairs data structure for processing a dataset for ROC curves. By design, the RT&E Center created test sets for calculating CMC Curves and a pairs dataset for calculating ROC curves.

ROC curves were calculated from datasets with an equal number of same pairs and different pairs to ensure that the dataset is balanced. For example, our LFW dataset implementation takes a text file containing a list of pairs of images and initializes a dataset structure. Each processing round consists of pairs of images that are of the same person, followed by pairs of images that are different. An ROC curve could then be created from pair results.

CMC Curves were calculated from datasets split into a gallery set and a probe or query set. The CMC gallery set was comprised of a randomly selected image for each identity in a given dataset. The CMC probe or query set was created with the remainder of the images for each identity limited to the mean value of images per identity.

In the process of creating these test sets for subsequent ROC and CMC curve calculations, the RT&E Center removed all images from each dataset that could not be processed by any of the preprocessing or feature extraction stages shown in Figure II-2–1. We refer to the resulting subsets of the original datasets as pruned datasets.

Similar to the Phase I effort (face detection), the software architecture uses classes to wrap both the algorithms and data sources. Since face recognition is inherently a one-to-many comparison application, the software includes internal data structures to store the representations of the two sets of images being compared. Since the goal of the assessment is to determine how well the recognition algorithms match images with the same identity, individual images can be on one or both sides (i.e., query, gallery) of the comparison, though individual images are never compared to themselves. The overall process flow for face recognition is provided in Figure II-2–1.

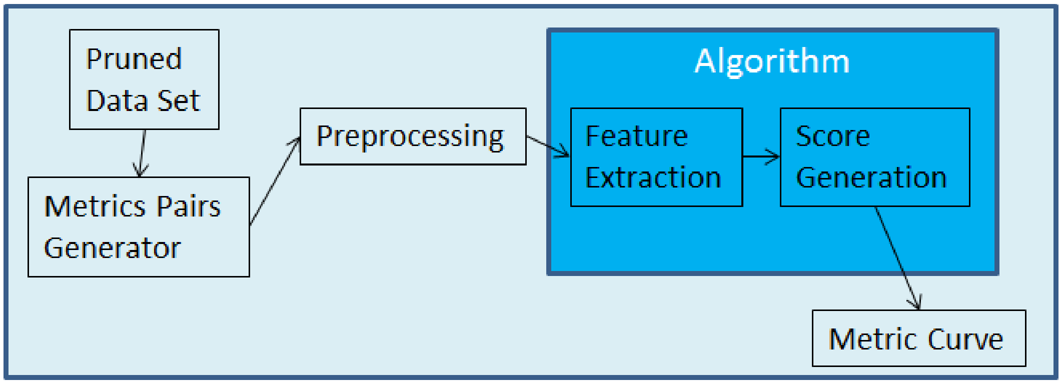


Figure II-2–1: Phase II Process Flow

Depending on the algorithm, the results from the face recognition algorithms can include similarity matrices for a set of query and gallery images or individual face feature vectors. These are used to generate ROC curves and CMC curves, the primary metrics for the assessment. Metrics are described in the Assessment Components, Section II-4.3.

1. Start with a dataset of images.
2. Run any dataset specific required preprocessing/pruning to remove invalid images.
3. Run feature extraction using a specific face-recognition algorithm to extract features from each image in pruned dataset.
4. Create datasets of image pairs for ROC curve and CMC curve analysis for a given pre-processor-algorithm-dataset combination.
5. Compare all images for both ROC test set and CMC test set using a preprocessor-algorithm.
6. Create ROC curve for pre-processor-algorithm-dataset combination from ROC test set results.
7. Create CMC curve for pre-processor-algorithm-dataset combination from CMC test set results.

Figure II-2–2: Conceptual Flow (sequence) of Image Recognition Processing

II-3 FACE RECOGNITION ALGORITHMS

Face recognition is usually a multi-step process: given an input image, face detection is performed to find the face; then preprocessing is performed to provide the correct face orientation; next, a low-dimensional representation is calculated. A similarity score can be calculated by finding the distance between the low-dimensional representation of the current face with that of the gallery faces. Face recognition is then based on the value of the similarity score (i.e., the distance value). Performing the assessment requires datasets with known (ground truth) matches and mismatches. In some cases, these are provided by the protocols constructed from the dataset and in some cases, the RT&E Center created protocol-like subsets from the single dataset by leveraging the availability of multiple images of the same identity.

Face recognition has two main methodologies: feature based (i.e., a set of values for hand-crafted features are individually calculated for the face), and deep learning based (i.e., the pixels comprising the image are processed as a unit to generate a feature set). This assessment includes both – PittPatt for feature-based and both OpenFace and CMU for deep learning processing. The latter two are based on convolutional neural network (CNN) technology. “Today’s top-performing face recognition techniques are based on convolutional neural networks. Facebook’s DeepFace […] and Google’s FaceNet […] systems yield the highest accuracy.” [18]

The CNN “predicts some probability distribution $\hat{p}$ and the loss function L measures how well $\hat{p}$ predicts the person’s actual identity i. DeepFace’s innovation comes from three distinct factors: the 3D alignment, a neural network structure with 120 million parameters, and training with 4.4 million labeled faces. Once the neural network is trained on this large set of faces, the final classification layer is removed and the output of the preceding fully connected layer is used as a low-dimensional face representation. [18]

The CMU face recognition software, which is CNN-based, was assessed by comparing performance with two other recognized algorithms: PittPatt, OpenFace. The PittPatt algorithm is discussed in the Phase I report and its application to face recognition is discussed below. OpenFace is an open source application library that is often referenced in the literature. ([18], [6], [14])

The assessment comprises three algorithms processing three different datasets. Because of the differences in dataset content and algorithm required data formats, the feature calculation (feature extractor) component of the algorithms had different input configurations. A preprocessing step is used to convert the image data to the format required for the feature extractor. Both the OpenFace and CMU algorithms include built-in preprocessors; however, PittPatt input requires preprocessing. To insure standardized input data, Dlib (a dependency of the native OpenFace software package) was used to orient the faces and standardize the image input size for CMU-algorithm (denoted Dlib CMU), OpenFace (Dlib is the built-in preprocessor), and PittPatt processing as outlined in Figure II-3–1. For the cases of Dlib CMU and Pittpatt, the use of Dlib preprocessing occurred prior to any native image preprocessing that each algorithm contained internally. In addition, the CMU algorithm was run without this preprocessing step (denoted Native CMU). This approach exercised the CMU-algorithm with two different preprocessing configurations for each dataset. The CMU-algorithm results preprocessed using Dlib provided a more standard input configuration for the comparison between the three algorithms while the Native CMU results provide a better indication of the absolute performance of the CMU-algorithm without the standardizing preprocessing step.

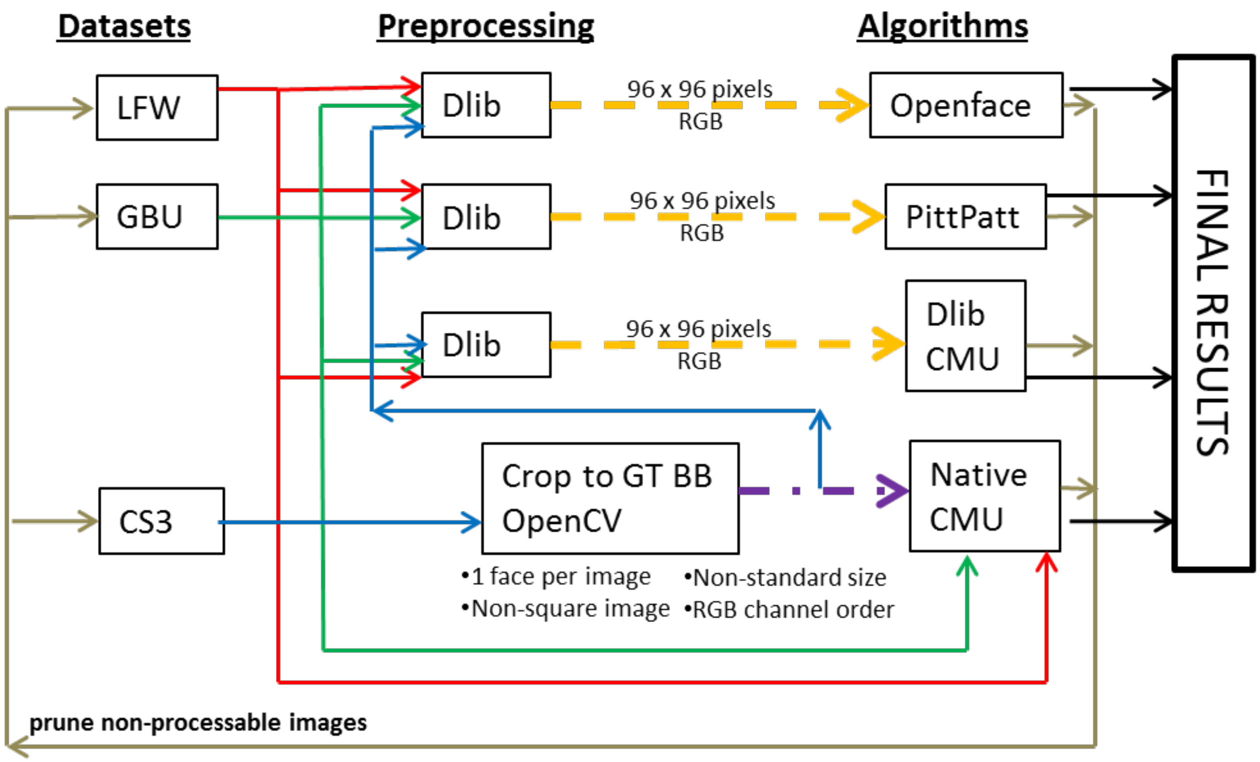


Figure II-3–1: Dataset Preprocessing Summary

II-3.1 CMU-Algorithm

The RT&E Center received the CMU Face Recognition software as source code with a user guide. From the Caffe prototype files within the source code, the RT&E Center recognized that the algorithm utilizes a convolutional neural network based deep learning technology to generate a feature set to describe a face and a custom metric to calculate the distance between two faces based on the individual feature sets of the faces.

The dataset preparation includes establishing a spatial correspondence between images. “This is commonly done by first locating the eyes in both images. Then the image may be adjusted by positioning, scaling, and rotating the face so that the eyes always fall at exactly the same position. The result of localization is typically a new smaller image, called a face image chip, created by re-sampling pixels in the original, such that eyes, mouth, and so forth appear at approximately the same pixel coordinates in every face chip.” [15] Other (possible) image preprocessing steps native to CMU may influence some of the dominant control parameters in face recognition, e.g., illumination normalization, zeroing the pixels at the face edges, and removing background clutter.

II-3.2 OpenFace

OpenFace is a Python and Torch (an open source machine learning library) implementation of face recognition constructed with deep neural networks; it is based on DeepFace and FaceNet ([17], [18]). The DeepFace contribution to face recognition capabilities included the alignment procedure to present the correct face orientation, the application of the CNN architecture and the very large dataset of labeled faces. FaceNet's contribution has four distinct factors: the triplet loss used to select training examples and facilitate training, their triplet selection procedure, extremely large training set, and extensive network architecture tuning via experimentation. Both systems have the high-level architecture shown in Figure II-3–2. The specific OpenFace implementation of this architecture is provided in Figure II-3–3.

OpenFace implements the same overall architecture as DeepFace and FaceNet; however, it has additional design goals of high accuracy with low training and prediction times. OpenFace is built on multiple existing components:

- Torch for network training, python libraries, OpenCV and scikit-learn for classification processing;
- Dlib pre-trained face detector for higher accuracy [18];
- Dlib affine transformation to provide correct face orientation for CNN;
- Novel representation - low-dimensional face representations for the faces in an image;
- Novel neural network training approach; and
- Euclidean distance for similarity.

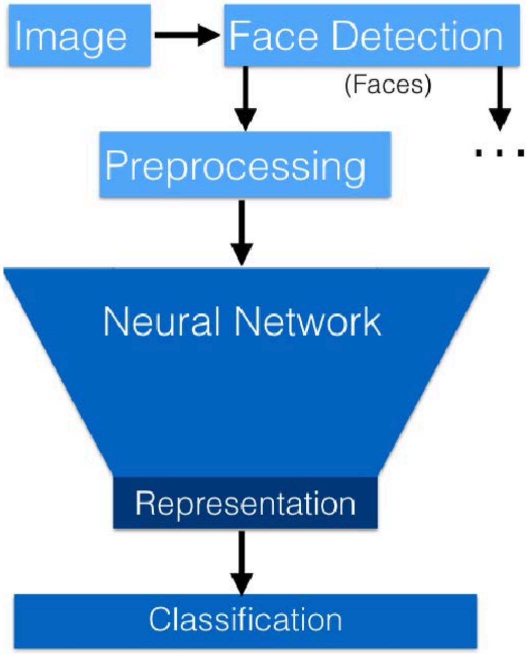


Figure II-3–2: Logic Flow for Face Recognition with a Neural Network [14]

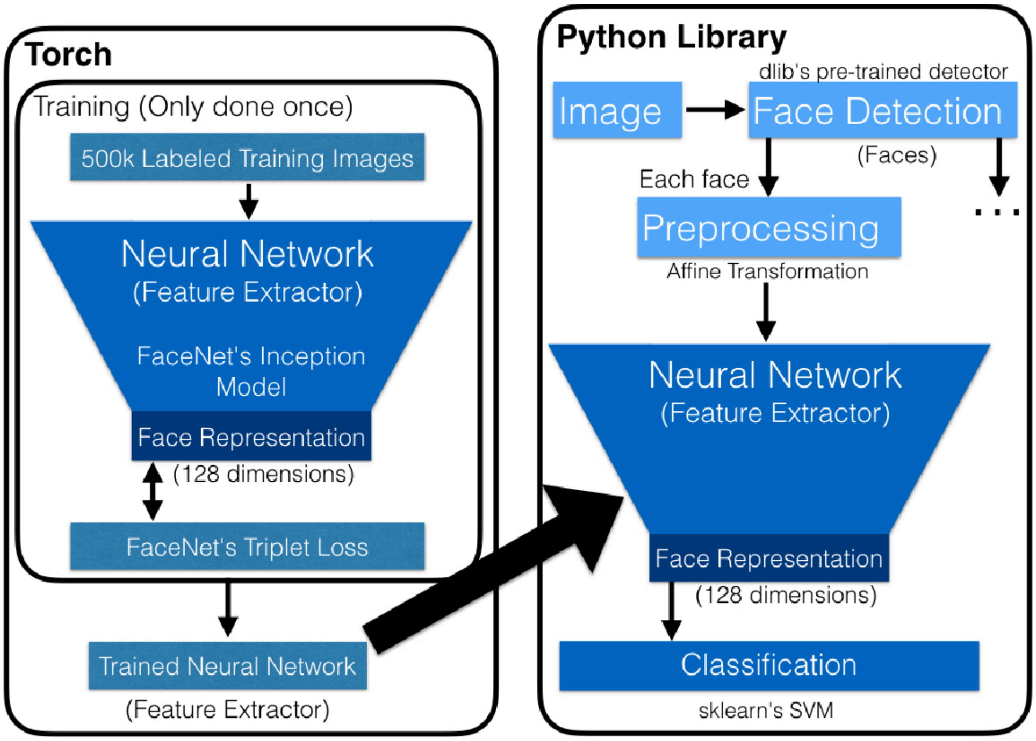


Figure II-3–3: OpenFace Architecture [14]

II-3.3 PittPatt

The RT&E Center acquired PittPatt V 5.2.2 from contacts at Carnegie Mellon University. This software does not have documentation but a limited example application was provided that identified the roles of two components: create galleries and compare galleries. The example summary includes:

- The software uses two sets of photographs:
  - One set (probes) contains the query images, i.e., the images to be recognized; and
  - The second set (gallery) contains the collection of already characterized or known images.
- Preprocessing the images; according to the format configurations identified in Table II-4–1.
- Reading the images from the gallery, extracting the feature set for each image and storing it in a file. This is the create galleries component.

- Executing the face matching, or compare galleries, component:
  - Draw a single image from the probe or query-images and calculate the associated features;
  - Compare the associated features, via a distance calculation, to each image in the gallery of known images;
  - Save the comparison results (similarity scores) as a row of elements of a matrix in results file; and
  - Repeat the steps for each query image.

II-4 ASSESSMENT COMPONENTS

The assessment task computes performance metrics of three recognition algorithms: CMU's algorithm, OpenFace (publically available), and PittPatt (a standard of an earlier generation) processing three datasets: GBU, LFW, and Challenge Set 3 (CS3). Table II-4–1 (and Tables I-4–1 and I-4–2) provides summary information for these datasets. Note that each dataset had to be pruned of any images that could not be successfully (i.e., generate a feature set) processed by all of the preprocessor-algorithm combinations (See Figures II-2–2 and II-3–1.)

As mentioned earlier, the assessment is based on three metrics: ROC curves, CMC curves and AUC (Area Under the Curve) for the ROC curves. The similarity scores are used to compute these metrics. After all comparisons have been made, the contents of the results file are processed by the ROC and CMC classes (of the processing framework) to generate the corresponding graphs. For the ROC curves, the maximum and minimum score values in the results file are used to bound the threshold range.

Dataset creation for face recognition processing and dataset characterization are discussed in Section II-4.1. Section II-4.2 discusses the individual datasets; Section II-4.3 discusses the metrics used to characterize the software performance, and Section II-4.4 discusses the metric values and their interpretation.

II-4.1 Test Set Construction

II-4.1.1 Test Set Preparation Summary

The summary metrics for face recognition are standards, i.e., ROC curves including AUC, CMC curves; however, the methodology to generate these values is impacted by the amount of data and the nature of the task. In particular, sets of images are compared to determine the identity of an unknown image by matching with an image of a known identity. Ultimately, the goal is to characterize the performance of the recognition software.

In the usual process, the recognition software computes a feature set of an unknown image that is compared with feature sets of identified images to determine possible matches; these results are an output of the process. For this assessment, all images have known identities and the identities are used in creating two sets of image pairs (matched images, non-matched images) that are compared. For ROC curves, the sets of image pairs are balanced by construction; for CMC datasets there is no need to balance the datasets since the known dataset just needs to contain a single image for each identity. The value of balanced versus imbalanced requirement for ROC curves, is that with the balanced case, you equally weight your true positive rate and false positive rate by giving equal opportunities to generate good true negatives and true positives.

II-4.1.2 ROC Curve

In this application, the face recognition algorithm results are comprised of the true positive rate (TPR) and the false positive rate (FPR). The two values are computed at multiple threshold values and the resulting set of data points comprise the ROC curve. The AUC value provides a single number summarizing the ROC curve. An AUC value of 1.00 is ideal since it represents the case of a TPR = 1.00 (100 percent) for all FPR values.

To create an ROC curve, one set contains matched pairs and the second contains unmatched pairs; each set contains multiple, but different, images of the same identity. After the feature sets have been created, the set of matched pair identities is processed to generate a matrix of similarity scores that are subsequently used to calculate (for a chosen threshold) the TPR; the matrix of scores from the set of mismatched identities is used to calculate the FPR for the corresponding threshold.

The features used to represent the images and how they are calculated varies with the algorithm software as discussed in Section I-3. The Euclidean distance measure was used to compare features in the OpenFace algorithm; the distance calculation method for the PittPatt detector is unknown but monotonic; and the score calculation method for CMU uses custom distance metric. The score interpretation for individual algorithms is dependent on the distance calculation implemented by the algorithm software. For example:

- OpenFace: score = 0 → images are very close; score = 4 → images far apart
- PittPatt: score = -3 → images are far apart; score = 17 → images close together

These values are used to compute the TPR and FPR values for each threshold that form the ROC curve, where:

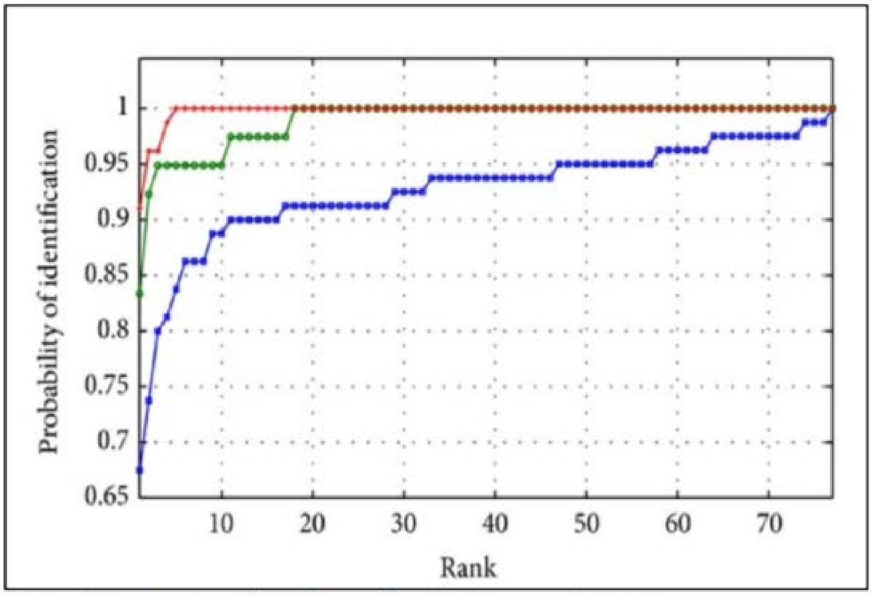
- TPR = TP / (TP + FN) interpreted as the percentage of positive matches correctly identified by the recognition algorithm; and
- FPR = FP / (FP + TN) interpreted as the percentage of positive matches in-correctly identified by the recognition algorithm.

II-4.1.3 Cumulative Matching Characteristic Curve

A CMC curve is a rank-based metric used to assess the accuracy of algorithms that produce an ordered list of possible matches. The CMC curve applies to only closed-set face recognition scenarios where the main purpose is to compare a query subject to a gallery of known subjects. A distance metric is calculated between the query image and all the gallery subjects. The distances or scores are then ranked from best to worst. The CMC score at rank N is the percentage of time that the correct match is in the top N matches. Thus, the CMC curve is obtained by plotting the corresponding percentage for every value of N. For example, in facial recognition the output of the algorithm would be a list of faces from the gallery-set, ordered (ranked) from most to least likely to be the unknown identity. Each set of scores is ordered (i.e., ranked) and the rank at which a true match occurs is noted. For each rank, the accuracy is computed as the percentage of times the true match is within that rank.

Imagine a CMC curve where the rank 10 accuracy is 50 percent. This means that the correct match will occur somewhere in the top 10, 50 percent of the time. In general, the better the algorithm, the higher the rank N CMC-percentage (assuming the algorithms have been tested on the same dataset).

Examples of CMC curves are shown in Figure II-4–1, where the top curve (in red) is considered better because on average the correct match is in the top five matches obtained from the gallery. Most of the CMC curves in Section II-4.4 appear to be continuous lines; however, this is a reflection of the low resolution of the x-axis where individual rank values are not discernable.



Note: The top curve in red provides the best performance.

Source: RT&E Center's Proposed Technical Approach.

Figure II-4–1: Example CMC Curves

For the CMC curves in this assessment, one subset of images contains a portion (i.e., the query set) of all images and the second subset (the gallery set) contains images from the complement set. Using the fact that each image has a known identity, the recognition algorithm is used to compute a score in a one-to-many comparison between each element of the query set and the entire gallery set. Given the scores and the known identities, the rank of a match is the order of the score (best value is Rank 1) at which the match occurred. Rank values are computed for each one-to-many matches and for each rank compute the percent of correct matches for that rank. The CMC data points (rank, percent of correct matches at or below rank) are calculated and the CMC curve plotted.

II-4.2 Datasets Discussion

Many of the datasets created to support face detection and/or recognition use images collected in controlled conditions. In this application, 'face recognition" is interpreted as deciding if two images containing faces are images of the same person (same identity). For example, given a pair of pre-existing face images – that is, images whose composition we had no control over – can the algorithm determine if they are images of the same person?

The contents of the three datasets selected for the assessment were pruned such that all remaining images could be used with each face recognition algorithm. Table II-4–1 provides summary information.

Table II-4–1: Dataset Summary for Face Recognition Task

Characteristic	LFW	GBU	CS3

Number of original images	13,233	Good = 2,170
		Bad = 2,170
		Ugly = 2,170
		All = 6,510
			68,714

Number of images after pruning	5,665	Good = 1,911
		Bad = 1,787
		Ugly = 1,484
		All = 5,182
			29,568

Number of faces after pruning	Same as number of images after pruning	Same as number of images after pruning	Same as number of images after pruning

Original number of unique identities	5,749	Good = 437 identities
		Bad = 437 identities
		Ugly = 437 identities
		All = 437 identities
			1,870 identities

Number of unique identities after pruning	3,293	Good = 428
		Bad = 425
		Ugly = 424
		All = 437
			1,780

Pruned number of image pairs for ROC analysis	2,999 match pairs/ 2,999 non-match pairs	Good = 2,649 match pairs/ 2,649 non match pairs
		Bad = 2,341 match pairs/ 2,341 non match pairs
		Ugly = 1,503 match pairs/ 1,503 non match pairs
		All = 6,493 match pairs/ 6,493 non match pairs
			321,592 match pairs/ 321,592 non-match pairs

Pruned number of images per identity	Varies from 1 to 14 images	Good – Varies from 1 to 8 images Bad – Varies from 1 to 8 images Ugly – Varies from 1 to 8 images All – Varies from 2 to 24 images	Varies from 1 to 17 images

II-4.2.1 Labeled Faces in the Wild

The LFW dataset [19] was created to support face recognition development by providing images that were not controlled for parameters such as position, pose, lighting, expression, background, camera quality, and occlusion. Instead, the dataset contains "a set of labeled face photographs spanning the range of conditions typically encountered by people in their everyday lives. The database exhibits "natural" variability in pose, lighting, focus, resolution, facial expression, age, gender, race, accessories, make-up, occlusions, background, and photographic quality. Despite this variability, the images in the database are presented in a simple and consistent format for maximum ease of use" [19]. All images were collected from the web as a result of the Viola-Jones detector processing and originally captured for other purposes. All have unique names, and have been processed to a standard size to provide some limits on the face presentations. Examples of matched and mismatched pairs are provided in Figures II-4–2 and II-4–3, respectively.

A pairs file was provided with the LFW datasets and the RT&E Center used this with the exception of excluded (i.e., pruned) images. Several images (6,212 of the total) could not be used because they could not be processed successfully by all three algorithms. Typically, the Dlib facial boundary algorithm or PittPatt pre-processing failed - mostly in PittPatt. In summary:

- Number of unique identities is 3293 out of 5,749 identities in original set;
- 2999 matched pairs and 2999 non match pairs were used for ROC curve generation
- Number of knowns for CMC curve calculation is 3293; and
- Number of unknowns for CMC is 2,308.

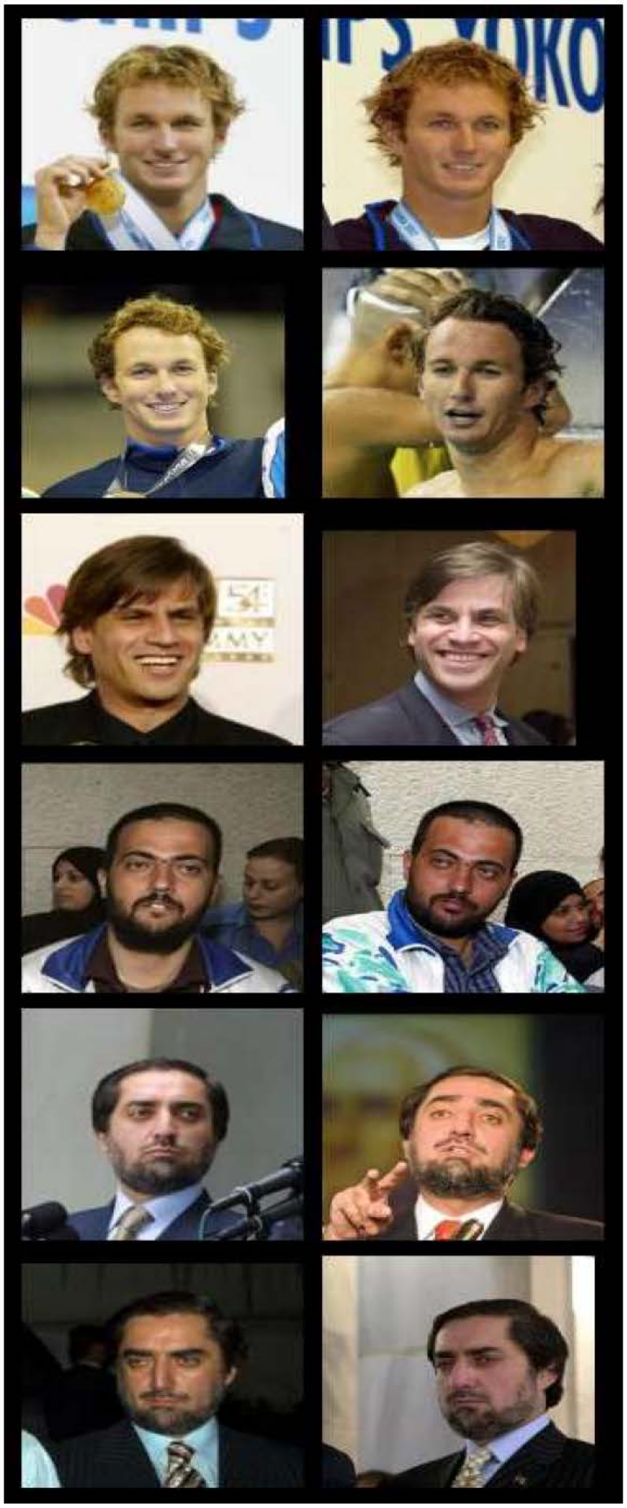


Figure II-4–2: Matched Pairs of Faces – LFW

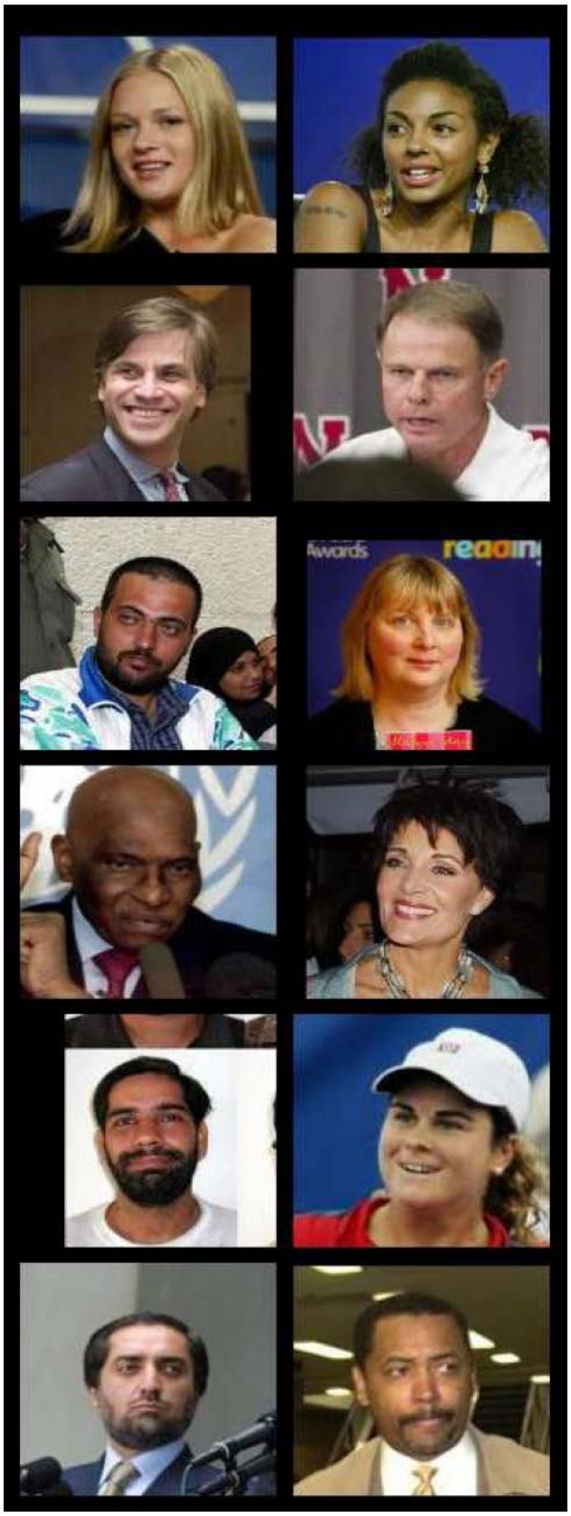


Figure II-4–3: Mis-matched Pairs – LFW

II-4.2.2 The Good, the Bad, and the Ugly (GBU)

The GBU dataset [6] was created to foster the development of face recognition algorithms for various levels of difficulty in still frontal face images. The dataset is composed of three partitions that are designed to range in recognition difficulty from easy to hard. The Good partition contains pairs of images that are considered easy to recognize. On the Good partition, the base true positive rate (TPR) is 0.98 at a false positive rate (FPR) of 0.001. The Bad partition contains pairs of images of average difficulty to recognize. For the Bad partition, the TPR is 0.80 at a FPR of 0.001. The Ugly partition contains pairs of images considered difficult to recognize, with a TPR of 0.15 at a FPR of 0.001. The base performance is from fusing the output of three of the top performers in the FRVT 2006 [6]. This provides test cases to measure performance across a range of difficulties. Across all three partitions, the GBU dataset controls for pose variation, subject aging, ambient lighting, camera and variations in a person's appearance and presentation to the camera. The dataset also limits the number of images per person and requires that pairs of images of a person be taken on different days. Taken together, the set of images captures the range of variability in less constrained images.

