NOVETTA

Amazon Rekognition

Quick-Look Biometric Performance Assessment

August 2017 | novetta.com | Copyright © 2017, Novetta, LLC. | COMPANY PROPRIETARY

OVERVIEW

Cloud-based face recognition (CBFR) makes automated face detection and matching accessible based on its negligible start-up costs, volume-based pricing model, and streamlined application programming interfaces (APIs). In contrast to traditional face recognition solutions – whose implementation costs, licensing models, and development effort can pose barriers to entry to newcomers – CBFR prioritizes low-cost, simplified deployment.

While cloud-based face recognition services have been commercially available since 2009, the launch of Amazon Web Services (AWS) Rekognition in November 2016 represented a significant milestone in the market. Leveraging widely-adopted elements of the AWS ecosystem, Rekognition increases the range of government and commercial applications and environments for which face recognition becomes a potential solution. Because AWS provides scalable, trusted computing environments for the majority of Fortune 500 companies and over 90% of Fortune 100 companies, Rekognition represents a more stable platform for developing FR-enabled solutions than services from less established companies. CBFR services such as Rekognition have the potential to complement or disrupt traditional face recognition solutions used for security, fraud detection, and identification applications such as watchlisting.

Decisions on adoption and integration of CBFR services are based on a wide range of considerations, including the following:

- **Biometric performance, including face detection and identification rates.** How accurate is the CBFR service (e.g. in matching a specific face from a large, enrolled gallery of images), and how does accuracy change based on gallery size and face image quality?
- **Security and compartmentalization of face image galleries.** Does the CBFR service deliver a similar level of data protection as on-premise or vendor-located solutions?
- **Support for advanced face recognition workflows.** Does the CBFR service enable or support capabilities and workflows commonly found in traditional face recognition solutions, such as identity-based case management, template updating, and match adjudication?

- **Scalability.** What limits does the CBFR service impose in terms of transaction volume, candidate lists, and gallery size?
- **Cost savings.** Does the use of CBFR drive significant cost savings in licensing, development labor, analyst labor, or other areas?
- **Extent of current AWS utilization.** Does the organization have robust competencies and/or investments in AWS?
- **Use of machine learning approaches to drive improved performance.** Is the CBFR capability based on machine learning or related approaches, and if so what is known about their adaptability to increasingly challenging face recognition applications?

This whitepaper focuses on the first of these considerations – core face recognition performance expressed as detection rates and identification rates.[1] To gain first-order insights into Amazon Rekognition performance, this paper present results from two separate evaluations. The first section compares Amazon Rekognition performance to that of face recognition tools commonly used in entry-level face processing applications. The second section compares Amazon Rekognition performance to that of CBFR services from Microsoft and Kairos. This assessment provides a baseline from which more advanced performance evaluations – e.g. at larger scale, or expressed as a function of throughput – can be conducted.

REKOGNITION PERFORMANCE RELATIVE TO BASELINE FACE RECOGNITION TOOLS

SCOPE

This part of the paper examines Rekognition's biometric performance in comparison to conventional, entry-level face recognition tools. These tools can be considered a baseline for assessing the relative performance of Rekognition. Face recognition tools evaluated relative to Rekognition are as follows:

- **PittPatt FR SDK 5.2.2**. Commonly used by USG agencies for rudimentary testing against unconstrained images. Acquired by Google.
- **Neurotechnology VeriLook 5.3[2].** Inexpensive face recognition software commonly used in commercial and government applications.
- **Neurotechnology FaceCell 1.2.** Legacy face recognition software developed for mobile phone applications.
- **OpenBR 0.5.0[3].** Open source face recognition implementation developed by MITRE.

These tools are commonly used in academic or low-value commercial settings where basic face recognition functionality such as detection or matching is necessary.

[1] *Rekognition offers computer-vision based capabilities such as object and scene analysis that are outside the scope of this paper.*

[2] *www.neurotechnology.com/verilook.html*

[3] *github.com/biometrics/openbr/releases*

USE OF THE REKOGNITION API FOR TESTING

Software development kits (SDKs) for Rekognition are available in Android, JavaScript, iOS, Java, .NET, Node.js, PHP, Ruby, and Python. Rekognition can also be used through the AWS Command Line Interface. In this assessment, we used Version 3.3.1 of the .NET SDK.

The Rekognition API can be used by making calls (via requests) to AWS along with applicable parameters. Rekognition processes each request and returns result data in a response. Detection API calls are used to locate faces in a submitted image and analyze detected faces. These two calls – DetectFaces and IndexFaces – return metadata that includes face location, detection confidence, facial landmarks, pose angles, quality scores, age/gender/emotion estimates, and a set of values indicating the presence of several features (e.g. moustache, beard, glasses). The IndexFaces API call stores the corresponding face feature vector in a specified face collection on AWS. Face metadata is also returned during the IndexFaces API call. In standard biometric technology terms, "index" is equivalent to enrollment, and "vector" is equivalent to template.

Two separate API calls can be used to match faces. CompareFaces performs a 1:1 comparison attempt, while SearchFaces performs a 1:N search attempt. In each case, the API call compares the largest face in the source or probe image with all faces in the target image(s). Each API call returns one or more similarity scores indicating the confidence that the faces match on a 0-100 scale. Higher scores indicate a stronger match. A similarity threshold can be specified for each API call; when used, scores greater than the threshold value are returned. CompareFaces returns face detection metadata for both the source and target images, while SearchFaces returns detection metadata for the source image only. The maximum number of matches returned by a SearchFaces API call is 4096.

In our testing, target face images were enrolled in a face collection for each dataset described below, and matching was executed using the SearchFaces call. In instances where a probe-target image pair generated multiple scores (due to the detection of multiple faces), only the highest similarity score was recorded. The 'FaceMatchThreshold' parameter was set to 0.0 and the 'MaxFaces' parameter was set to 4096 to maximize the number of returned results. Because the match threshold was effectively disabled, any result not returned by the SearchFaces API call was assigned a score of 0.0.

TEST DATA

Rekognition biometric performance was evaluated using two face datasets: FRGC-1000 and FRAME-ID.

- FRGC-1000 is a subset of a NIST dataset[4] that contains frontal face images, including a significant percentage with background noise.
- FRAME-ID is a subset of images from a larger Novetta dataset called SOCIAL-ID (Sanitized Online Collection and Identity Analysis Library - Image Dataset). SOCIAL-ID contains images from social networking services and photo sharing sites, including Facebook, Flickr, and Instagram.

[4] *www.nist.gov/itl/iad/ig/frgc.cfm*

Dataset characteristics are shown in Table 1 and sample images are shown in Figure 1 and Figure 2.

DATASET	DESCRIPTION	IMAGES	SUBJECTS	GENUINE PAIRS	IMPOSTER PAIRS
FRGC-1000	Conventional face images with frontal pose and varied illumination and expression. Primarily white and Asian males and females.	1,000	376	661	249,339
FRAME-ID	SNS-style face images with varied pose, expression, and illumination. ~95% of subjects are white.	500	50	1,250	61,250

Table 1: Evaluation Datasets

FRAME-ID contains 50 subjects (30 white male, 20 white female) with 10 images per subject. Images were manually reviewed to ensure that each subject's face was the only face present in an image, was not obstructed by glasses or hats, and was not cut off by the image border. FRAME-ID was created for use in testing the performance of face recognition algorithms on SNS image data. As such, images have varied face pose angles and inter-ocular distances (IOD), as described in Table 2.

	IOD	YAW	PITCH	ROLL
MEAN	63.33	9.70°	5.16°	7.31°
MEDIAN	54.18	8.26°	4.54°	5.70°
STANDARD DEVIATION	33.49	6.98°	3.68°	6.38°
MINIMUM	17.31	0.00°	0.01°	0.00°
MAXIMUM	276.89	37.94°	19.38°	51.02°

Table 2: FRAME-ID IOD, Yaw, Pitch, and Roll Data

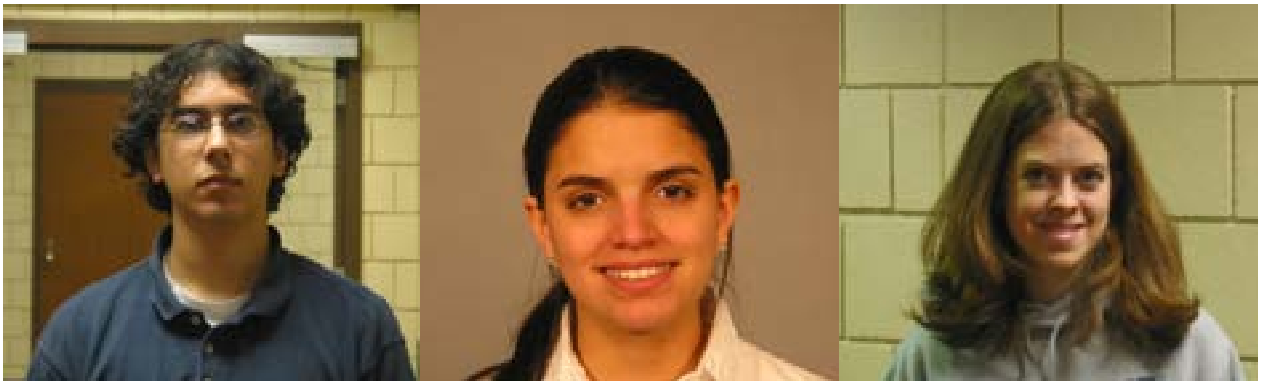


Figure 1: Representative FRGC-1000 Images

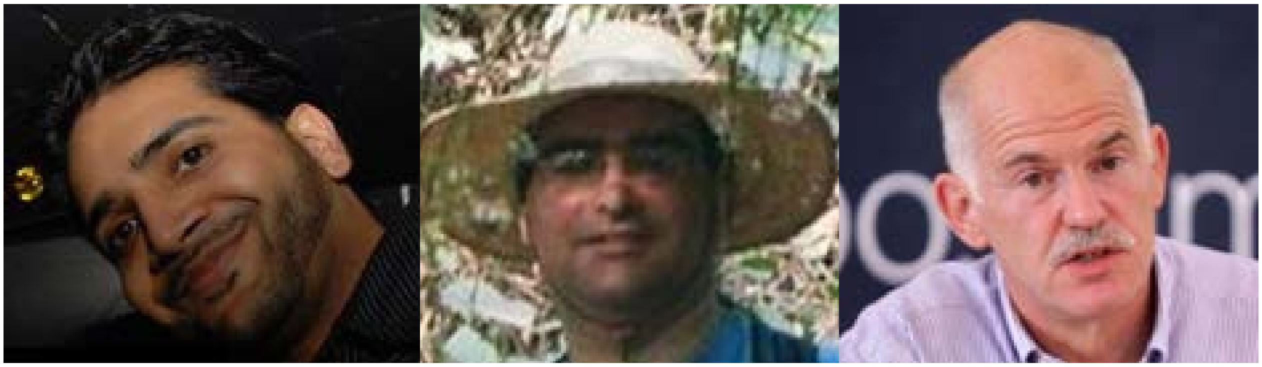


Figure 2: Representative FRAME-ID Images

ENROLLMENT RESULTS (FRGC-1000 AND FRAME-ID)

Failure to enroll (FTE) rate was calculated based on the number of target images that failed enrollment plus the number of probe images that failed face detection or feature extraction, all divided by the total number of face images (1000).