The selection criteria for the partitions results in the following properties: an image is only in one partition; and there are the same number of face pair matches in each partition and the same number of non-match pairs between any two subjects. This implies that any difference in performance between the partitions is not a result of different people. The difference in performance is a result of the different conditions under which the images were acquired. Figure I-4-1 is an example of matching face pairs from each of the partitions.

The GBU protocol dictates that "each of the GBU query and gallery sets contain 1,085 images for 437 distinct identities. The distribution of image counts per person in both the query and gallery sets are 117 subjects with 1 image; 122 subjects with 2 images; 68 subjects with 3 images; and 130 subjects with 4 images. In each partition, there are 3,297 face pair matches and 1,173,928 non-matched face pairs" [6]. Instead of adhering to the protocol that GBU specifies, the RT&E Center produced face pair matches by exhaustively computing the maximum number of pairs from the set of pruned images on a partition-by-partition basis. The center then created a random selection of non-match face pairs to match the total number of match face pairs in each partition. See Table II-4–1 for totals. Examples of face pairs are provided in Figure I-4-1 and [6].

The GBU Challenge Problem provided a baseline algorithm that included a discussion of the image preprocessing required before the similarity score calculation. Steps include starting with a frontal image, and extracting a cropped image of "128 by 128 pixels with the centers of the eyes spaced 64 pixels apart" [6].

The recognition algorithms compute a similarity score between all possible pairs of images in the query and gallery sets. The distribution of the match similarity scores for each partition, computed using the baseline algorithm, is compared with the distribution of non-match scores in Figure II-4–4 [6].

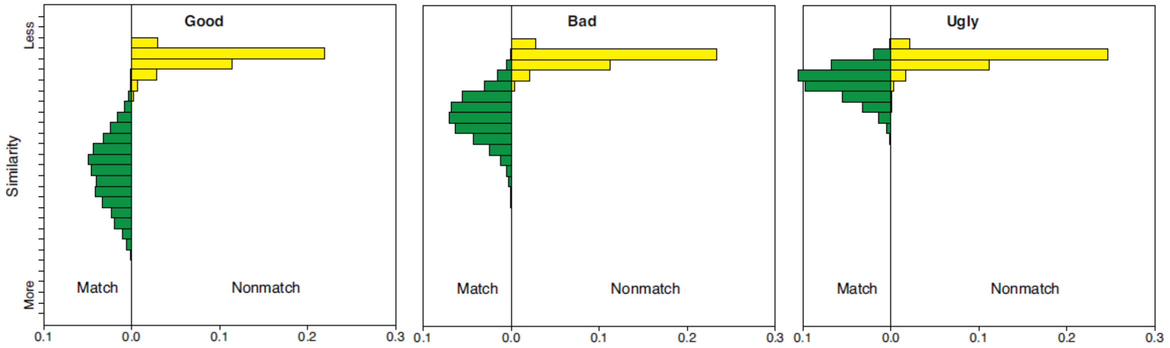


Fig. 5. Histogram of the match and non-match distributions for the Good, the Bad, & the Ugly partitions. The green bars represent the match distribution and the yellow bars represent the non-match distribution. The horizontal axes indicate relative frequency of similarity scores.

Figure II-4–4: Histograms of Similarity Scores for GBU Partitions

In addition to the histograms in the above figure, the reference also provides ROC curves for the three partitions processed with the baseline algorithm. The TPR values for an FPR = 0.001 for the three partitions are provided in Figure II-4–5. In this figure, note that TPR is equivalent to verification rate and FPR is equivalent to false accept rate.

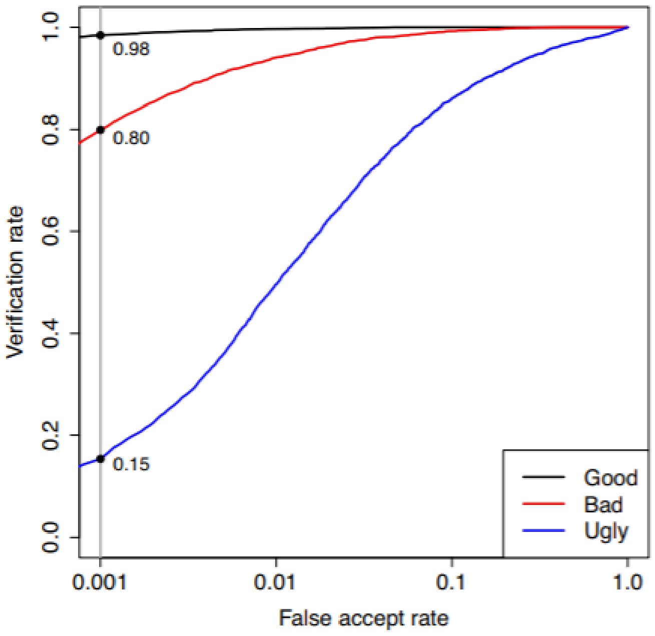


Figure II-4–5: ROC Curve for Three Partitions [6]

II-4.2.3 Challenge Set 3 (CS3)

The CS3 dataset, an extension of IJB-B dataset, is discussed in some detail in the Phase I section of this document; details are provided in Table I-4-1 and Table II-4–1. It was created to support development of unconstrained face recognition, i.e., "... the ability to perform successful face detection, verification and identification regardless of subject conditions (pose, expression, occlusion) or acquisition conditions (e.g., illumination, standoff)" [5].

The total number of faces available after pruning in the CS3 dataset is 29,568. The RT&E Center selected the 1,870 dataset protocol of original identities for the assessment because it was the protocol that had the most identities; images from 1,780 of the identities were actually used (Table II-4–1). The CS3 dataset preprocessing was done two ways (Figure II-3–1). The initial preprocessing used OpenCV to extract a single, non-standard face per image as specified by the protocol. The extracted face is then processed with Native CMU or Dlib CMU.

Because the CS3 dataset did not provide the needed pairs subsets, and because the subset of the CS3 dataset used for CMC curves contained multiple instances of the same identity as per the distribution provided in Figure II-4–6, the RT&E Center created a same-pairs dataset comprised of up to 17 images with same identity; and, for the different-pairs dataset, the process started with the same pairs list and for the left image of each pair a right image was selected from the set of different identity images. As a result, the number of different-pairs for each identity is the same as the number of same-pairs for the identity. The limit was set to 17 because it is the mean of the number of images with same-pair identities (see Figure II-4–6). This approach is hypothesized to limit one identity from skewing classification results.

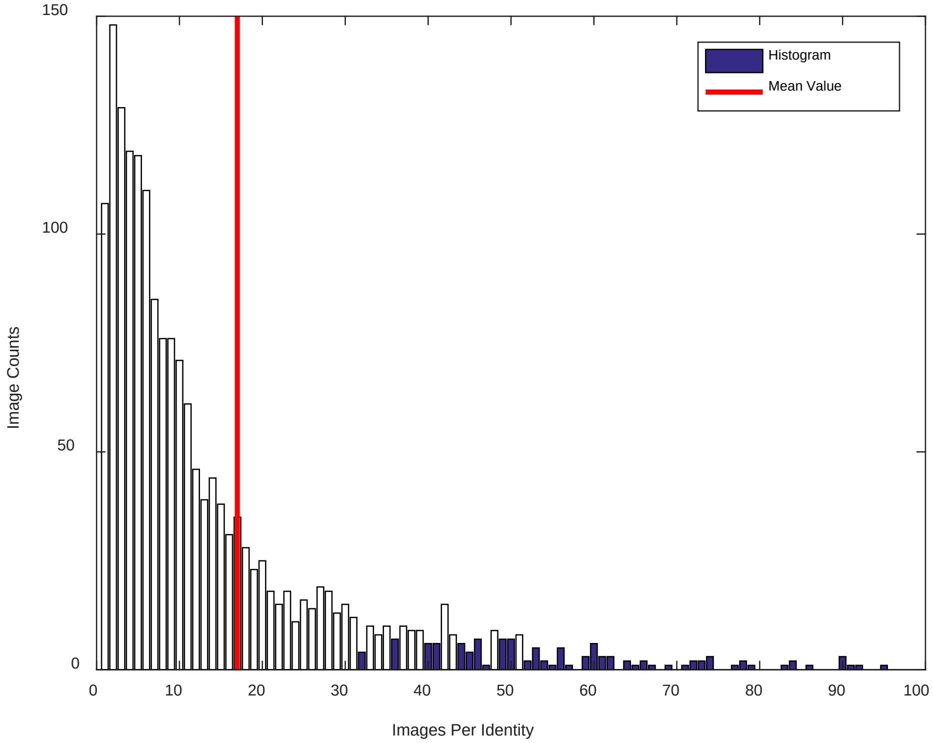


Figure II-4–6: Histogram of Matches per Identity with Outliers Above 100 Omitted

The CMC curves (Section II-4.1.3) are computed from a subset of the data. In particular, the gallery (i.e., knowns) faces were selected by taking a single instance of every identity. The probe or query (unknowns) faces were comprised of *all* of the remaining images.

Several images could not be used—37,901 of them. The Dlib facial boundary algorithm or PittPatt preprocessing failed to identify a face in these images. In forming the sets of images used in the assessment, images were selected in the order in which they appeared in the CS3 dataset. After the sets of images were formed, all possible matches were used to calculate performance. The number of matched and non-matched pairs is provided in Table II-4–1.

II-4.3 Metrics

The approach for the assessment is to compare the three algorithms using a suite of metrics: ROC curves, CMC curves, and summary values such as area under curve (AUC). The PittPatt algorithm is included as a baseline of earlier technology and the OpenFace algorithm because it is representative of recent technology. These two provide a comparison for the CMU algorithm.

For this assessment, the comparison consisted of three metrics:

1. ROC curves computed using algorithm confidence values – Individual algorithms performance using ROC curves will be graphically compared across datasets and results for an individual algorithm will be compared across all datasets.
2. CMC results compared across preprocessor-algorithm combinations using histograms of rank location of “correct” image.
3. Summary values (AUC of ROC) for overall comparison between algorithms.

ROC Curves are discussed in more depth in Section II-4.1.2. In this application, they are used to characterize performance of algorithms to recognize a face and then match the face to another face.

Similarly, the CMC curves are discussed in Section II-4.1.3. In this application, the CMC characterizes how well the preprocessor-algorithm combination results match images in a gallery of images that include known identities.

II-4.4 Results

Overview of Results

The OpenFace and CMU algorithms are based on deep learning technology. The PittPatt results are compared to the others to show relative performance of the component feature-based approach to the deep learning approach. The assessment implemented two sets of preprocessing configurations:

- Dlib preprocessing applied to each algorithm input. This set includes the Pittpatt, Dlib CMU, and OpenFace algorithms
- Only the native preprocessing of a given algorithm. This set includes the OpenFace and Native CMU algorithms.

This provides results for both a common initial preprocessed input (using Dlib) and a preprocessed input tuned to the algorithm (i.e., OpenFace, Native CMU).

ROC curves provide a measure of overall performance of a face preprocessor-algorithm combination since it provides a graphical presentation of the algorithm's ability to correctly recognize faces with the associated false positive rate. The results include direct comparison of ROC curves and associated AUC values for the three Dlib preprocessing-algorithm combinations for all three datasets. The results for the Native CMU ROC curves and corresponding AUC values are also provided for comparison with OpenFace results.

Another perspective on face recognition capabilities is summarized by the CMC curves, which present a more cumulative view of an algorithm's ability to match images. These results are also provided both as multiple CMC curves for each dataset and single CMC curves for preprocessor-algorithm-dataset combinations. Also, direct comparison of CMC curves between preprocessor-algorithm-dataset combinations using a tabularized version of the curves is provided.

For many of the ROC curves the performance is very good; consequently, graphs with expanded scales in the vicinity of (0, 1) are provided for improved resolution. Similarly, for some of the CMC curves, a graph with expanded scales in the region of rank equal zero is provided in Appendix II-B for more resolution.

II-4.4.1 ROC Curve Metric

The ROC curve results for preprocessor-algorithm-dataset combinations are provided in Figures II-4–7 through II-4–9 and Table II-4–2. The individual ROC curves are presented in Appendix II-A for each combination.

Table II-4–2: Preprocessor-Algorithm-Dataset AUC-ROC Summary

Dataset	Dlib CMU	OpenFace	PittPatt	Native CMU
LFW	0.9703	0.9913	0.8992	0.9985
Aggregate GBU *GBU Partitions*	0.9778 *G - 0.9941* *B – 0.9798* *U – 0.9484*	0.9725 *G – 0.9898* *B – 0.9749* *U – 0.9426*	0.9200 *G – 0.9859* *B – 0.9273* *U – 0.8096*	0.9933 *G - 0.9980* *B - 0.9930* *U - 0.9841*
CS3_1870Bal	0.9646	0.9834	0.9364	0.9970
Spread (%)	1.3	1.9	4.1	0.5

All four preprocessor-algorithm combinations have a region of low false positive rate and moderate true positive rate (see Figures II-4–7 through II-4–9). The AUC values associated with the ROC curves are summarized in Table II-4–2. In addition, a spread in AUC values (ratio of highest to lowest AUC minus 1.0) expressed as a percentage is provided. Small spreads indicate similar AUC values. Except for two PittPatt-dataset AUC exceptions, all algorithms have AUC-ROC values greater than or equal to 0.9.

The Native CMU algorithm has the highest AUC results across all three datasets; and OpenFace is superior to Dlib CMU – the third best (an exception is the Dlib CMU – Aggregate GBU combination). Pittpatt is universally the worst performer. Additionally and notably, the AUC values provided in Table II-4–2 are consistent with the designed difficulty of each of the GBU partitions. That is, across all preprocessor-algorithm combinations, AUC values are highest for the GBU-G partition and lowest for the GBU-U partition.

Observations for the Dlib-algorithm-dataset combinations include:

- PittPatt is generally inferior to OpenFace and Dlib CMU in AUC values and spreads.
- The Dlib CMU AUC values are marginally inferior to those of OpenFace by a few percent in two of three cases and marginally superior in one case.
- The spreads are approx. 50 percent larger for OpenFace than Dlib CMU.

Observations for the native preprocessor (i.e., OpenFace-dataset and Native CMU-dataset combinations include:

- The Native CMU algorithm has marginally higher AUC values than OpenFace for all three datasets. The differences are on the order of 1 percent.
- The AUC spread across the three datasets is considerably smaller for Native CMU than OpenFace. The difference in spread is on the order of a factor of 4.
- Similar results to the previous two bullets hold for the GBU partitions, i.e., the Native CMU values are larger (by a few percent) than for OpenFace and the spread is considerably smaller.
- The general pattern of Native CMU performing best (lowest percentage spread) and PittPatt worst is repeated.

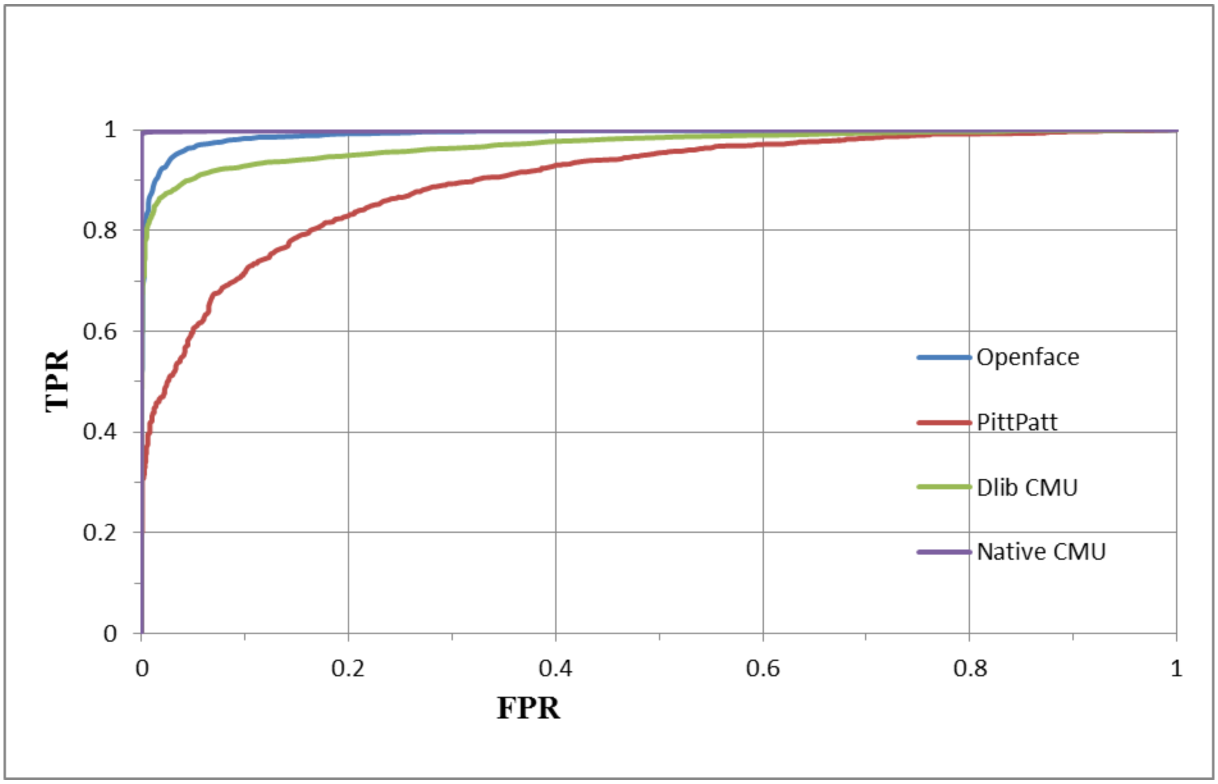


Figure II-4–7: Preprocessor-Algorithm Combinations on LFW Dataset with ROC Curve

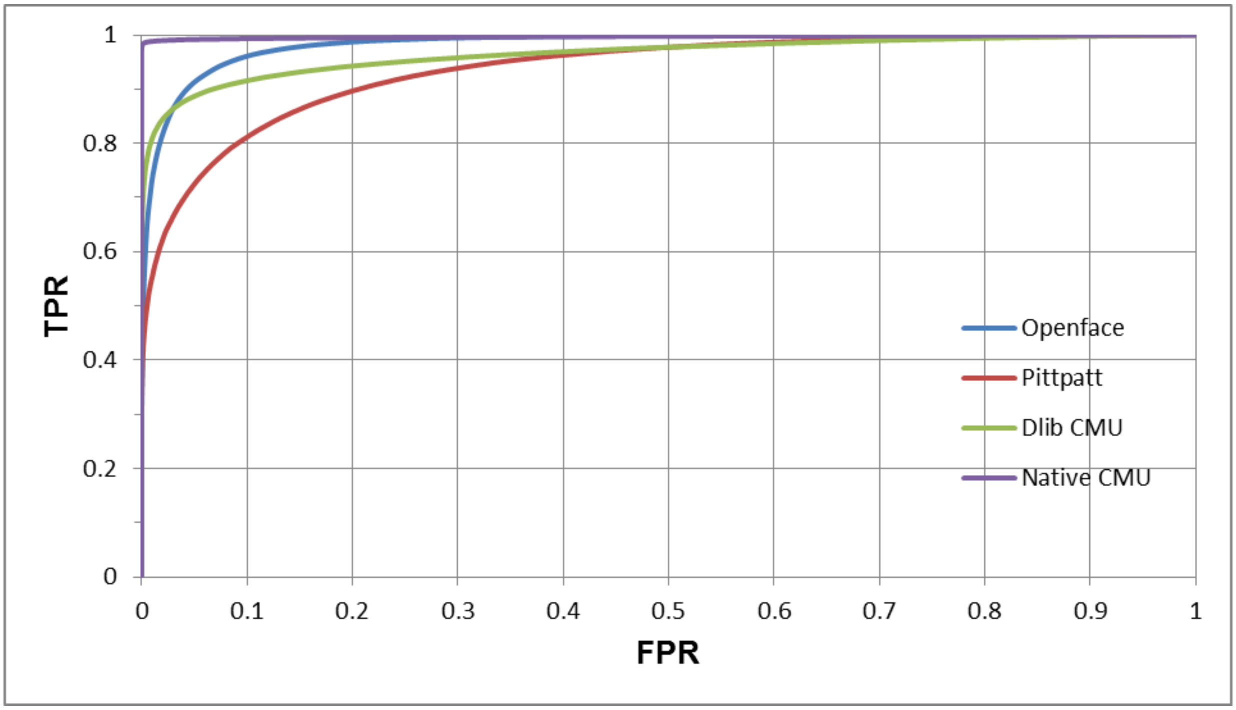


Figure II-4–8: Preprocessor-Algorithm Combinations on CS3 Dataset with ROC Curve

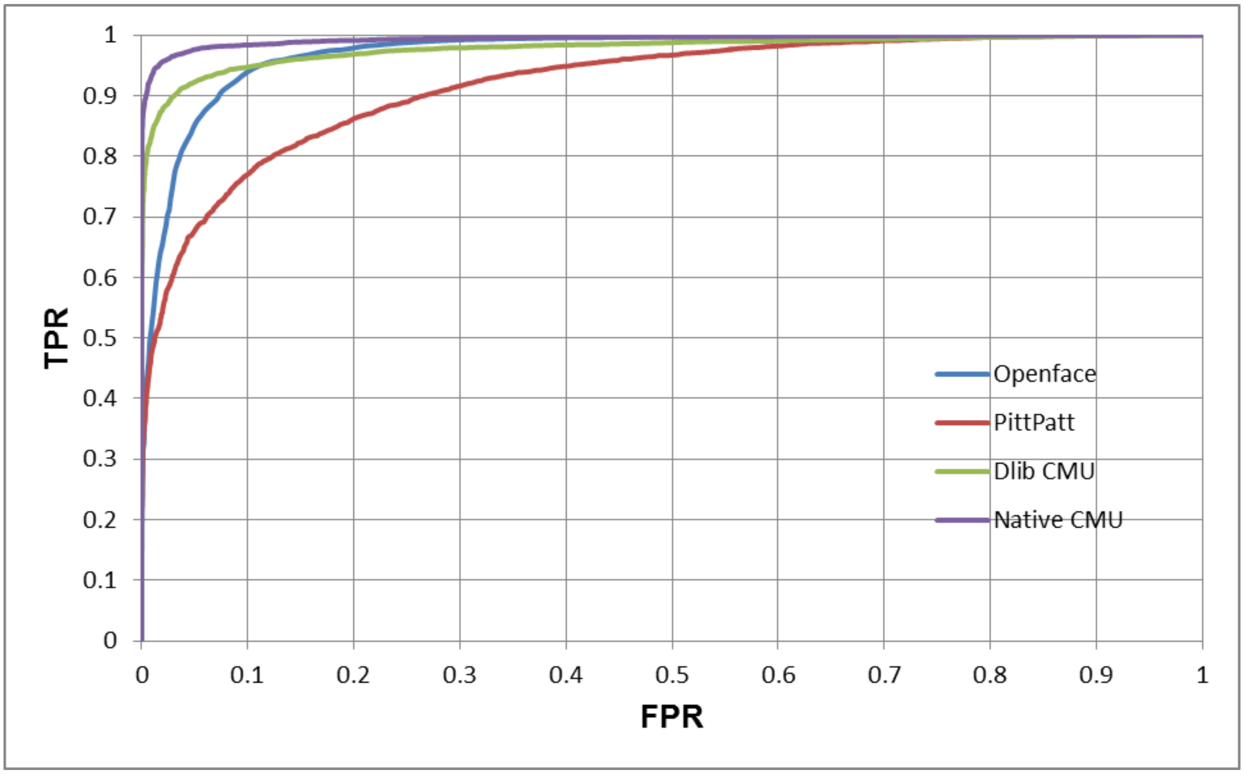


Figure II-4–9: Preprocessor-Algorithm Combinations on GBU Dataset with ROC Curve

II-4.4.2 CMC Curve Metric

The summary results are presented in multiple formats:

- One CMC curve for each of three datasets for each preprocessor-algorithm-dataset combination (Figures II-B–1 through II-B–22; note the GBU dataset results are presented for the average (across three partitions) ROC and CMC curves);
- One plot for each dataset containing CMC curves for each preprocessor-algorithm combination (Figures II-4–10 through II-4–12; and
- A rank-percentage table to compare/contrast CMC curve shapes (Table II-4–3).

Comparison of the individual and combined CMC curves follows the same approach as the ROC curve analysis; however, an AUC value is not used to characterize CMC curves. Since the ideal behavior for these curves is to start at a high value (in the Rank = 1 region), Table II-4–3 shows rank value at various percentages to allow easy comparison across algorithms and datasets.

The CMC curves follow the same general performance trend as the ROC curves – the Native CMU preprocessor-algorithm combination has superior performance compared to all of the other combinations. The Dlib CMU algorithm has comparable performance to OpenFace, and all three are superior to PittPatt. The same pattern appears in the percentage-at-rank table (Table II-4–3). In particular, the percentage at Rank =1 of Native CMU is between 20 and 30 points higher than for Dlib CMU and OpenFace, the next highest performing algorithms. The percent values of Dlib CMU are 0 to 20 points higher than OpenFace and OpenFace is 10 to 40 points higher than PittPatt.

Table II-4–3: CMC Curve Summaries

	LFW					GBU					CS3			
Rank at	OpenFace	PittPatt	Dlib CMU	Native CMU		OpenFace	PittPatt	Dlib CMU	Native CMU		OpenFace	PittPatt	Dlib CMU	Native CMU
~ 10%	1	1	1	1		1	1	1	1		1	1	1	1
~ 20%	1	1	1	1		1	1	1	1		1	1	1	1
~ 30%	1	6	1	1		1	1	1	1		1	1	1	1
~ 40%	1	18	1	1		1	1	1	1		1	7	1	1
~ 50%	1	47	1	1		1	3	1	1		2	23	1	1
~ 60%	1	110	1	1		1	8	1	1		5	64	1	1
~ 70%	2	235	4	1		2	18	1	1		13	144	5	1
~ 80%	7	457	15	1		5	40	2	1		32	294	26	1
~ 90%	28	958	116	1		13	88	8	1		90	603	188	1
~ 99%	604	2529	2251	2		83	257	195	52		485	1387	1389	458
~ 100%	2832	3292	3292	2826		259	383	424	411		1389	1779	1778	1757

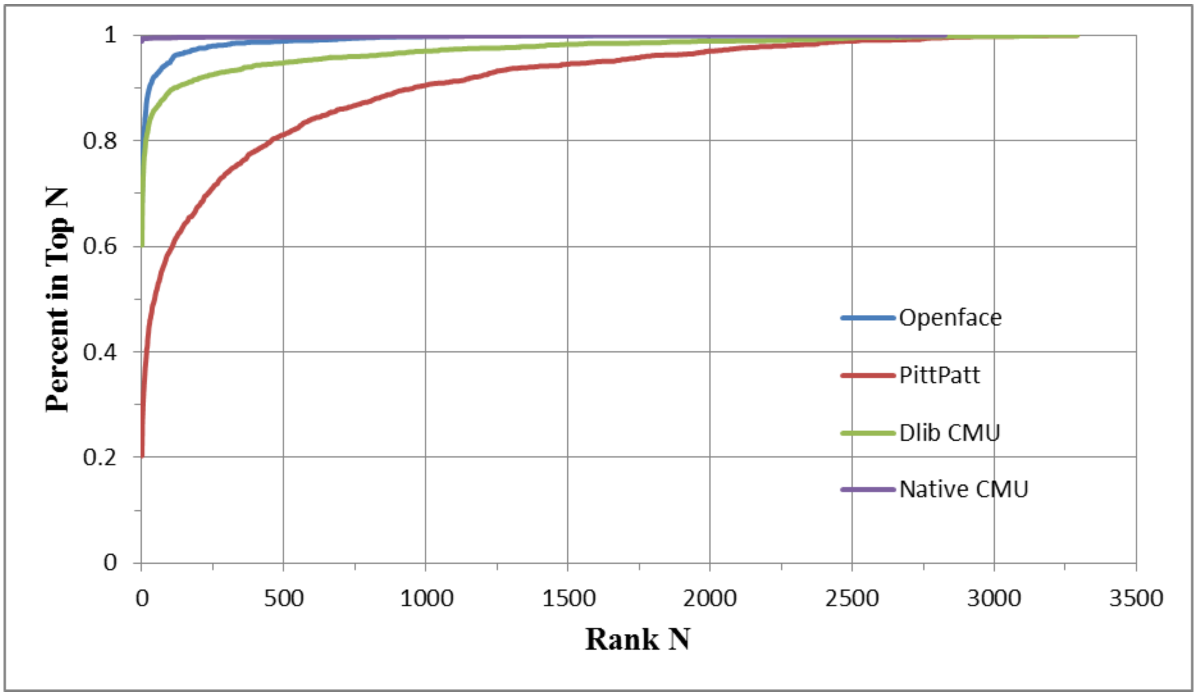


Figure II-4–10: Preprocessor-Algorithm Combinations on LFW Dataset with CMC Curve

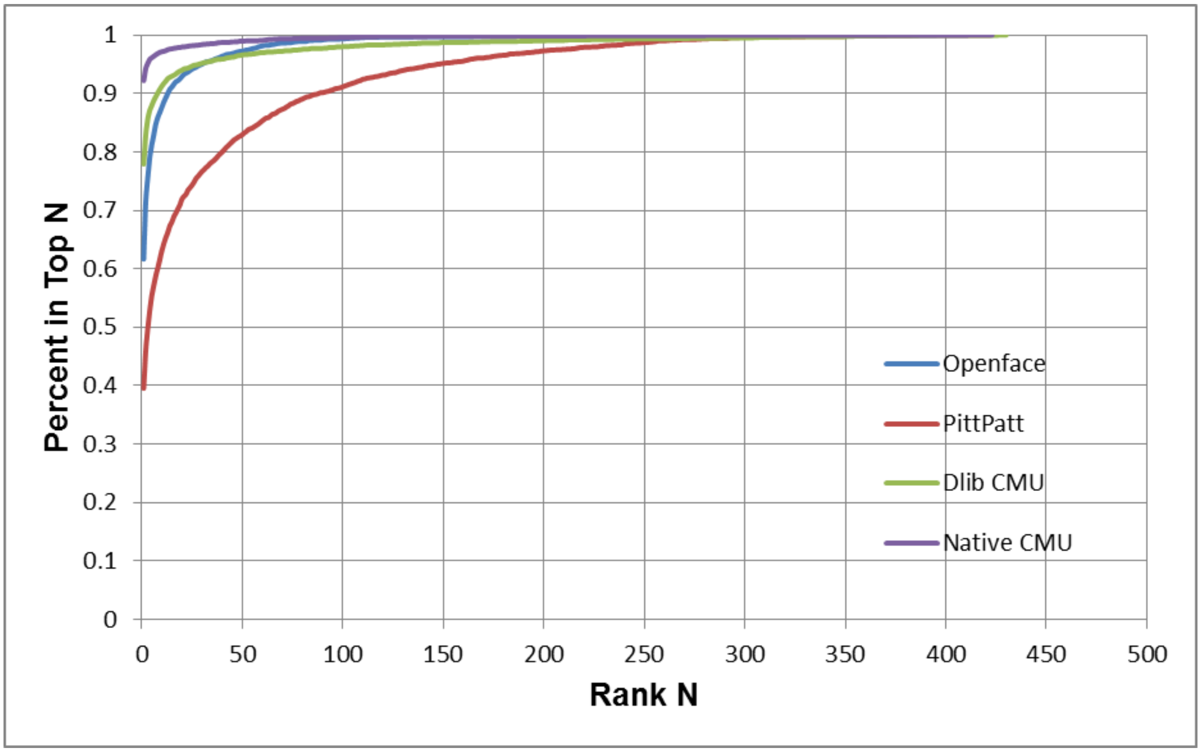


Figure II-4–11: Preprocessor-Algorithm Combinations on CS3 Dataset with CMC Curve

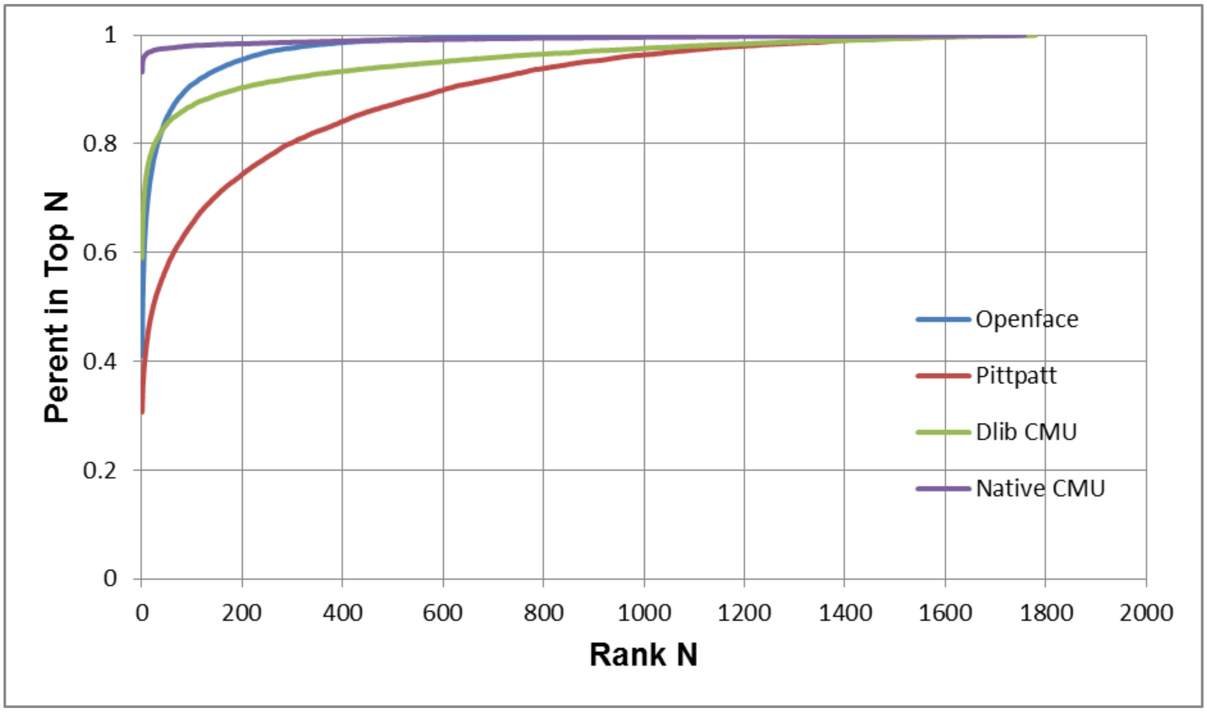


Figure II-4–12: Preprocessor-Algorithm Combinations on CS3 Dataset with CMC Curve

II-5 CONCLUSIONS

This assessment of face recognition algorithms compared results from four preprocessor-algorithm combinations (comprised of two preprocessors and three algorithms) processing three datasets using two primary metrics: ROC curves, CMC curves. The four combinations were used to address two assessment objectives: compare three algorithms receiving input from the same Dlib preprocessor; and compare two algorithms with native (i.e., built-in) preprocessing. The results found that for both the ROC and CMC Curves, the two algorithms based on CNN implementations (CMU, OpenFace) significantly out-performed the algorithm based on manually crafted features (PittPatt).

Practical applications of face recognition technology will likely require capabilities with high true positive rate and low false positive rate. This is the region in the vicinity of (0,1) for ROC curves and (rank 1, 100 percent) for CMC curves. A distinguishing characteristic of the best performing algorithms is their large values on the left-hand side of their respective curves. Consistently, for assessment of the native preprocessing using the AOC-ROC metric, the Native CMU algorithm outperformed OpenFace by a few percent (see Table II-4–2). For assessing the fixed preprocessor (Dlib) combinations on all datasets, PittPatt significantly lagged performance of both OpenFace and Dlib CMU, which were more comparable with each other. In particular, in the AUC-ROC analysis, OpenFace is superior to Dlib CMU in two out of three cases, and in all cases, the differences are on the order of 2 percent or less.

The same trend holds in the results using the CMC Curve (see Table II-4–3). For the native preprocessor cases across all datasets, the maximum percentage to which Rank = 1 extends is 30 to 50 points higher for Native CMU than OpenFace. For the fixed preprocessor case, the maximum percentage to which Rank = 1 extends is zero to 20 points higher for Dlib CMU than OpenFace. OpenFace is 10 to 40 points higher than PittPatt.

The results for both metrics show the Native CMU algorithm is superior; however, in the case of the ROC curve analysis, Native CMU is only higher by an AUC of 0.02 or less. The relative superiority of Native CMU is larger for the CMC analysis.

II-6 ACRONYMS & ABBREVIATIONS

AUC	Area under the curve
BB	Bounding Box
CMC	Cumulative Matching Characteristic
CMU	Carnegie Mellon University
CNN	Convolutional Neural Networks
CS3	Challenge Set 3
FN	False negative
FP	False positive
FPR	False positive rate
FRVT2006	Face Recognition Vendor Test 2006
GBU	The Good, the Bad and the Ugly
GT	Ground Truth
IARPA	Intelligence Advanced Research Projects Agency
IJB	IARPA Janus Benchmark
JHU/APL	Johns Hopkins University Applied Physics Laboratory
LEA	Law Enforcement Agency
LFW	Labeled Faces in the Wild
NIJ	National Institute of Justice
PR	Precision/Recall
R&D	Research and Development
ROC	Receiver Operating Characteristic
RT&E Center	Research, Test, and Evaluation Center
TP	True positive
TPR	True positive rate

APPENDIX II-A: INDIVIDUAL ROC CURVES

Appendix II–A: Figures

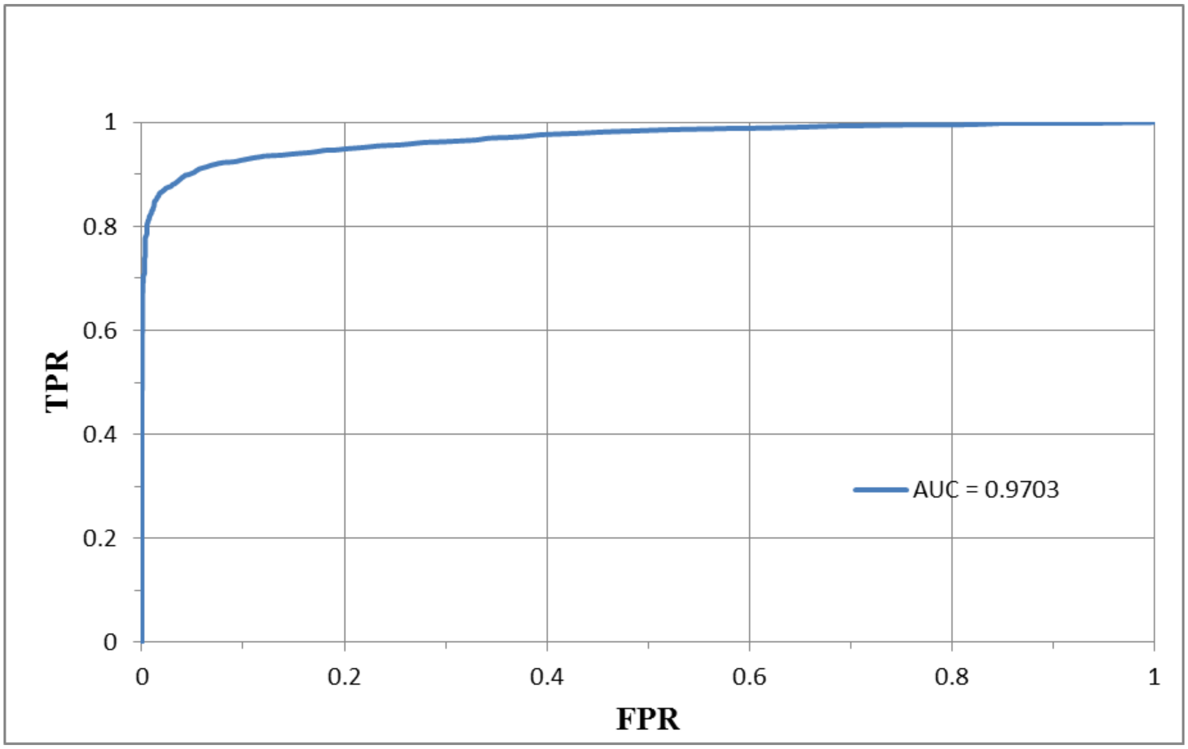


Figure II-A–1: Dlib CMU for LFW Dataset with ROC Curve

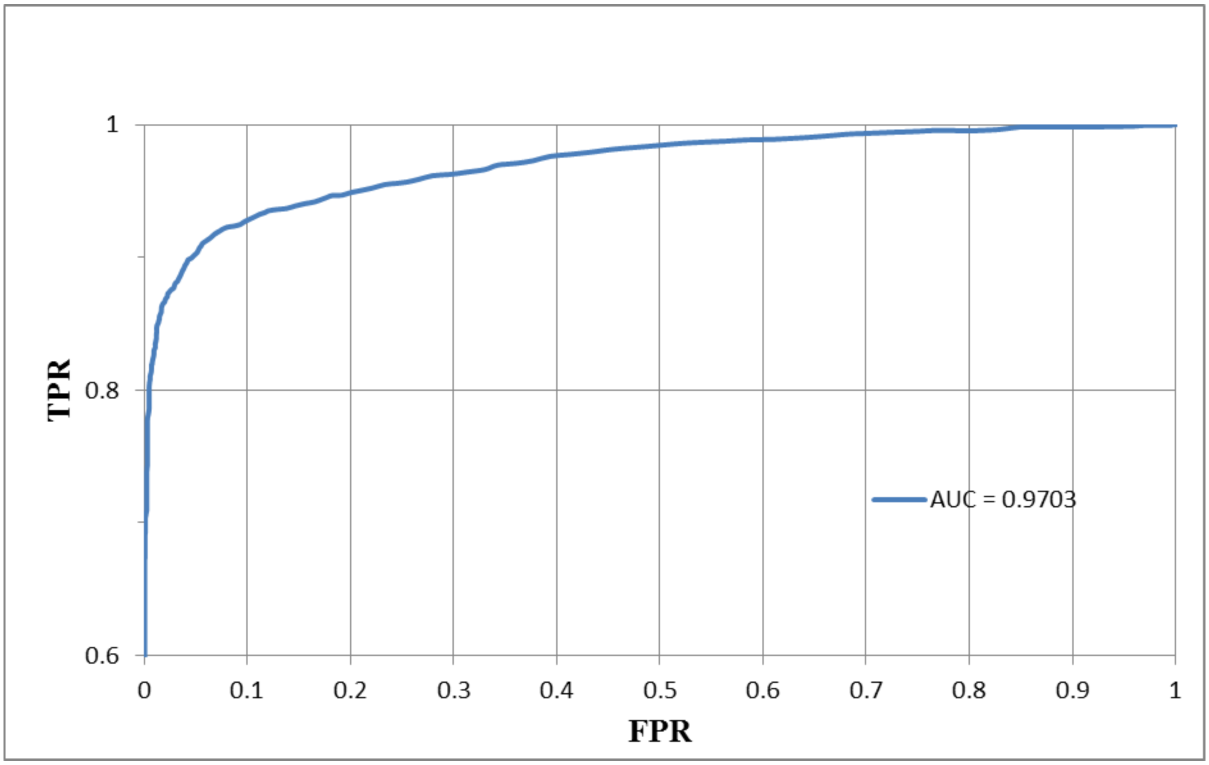


Figure II-A–2: Dlib CMU for LFW Dataset with ROC Curve

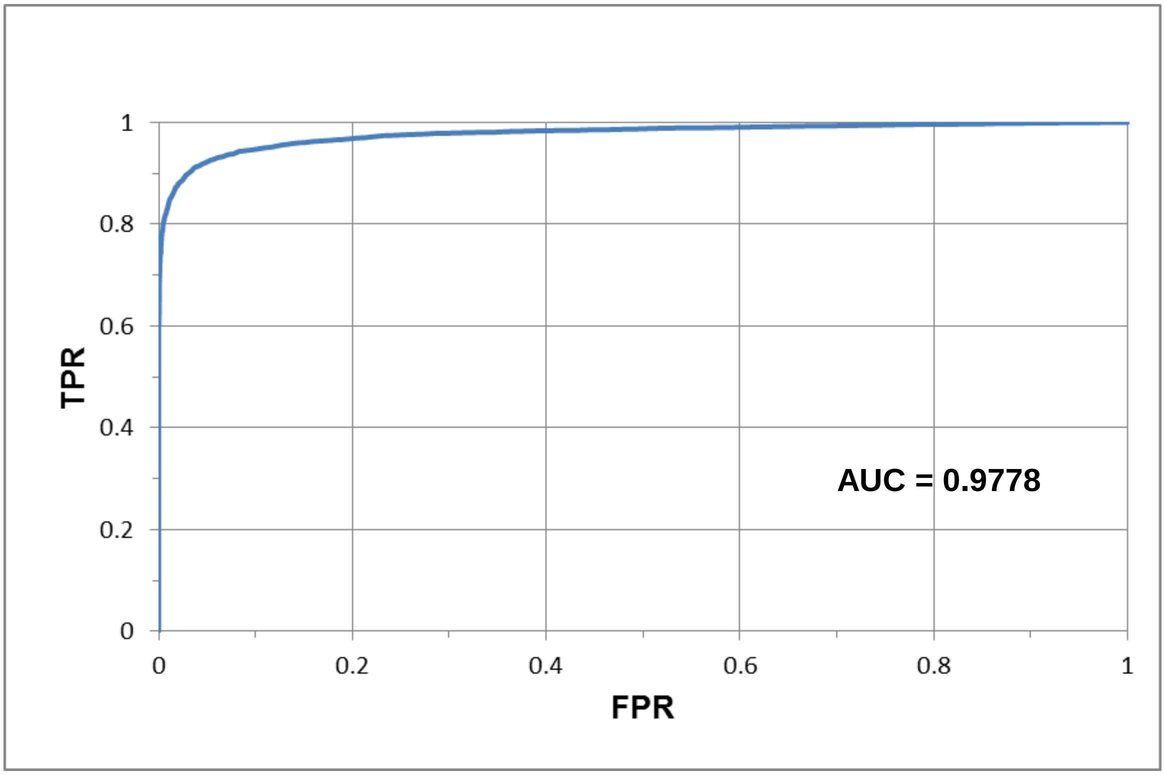


Figure II-A–3: Dlib CMU on GBU Dataset with ROC Curve

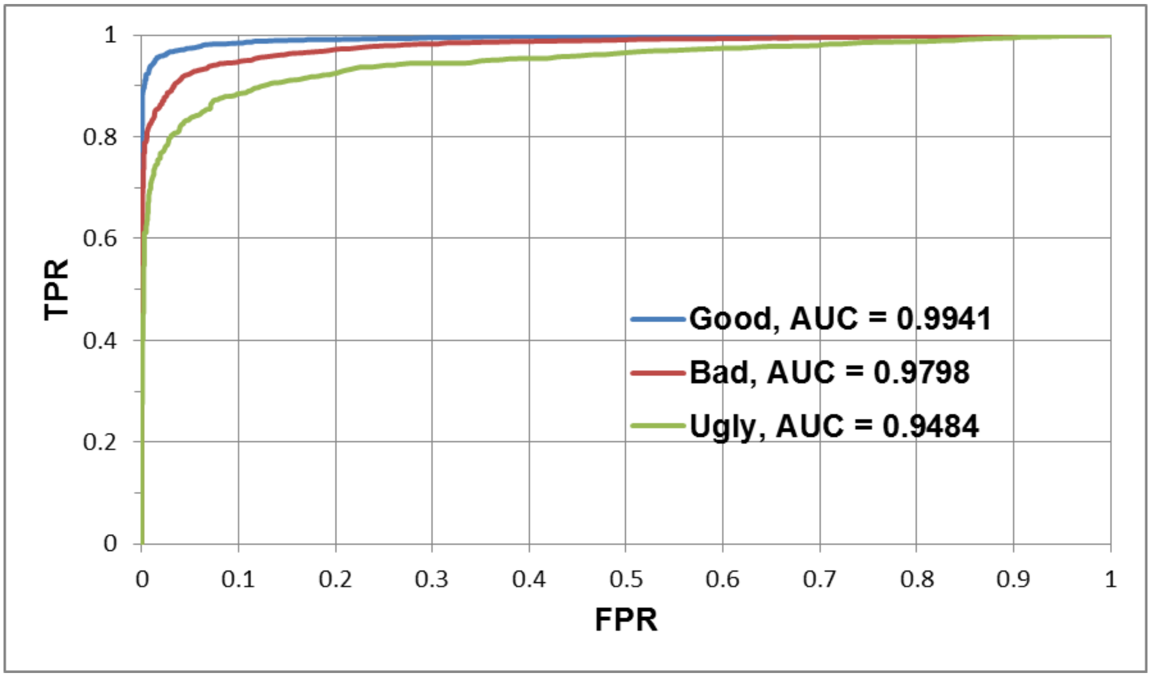


Figure II-A–4: Dlib CMU on GBU Partitions with ROC Curve

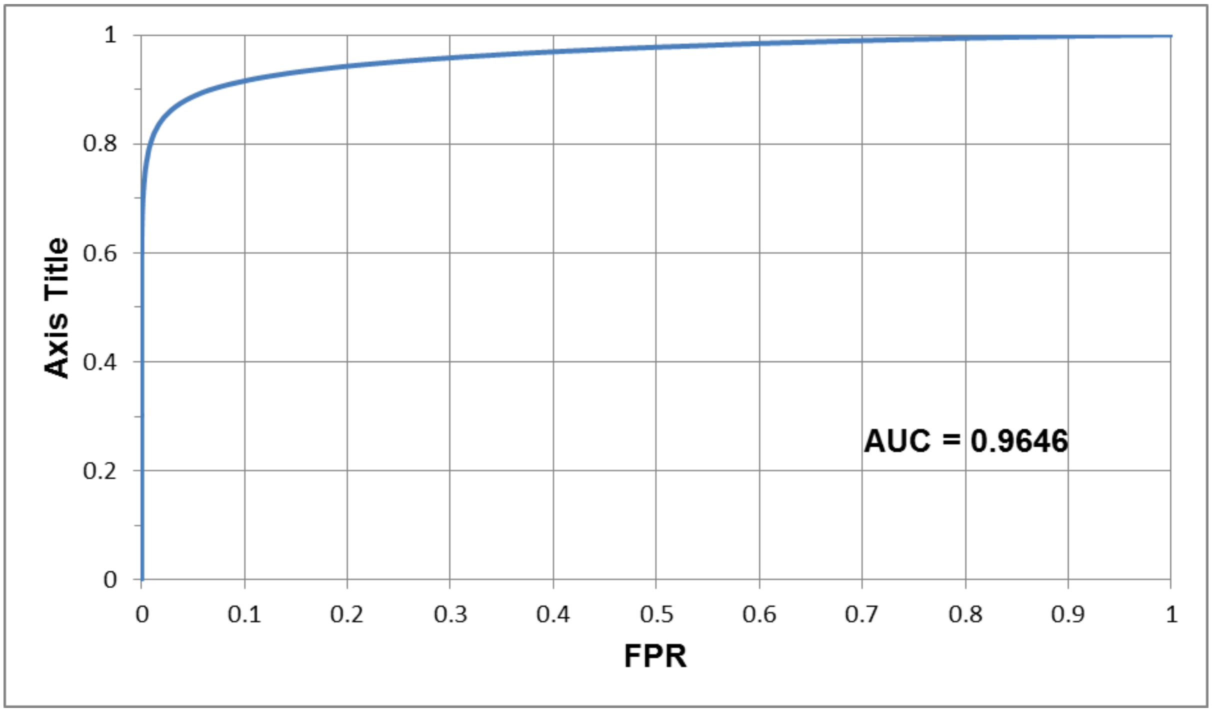


Figure II-A–5: Dlib CMU on CS3 Dataset with ROC Curve

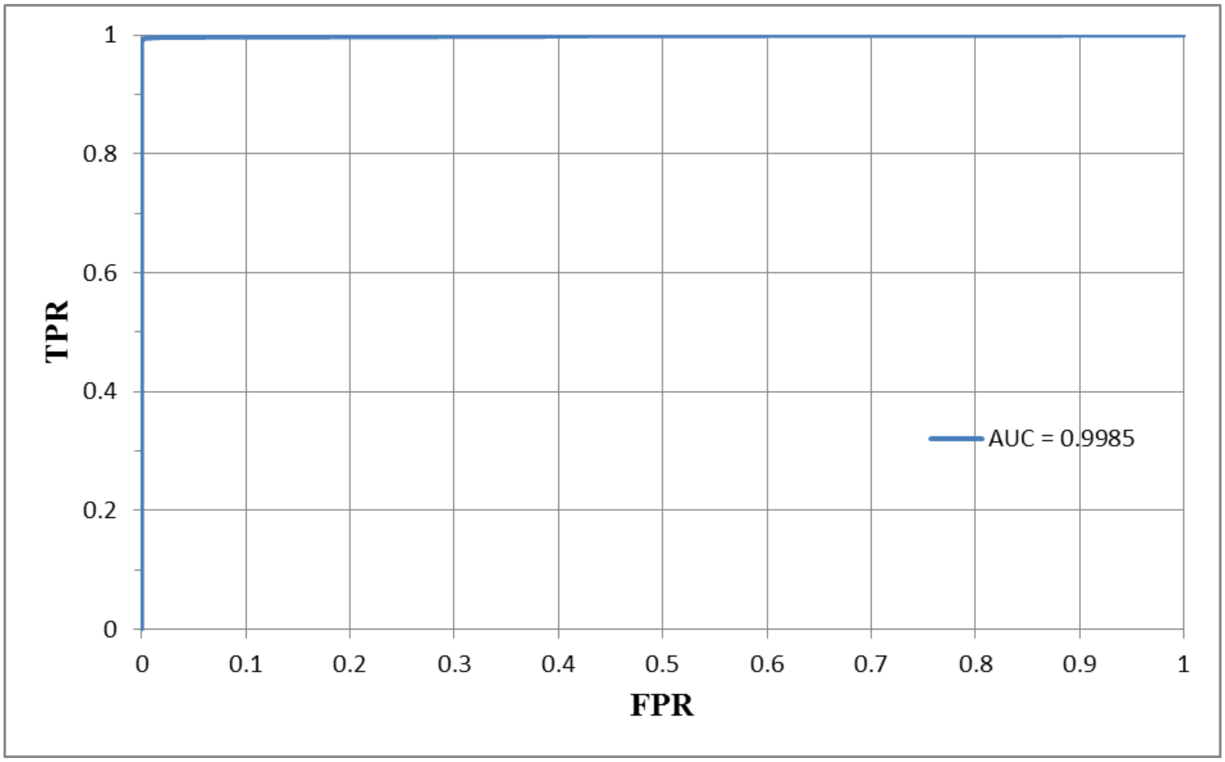


Figure II-A–6: Native CMU on LFW Dataset with ROC Curve

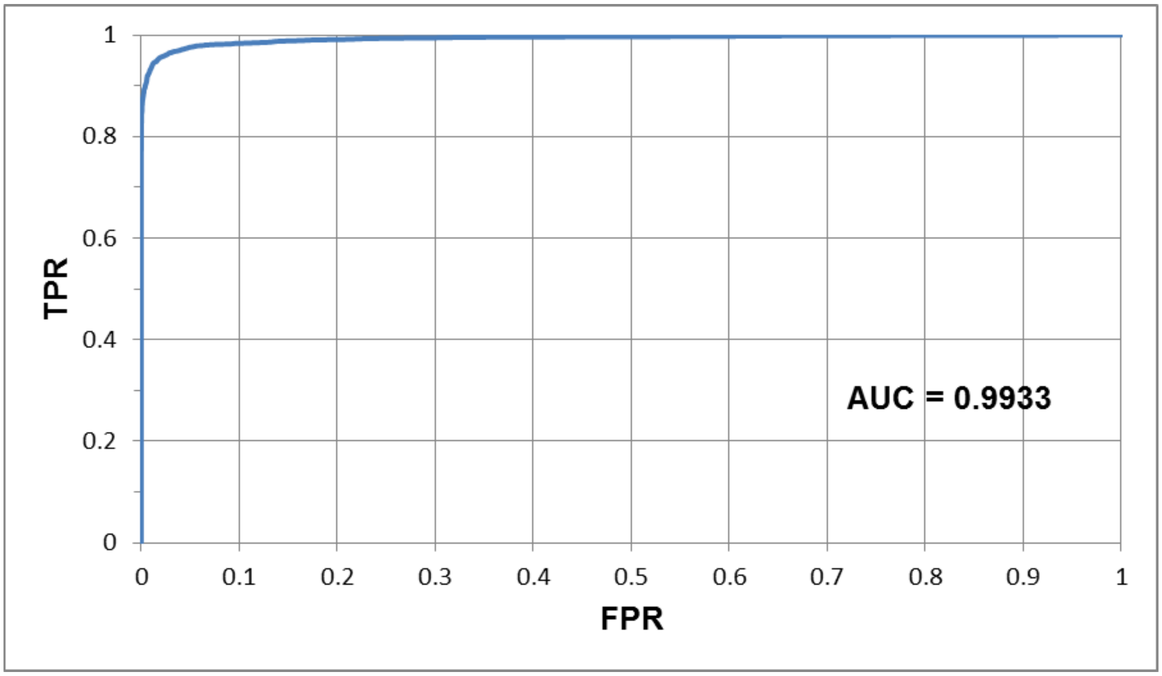


Figure II-A–7: Native CMU on GBU Dataset with ROC Curve

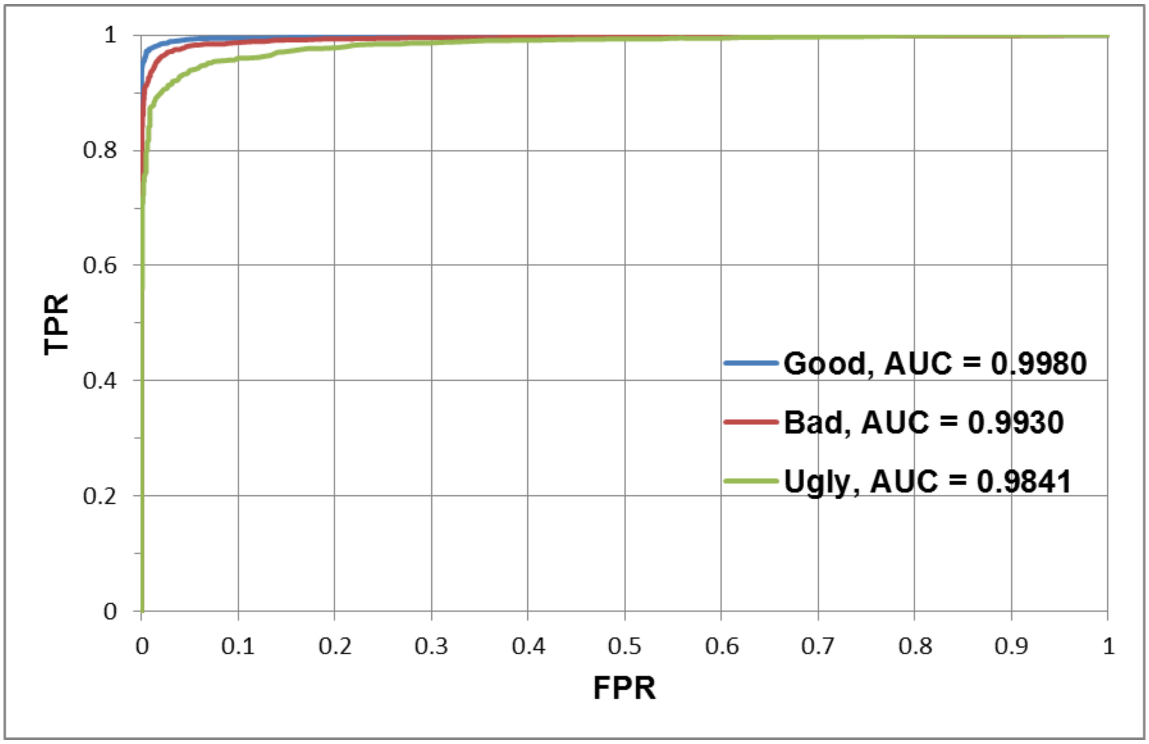


Figure II-A–8: Native CMU on GBU Partitions Dataset with ROC Curve Metric

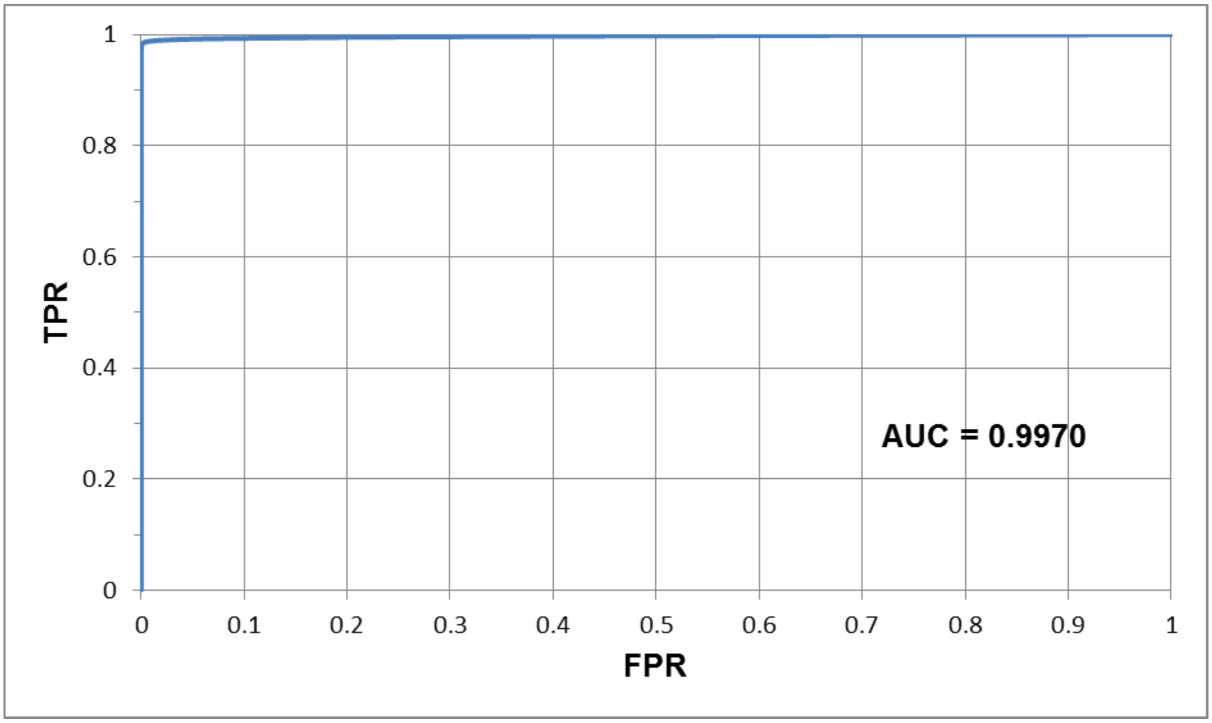


Figure II-A–9: Native CMU on CS3 Dataset with ROC Curve

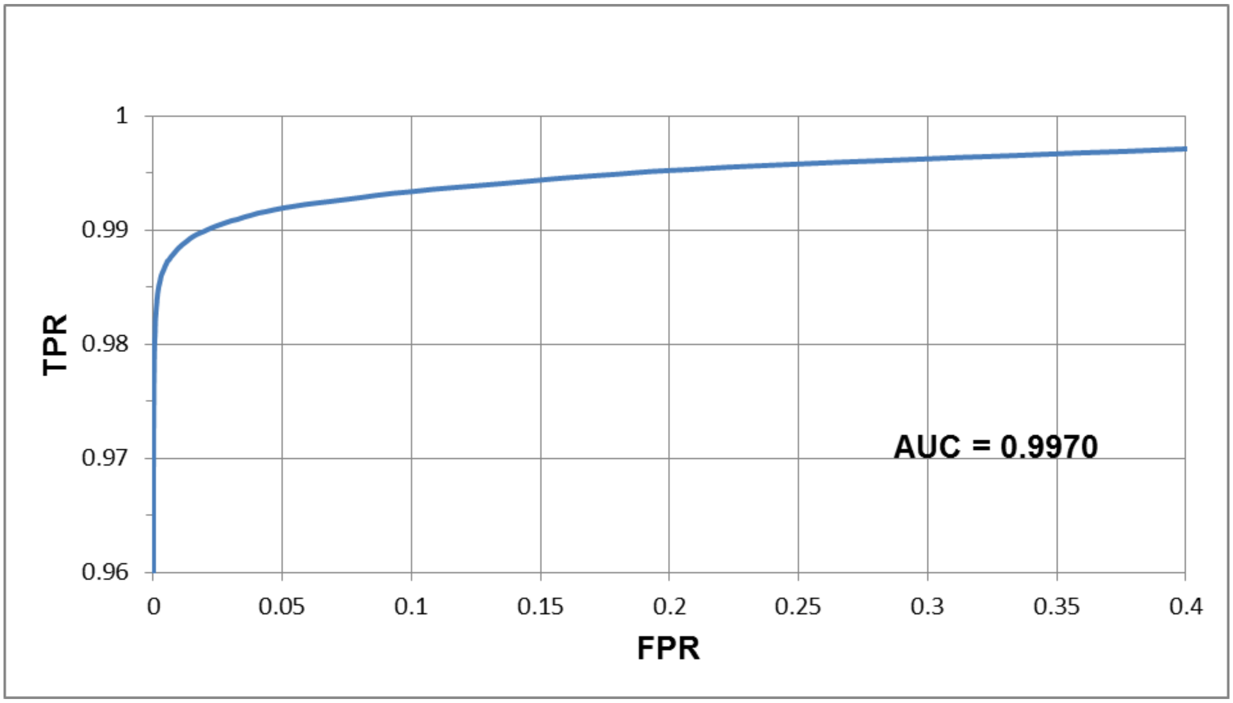


Figure II-A–10: Native CMU on CS3 Dataset with ROC Curve (Zoomed View)