Enrollment results for the FRGC-1000 dataset are shown in Table 3.

	FACECELL	OPENBR	PITTPATT	VERILOOK	REKOGNITION
IMAGES	1000	1000	1000	1000	1000
ENROLLED	978	995	998	1000	1000
FTE%	2.20%	0.50%	0.20%	0.00%	0.00%

Table 3: FRGC-1000 Enrollment Results

With the exception of FaceCell, FTE for baseline FR tools was low. Rekognition and VeriLook enrolled all FRGC-1000 images, while PittPatt and OpenBR FTE were below 1%.

Enrollment results for the FRAME-ID dataset are shown in Table 4.

	FACECELL	OPENBR	PITTPATT	VERILOOK	REKOGNITION
IMAGES	500	500	500	500	500
ENROLLED	306	375	492	499	498
FTE%	38.80%	25.00%	1.60%	0.20%	0.40%

Table 4: FRAME-ID Enrollment Results

FRAME-ID, being comprised of relatively unconstrained faces, generated more interesting FTE results than FRGC-1000.

Rekognition and VeriLook had the lowest FTE rates with the SNS-style dataset, failing on only 2 images and 1 image respectively. PittPatt also performed well, with a 1.6% FTE rate (8 images). FaceCell and OpenBR struggled on this particular dataset, with a significantly higher FTE than the other three algorithms (38.8% and 25.0% respectively). Examples of images that failed enrollment are shown in Annex A.

These results demonstrate that Rekognition face detection performance meets or exceeds that of baseline FR tools, which is to be expected – like most newer FR tools, Rekognition is designed to process unconstrained images. We did not observe any false detections in which a non-facial region of an image is detected as a face, through this is known to be an issue for face recognition tools when processing complex or noisy images (e.g. with graffiti). More exhaustive testing is necessary to examine Rekognition performance with images known to generate false detections in other Novetta testing.

1:1 MATCHING RESULTS (FRGC-1000 AND FRAME-ID)

Images that successfully enrolled were used for matching. FRGC-1000 and FRAME-ID were each divided into probe and gallery sets by random selection. Results are shown on a 1:1 (i.e. verification) basis, as opposed to a 1:N (i.e. identification) basis. This is because not all face recognition tools included in this baseline evaluation are suited for 1:N identification. False match rate (FMR) and false non-match rate (FNMR) were calculated based on the comparison scores for each SDK.

1:1 FRGC-1000 MATCHING RESULTS

FNMR at primary FMR values are shown for each face recognition tool in Table 5. Lower values indicate more robust performance. Blank cells indicate that there were no observations at a given FMR.

FALSE MATCH RATE	FACECELL	OPENBR	PITTPATT	VERILOOK	REKOGNITION
10.00%	-	10.3%	1.8%	45.3%	-
01.00%	-	25.3%	11.2%	62.7%	-
0.10%	**10.3%**	37.4%	24.6%	-	50.4%
0.01%	18.5%	43.7%	37.6%	-	53.7%

Table 5: FNMR at Primary FMR Values for FRGC-1000 Dataset

FMR and FNMR across the full range of performance are shown as detection error tradeoff (DET) curves in Figure 3. Figures lower and to the left indicate more robust performance.

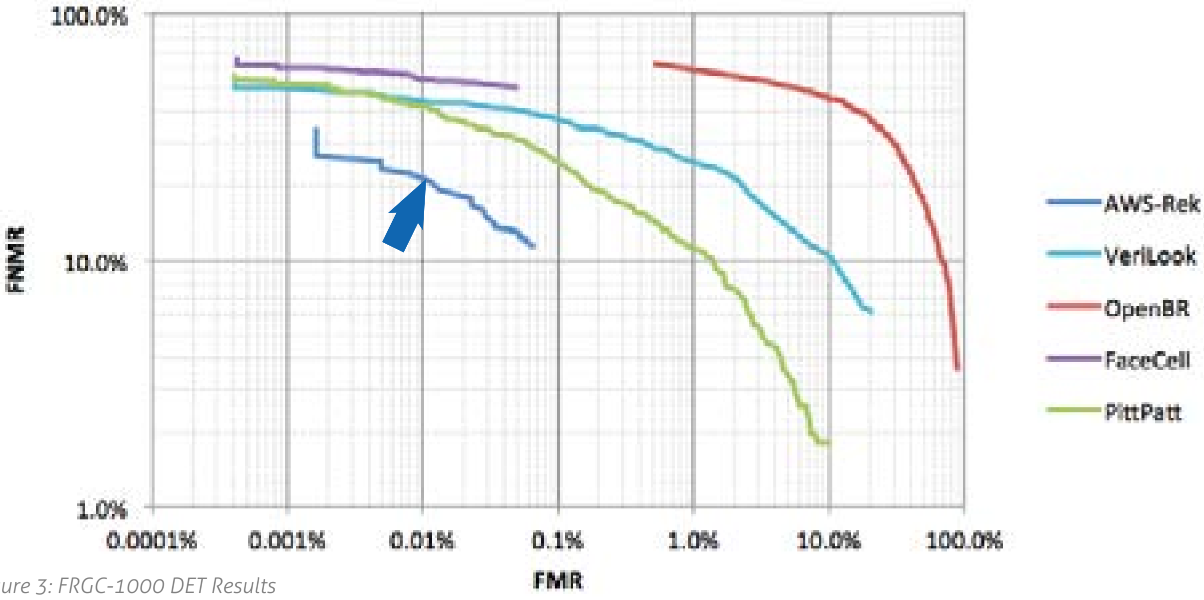


Figure 3: FRGC-1000 DET Results

Rekognition was the most accurate face recognition tool against FRGC-1000 data by a significant margin. At primary FMR operating points, Rekognition FNMR was half that of the second best algorithm, PittPatt. The open-source algorithm (OpenBR) and mobile SDK (FaceCell) predictably had the worst performance.

1:1 FRAME-ID MATCHING RESULTS

FNMR at primary FMR values are shown for each face recognition tool in Table 6. Blank cells indicate no observations at a given FMR. FaceCell did not generate usable match results.

FALSE MATCH RATE	AWS RE-	VERILOOK	PITTPATT	OPENBR
10.00%	-	50.1%	14.2%	50.1%
01.00%	-	82.8%	38.3%	68.2%
0.10%	-	95.6%	60.0%	-
0.01%	20.2%	98.5%	78.2%	-

Table 6: FNMR at Primary FMR Values for FRAME-ID Dataset

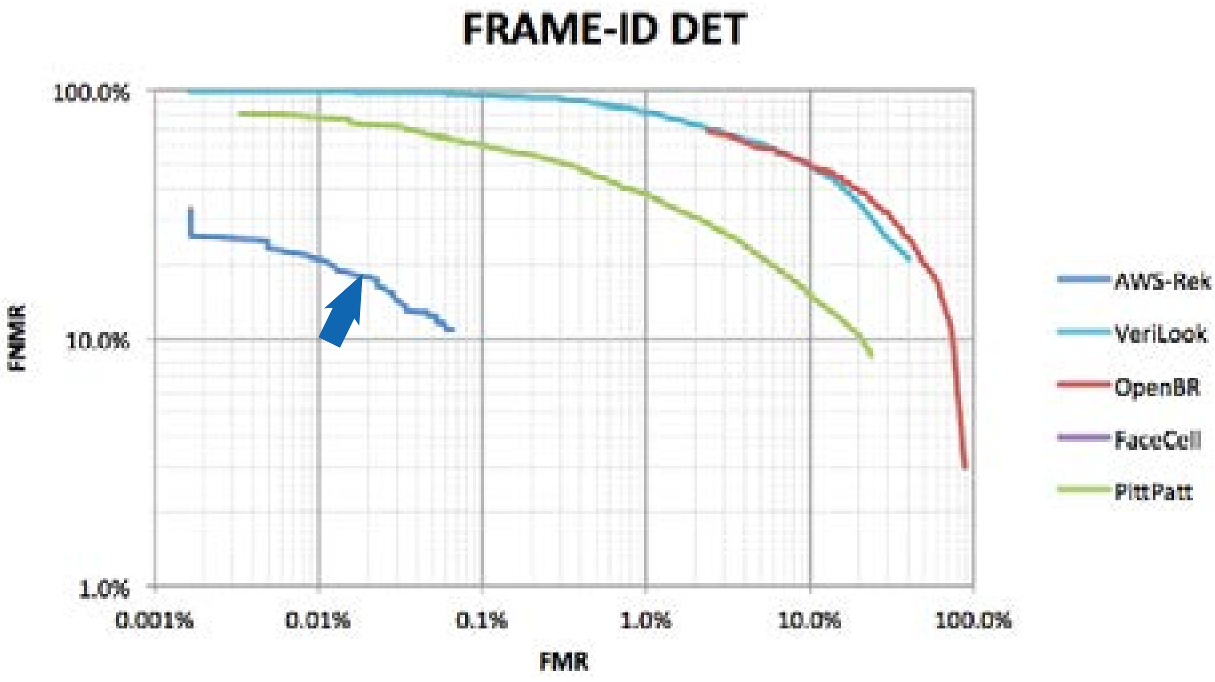


Figure 4: FRAME-ID DET Results

FMR and FNMR across the full range of performance are shown as DET curves in Figure 4.

The DET curves again highlight the strength of Rekognition's performance, which outperformed the second most accurate algorithm, PittPatt, by a factor of nearly 4.

CONCLUSIONS & OBSERVATIONS

Detection and matching results show that Rekognition performance far outpaces that of baseline FR tools for both constrained and unconstrained images. Initial analysis shows that Rekognition appears to minimize FMR at the expense of FNMR. At the lowest possible match threshold (0), Rekognition FMR is only 0.17% for FRGC-1000 and 0.07% for FRAME-ID. This explains why no FNMR values appear at lower FMR values in the tables above.

If Amazon were to relax this floor or recalibrate their scores, a developer could likely reduce FNMR values by accepting an increase in FMR – a tradeoff that many users might find acceptable for certain applications.

REKOGNITION PERFORMANCE RELATIVE TO COMPETING CLOUD-BASED FR SERVICES

OVERVIEW & SCOPE

To assess Rekognition performance relative to representative CBFR services, a comparative test was executed between Rekognition, Microsoft Cognitive Services, and Kairos.

- Microsoft Cognitive Services, formerly known as Project Oxford, is a collection of APIs that provide capabilities built upon Microsoft expertise in machine learning. One such service, Face API, offers face detection, analysis, recognition, and grouping. Face API was launched in March 2016.

- Kairos is an artificial intelligence company founded in 2010 that specializes in face recognition based on machine learning. The Kairos API offers capabilities in face detection, analysis, and recognition. Kairos also has novel capabilities in emotion analysis from video and face tracking in video (although the latter is only available with their offline SDK).