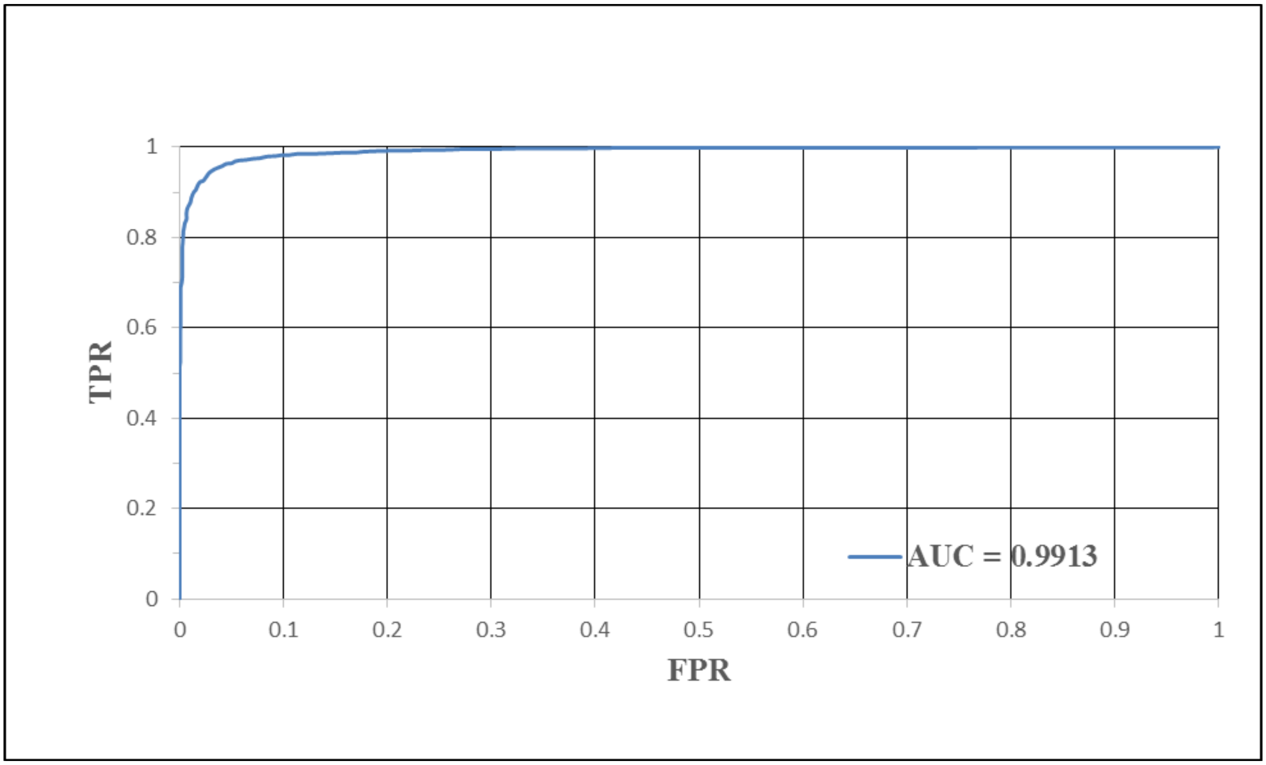


Figure II-A–11: OpenFace on LFW Dataset with ROC Curve

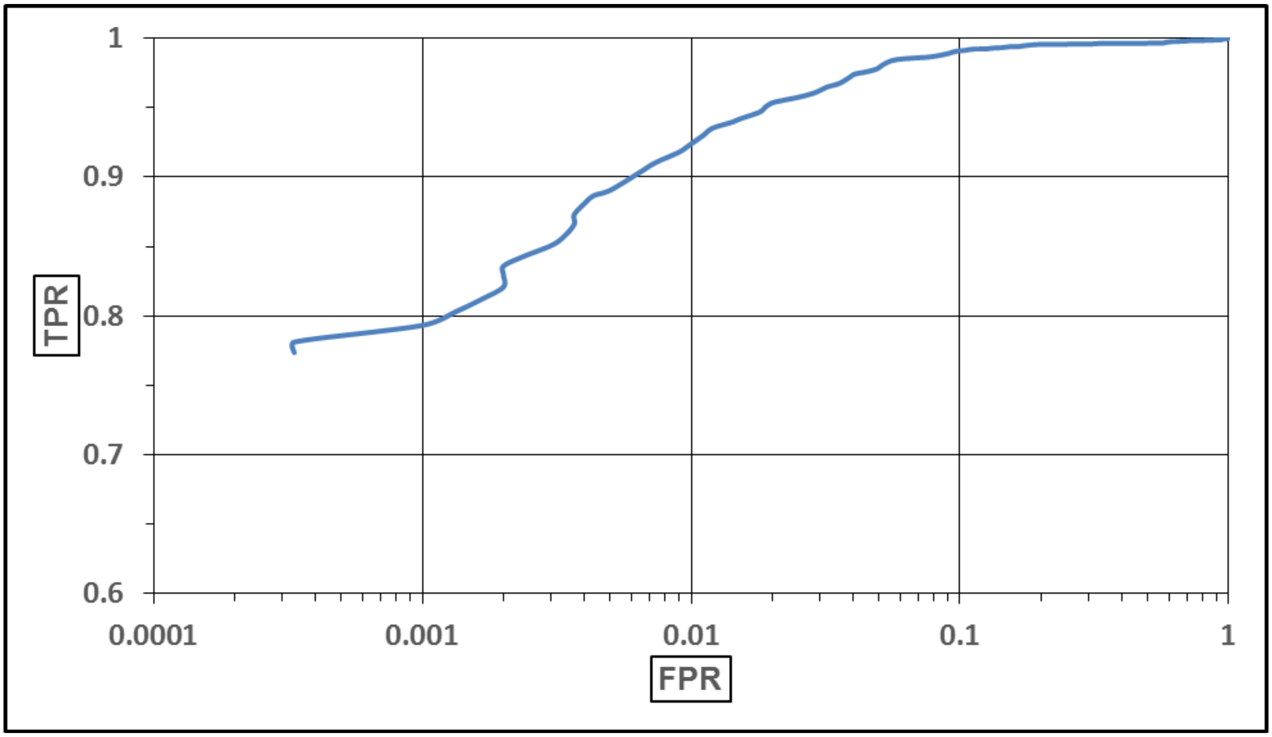


Figure II-A–12: OpenFace on LFW Dataset with ROC Curve

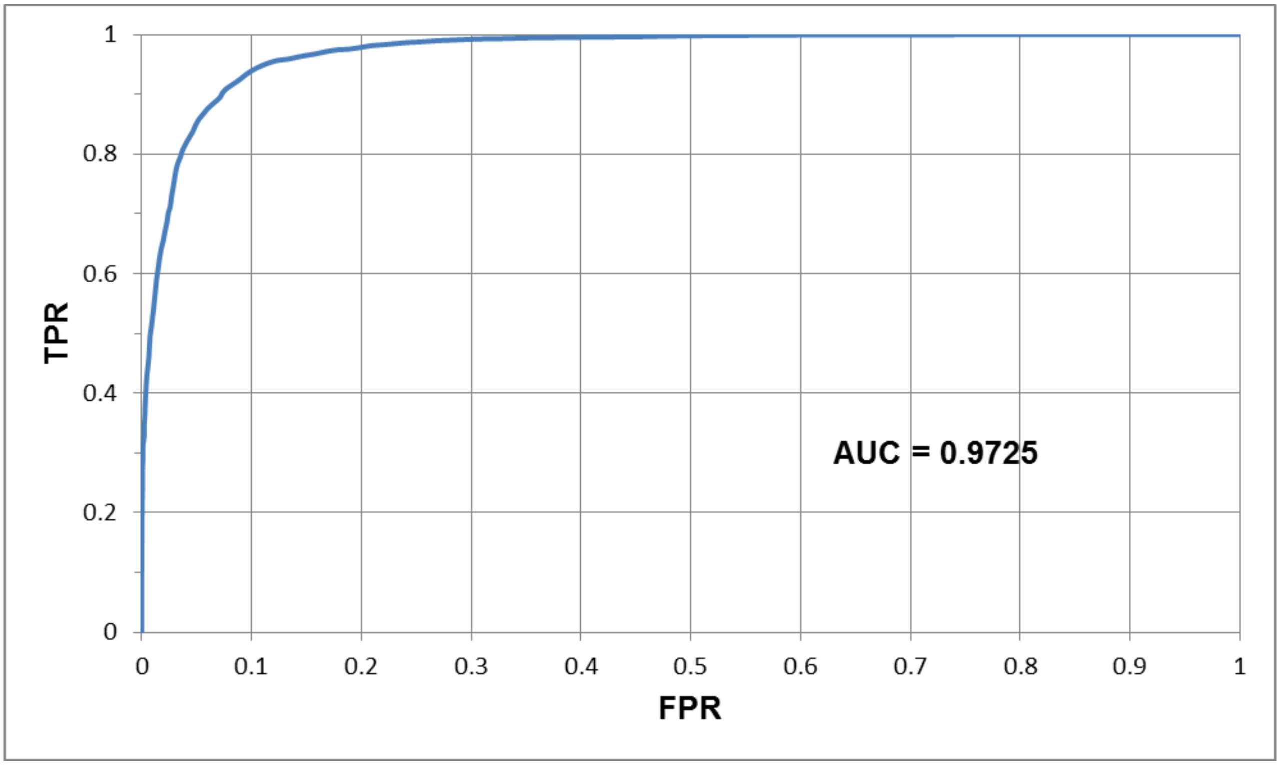


Figure II-A–13: OpenFace on GBU Combined Dataset with ROC Curve

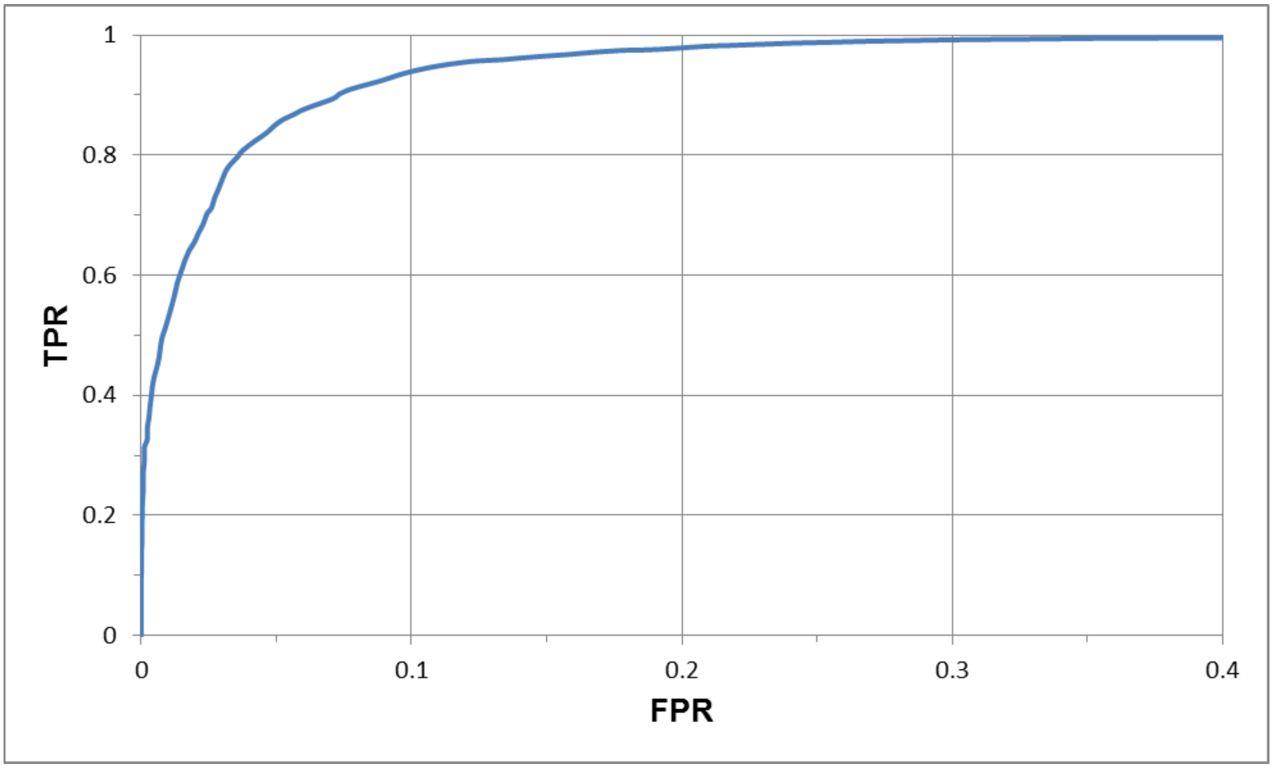


Figure II-A–14: OpenFace on GBU Combined Dataset with ROC Curve (Zoomed View)

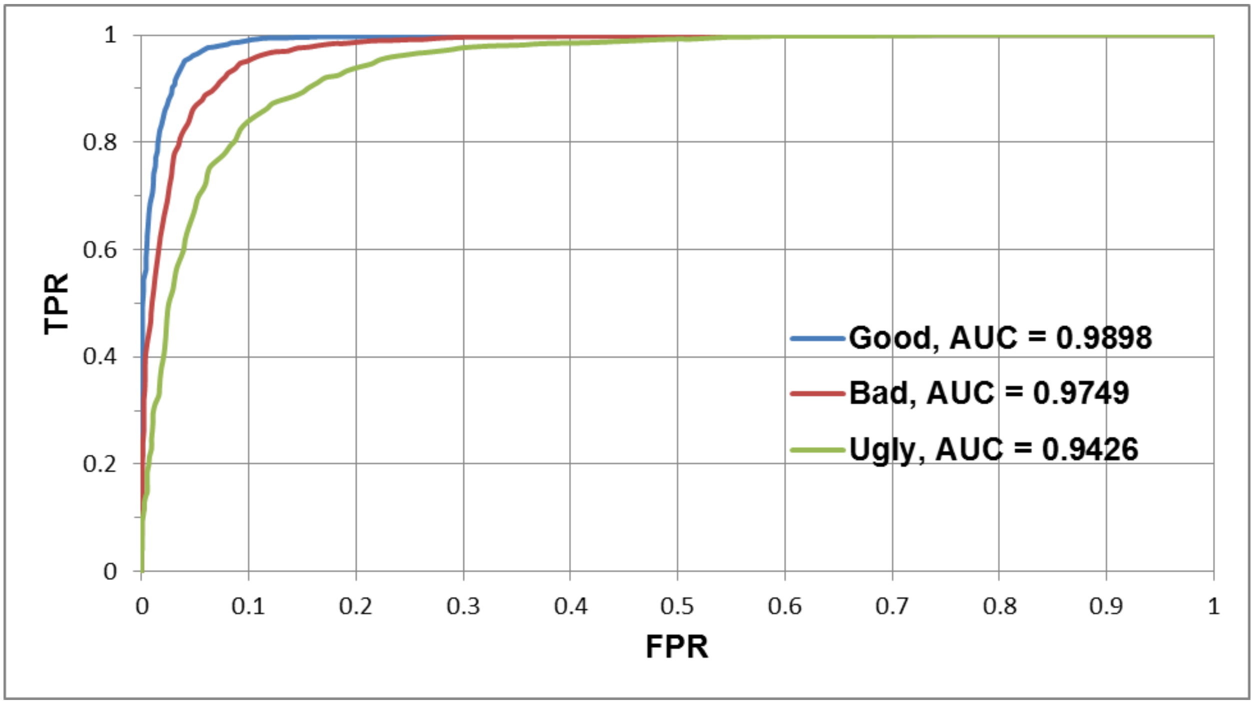


Figure II-A–15: OpenFace on GBU Partitions Dataset with ROC Curve

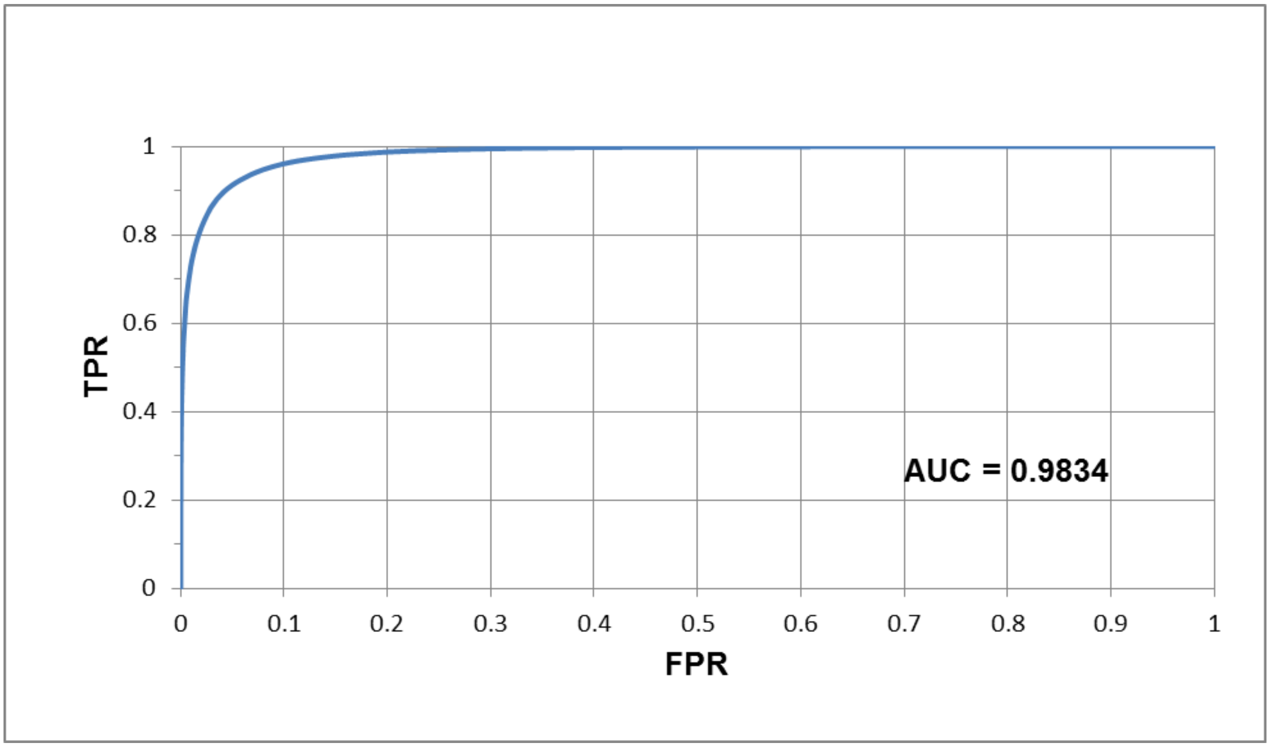


Figure II-A–16: OpenFace on CS3 Dataset with ROC Curve

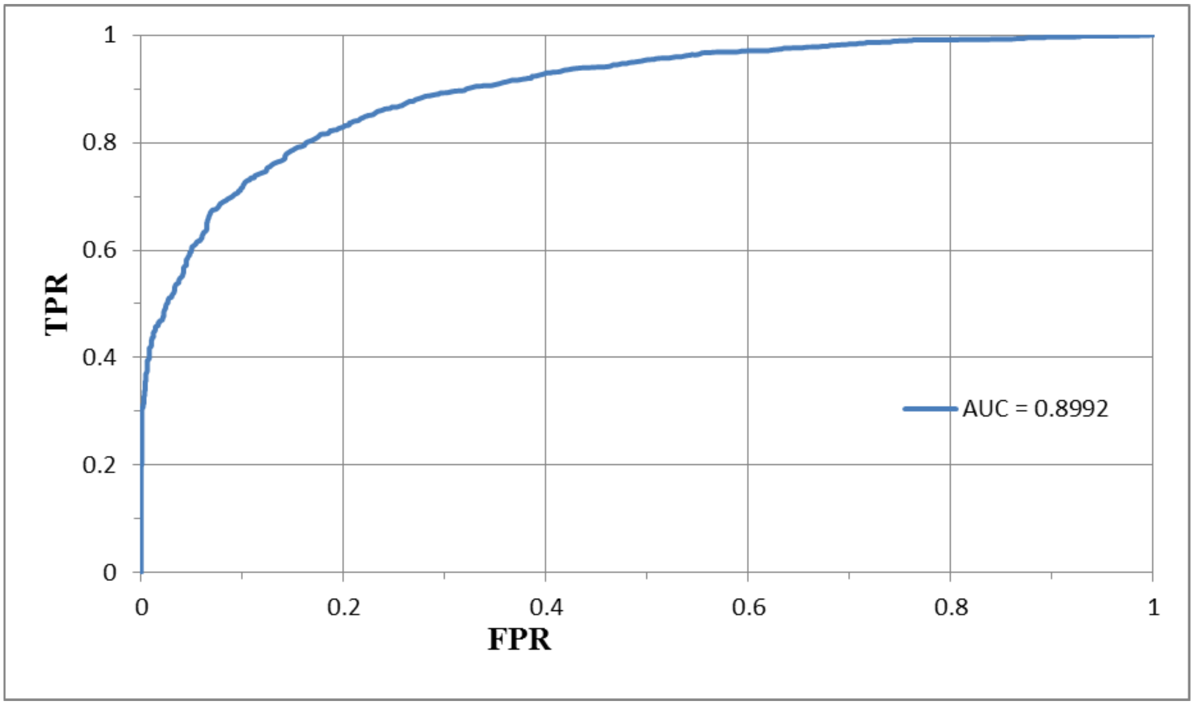


Figure II-A–17: PittPatt on LFW Dataset with ROC Curve

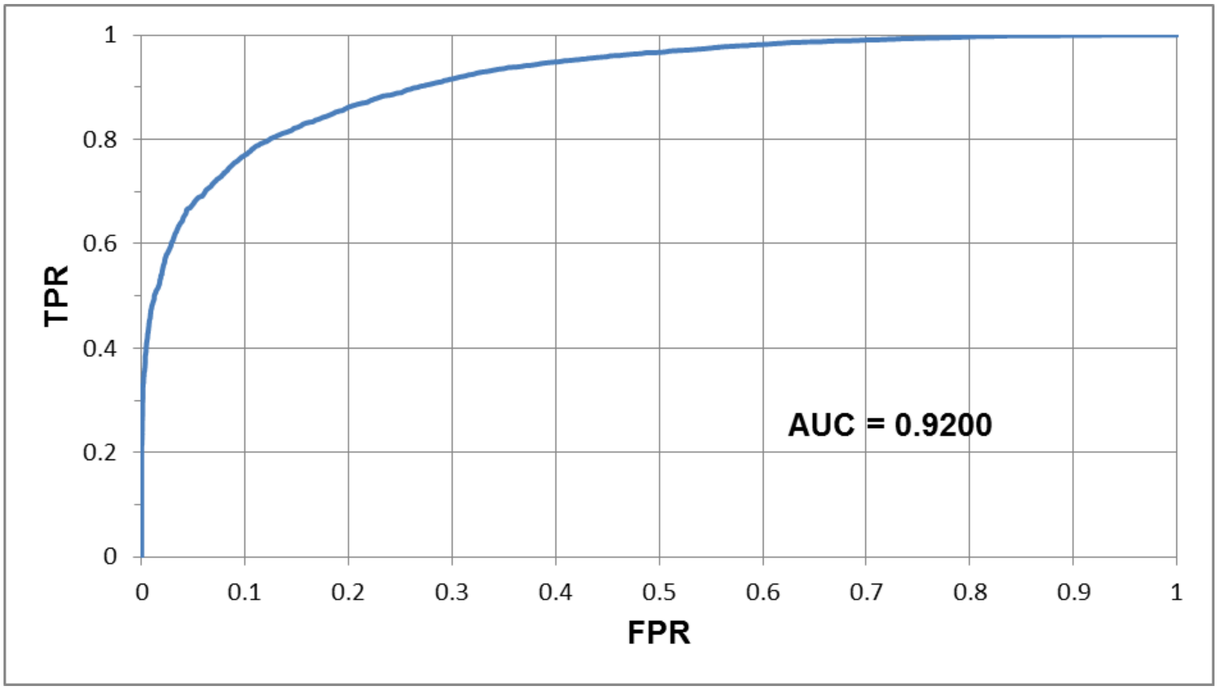


Figure II-A–18: PittPatt on GBU Combined Dataset with ROC Curve

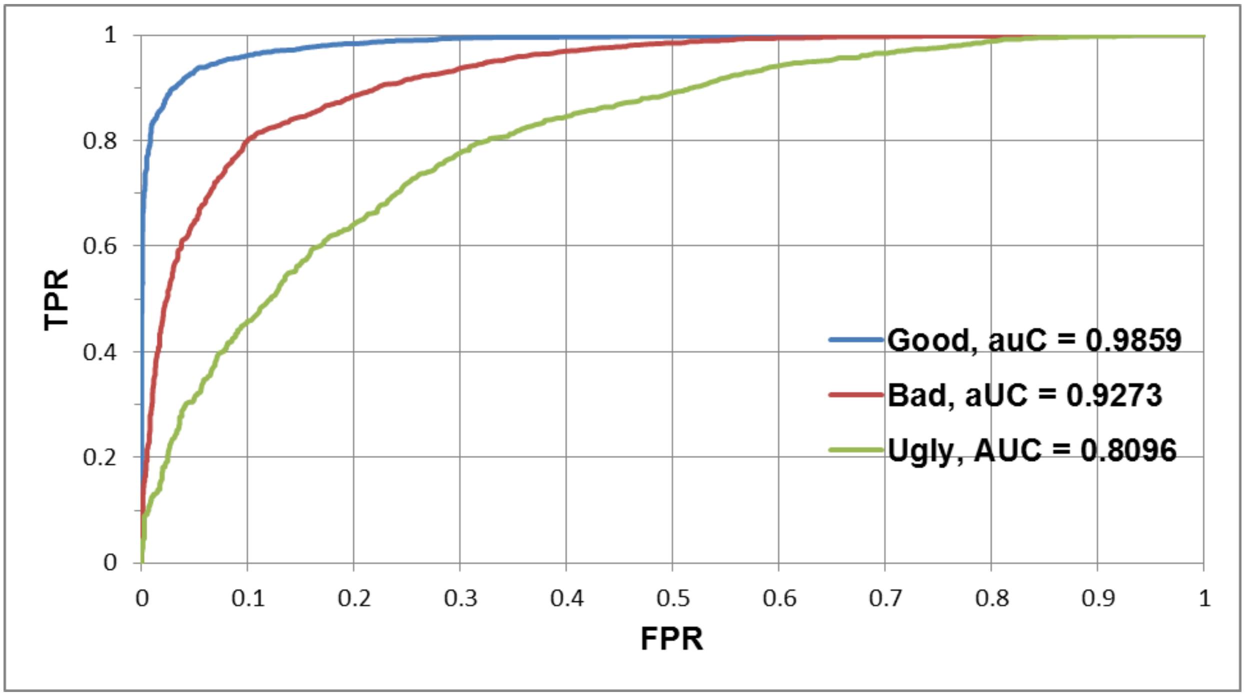


Figure II-A–19: PittPatt on GBU Partitions with ROC Curve

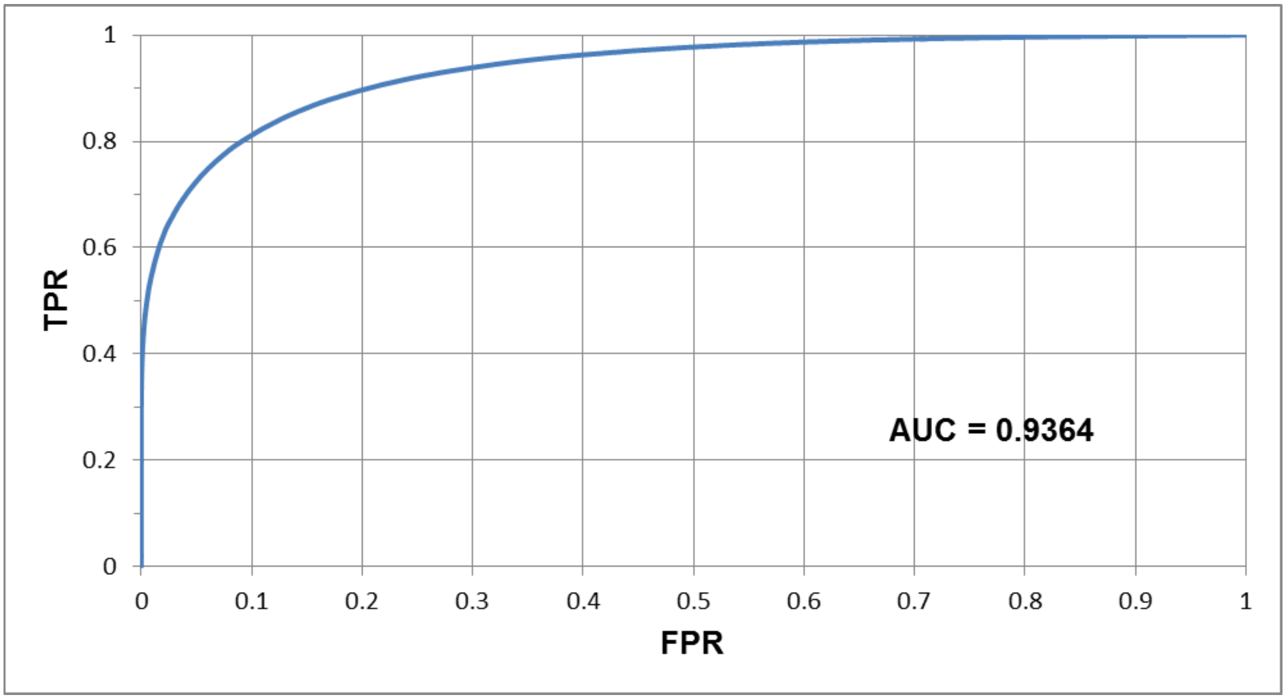


Figure II-A–20: PittPatt on CS3 Dataset with ROC Curve

APPENDIX II-B: INDIVIDUAL CMC CURVES

Appendix II–B: Figures

(Note: percent values (Y-axis) have been scaled to range [0,1])