These services were selected based on ease of implementation; a more comprehensive performance evaluation to include an additional ~10 CBFR services is in the planning stages.

DEVELOPMENT OF TEST APPLICATIONS

A test application for the Amazon Rekognition API was developed using the .NET SDK. For each dataset, the test application created a "face collection" (gallery) and attempted enrollment of all gallery images. The Rekognition identification API call, SearchFacesByImage, was used to implement matching; this permitted the comparison of a detected probe face against the gallery. Two API parameters were specified in order to return all results: "FaceMatchThreshold" (set to 0.0) and "MaxFaces" (set to 4096).

A test application for the MCS Face API was developed using the .NET SDK. For each dataset, the application created a "face list" (gallery) and attempted enrollment for all gallery images. The Find Similar API call was used in matchFace mode to implement matching. This permitted the comparison of a detected probe face against face lists without enforcement of a match threshold. In testing, the scores returned by the Find Similar API call were identical to those returned by Identify API call, which can only return a maximum of 5 results. The Find Similar API call was configured to return a maximum of 1000 match results. Because this value was in excess of our gallery sizes, it ensured all results would be returned.

A test script for the Kairos API was developed using a third-party Python SDK (kairos-face-sdk-python). For each dataset, the application created a gallery and attempted enrollment for all gallery images. The Kairos identification API call, Recognize, was used to implement matching. This permitted the comparison of a probe image against the galleries. Two API parameters were specified to return all results: "threshold" (set to 0.0) and "max_num_results" (set to 1000).

CALCULATING 1:N RESULTS

For each vendor, gallery images were enrolled for both FRGC-1000 and FRAME-ID datasets. The gallery contains 500 images for FRGC-1000 and 250 images for FRAME-ID. Each dataset has the same number of probe samples. Each probe was used to search the corresponding gallery, with a list of potential matches returned in order of descending match score.

Matching performance is measured as true-positive identification rate (TPIR). True-positive identification rate is the percentage of identification attempts for which a probe's matching gallery sample is returned in the top N ranked results. For example, Rank-1 refers to the percentage of probes that return their matching gallery sample as the highest scoring result (i.e. the 1st result). Rank-10 refers to the percentage of probes that return their matching gallery sample in the top 10 highest-scoring results. TPIR provides insight into how likely a probe is to return the correctly matching gallery sample when N results are returned. TPIR is dependent on gallery size and should not be used to extrapolate performance on larger or smaller galleries.

ENROLLMENT RESULTS (FRGC-1000 AND FRAME-ID)

Enrollment failure rates were calculated based on the number of target images that failed enrollment and the number of probe images that failed face detection or feature extraction. Enrollment results for the FRGC-1000 dataset are shown in Table 7.

	REKOGNITION	**MICROSOFT**	**KAIROS**
IMAGES	1000	1000	1000
SUCCESS	1000	999	987
FAILURE %	0.00%	0.10%	1.30%

Table 7: FRGC-1000 Enrollment Results

Rekognition was able to process all images and Microsoft failed on a single image. Kairos had a slightly higher failure rate, with 12 failed gallery images and 1 failed probe image. Images that failed Kairos gallery enrollment reported the same exception: "too many faces in image." The Kairos API will not automatically enroll the largest face when multiple faces are detected, and each submitted image must contain only one face. The Kairos failure rate would be closer to Rekognition and Microsoft under a workflow where face detection and image cropping are performed prior to enrollment submission.

Enrollment results for the FRAME-ID dataset are shown in Table 8.

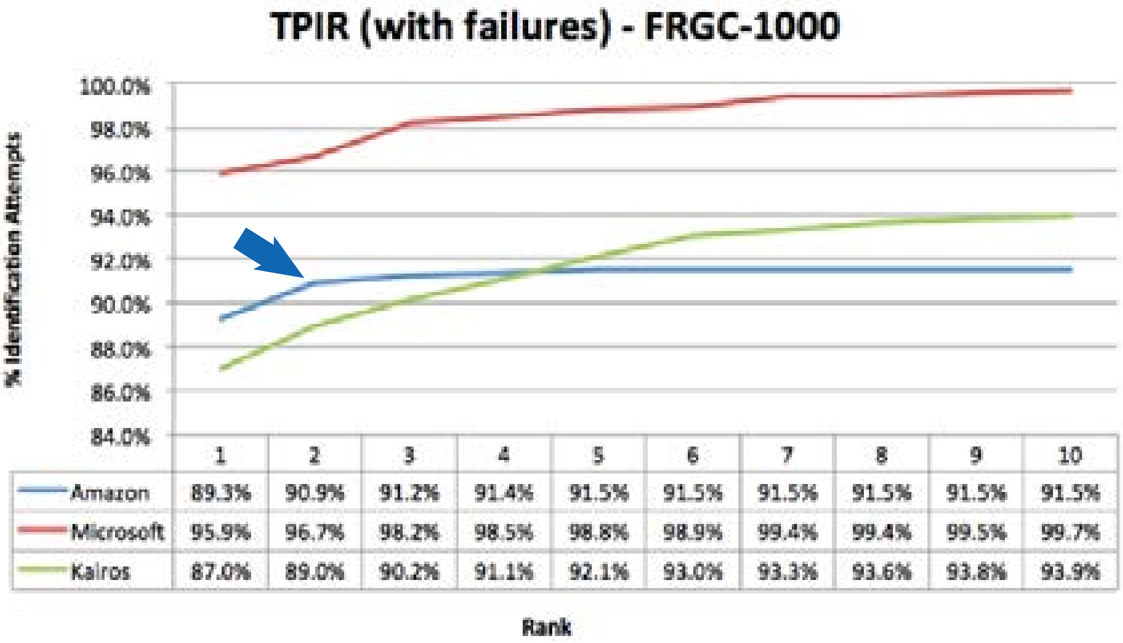
	REKOGNITION	**MICROSOFT**	**KAIROS**
IMAGES	500	500	500
SUCCESS	498	476	473
FAILURE %	**0.40%**	4.80%	5.40%

Table 8: FRAME-ID Enrollment Results

Rekognition was able to process all but 2 images. Microsoft and Kairos both encountering a higher proportion of failures, with 24 and 27, respectively. Unlike the multi-face images that failed Kairos for FRGC-1000, none of Kairos' FRAME-ID image failures were caused by multi-face detection. 10 images failed both Microsoft and Kairos. The images responsible for failures were largely low-resolution and/or exhibited challenging pose, illumination, and expression.

1:N MATCHING RESULTS (FRGC-1000 AND FRAME-ID)

TPIR on the FRGC-1000 dataset is shown in Figure 5. Higher horizontal lines indicate more robust performance.



	1	2	3	4	5	6	7	8	9	10
Amazon	89.3%	90.9%	91.2%	91.4%	91.5%	91.5%	91.5%	91.5%	91.5%	91.5%
Microsoft	95.9%	96.7%	98.2%	98.5%	98.8%	98.9%	99.4%	99.4%	99.5%	99.7%
Kairos	87.0%	89.0%	90.2%	91.1%	92.1%	93.0%	93.3%	93.6%	93.8%	93.9%

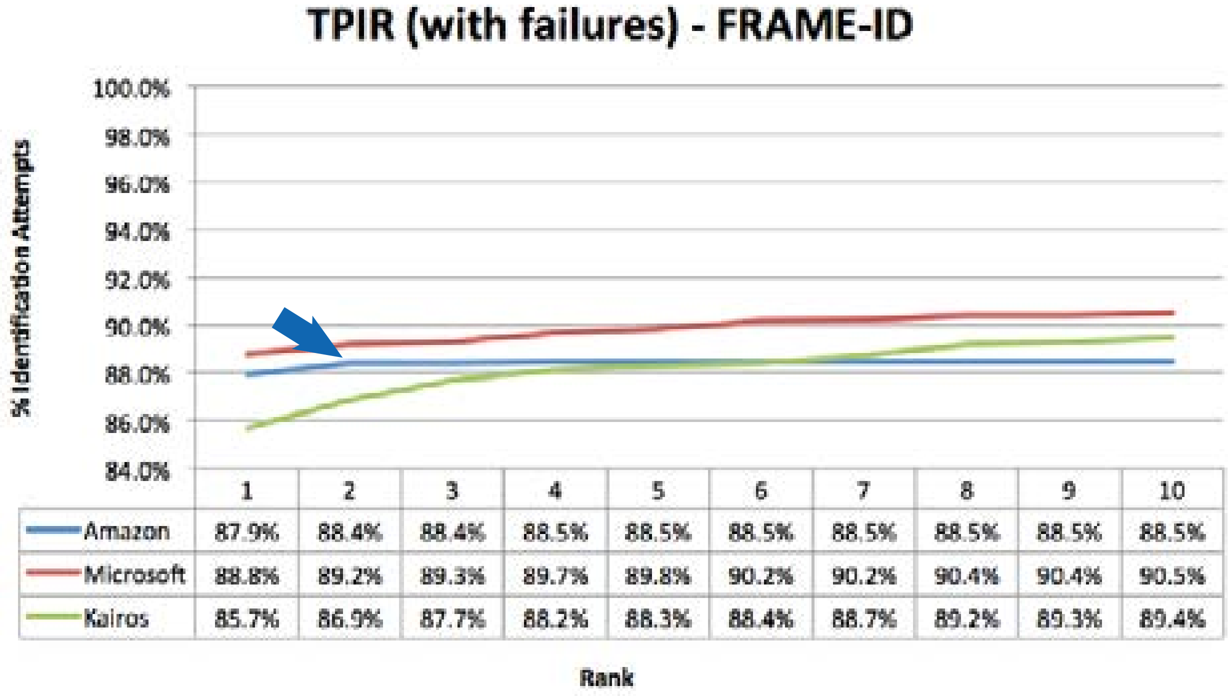
Figure 5: FRGC-1000 TPIR for Cloud-Based FR Services

Images whose faces could not be enrolled or detected are included in TPIR. This was necessary to avoid penalizing services that attempt to enroll and detect as many images as possible. Future testing will analyze the relationship between image quality, enrollment rates, and search results.

Microsoft displayed the strongest TPIR performance on FRGC-1000 data, with the correct gallery template returned with the highest score for 95.9% of probes (Rank-1). Additionally, the correct gallery template was returned in the top 10 results (Rank-10) for 99.7% of probes. Rekognition slightly outperforms Kairos at Rank-1 through Rank-4, but Kairos performs better at Rank-5 through Rank 10.

Rekognition's performance is likely a consequence of a behavior observed above in which their matching logic appears to be calibrated at an excessively strict tolerance to prevent false matches. This can be seen in their score distributions, where over 99.8% of their impostor comparisons generate the minimum possible score (0.0), compared to only 14.6% for Microsoft and 0.5% for Kairos. Similarly, 8.5% of Rekognition genuine comparisons produced the minimum possible score of 0.0, but no genuine comparisons generated the minimum score for either Microsoft or Kairos.