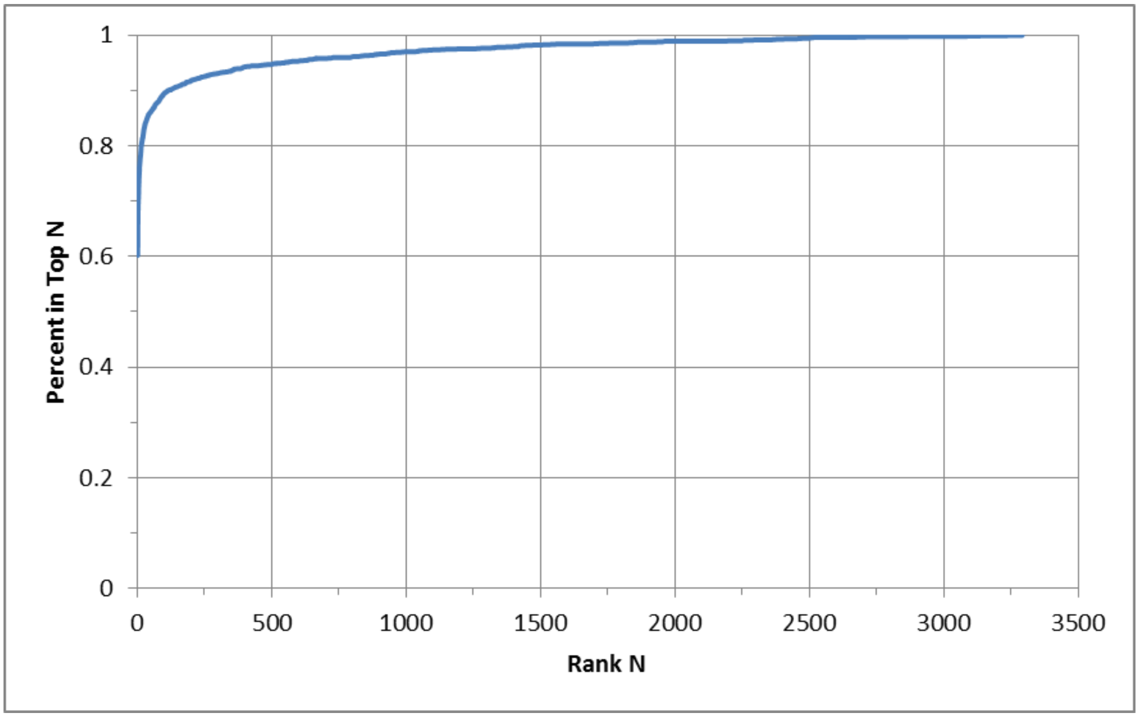


Figure II-B–1: Dlib CMU on LFW Dataset with CMC Curve

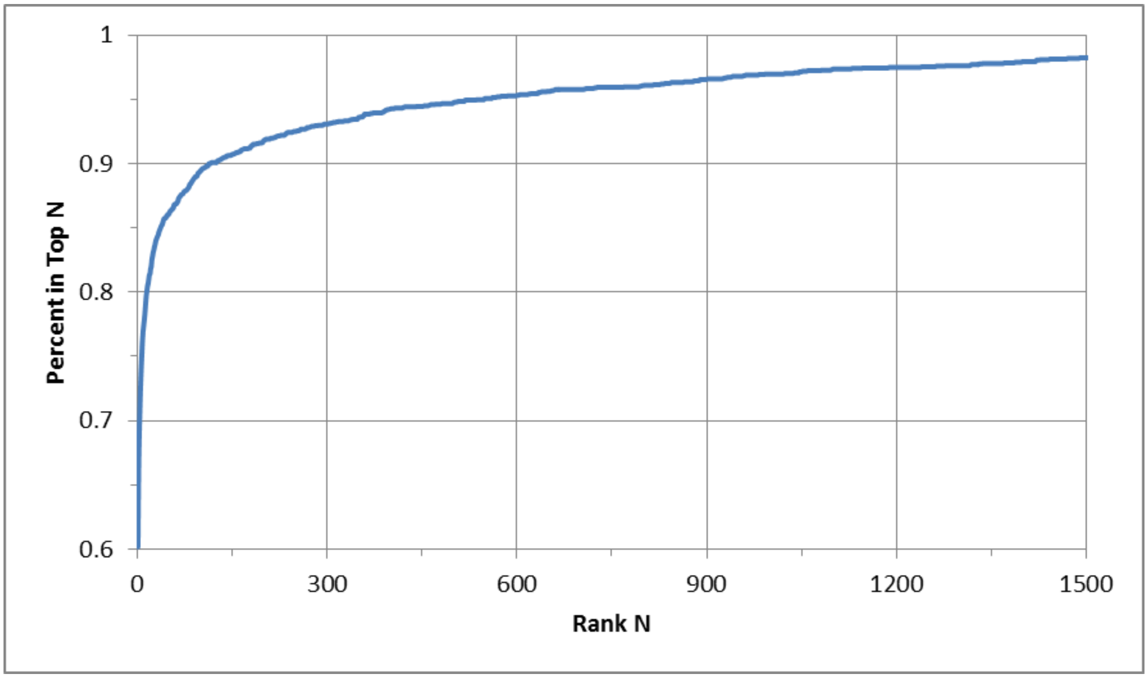


Figure II-B–2: Dlib CMU on LFW Dataset with CMC Curve (Zoomed View)

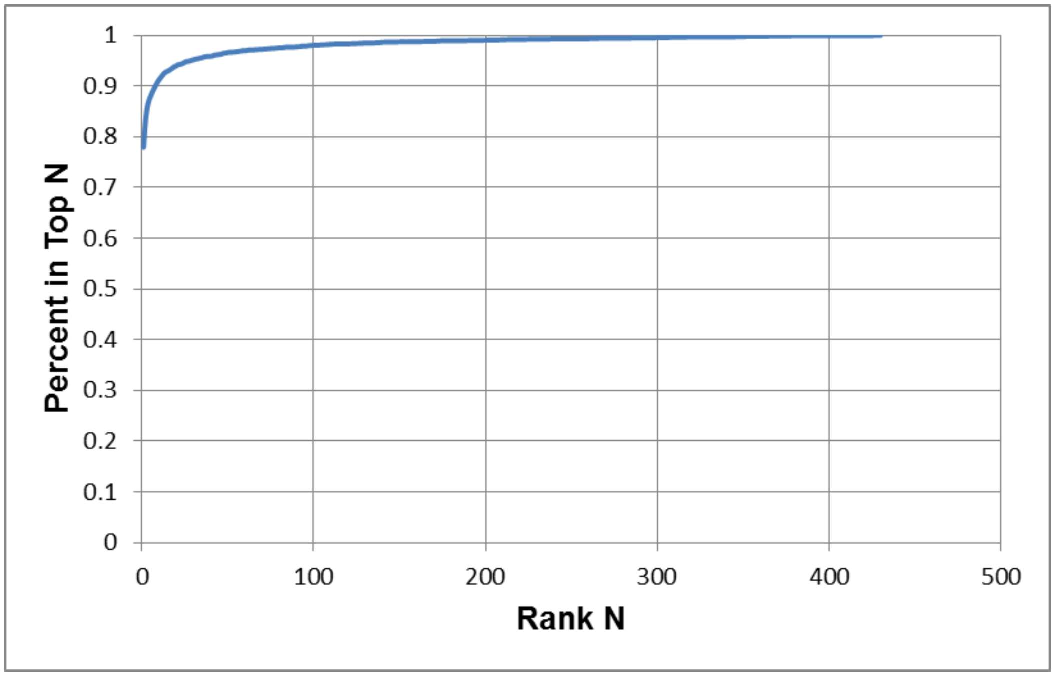


Figure II-B–3: Dlib CMU on GBU Combined Dataset with CMC Curve

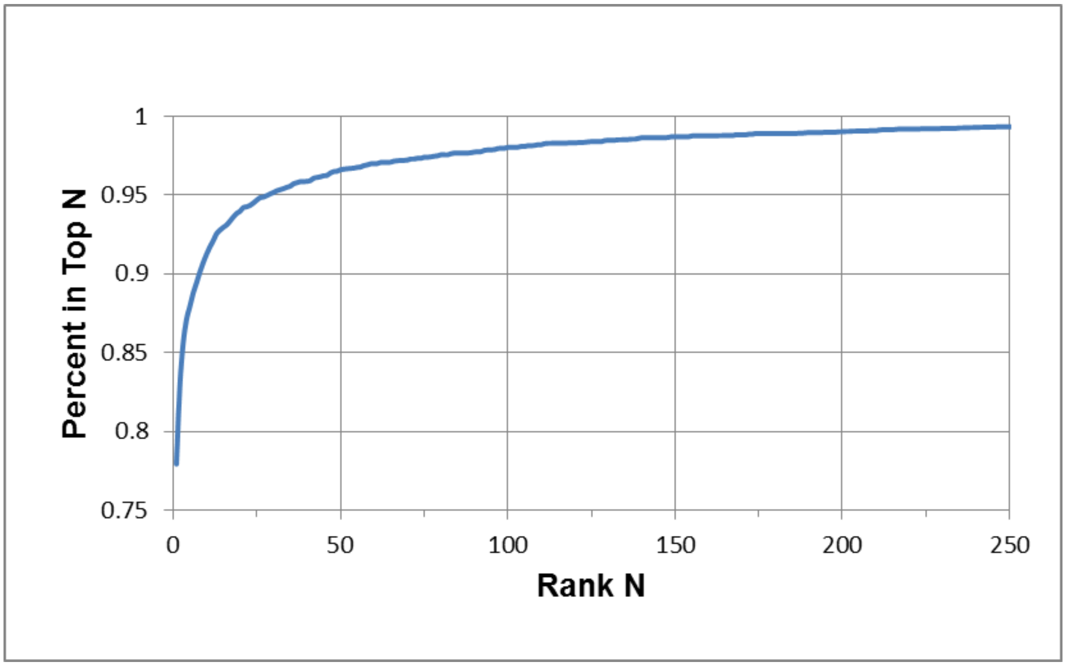


Figure II-B–4: Dlib CMU on GBU Combined Dataset with CMC Curve

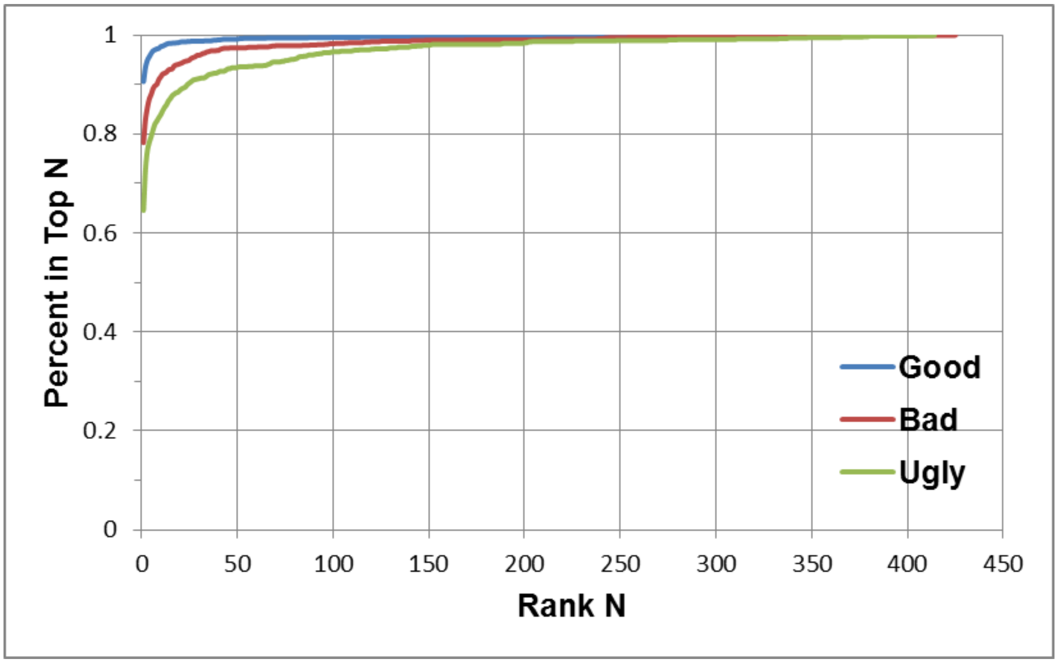


Figure II-B–5: Dlib CMU on GBU Partitions Dataset with CMC Curve

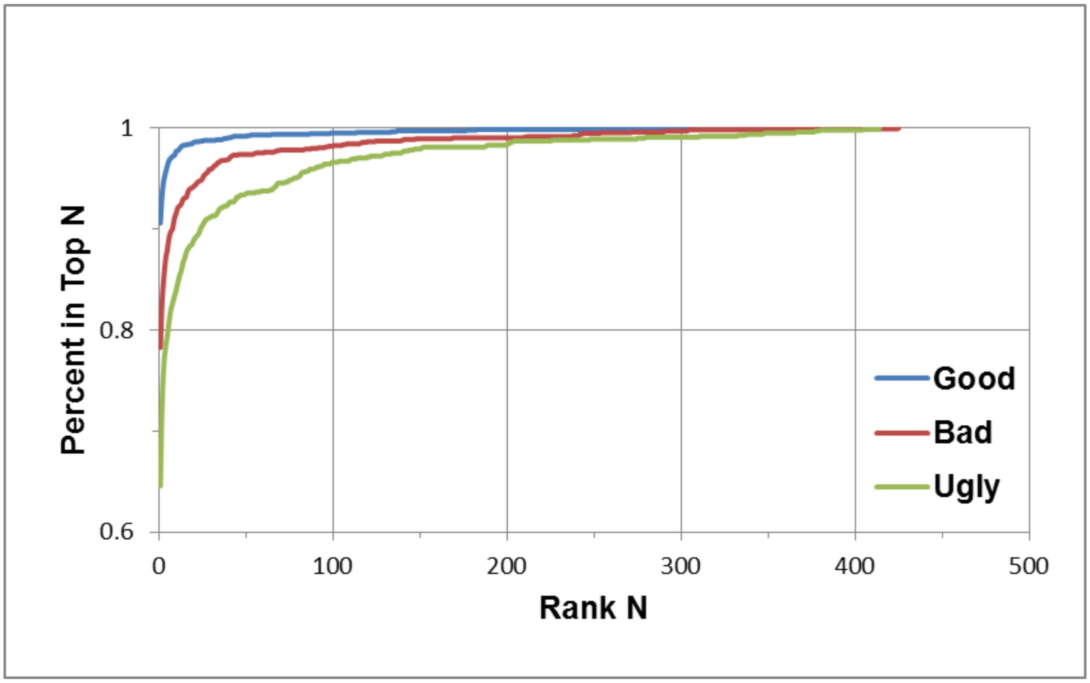


Figure II-B–6: Dlib CMU on GBU Partitions Dataset with CMC Curve

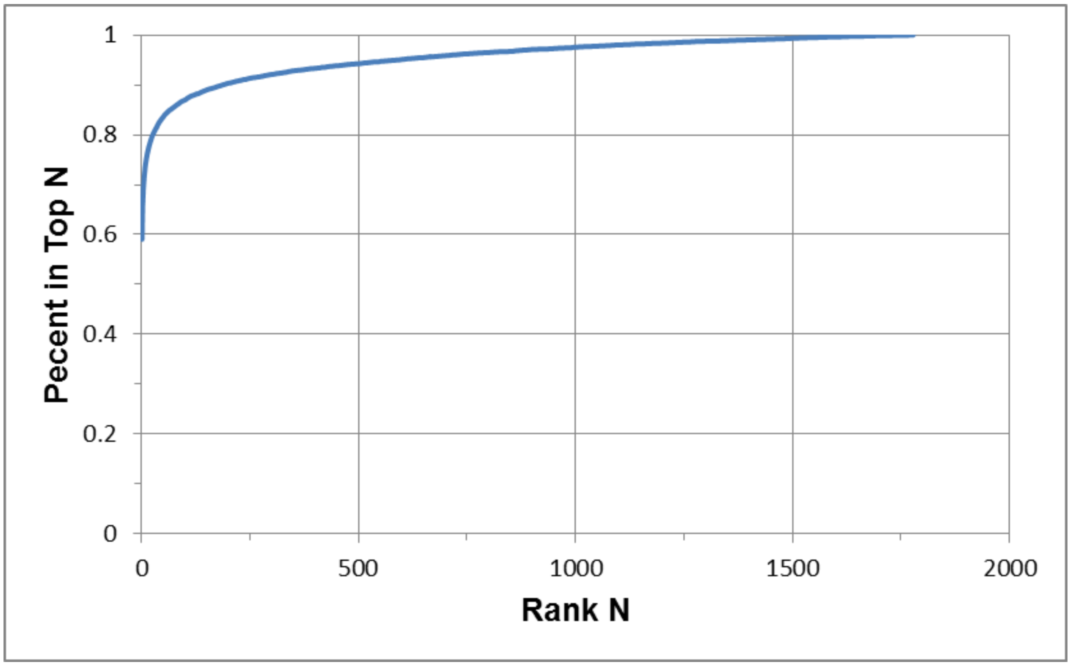


Figure II-B–7: Dlib CMU on CS3 Dataset with CMC Curve

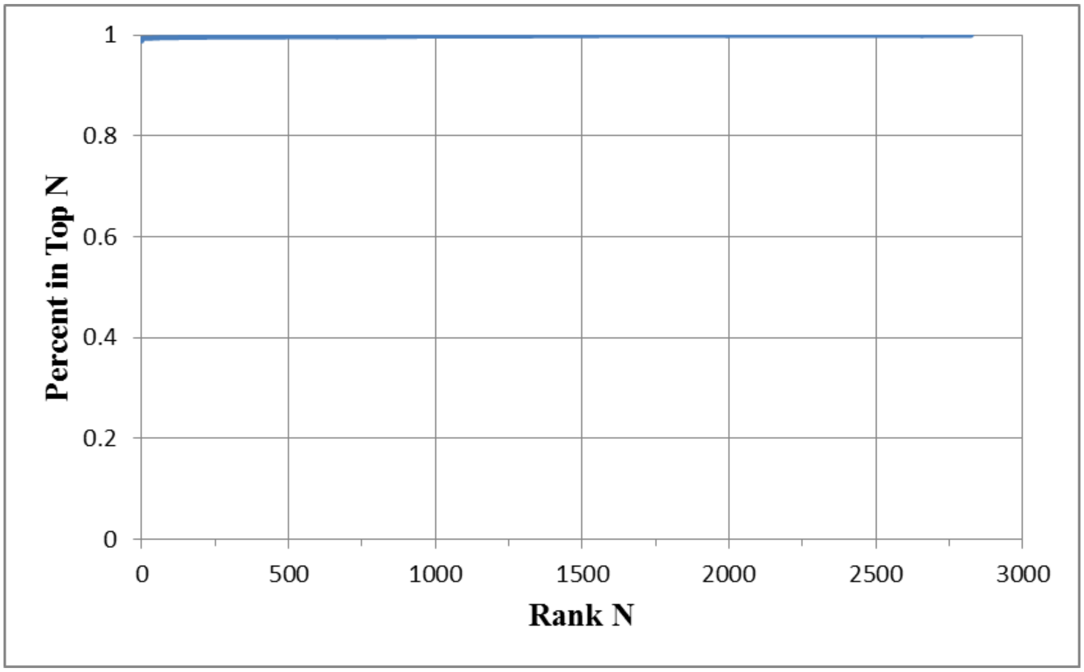


Figure II-B–8: Native CMU on LFW Dataset with CMC Curve

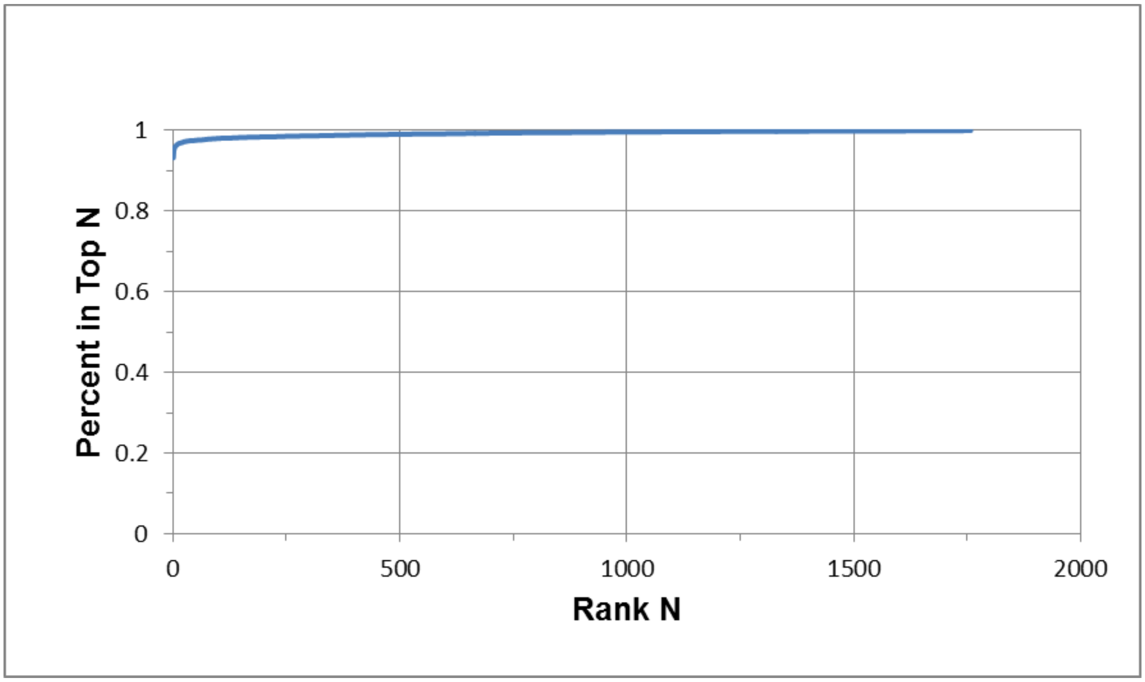


Figure II-B–9: Native CMU on CS3 Dataset with CMC Curve

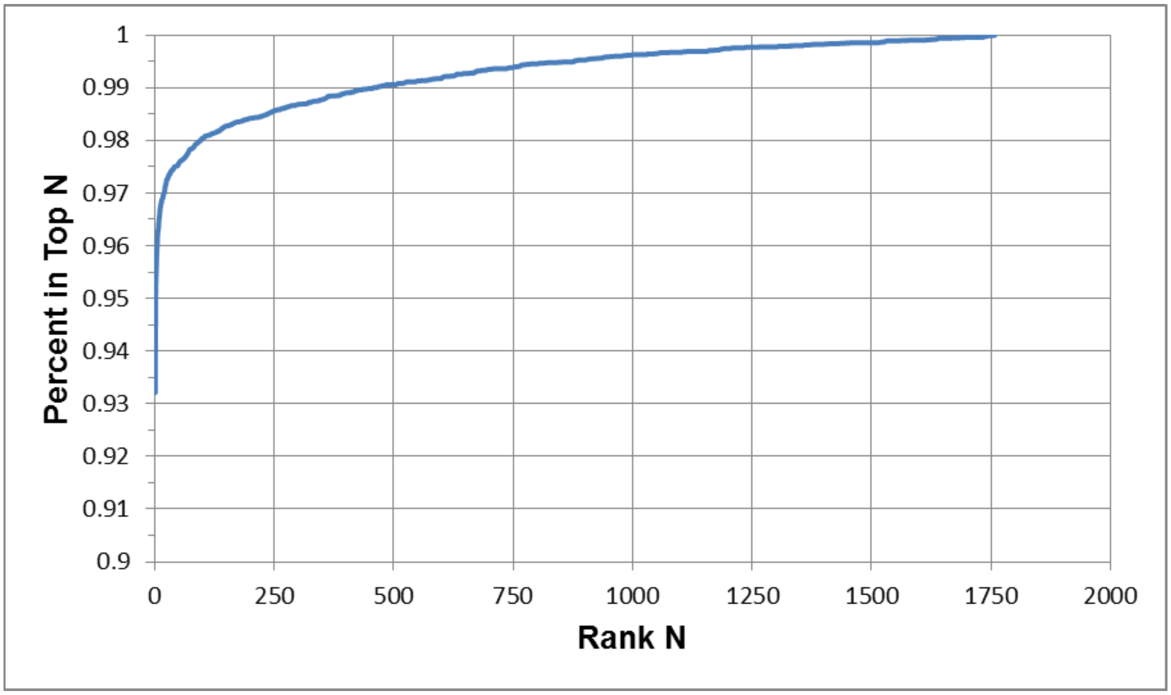


Figure II-B–10: Native CMU on CS3 Dataset with CMC Curve

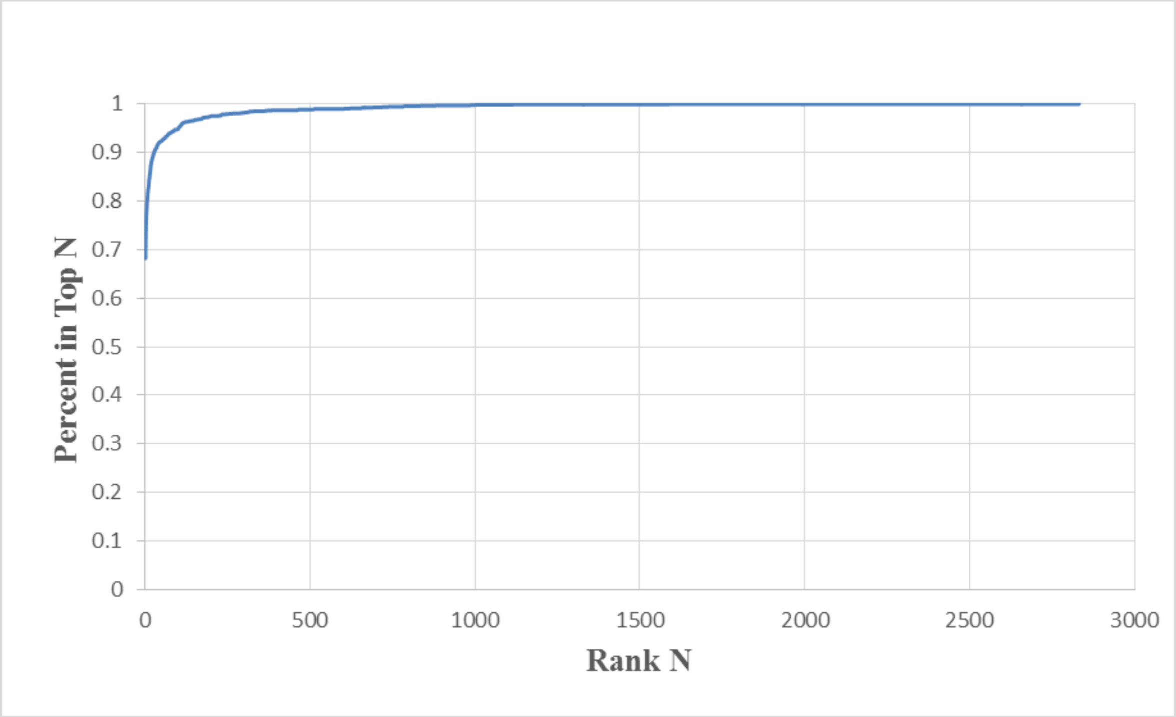


Figure II-B–11: OpenFace on LFW Dataset with CMC Curve

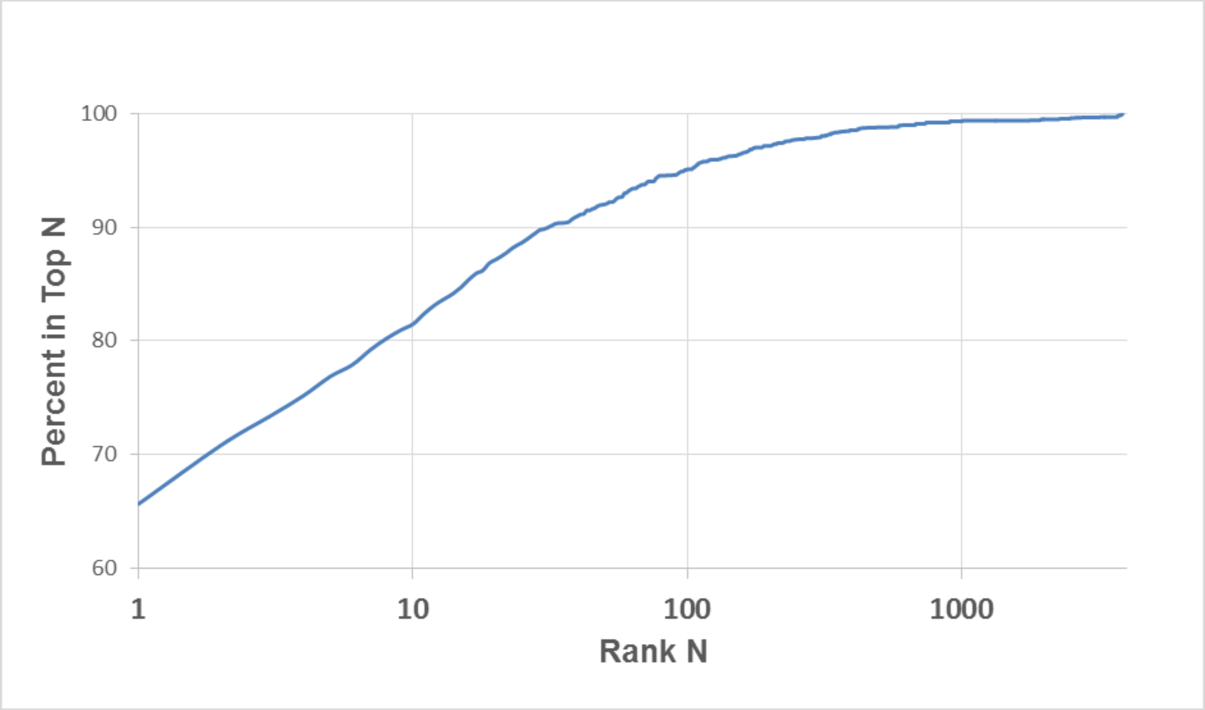


Figure II-B–12: OpenFace on LFW Dataset with CMC Curve

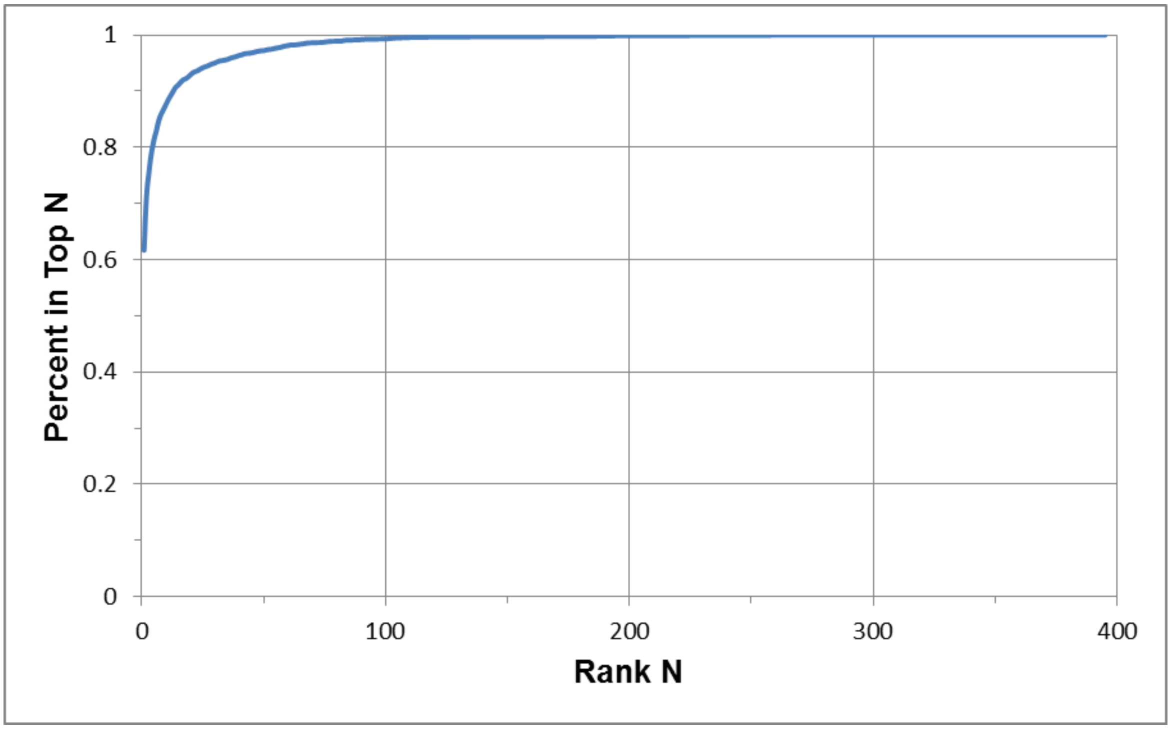


Figure II-B–13: OpenFace on GBU Combined Dataset with CMC Curve

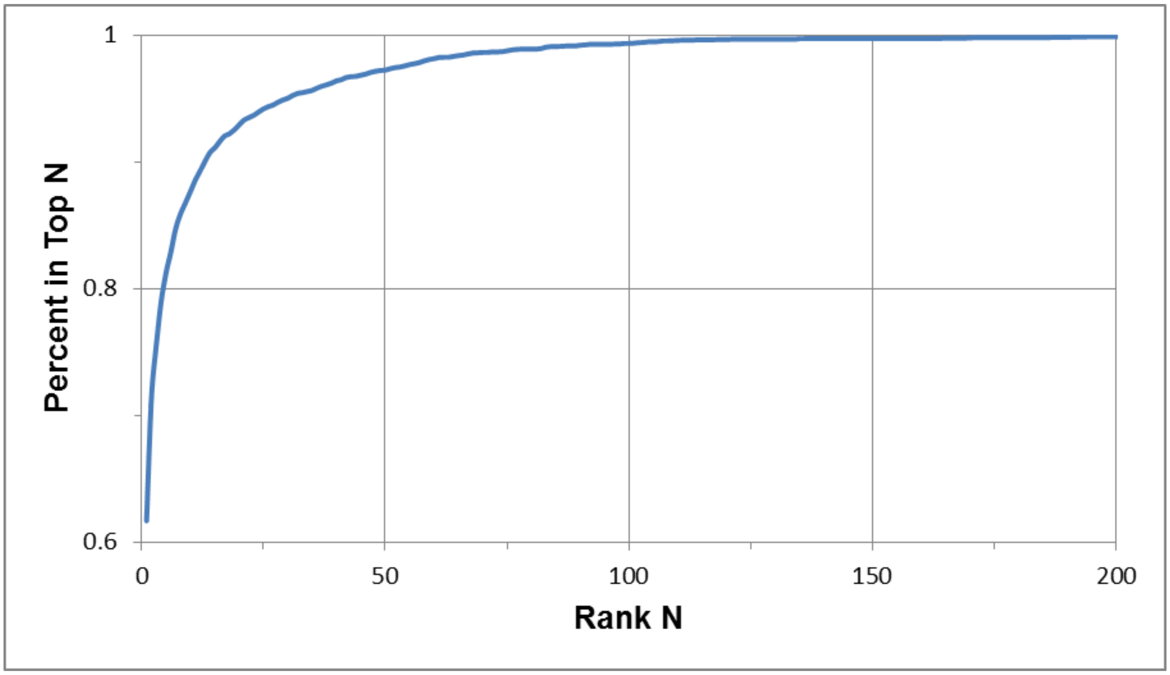


Figure II-B–14: OpenFace on GBU Combined Dataset with CMC Curve

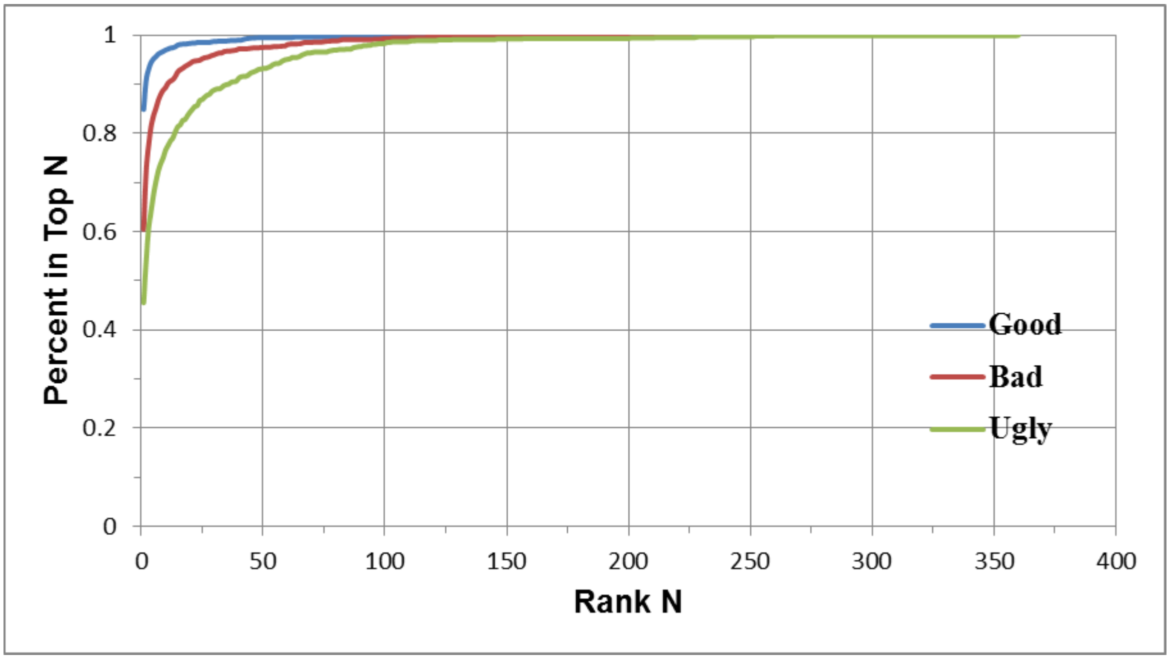


Figure II-B–15: OpenFace on GBU Partitions Dataset with CMC Curve

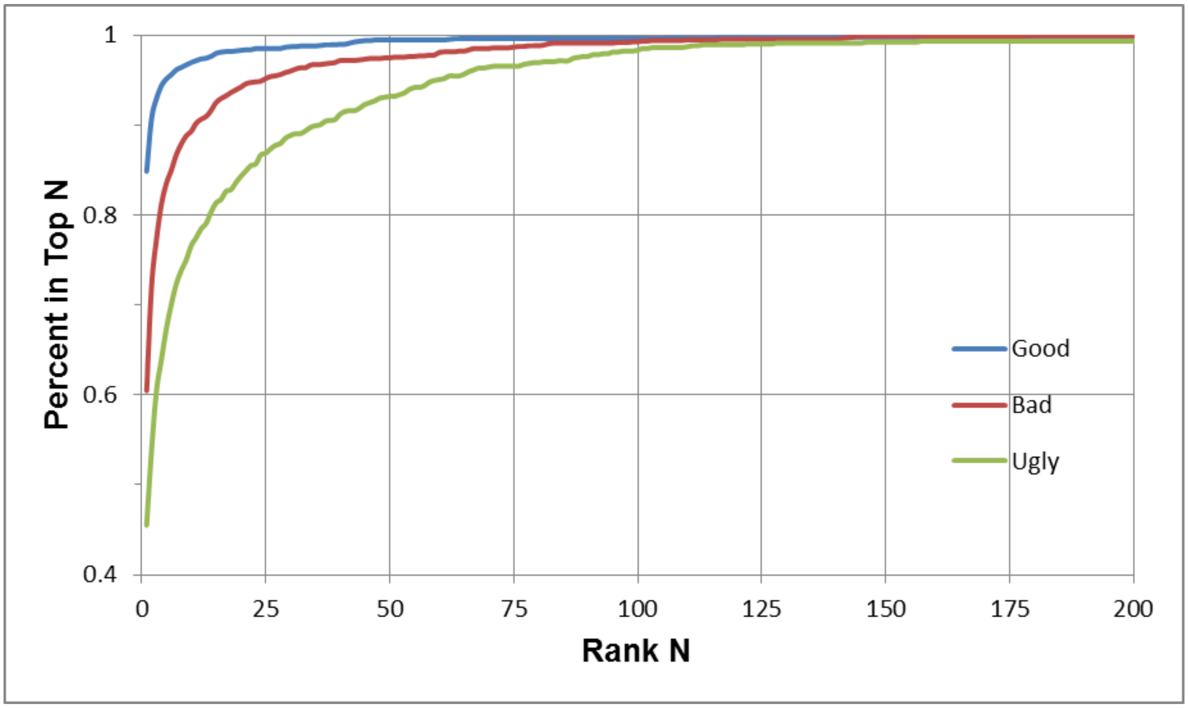


Figure II-B–16: OpenFace on GBU Partitions Dataset with CMC Curve

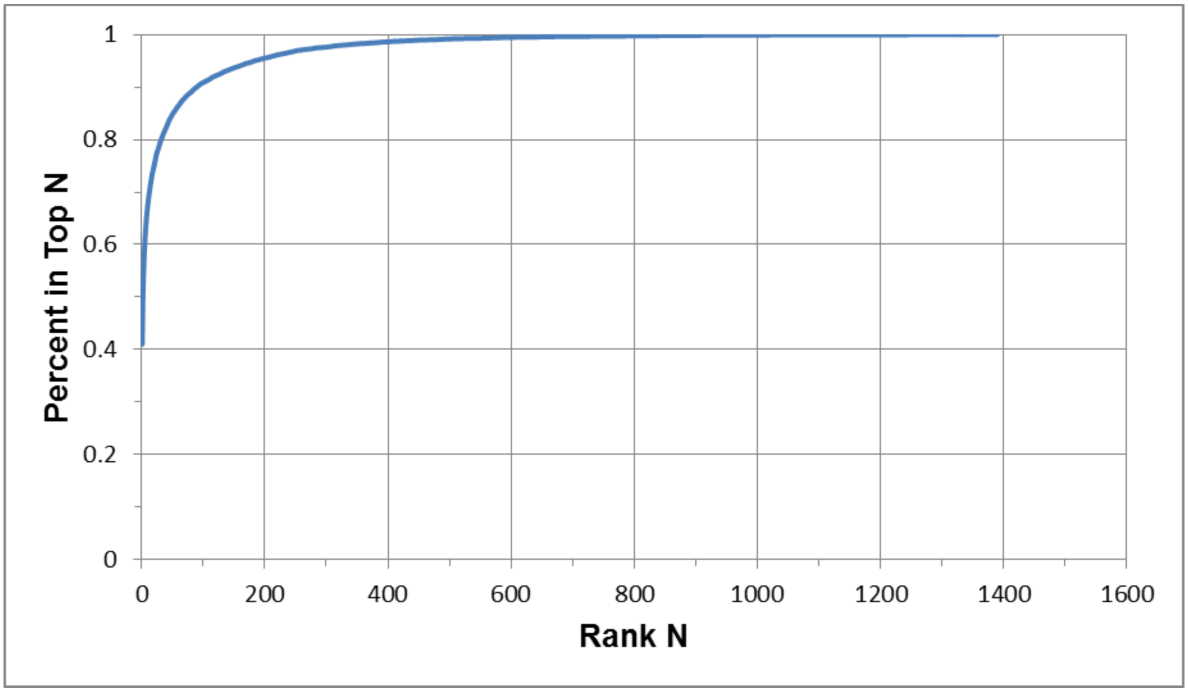


Figure II-B–17: OpenFace on CS3 Dataset with CMC Curve

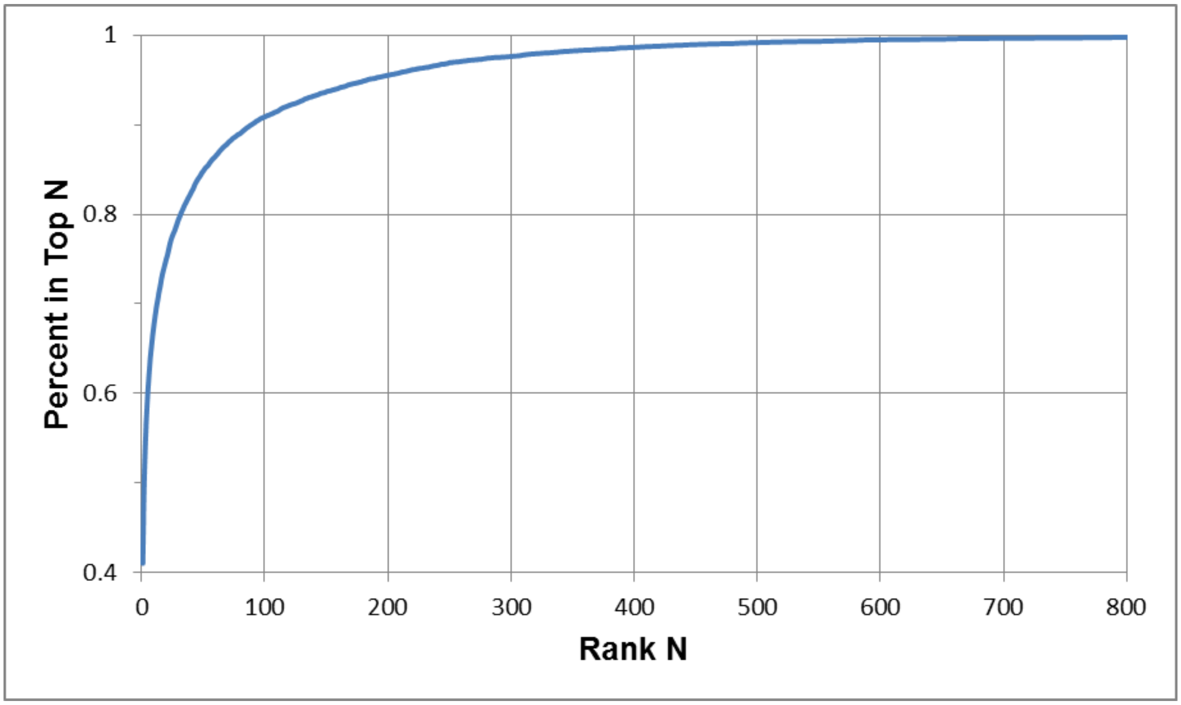


Figure II-B–18: OpenFace on CS3 Dataset with CMC Curve

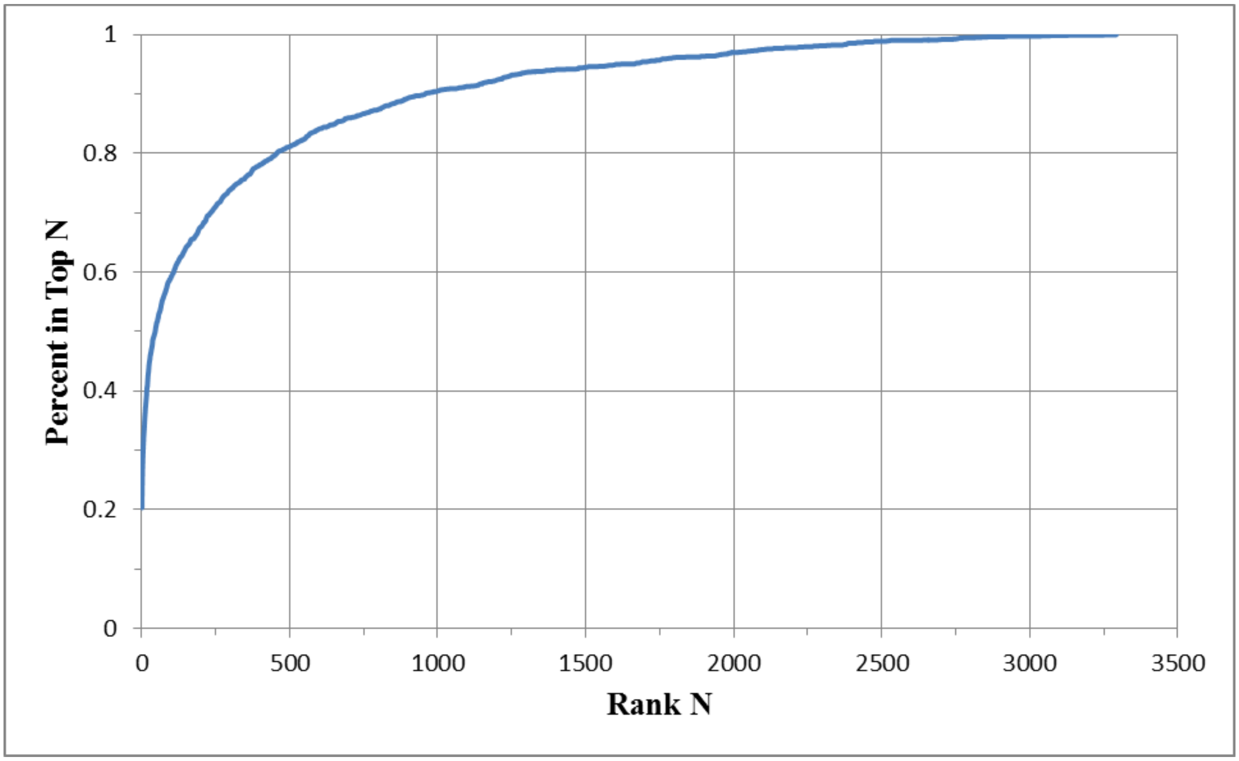


Figure II-B–19: PittPatt on LFW Dataset with CMC Curve

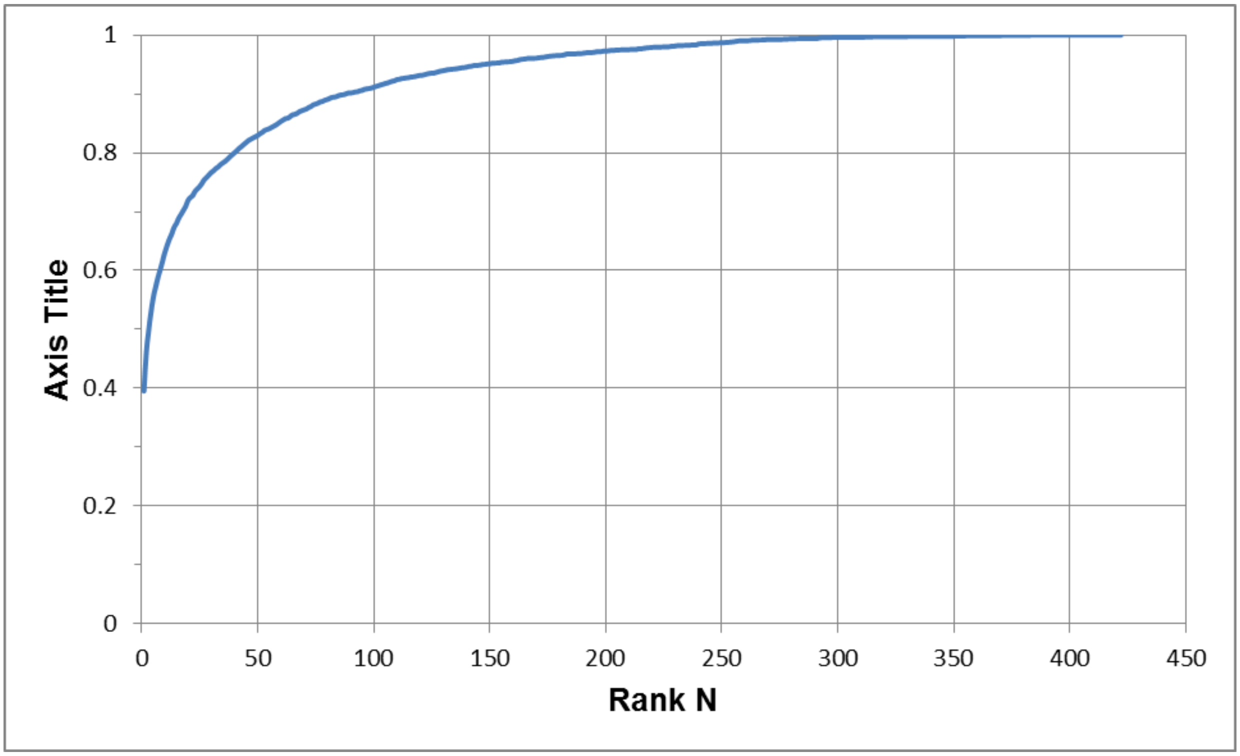


Figure II-B–20: PittPatt on GBU Combined Dataset with CMC Curve

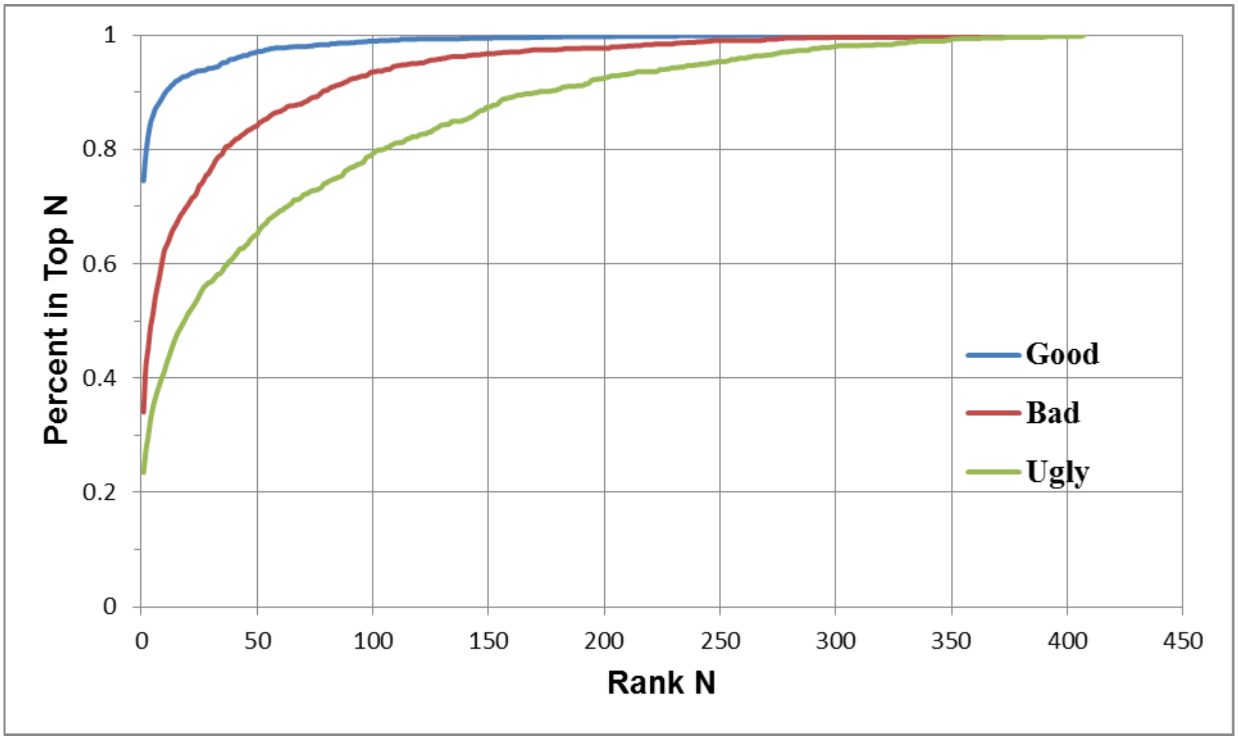


Figure II-B–21: PittPatt on GBU Partitions Dataset with CMC Curve

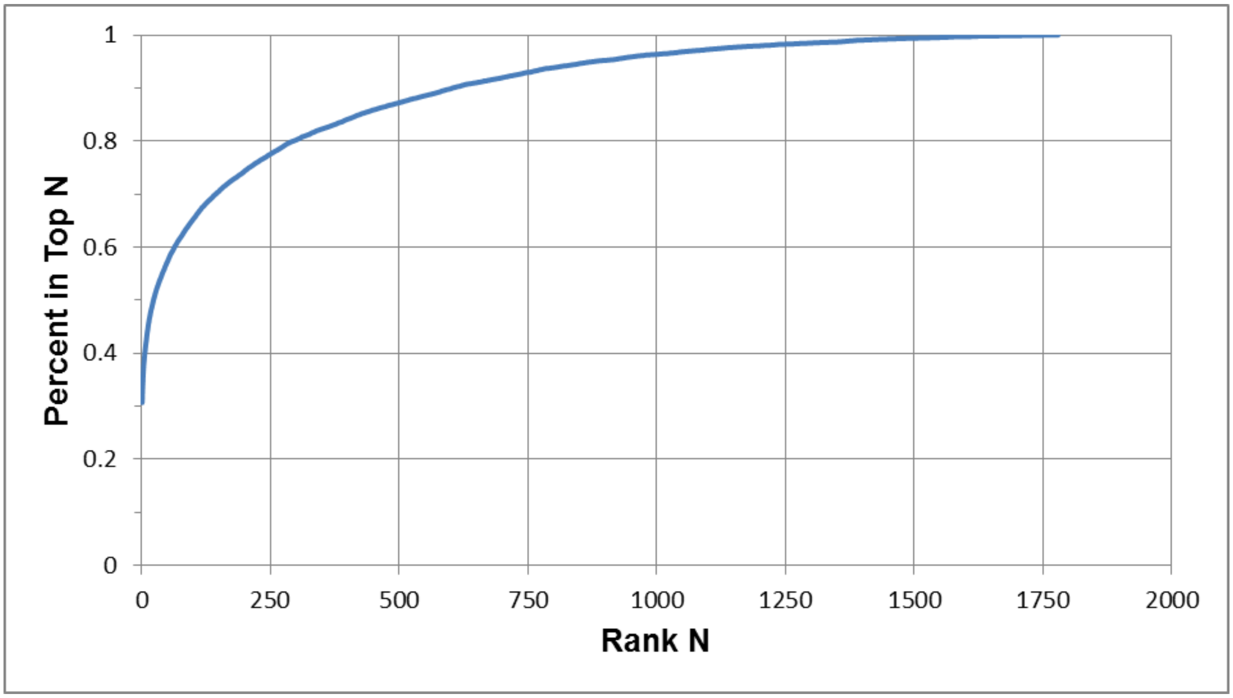


Figure II-B–22: PittPatt on CS3 Dataset with CMC Curve

JOHNS HOPKINS
APPLIED PHYSICS LABORATORY

11100 Johns Hopkins Road • Laurel, Maryland 20723-6099

AOS-18-1041

NIJ RT&E Center Project 15-FR

November 2018

NIJ FACE ALGORITHM ASSESSMENT PHASE III – PERIOCULAR FACE RECONSTRUCTION

Version 1.0

Authors: Richard L. (DJ) Waddell, et al.

Prepared for:

NIJ | National Institute of Justice

STRENGTHEN SCIENCE. ADVANCE JUSTICE.

Prepared by:

The National Criminal Justice Technology Research, Test, and Evaluation Center
The Johns Hopkins University Applied Physics Laboratory
11100 Johns Hopkins Rd.
Laurel, MD 20723-6099

Task No.: FGSGJ
Contract No.: 2013-MU-CX-K111/115912

This project was supported by Award No. 2013-MU-CX-K111, awarded by the National Institute of Justice, Office of Justice Programs, U.S. Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication/program/exhibition are those of the author(s) and do not necessarily reflect those of the Department of Justice.

PHASE III-CONTENTS

PHASE III-FIGURES

PHASE III-TABLES

PHASE III-EXECUTIVE SUMMARY

The National Criminal Justice Technology Research, Test & Evaluation Center (RT&E Center), operated by the Johns Hopkins University Applied Physics Laboratory, conducted an assessment of the periocular-based face reconstruction software developed by the Carnegie Mellon University. The software, Dimensionally Weighted Kernel Singular Value Decomposition (DWK-SVD), is based on a linear transformation that maps an image of a periocular region to an image of a full face (reconstructed image); the linear transformation is a solution to the optimization problem framed by CMU that relates the periocular region to the full face. For comparison, CMU also provided a solution based on Principal Component Analysis (PCA), and, as a demonstration, both solutions were applied to an established dataset. No other solutions were compared because no other relevant open source algorithms were identified. Analogous to Phase I & II assessments, the Phase III assessment computed metrics for both solutions and drew comparisons to assess relative performance of each. Additionally, the assessment used two different face matchers, Kernel Class-dependence Feature Analysis (KCFA) and PittPatt [1], to demonstrate the quality of the reconstructed images. The comparison is based on a suite of metrics calculated for a subset of the NIST-sponsored Face Recognition Grand Challenge dataset and a subset of the Cropped Yale dataset.

Standard metrics used to compare periocular-based face reconstruction algorithm performance include the receiver operating characteristic (ROC) curve and the peak signal-to-noise ratio values. The ROC curves are further characterized by the area under the ROC curve and the equal error rate point.

The RT&E Center results are identical in four of the six cases generated by CMU. The two cases with different results were investigated; however, the differences could not be resolved. The results also show that DWK-SVD approach outperforms that of PCA in reconstructing a face based on just its periocular region.

III-1 INTRODUCTION

In September of 2013, the Johns Hopkins University Applied Physics Laboratory (JHU/APL) was selected by the U.S. Department of Justice, National Institute of Justice (NIJ) to establish the National Criminal Justice Research, Test, and Evaluation Center (RT&E Center) within the National Law Enforcement and Corrections Technology Center System. The RT&E Center has been tasked to perform an assessment of the Carnegie Mellon University facial image processing algorithms using the technical approach documented in the RT&E Center's Proposed Technical Approach for Project 15-FR [1].

This effort includes three phases. The first phase assessed the Carnegie Mellon University software capability for face detection and the results are documented in the Phase I part of this report. Similarly, the face recognition capabilities of Carnegie Mellon University software are discussed in the Phase II portion. This portion documents the software capabilities for a technique called periocular reconstruction – a technique where an image of the periocular region of a face is linearly transformed to a "reconstructed" image of a full face.

III-1.1 Role of Periocular Reconstruction

"Periocular region-based human identification offers advantages over full face biometrics as it is least affected by expression variations, aging effects … and the changes due to growth of male facial hair. Moreover, full face recognition performance degrades in the presence of pose variations whereas the periocular region-based identification may perform better in the case of extreme pose changes when only one eye is completely visible" [20]. See Figure III–1 for examples.

Carnegie Mellon University developed a tool to reconstruct the whole face from an image of the periocular region. In this way, various face matchers can be used to perform matching on previously unusable faces. The experiments in this assessment aim to show that the reconstruction method generates a full-face image that is a close approximation to the true full-face image.