FRAME-ID TPIR is shown in Figure 6.



	1	2	3	4	5	6	7	8	9	10
Amazon	87.9%	88.4%	88.4%	88.5%	88.5%	88.5%	88.5%	88.5%	88.5%	88.5%
Microsoft	88.8%	89.2%	89.3%	89.7%	89.8%	90.2%	90.2%	90.4%	90.4%	90.5%
Kairos	85.7%	86.9%	87.7%	88.2%	88.3%	88.4%	88.7%	89.2%	89.3%	89.4%

Figure 6: FRAME-ID TPIR Cloud-Based FR Services

In contrast to the conventional FRGC-1000 dataset, all three algorithms performed similarly with TPIR on the SNS-style FRAME-ID dataset. Microsoft edged out Rekognition in Rank-1 identification by 0.9%, with Kairos trailing Rekognition by 1.2%. Performance was also similar at Rank-10. Microsoft remained the strongest with 90.5% of identification attempts returning the matching gallery sample in the top 10 results, followed by Kairos at 89.4% and Rekognition at 88.5%. Given the size of the datasets, these results are essentially indistinguishable.

CONCLUSIONS & OBSERVATIONS

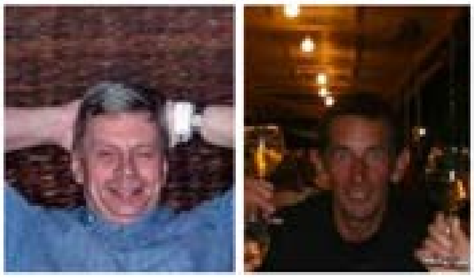
All three algorithms examined for the identification test demonstrated robust performance. Microsoft performed best with identification, demonstrating a stronger TPIR on conventional data and a marginally stronger TPIR on SNS-style data. Kairos displayed the weakest performance on both datasets, although it outperformed Rekognition at higher ranks for TPIR.

As discussed above, Rekognition's DET performance may be attributable to the manner in which low-scoring comparisons are quantized. In most cases, if the genuine comparisons had scored marginally higher (e.g. 1.0 instead of 0.0) identification performance would have likely been competitive with Microsoft. This assumes that calibration is the culprit; it may also be the case that there is some subset of images that provide trouble for Rekognition's face recognition algorithm. But given their distribution of match scores and the observed cap on FAR, this seems to be the likely cause. Results suggest that with improvements to its calibration, Rekognition performance could improve further, broadening the range of viable implementations.

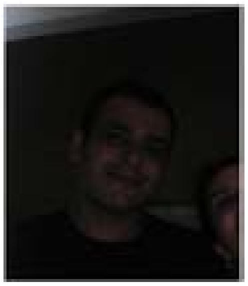
ANNEX A:
FRAME-ID IMAGES THAT FAILED ENROLLMENT / EXTRACTION

FRAME-ID images that failed to enroll are shown below.

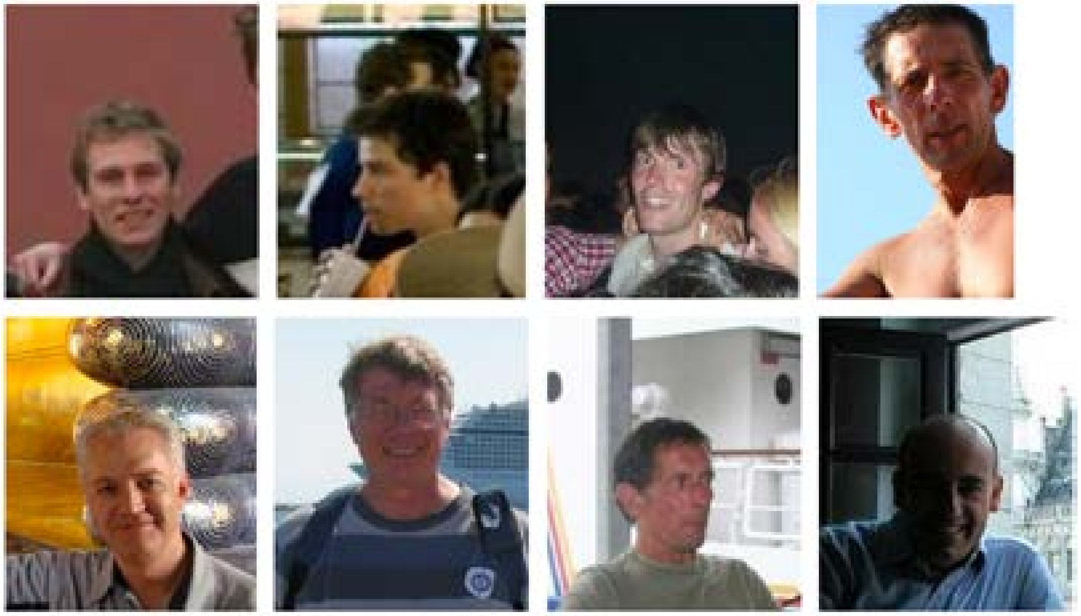
Rekognition (2 FTE on FRAME-ID images)