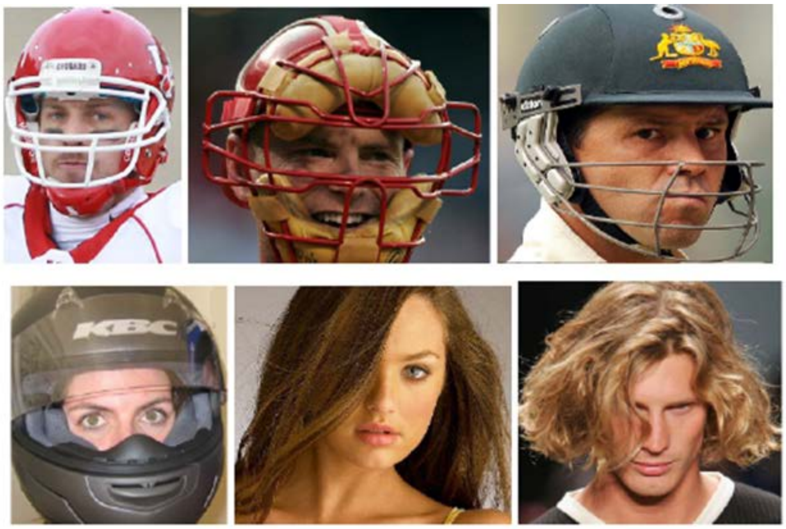


Figure III–1: Example Applications in Which Periocular Biometrics May Be More Effective Than Full-face Biometrics [20]

III-1.2 Goals of Assessment

The primary goal of this assessment was to replicate the Carnegie Mellon University results documented in [21]. Those results show relative performance of the Dimensionally Weighted Kernel Singular Value Decomposition (DWK-SVD) approach for reconstructing full-face images from periocular regions in comparison with a benchmark (PCA-based) approach, in particular, the results demonstrate the DWK-SVD software, in combination with a face matcher, enables matching periocular images with the corresponding full-face image. The experiments in this assessment quantify the quality of the reconstructed images to show that the software generates a reconstructed full-face image that is faithful to the true full-face image and that the reconstructed images can be used with existing face matchers.

A secondary goal was to demonstrate similar performance trends for DWK-SVD and PCA on an additional dataset.

III-2 BACKGROUND

A biometric system comprises a physical or behavioral trait of a person through which he or she can be recognized uniquely. The first two phases of this assessment focused on face detection and face recognition and these areas continue to improve; however, recognition/identification based on a full face is susceptible to weak performance since an individual face changes with age, facial hair and expression. The periocular region of the face is least susceptible to these changes; consequently, the academic community is working to identify a person based on the periocular region. Periocular (peripheral area of ocular) region refers to the immediate vicinity of the eye, including eyebrow and lower eye fold as depicted in Figure III–2. Classification and recognition through the periocular region show significant accuracy [22]; Figure III–3 illustrates a conceptual model of a biometric system that employs the periocular region as a trait for recognition.

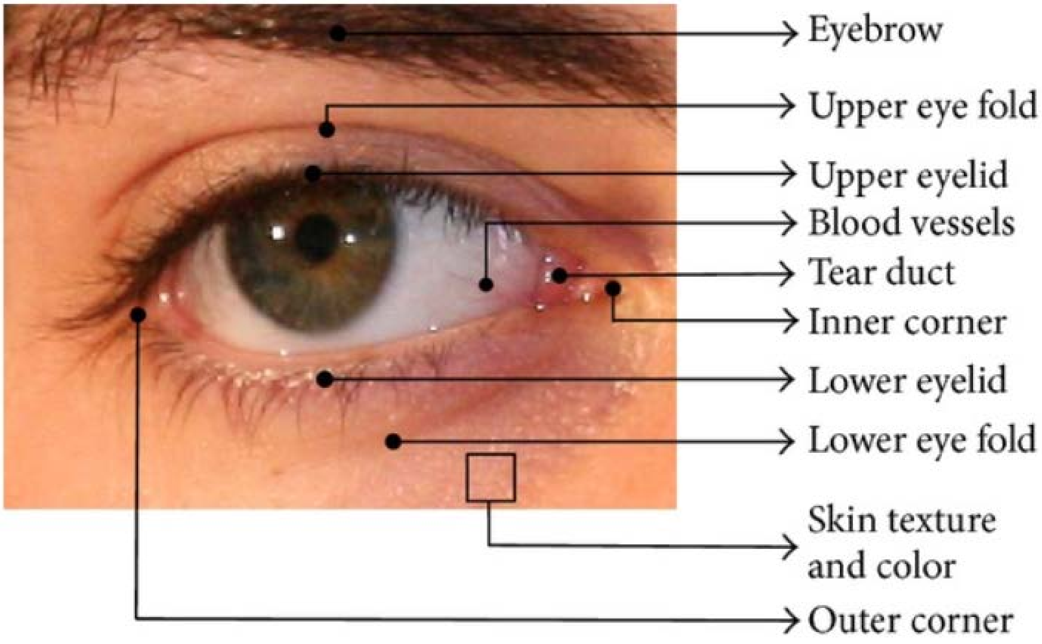


Figure III–2: Important Features from a Periocular Image [22]

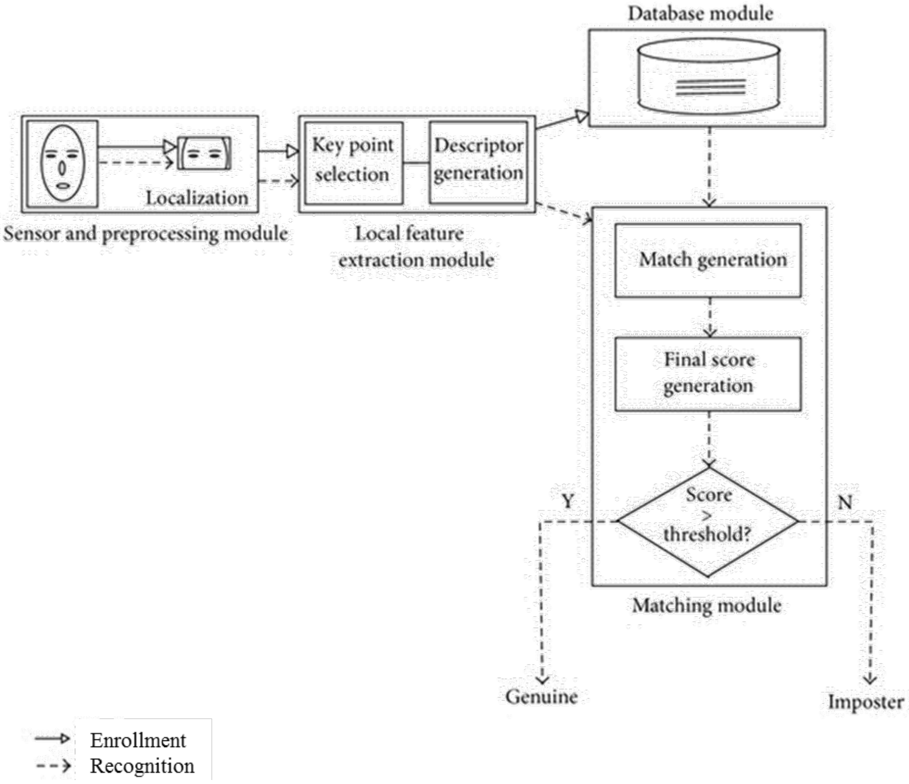


Figure III–3: Model of Periocular Biometric System [22]

III-3 PERIOCULAR RECONSTRUCTION ALGORITHMS

III-3.1 DWK-SVD and PCA Periocular Reconstruction Algorithms

The CMU approach to create a face representation and subsequent reconstruction based on the periocular region of the face is comprised of two components: a sparse set of features selected to represent the periocular region or full face; and a sparse array of coefficients (dictionary) that represent all faces. The central idea is to create the dictionary entries with the periocular related features over-weighted compared to the entries for the rest of the face. The subsequent reconstruction step uses the sparse representation of the periocular region and the full-face dictionary to retrieve the full-face image. See Appendix III-A for the mathematical formulation taken from the references ([21], [23]).

As described in [23], the fundamental problem is to define an optimization framework so the spatial relationships between the periocular features and the full-facial features are maintained, and the reconstruction error for the periocular region is penalized more than for the full face because the reconstruction should not, ideally, alter the original periocular region.

“Keeping in mind the issues related to the dictionary learning, we arrive at the problem of jointly optimizing the learning procedure for the two goals. The first is to learn a dictionary of whole faces so as to include prior knowledge about the spatial relationships between the facial features and the periocular features. The second is to obtain a dictionary in which the reconstruction error for the periocular region is penalized more than the entire face and both are jointly minimized for the same sparse coefficients” [23]. This is accomplished by sharing the approximation coefficients (dictionary entries) for the periocular region and the full face and weighting the dimensions corresponding to the periocular region more than the remaining dimensions.

At the direction of NIJ, the RT&E Center acquired from Carnegie Mellon University the software suite comprising two linear transformations, the dictionaries, (DWK-SVD, PCA), a toolbox to create sparse sets of coefficients to represent an image of a periocular region for both transformations, and two face matchers (KCFA, PittPatt). Note that the DWK-SVD mapping was trained on half a million mug shots – not Face Recognition Grand Challenge (FRGC) data. For this assessment, the university provided the cropped and resized image set from the FRGC; and, the RT&E Center generated a second set of images from the Cropped Yale dataset. The RT&E Center used both the DWK-SVD and PCA-based transformation software to calculate the probe image (i.e., reconstructed image) and used both the KCFA and PittPatt face matchers to calculate the similarity scores between the reconstructed images against all images in the cropped full-face images from the FRGC dataset. The reconstructed images were further compared using the peak signal-to-noise ratios (PSNR) metric. The workflows from dataset creations through metrics calculation for the KCFA and PittPatt matcher paths are shown in Appendix III-B.

Periocular reconstruction is characterized by the fidelity in hallucinating the original face from the reconstructed face. The university's process for periocular reconstruction has two steps: project the periocular region of the image to a lower dimension representation; and map the lower dimensional representation of the periocular region to a representation of the entire face (reconstruction). The fidelity of the process is gauged by the performance of a face matcher comparing the reconstructed image with a gallery of images containing the original. ROC curve analysis is used to gauge performance. In particular, the area under the curve (AUC) and equal error rate (EER) values of each ROC curve create a means to numerically rank performance.

In this assessment, two approaches were used to reconstruct a face from the periocular region. The CMU approach projects the image to a sparse representation followed by the mapping step using the DWK-SVD derived dictionary of representations of full-facial images. Finally, the representation of the full face is matched to a gallery of images using the KCFA and PittPatt face matcher algorithms.

The second approach (the benchmark) projects the array of pixels to a higher dimension and extracts the top 40 eigenvectors of the array. A subsequent mapping step using the PCA-derived dictionary of representations generates the reconstructed image. Finally, the representation of the full face is matched to a gallery of images using the KCFA and PittPatt face matcher algorithms.

The reconstructed images were compared to all the original images using both face matchers (PittPatt, KCFA) for the FRGC dataset; however only the PittPatt matcher was used for the Cropped Yale dataset. KCFA was not used because there were not enough images in Cropped Yale to create a training subset for KCFA matched filter training and a test set for cross validation with those trained filters. The comparisons were done using three sets of images: the original images (no reconstruction), reconstructed images using DWK-SVD and the reconstructed images using PCA. Each comparison process generates a set of TP, FP, TN and False negative (FN) values needed to generate a ROC curve. In all cases for the FRGC dataset, the reconstructed image PittPatt feature vectors were successfully generated. However, for the Cropped Yale dataset, many face images produced no face detections for subsequent use by the PittPatt face matcher. The PittPatt face matcher failed (threw an error flag) for the offending face images. These images were not counted as a positive or negative sample, but simply excluded from ROC analysis. Consequently, the subsets of the Cropped Yale dataset used for ROC analysis differed for each of the three sets of images described above (see Table III–1).

III-3.2 KCFA Matcher Algorithm

Face recognition is often done by comparing feature vectors calculated in the image space for the unknown and gallery images and calculating similarity based on those features. This is the approach of the software assessed in Phase II. Kernel based Class-Dependence Feature Analysis (KCFA) is based on feature vectors in the Discrete Fourier Transform space. Each identity to be recognized has a corresponding correlation filter (considered as an analogue to a dictionary element in the DWK-SVD methodology). The output feature vector for an unknown image is calculated as the inner product of the correlation filters and the unknown image. Then classification reduces to finding the best similarity between the output feature vector and a gallery image using a matching algorithm. The KCFA matcher was used to measure the similarity of the DWK-SVD reconstructed face to the dictionary entries [24].

III-3.3 PittPatt Matcher Algorithm

The RT&E Center acquired the PittPatt binary from contacts at Carnegie Mellon University at the direction of NIJ. This software does not have documentation but a limited example application was provided that identified two processing steps for face recognition: create galleries and compare galleries. The compare galleries capability was used to match the probe image representation to the gallery of original images. In summary, PittPatt uses two sets of photographs: a set of images to be recognized (unknowns or probes) and a gallery of known or target images. The images are preprocessed using a Matlab script from the university, which resizes them from 32 x 32 pixels to 340 x 340 pixels and centers them on a 700 x 700 pixel black background to maximize probability of face detection. The matching process consists of drawing an image from the set of probe images, calculating the reconstructed face from only the periocular region using DWK-SVD, calculating similarity (distance metric) with target gallery images, and saving similarity scores as a row in a comparison matrix. The process repeats this methodology for each unknown image.

After all comparisons have been made, the ROC Curve (Sec II-4.2.2) is calculated using the contents of the comparison matrix; the contents of the matrix are processed by the CMU's ROC class to generate the ROC curves using the maximum and minimum score values to fix the threshold range.

III-4 ASSESSMENT COMPONENTS

III-4.1 FRGC Dataset

The RT&E Center's assessment of the DWK-SCD periocular reconstruction software used a common dataset to Carnegie Mellon University's own internal assessment of DWK-SVD, i.e., a subset of the Face Recognition Grand Challenge dataset ver. 2.0 (FRGC). In addition, a subset of the Cropped Yale dataset was tested. The FRGC dataset was designed to provide a capability to support an order of magnitude improvement, i.e., a 98% True Positive Rate (TPR) (2% error rate) and a False Positive Rate (FPR) of 0.1%, in performance from the state of the art in 2005 [25]. The verification rate provided in [25] is the same calculation as the TPR used in the ROC analysis. The purpose of the FRGC dataset is to facilitate the development of new algorithms that use the additional information inherent in high-resolution images. Earlier image sets had 40 to 60 pixels between the centers of the eyes; high-resolution images have 250 pixels between the centers of the eyes on average.

The data was collected over one academic year on a weekly session basis in order to introduce variability for individual subjects; each session included seven images: four still images under controlled lighting, two still images under uncontrolled lighting, and one 3D image. The validation dataset is comprised of images from 466 subjects collected in 4,007 subject sessions.

Summary notes include [25]:

- The controlled images were taken in a studio setting, are full frontal facial images taken under two lighting conditions (two or three studio lights) and with two facial expressions (smiling and neutral);
- The uncontrolled images were taken in varying illumination conditions and also have two facial expressions (smiling and neutral); and
- The 3D images were also taken under controlled illumination conditions.

The FRGC ver. 2.0 dataset "has three components, the first is the generic training set that contains both controlled and uncontrolled images of 222 subjects, and a total of 12,776 images. Second, the target set containing 466 different subjects with a total of 16,028 images. Lastly, the probe set containing the same 466 subjects as in the target set, with half as many images for each person as in the target set, bringing the total number of probe images to 8,014 (Note: the probe image set was not used in this assessment). Image examples from the FRGC database are shown in Figure III–4. The target set contains images of the highest quality and are under the most controlled environment. The FRGC creators recommend testing the periocular-based full face reconstruction algorithm on the target set of the FRGC dataset; examples are shown on the left portion of Figure III–4. The FRGC creators decided it is optional to test on the less controlled probe set of the FRGC dataset which is shown on the right portion of Figure III–4" [21].

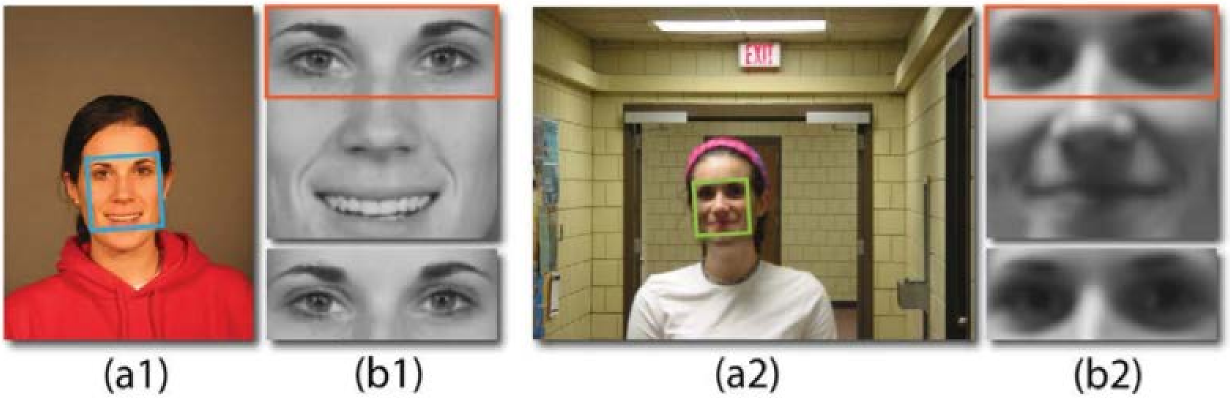


Figure III–4: Example Image from the FRGC Dataset: (a1,a2) Controlled and Uncontrolled Still of the Same Subject, (b1,b2) Cropped Full Face and Periocular Region

The FRGC data was preprocessed to provide frontal face images by Carnegie Mellon. The eye locations, provided as ground truth data, of the subject are used to anchor the face orientation. The images are aligned, cropped and resized to 32 x 32 pixels. The top portion of size 13 x 32 is considered the periocular region of the subject. This periocular region along with the visibility mask specified by the end user is then fed into the reconstruction method to generate the full 32 x 32 image from the periocular region [21].

The FRGC distribution defined multiple experiments; Carnegie Mellon adhered to the FRGC Experiment 1 protocol, which involves 1-to-1 matching of the 16,028 controlled target images (original full faces) to the reconstructed counterparts based on the periocular regions (~256 million pairwise face match comparisons). For this experiment, the RT&E Center identified the normalized cosine distance (NCD) to compute the similarities between images.

$$d(\mathbf{x}, \mathbf{y}) = 1 - \frac{\mathbf{x} \cdot \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

Figure III–5: Normalized Cosine Distance Equation

The result of the comparison is a similarity matrix with the size of 16,028 x 16,028, whose $(i,\mathrm{j})\mathrm{t}h$ entry is the NCD between the feature vector of query image i and gallery image j. In the case of FRGC Experiment 1, the query set and gallery set are the same. The performance is analyzed using verification rate (or TPR) at 1% (0.01) FPR, EER and the receiver operating characteristic (ROC) curves.

III-4.2 Cropped Yale Dataset

In addition to the FRGC dataset, the RT&E Center evaluated the face reconstruction algorithm on a subset of the cropped version of the Extended Yale Face Database B [26]. For brevity, this dataset is referred to as the "Cropped Yale" dataset throughout this assessment. The dataset is comprised of 64 images (with varying illumination conditions) for each of 38 subjects with a single pose (full frontal since that is what FRGC uses). The illumination conditions provided light from 64 locations over a short time (2 sec) to enable a stable facial image while filming. Additionally, there is an ambient light image for each identity.

Example images are shown in Figure III–6. The acquired images are 8-bit (gray scale). The dataset curators manually aligned, cropped, and then re-sized all test image data used in the dataset to 168 x 192 pixels per image. The RT&E Center cropped the bottom 24 rows of pixels of each 192 x 168 image to make 168 x 168 pixels per image and then resampled the images to 32 x 32 pixels.

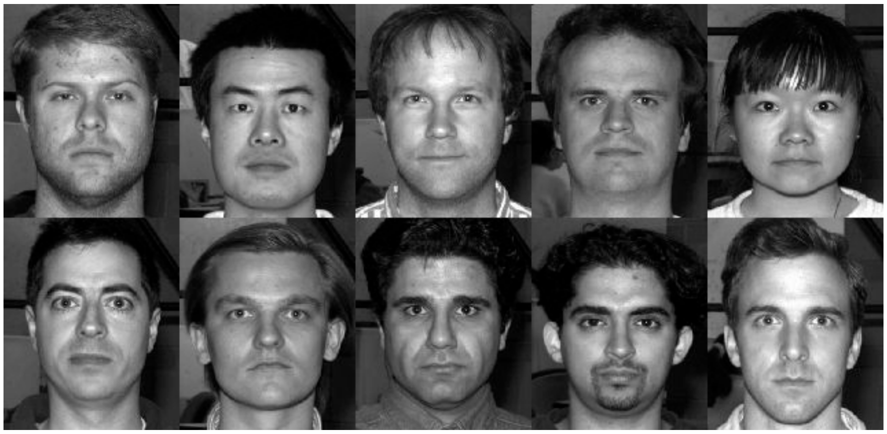


Figure III–6: Example Images from Cropped Yale Dataset

The curators of the dataset omitted a select few images that were corrupted during data acquisition. As a result, 2,452 images, instead of 2,470, comprise the dataset. Additionally, many of the illumination conditions inhibited face detection, so only a subset of the Cropped Yale dataset images was actually used to generate the corresponding results in Figure III–16. See Table III–3 for exact image counts for each comparison and the corresponding percentages of the original dataset used for ROC analysis. By contrast, there were no missed face detections with the FRGC dataset. Figure III–7 and Figure III–8 contrast reasonable versus difficult illumination conditions and the corresponding ramifications in reconstructed images from the periocular region of the original image.

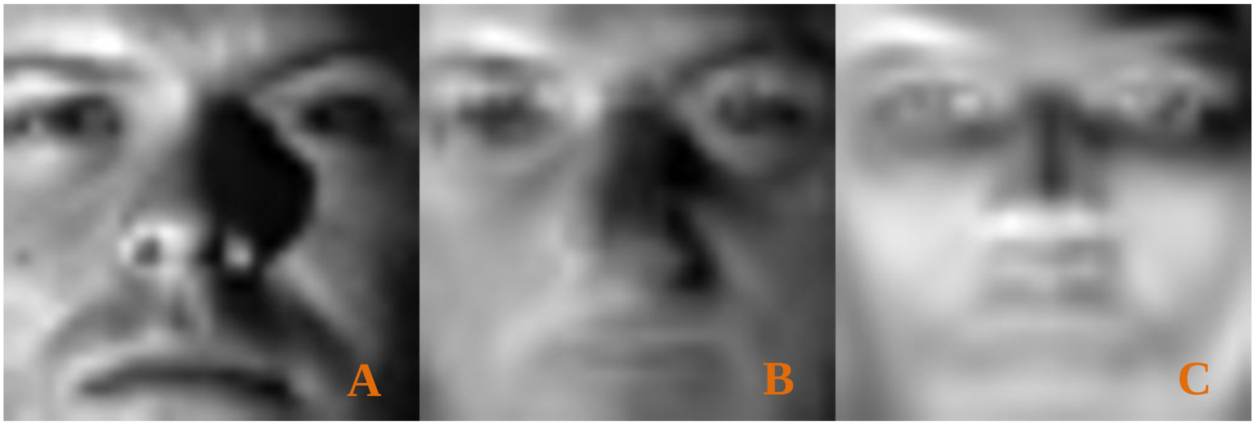


Figure III–7: Reasonable Illumination Condition for (A) Original Cropped Yale Dataset Image, (B) DWK-SVD Reconstructed Image, and (C) PCA Reconstructed Image

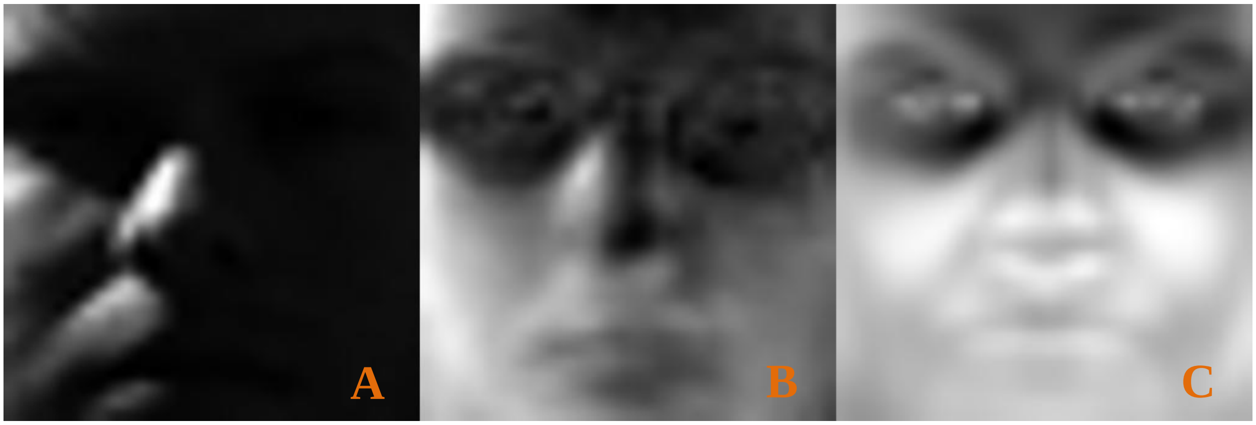


Figure III–8: Difficult Illumination Condition for (A) Original Cropped Yale Dataset Image, (B) DWK-SVD Reconstructed Image, and (C) PCA Reconstructed Image

Table III–1: Subsets of Cropped Yale Dataset Used by Various Reconstruction Algorithms

Methods applied on Cropped Yale dataset	Number of images used for ROC analysis	Percentage of total dataset presented to PittPatt Matcher	Resulting number of comparisons for ROC curve points
PittPatt Original vs Original	1020	41.6%	1,040,400
PittPatt Original vs DWK-SVD	873	35.6%	762,129
PittPatt Original vs PCA	716	29.2%	512,656

III-4.3 Metrics

The RT&E assessment measured periocular reconstruction fidelity and associated face verification using the same metrics as CMU: PSNR and ROC analysis.

For the periocular reconstruction fidelity component, the original and reconstructed faces were used to compute the mean and standard deviation of the PSNR between the original image and the reconstructed image in each pair.

For the face verification component, the RT&E Center used three different matching approaches; each approach was executed with two different matching algorithms (KCFA, PittPatt). The first is to exhaustively compute the similarity measures between all original images; the second is to compute similarity between all face images reconstructed using the DWK-SVD capability and the original face images; and the third is to compute the similarities of the face images reconstructed using the PCA capability and the original images. The matching results from each similarity calculation are used to generate an ROC curve with corresponding EER value.

III-4.3.1 Peak Signal-to-Noise Ratio

After periocular reconstruction, the PSNR, an indication of reconstructed face fidelity, between the full-face reconstruction and original face images, was calculated between each original face image I and each corresponding reconstructed face image I' as follows:

$$\mathrm{PSNR} = 10\log_{10}\left(\frac{255^2}{\mathrm{MSE}}\right) = 10\log_{10}\left(\frac{255^2}{\frac{1}{mn}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - I'(i,j)]^2}\right)$$

Figure III–9: Peak SNR Equation [21]

III-4.3.2 Receiver Operating Characteristic (ROC) Curve

A ROC curve for a binary classifier plots the TPR versus the FPR of the classifications; it is used as a measure of matching verification. These curves were used in the Phase II section of this report to characterize the performance of algorithms in identifying faces. In this application, they are used to characterize performance of algorithms to match the reconstructed face to the original face.

The curves were calculated using a Matlab routine provided by the university. In summary, a feature-based representation of each of the two faces to be compared is calculated, and a cosine-based distance metric is used to quantify the similarity (i.e., distance) between the two representations. A threshold value is applied to the distance to determine matches (TPs and FPs) and non-matches (TNs and FNs); and, varying the threshold produces the set of FPR and TPR that comprise the ROC curve.

The ROC curve is characterized with two parameters: AUC and the point of EER:

- AUC – The ROC curve points lie in the range [0.1] for both FPR and TPR, and AUC values in the proximity of 1.0 are desired.
- EER – The EER is associated with ROC curves, in particular it identifies the point on the ROC curve where the FPR (aka false accept rate) equals the false reject rate (= 1 – true positive rate) as shown in Figure III–10; alternately where FPR + TPR = 1. EER points with lower FPRs are better. Because of the limited amount of data, when the term (FPR + TPR – 1) <= threshold, analysts found that thresholds of 0.001 and 0.01 did not yield an EER point; however, a threshold value of 0.1 did.

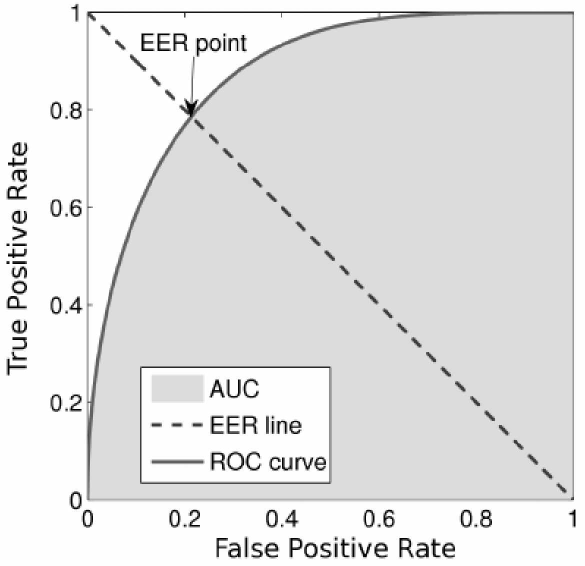


Figure III–10: Illustration of EER Point for ROC Curve

III-4.4 Results

The primary goal of this assessment is to replicate the results reported by CMU for the FRGC dataset, in particular the reconstruction fidelity results as measured by the PSNR metric and the true positive rate (TPR) measured as ROC curves as reported in [21].

Figure III–11: Distributions for DWK-SVD and PCA Face Representation Methods with PSNR Metric (FRGC Dataset)

The distribution of PSNR values for both the PCA baseline and DWK-SVD representation approaches are provided in Figure III–11 for the FRGC dataset. The histograms show both a "tighter" distribution and larger values for DWK-SVD than PCA. Based on a visual inspection (corresponding data was unavailable), these results compare favorably with the analogous results reported in Table 1 of [21].

The corresponding ROC curves for the FRGC dataset are presented in Figures III–12 through III–15 for both the PittPatt and the KCFA face matchers; and the ROC curves for the Cropped Yale dataset using the PittPatt face matcher are presented in Figure III–16. The KCFA matcher was not used for the Cropped Yale dataset because there were not enough images to train the matched filters. The associated AUC values for each curve are provided in the annotation. Each set of ROC curves is presented on the customary linear-linear plot and a log-linear plot (when appropriate) to provide an expanded view in the vicinity of FPR equal zero.