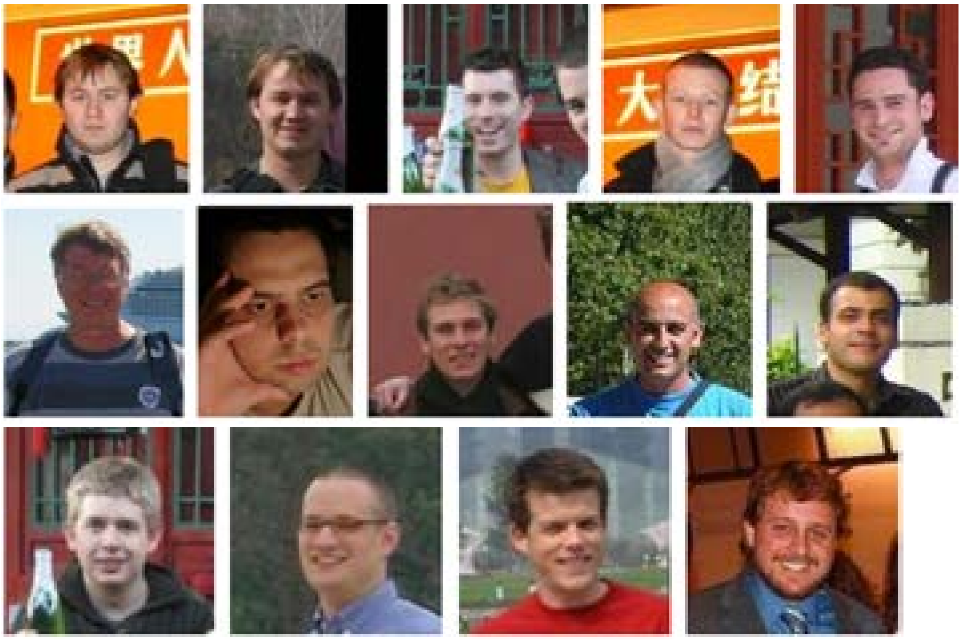
VeriLook (1 FTE on FRAME-ID images)



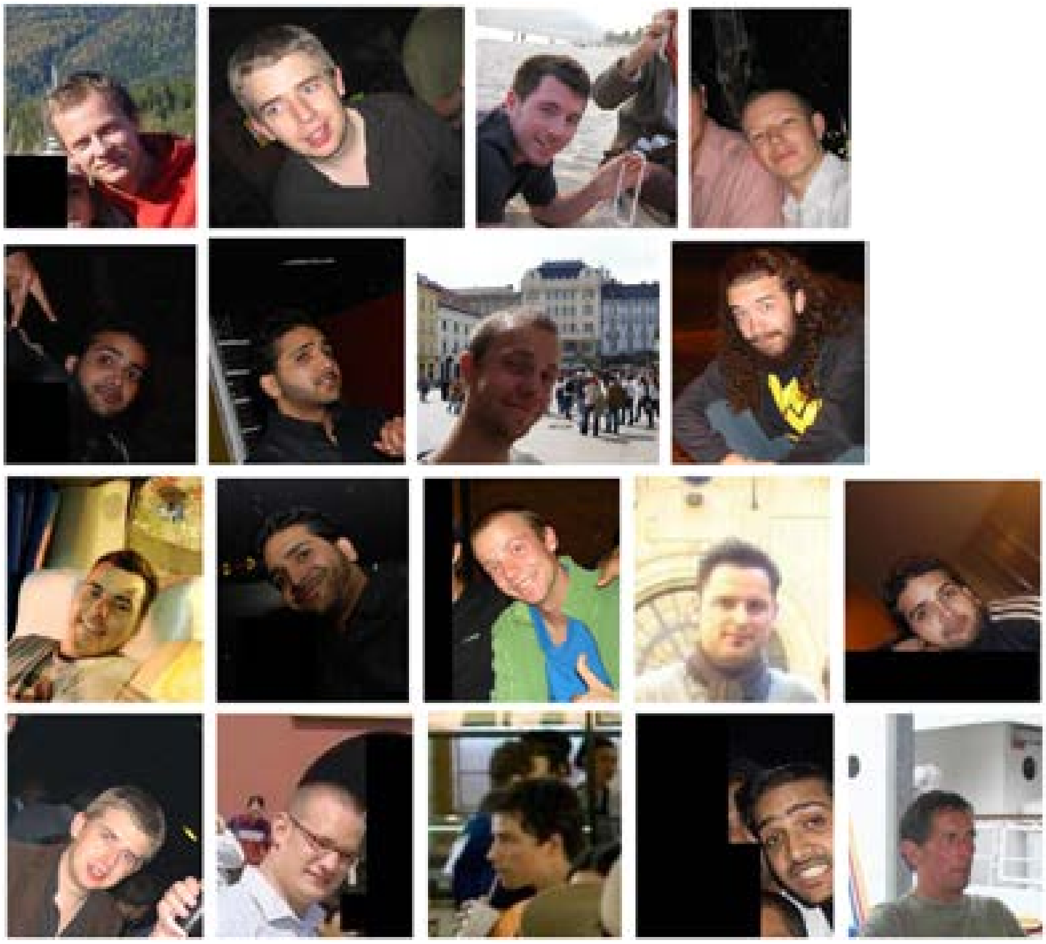
PittPatt (8 FTE in FRAME-ID Images)



Microsoft (14 of 24 FTE in FRAME-ID Images shown)



Kairos (17 of 27 FTE shown)



NOVETTA

From Complexity to Clarity.

Headquartered in McLean, VA with over 700 employees across the US, Novetta has over two decades of experience solving problems of national significance through advanced analytics for government and commercial enterprises worldwide. Grounded in its work for national security clients, Novetta has pioneered disruptive technologies in four key areas of advanced analytics: data, cyber, open source/media and multi-int fusion. Novetta enables customers to find clarity from the complexity of 'big data' at the scale and speed needed to drive enterprise and mission success. Visit www.novetta.com for more information.

7921 Jones Branch Dr, Suite 500
McLean, VA 22102

(571) 282-3000
novetta.com

@novettasol
novetta