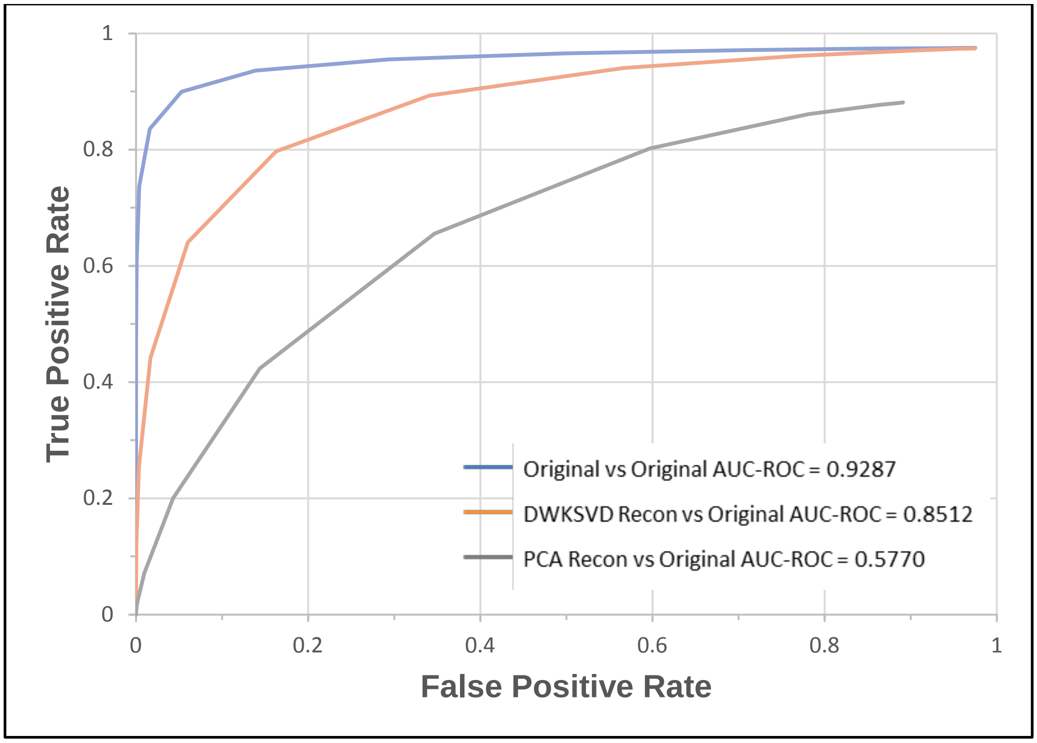


Figure III–12: FRGC Reconstruction and Matching, PittPatt Face Matcher, with ROC Curve (linear-linear)

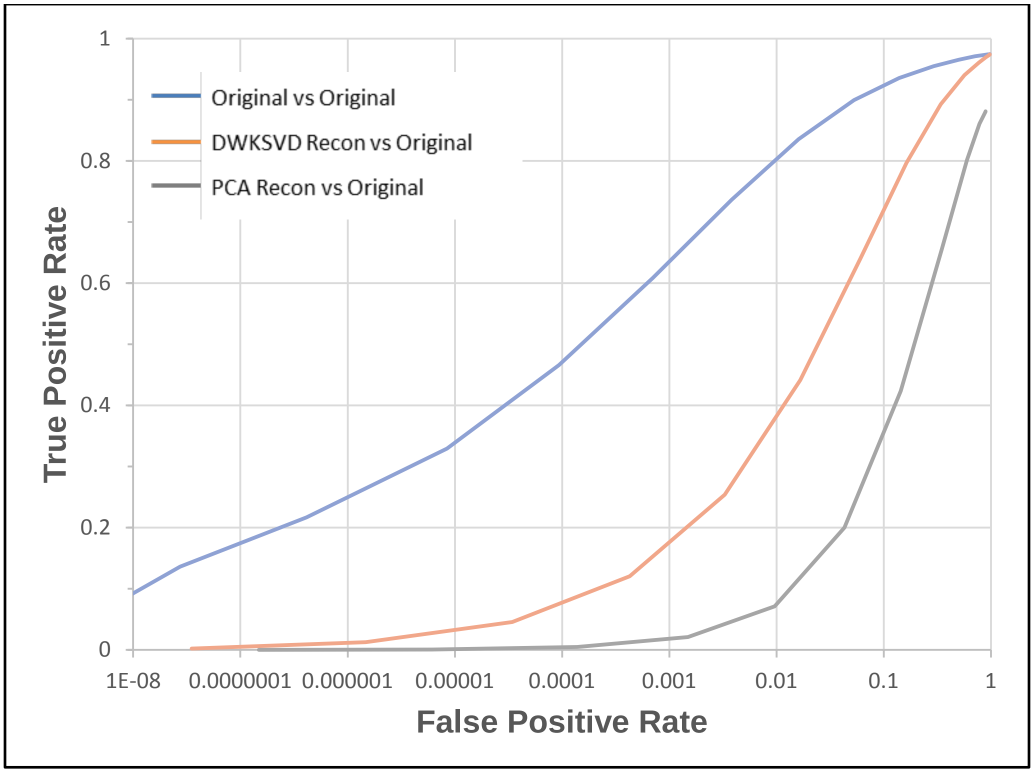


Figure III–13: FRGC Reconstruction and Matching, PittPatt Face Matcher, with ROC Curve (log-linear)

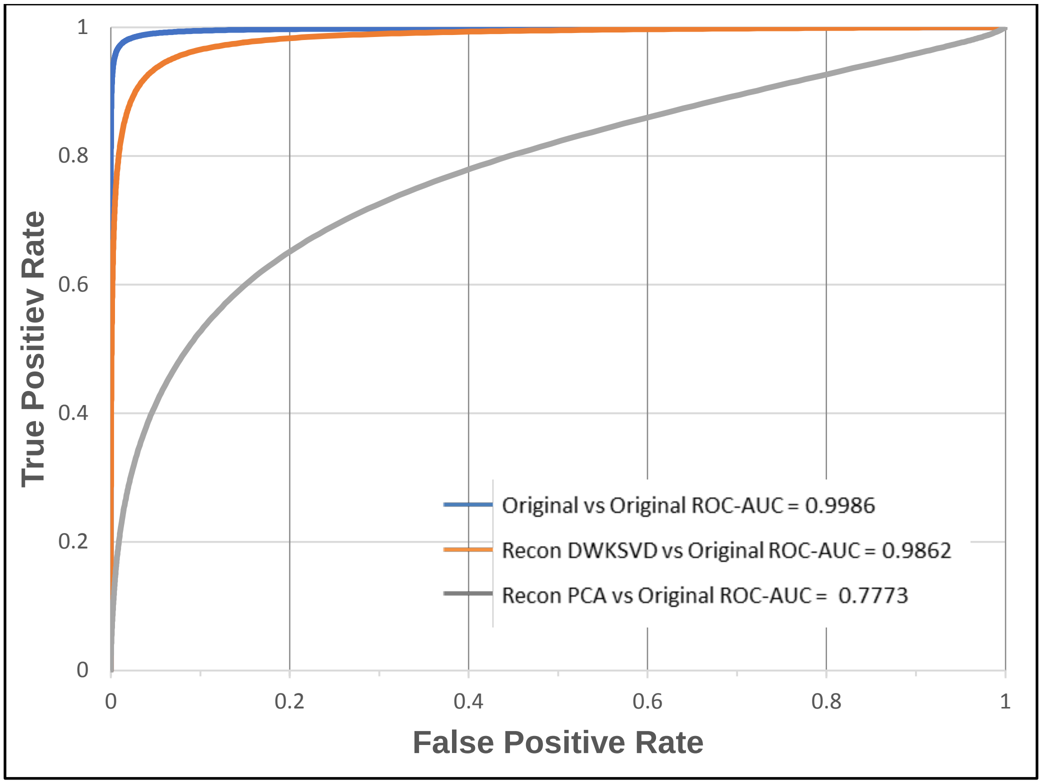


Figure III–14: FRGC Reconstruction and Matching, KCFA Face Matcher, with ROC Curve (linear-linear)

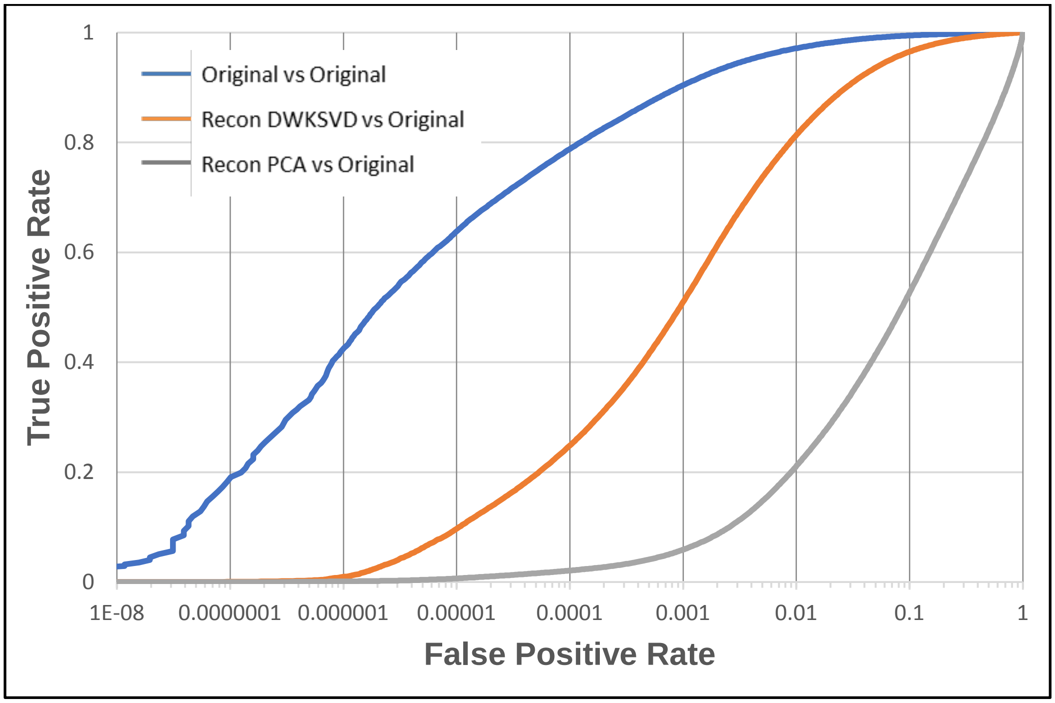


Figure III–15: FRGC Reconstruction and Matching, KCFA Face Matcher, with ROC Curve (log-linear)

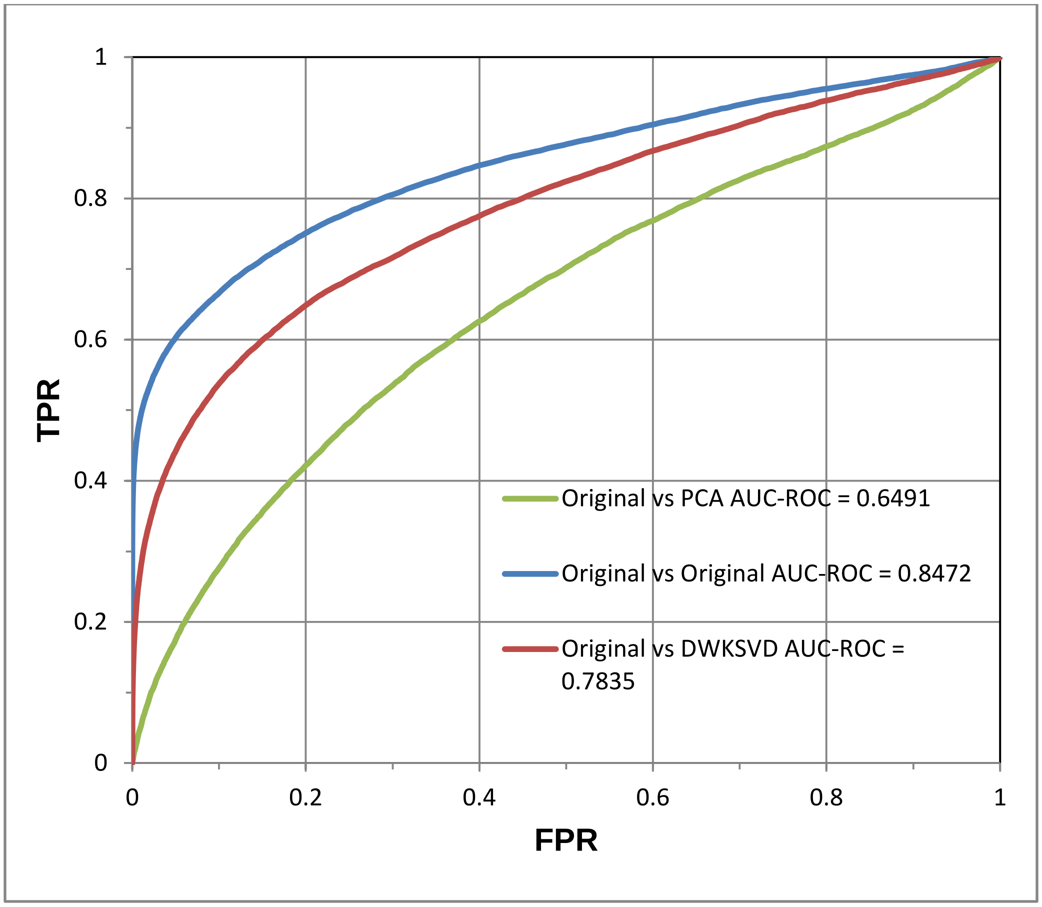


Figure III–16: Cropped Yale Reconstruction and Matching, PittPatt Face Matcher, with ROC Curve

As discussed in the Metrics section, the EER is a point on the ROC curve where the FPR and the false rejection rate (i.e., 1 – TPR) are equal. The lowest value is the most desirable. Both the university's results and the RT&E results are reported in Table III–2. The results are identical for all three cases of KCFA matcher and one case for the PittPatt matcher. The RT&E center results for the other two PittPatt cases deviated from the university’s. These discrepancies were investigated, but no satisfactory rationale was found for their differences. Observations include:

- All results from the KCFA matcher are lower (i.e., better) than the corresponding results for the PittPatt matcher; and
- Regardless of which matcher is used, the lowest error rates are for comparing the original faces with themselves (as expected), followed by comparing the reconstructed face using the DWK-SVD to the original, and the largest error rates result from the PCA reconstruction methodology.

The EER results for the corresponding analysis of the cropped Yale dataset are provided in Table III–3. The same general trend of the DWK-SVD reconstruction methodology out-performing the PCA is repeated.

The EER is a summary metric for ROC curves; the other summary metric is the AUC. The ROC curves and AUC values provided in Figures III–12 through III–15 show the DWK-SVD representation and KCFA face matcher AUC value to be 26.9% larger than the PCA-KCFA combination.

Table III–2: Assessment EER Results vs. CMU Results (FRGC Dataset)

Methods	AUC RT&E Center	EER RT&E Center	EER CMU Results [21]
KCFA Original vs Original	0.9986	0.014	0.014
KCFA Recon. DWK-SVD vs. Original	0.9862	0.056	0.056
KCFA PCA vs Original	0.7773	0.279	0.279
PittPatt Original vs Original	0.9287	0.053	0.083
PittPatt Recon. DWK-SVD vs. Original	0.8512	0.163	0.188
PittPatt Recon. PCA vs Original	0.5770	0.347	0.347

Table III–3: Assessment EER Results (Cropped Yale Dataset)

Methods	AUC RT&E Center	EER RT&E Center
PittPatt Original vs Original	0.8472	0.2299
PittPatt DWK-SVD vs Original	0.7835	0.2853
PittPatt PCA vs Original	0.6491	0.3815

III-5 CONCLUSIONS

The primary goal of this assessment was to replicate the results reported by CMU in the comparison of the performance of the DWK-SVD face reconstruction capability against a baseline PCA approach on the FRGC dataset. This verification used two main metrics, PSNR and ROC curves, and two ROC curve characteristics (EER, AUC). As a secondary objective, the same methodology was applied to a second dataset – Cropped Yale.

The RT&E results for the FRGC dataset and the corresponding Carnegie Mellon University results for EER are presented in Table III–2. The results are identical for four of the six cases. The two cases of different results were investigated. In all cases, the RT&E Center used the software and data components that the university stated they used; however, the differences could not be resolved. In the future, an investigation needs to occur to identify the source of this discrepancy.

The comparisons found for the FRGC dataset:

- The PSNR comparison (Figure III–8) was (visually) judged to have a larger mean and a narrower distribution than the PCA distribution;
- The EER comparison found the KCFA/DWK-SVD face matcher-representation methodology to be 20.1% of the KCFA/PCA methodology value;
- The EER comparison also found the KCFA face matcher, in combination with all the image representations, to have lower EER values than the PittPatt matcher in combination with any representation; and
- The AUC-ROC curve comparison found the DWK-SVD representation to be larger (26.9% for KCFA matcher, 47.5% for PittPatt matcher) than the PCA representation.

The corresponding results for the Cropped Yale dataset are as follows:

- The EER comparison found the PittPatt-DWK-SVD face matcher-representation methodology to be 74.8% of the PittPatt-PCA methodology;
- The AUC-ROC curve comparison found the DWK-SVD representation to be 20.7% larger than the PCA representation.

The illumination variability of the Cropped Yale dataset may or may not be relevant to end users of CMU’s algorithm, but to productize the algorithm itself, it should be retrained on a larger dataset of representative images of the end use case. Furthermore, a firm preprocessing stage to detect the periocular region would be required as input to the reconstruction algorithm.

III-6 ACRONYMS & ABBREVIATIONS

AUC	Area under the curve
CMU	Carnegie Mellon University
DWK-SVD	Dimensionally Weighted Kernel Singular Value Decomposition
EER	Equal Error Rate
FN	False negative
FP	False positive
FPR	False Positive Rate
FRGC	Face Recognition Grand Challenge
JHU/APL	Johns Hopkins University Applied Physics Laboratory
KCFA	Kernel Class-dependence Feature Analysis
NIJ	National Institute of Justice
NIST	National Institute of Standards and Technology
PCA	Principal Component Analysis
PSNR	Peak Signal-to-Noise Ratio
R&D	Research and Development
ROC	Receiver Operating Characteristic
RT&E Center	Research, Test, and Evaluation Center
TP	True positive
TPR	True Positive Rate

APPENDIX III-A: DIMENSIONALLY WEIGHTED KERNEL SINGULAR VALUE DECOMPOSITION (DWK-SVD) PROBLEM FORMULATION

Datasets of face images are large so representations of faces are created to facilitate processing. This reduces the dimensionality of "representing" a face; consequently, the size of the archived data is smaller and the complexity of the face reconstruction is reduced to matrix multiplication.

The CMU algorithm applies a face representation approach to both face regions. The intent is to have the two representations overlap so that if you have the periocular representation you can retrieve the full-face representation, and from that, hallucinate the image of the full face.

CMU provided the RT&E Center with both a dictionary mapping and the preprocessed Face Recognition Grand Challenge (FGRC) dataset to be processed to enable the assessment on that dataset. The mathematical formulation of the joint optimization of the full face and periocular region (defined by subscript lambda) of the face combines learning the elements of this dictionary (**D**), the face features (**X**), and face instances (**Y**) using the parameter (Beta) to weight the fidelity of the reconstruction in the periocular region. The individual minimization problems are solved jointly to solve for **D** and $\mathbf{D}_\Lambda$. [23]

$$\underset{\mathbf{D},\mathbf{D}_\Lambda,\mathbf{X}}{\arg\min} \left\| \begin{pmatrix} \mathbf{Y} \\ \sqrt{\beta}\mathbf{Y}_\Lambda \end{pmatrix} - \begin{pmatrix} \mathbf{D} \\ \sqrt{\beta}\mathbf{D}_\Lambda \end{pmatrix} \mathbf{X} \right\|_F^2 \text{ subject to } \quad \forall i, \|\mathbf{x}_i\|_0 \leq K$$

Figure III-A–1: Optimization Problem for Dictionary [21]

Given the representation, the RT&E Center used two face matchers to hallucinate the full image: Kernel Class-dependence Feature Analysis (KCFA), PittPatt (see Section III-3).

APPENDIX III-B: KCFA MATCHER FLOWCHART

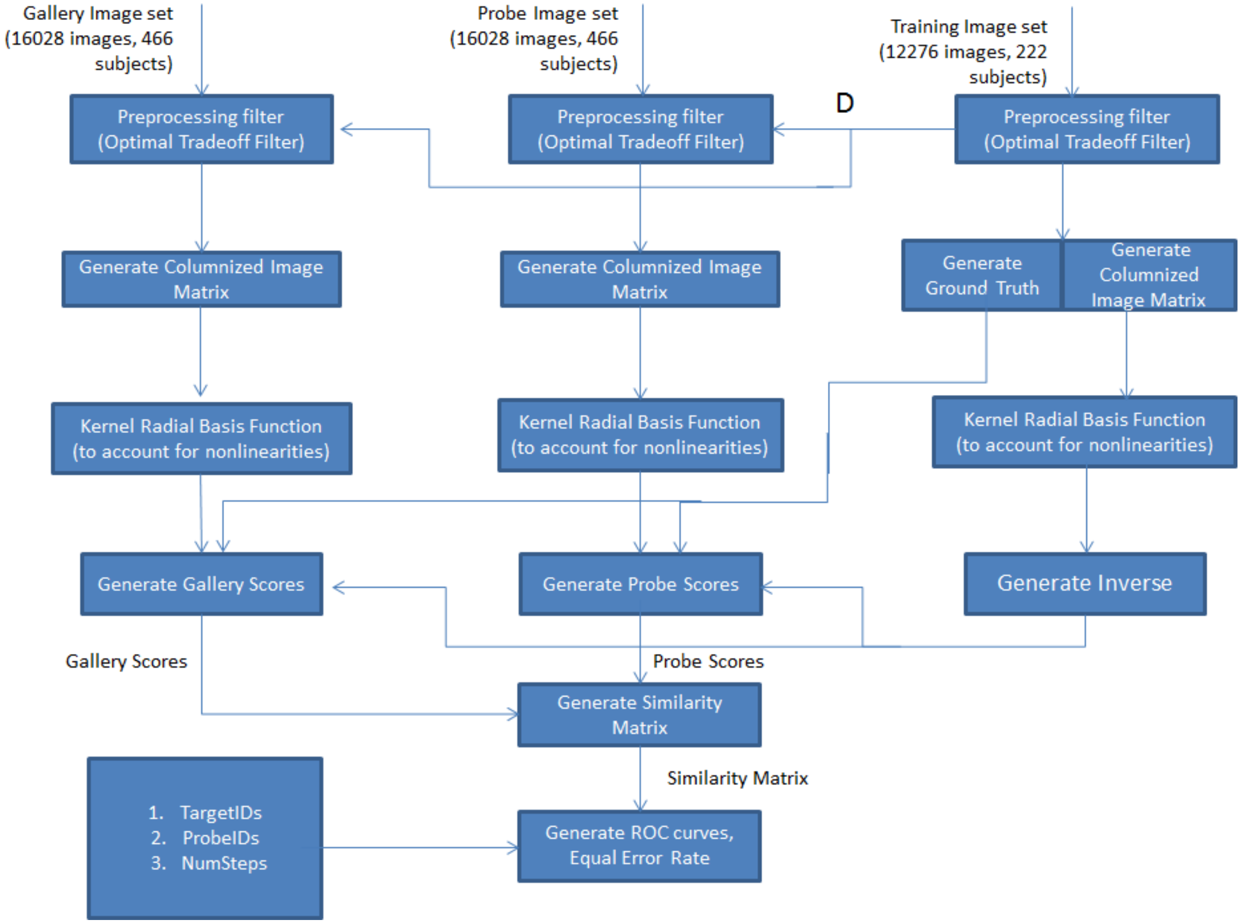

Gallery Image set (16028 images, 466 subjects)
Probe Image set (16028 images, 466 subjects)
Training Image set (12276 images, 222 subjects)
D
Preprocessing filter (Optimal Tradeoff Filter)
Preprocessing filter (Optimal Tradeoff Filter)
Preprocessing filter (Optimal Tradeoff Filter)
Generate Columnized Image Matrix
Generate Columnized Image Matrix
Generate Ground Truth
Generate Columnized Image Matrix
Kernel Radial Basis Function (to account for nonlinearities)
Kernel Radial Basis Function (to account for nonlinearities)
Kernel Radial Basis Function (to account for nonlinearities)
Generate Gallery Scores
Generate Probe Scores
Generate Inverse
Gallery Scores
Probe Scores
Generate Similarity Matrix
Similarity Matrix
1. TargetIDs
2. ProbeIDs
3. NumSteps
Generate ROC curves, Equal Error Rate

Figure III-B–1: Kernel Class-dependence Feature Analysis (KCFA) Matcher Flowchart [23] [24] [27]

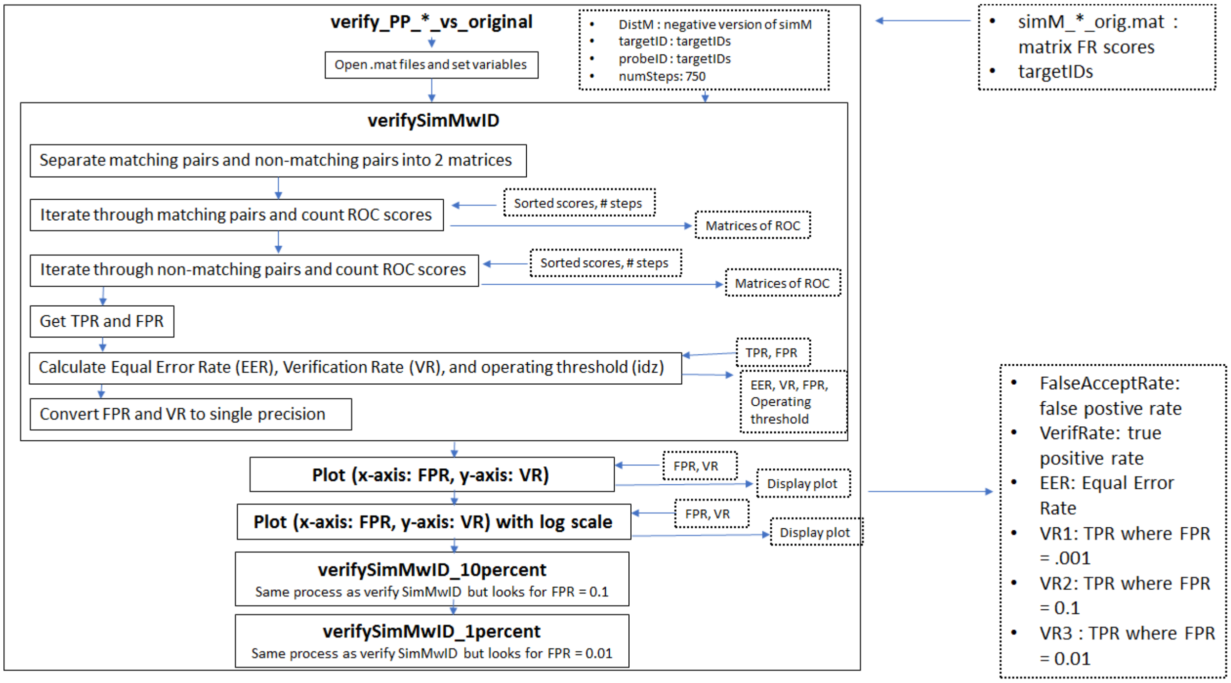

verify_PP_*_vs_original
Open .mat files and set variables
DistM : negative version of simM
targetID : targetIDs
probeID : targetIDs
numSteps: 750
simM_*_orig.mat : matrix FR scores
targetIDs
verifySimMwID
Separate matching pairs and non-matching pairs into 2 matrices
Iterate through matching pairs and count ROC scores
Sorted scores, # steps
Matrices of ROC
Iterate through non-matching pairs and count ROC scores
Sorted scores, # steps
Matrices of ROC
Get TPR and FPR
Calculate Equal Error Rate (EER), Verification Rate (VR), and operating threshold (idz)
TPR, FPR
EER, VR, FPR, Operating threshold
Convert FPR and VR to single precision
FalseAcceptRate: false postive rate
VerifRate: true positive rate
EER: Equal Error Rate
VR1: TPR where FPR = .001
VR2: TPR where FPR = 0.1
VR3 : TPR where FPR = 0.01
Plot (x-axis: FPR, y-axis: VR)
FPR, VR
Display plot
Plot (x-axis: FPR, y-axis: VR) with log scale
FPR, VR
Display plot
verifySimMwID_10percent
Same process as verify SimMwID but looks for FPR = 0.1
verifySimMwID_1percent
Same process as verify SimMwID but looks for FPR = 0.01

Figure III-B–2: PittPatt Flow Chart

REFERENCES

[1] RT&E Center, "Proposed Technical Approach NIJ RT&E Center," February 2017.

[2] Nechyba, M.C., L. Brandy, and H. Schneiderman, "PittPatt Face Detection and Tracking for the CLEAR 2007 Evaluation," *CLEAR 2007 and RT 2007*, part of the Lecture Notes in Computer Science book series, 4625, 2008, pp. 126–137.

[3] Hu, P. and D. Ramanan. "Finding Tiny Faces," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 1522–1530, https://doi.org/10.1109/CVPR.2017.166.

[4] Redmon, J. and A. Farhadi, "YOLO9000: Better, Faster, Stronger," 2016, http://pjreddie.com/yolo9000.

[5] Whitelam, Cameron, et al., "IARPA Janus Benchmark-B Face Dataset," *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 592–600, https://doi.org/10.1109/cvprw.2017.87.

[6] Phillips, P.J. et al., "An Introduction to the Good, the Bad, & the Ugly Face Recognition Challenge Problem," 2011, https://www.nist.gov/sites/default/files/documents/itl/iad/ig/05771424.pdf.

[7] Yang, S., P. Luo, C. Change Loy, and X. Tang, "WIDER FACE: A Face Detection Benchmark," http://mmlab.ie.cuhk.edu.hk/projects/WIDERFace/.

[8] Everingham, M., L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes (VOC) Challenge," *International Journal of Computer Vision*, Vol. 88, 2010, pp. 303–338. https://doi.org/10.1007/s11263-009-0275-4.

[9] "Everingham, M., S.M.A. Eslami, L. Van Gool, C.K.I. Williams, J. Winn, and A. Zisserman, "The Pascal Visual Object Classes Challenge: A Retrospective," *International Journal of Computer Vision*, Vol. 111, 2015, pp. 98–136, https://doi.org/10.1007/s11263-014-0733-5.

[10] Pedregosa, F. et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, Vol. 12, 2011, pp. 2825–2830.

[11] Manning, C.D., P. Raghavan, and H. Schütze, "Evaluation of Ranked Retrieval Results," *Introduction to Information Retrieval*, by Cambridge University Press, 2008, https://nlp.stanford.edu/IR-book/html/htmledition/evaluation-of-ranked-retrieval-results-1.html.

[12] "F1 Score," Wikipedia, https://en.wikipedia.org/wiki/F1_score, accessed on June 26, 2018.

[13] Cheney, J., B. Klein, A.K. Jain, and B.F. Klare, "Unconstrained Face Detection: State of the Art Baseline and Challenges," *International Conference on Biometrics (ICB)*, Phuket, 2015, pp. 229–236, https://doi.org/10.1109/icb.2015.7139089.

[14] Amos B., B. Ludwiczuk, and M. Satyanarayanan, "OpenFace: A general-purpose face recognition library with mobile applications," June 2016, https://www.cs.cmu.edu/~satya/docdir/CMU-CS-16-118.pdf.

[15] Givens, G.H., J.R. Beveridge, P.J. Phillips, B. Draper, Y.M. Lui, and D. Bolme, "Introduction to Face Recognition and Evaluation of Algorithm Performance," *Computational Statistics and Data Analysis*, Vol. 67, November 2013, pp. 236–247.

[16] Schroff, F., D. Kalenichenko, and J. Philbin, "FaceNet: A Unified Embedding for Face Recognition and Clustering," March 2015.

[17] Face Recognition Homepage, http://www.face-rec.org/ accessed on July 11, 2018.

[18] Amos, B., B. Ludwiczuk, and M. Satyanarayanan, "OpenFace: A general-purpose face recognition library with mobile applications," June 2016, https://www.semanticscholar.org/paper/OpenFace-A-general-purpose-face-recognition-librar-Amos-Ludwiczuk/82e66c4832386cafcec16b92ac88088ffd1a1bc9.

[19] Huang, G.B., M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," https://www.google.com/search?q=labeled+faces+in+the+wild+studying&ie=utf-8&oe=utf-8&client=firefox-b-1.

[20] Uzair, M., A. Mahmood, A. Mian, and C. Mcdonald, (2013). Periocular Biometric Recognition using Image Sets. *Proceedings of IEEE Workshop on Applications of Computer Vision*. 10.1109/WACV.2013.6475025.

[21] "Periocular-Based Full Face Hallucination and Matching Test Plan."

[22] "Optimized Periocular Template Selection for Human Recognition" at https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3747475/figure/fig2/

[23] Juefei-Xu, F., D.K. Pal, and M. Savvides, "Hallucinating the Full Face from the Periocular Region via Dimensionally Weighted K-SVD" in *Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2014 IEEE Conference on, June 2014

[24] Kumar, B. V. K. Vijaya, M. Savvides, and C. Xie, "Correlation Pattern Recognition for Face Recognition," in *Proceedings of the IEEE*, vol. 94, no. 11, pp. 1963-1976, Nov. 2006.

[25] "Overview of the Face Recognition Grand Challenge" at https://www.researchgate.net/publication/4156256_Overview_of_the_Face_Recognition_Grand_Challenge

[26] Athinodoros Georghiades, Peter Belhumeur, and David Kriegman's paper, "From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose", PAMI, 2001. Dataset available at http://vision.ucsd.edu/~leekc/ExtYaleDatabase/ExtYaleB.html

[27] C. Xie, M. Savvides, and B. V. K. Vijaya Kumar, B "Kernel Correlation Filter Based Redundant Class-Dependence Feature Analysis (KCFA) on FRGC2.0 data" *in Proc. 2nd Int. Workshop Analysis Modeling of Faces Gestures* (AMFG 2005) held in conjunction with ICCV 2005, Beijing, 2005.

[28] Mathias M., Benenson R., Pedersoli M., Van Gool L. (2014) Face Detection without Bells and Whistles. In: Fleet D., Pajdla T., Schiele B., Tuytelaars T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8692. Springer, Cham