



Face Detection: Accuracy on the WIDER FACE Benchmark

Measured 28 October 2026 · iApp Technology Co., Ltd., Bangkok, Thailand

Executive summary. The iApp Face Detection API finds the faces in a photograph and returns a box and a confidence score for each. We measured the production service on the full validation set of WIDER FACE, the standard benchmark for face detection in everyday scenes from portraits to dense crowds: 3,226 images, scored with the official evaluation code. The service returns only faces it is highly confident about, and on those it is precise: 97.9% to 98.9% of the boxes it returned were correct faces. It found 85.6% of the 5,109 faces at least 64 pixels tall (95% CI [84.3, 86.9]) and 66.3% of faces 32 to 63 pixels tall, but few smaller ones. Its average precision is 80.5% on the Easy subset, 72.6% on Medium and 37.0% on Hard, which counts faces down to a few pixels; 67.6% of the annotated faces are under 32 pixels tall. For portraits, selfies and identity documents, the use the API is built for, the figures for larger faces are the ones that apply.

1 Introduction

Face detection is the first step of every face service: verification, recognition and liveness checks all start by locating the face. The iApp API offers two endpoints: `/single` returns the most prominent face, and `/multi` returns every face it detects with a confidence score. This paper measures `/multi` on a public benchmark whose images range from single portraits to crowds of several hundred people, so that the result can be read for the image sizes and scenes a customer actually has.

2 Test Data

WIDER FACE [1] contains 32,203 images in 61 event categories, with faces annotated down to a few pixels. Its validation set of 3,226 images has public annotations; the test-set annotations are withheld. The official evaluation defines three nested subsets by difficulty, which differ in how small, occluded or blurred the faces are: Easy (7,211 faces), Medium (13,319) and Hard (31,958). The dataset is published under the Creative Commons BY-NC-ND 4.0 licence; we used it for evaluation only, and no image is reproduced here.

3 Method

Service. Every validation image was sent once, as distributed, to the production detection service that the public `/v3/store/ekyc/face-detection` endpoints call, one request at a time. A seeded sample of 300 images was also sent through the public API to confirm that it returns the same boxes (Section 6).

Operating point. The service returns a face only when its confidence exceeds a fixed threshold; the lowest score among the 11,998 returned boxes was 0.950. Faces below the threshold are not returned, so the precision–recall curve ends where the service’s threshold lies; average precision (AP) counts the missing part of the curve as zero.

Scoring. We ported the official WIDER FACE evaluation code to Python line by line: intersection over union of at least 0.5, scores normalised over the whole set, 1,000 score thresholds, and faces outside the subset being scored ignored rather than counted as false detections. On the ground-truth boxes themselves, with random scores, the port gives an AP of 99.99% on every subset. We also report recall and precision at

the service's operating point, and the share of annotated faces found, by face height in pixels (faces marked invalid in the annotations excluded).

4 Results

Table 1: WIDER FACE validation, official evaluation. Recall and precision are at the service's operating point (every returned box).

Subset	Faces	AP (%)	Recall (%)	Precision (%)
Easy	7,211	80.5	80.8	97.9
Medium	13,319	72.6	72.8	98.7
Hard	31,958	37.0	37.1	98.9

Table 2: Annotated faces found ($\text{IoU} \geq 0.5$), by face height.

Face height (pixels)	Faces	Found	Found (%)
under 16	15,266	110	0.7
16–31	11,184	2,376	21.2
32–63	7,553	5,011	66.3
64–127	3,191	2,607	81.7
128 or more	1,918	1,766	92.1

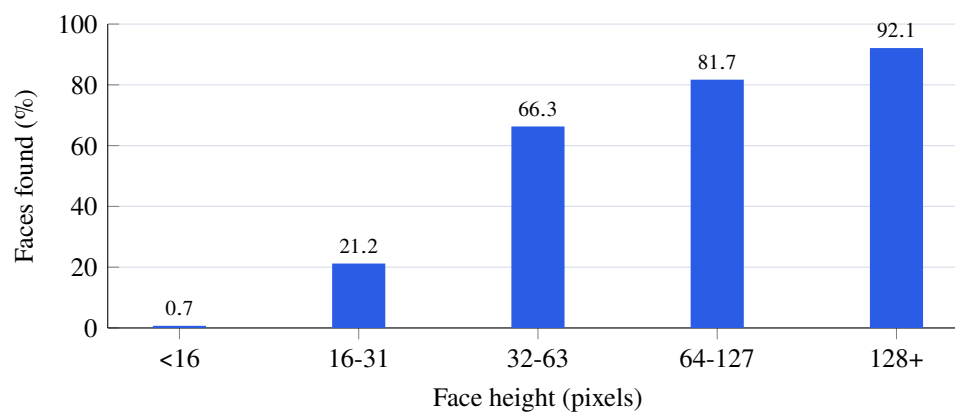


Figure 1: Share of annotated WIDER FACE validation faces the service found, by face height.

Table 3: Service replies over the 3,226 images.

Reply	Images
Faces returned	2,918
No face found (HTTP 423)	308
Processing time limit exceeded (HTTP 561)	0
Other errors	0
Most faces returned for one image	101

5 Discussion

Few false detections. At the service's operating point, 97.9% of the returned boxes on the Easy subset, and 98.9% on the whole set, matched an annotated face. A caller can use every returned box without further filtering.

Size decides recall. The service found 85.6% of the 5,109 faces at least 64 pixels tall and 66.3% of those 32 to 63 pixels tall; below 32 pixels it found 9.4%. Faces under 32 pixels make up 67.6% of the annotated faces, which explains the AP of 37.0% on the Hard subset. The service returned no face for 308 of the 3,226 images; in 129 of them every annotated face is under 32 pixels tall.

What the result means for callers. For a selfie, an identity-card photograph or a portrait, in which the face fills a large part of the frame, the relevant figure is the recall for large faces. The service is not designed for counting people in crowds; `/multi` returned at most 101 faces for one image.

The claim the evidence supports: on WIDER FACE validation the service returned almost no false faces (precision 98.9%), found 85.6% of faces at least 64 pixels tall, and reached an AP of 80.5% on the Easy subset.

6 Production Performance

We sent a seeded sample of 300 images to the public endpoint `/v3/store/ekyc/face-detection/multi` at `api.iapp.co.th`. It returned the same number of boxes and the same reply as the production service for 300 images; the largest difference in any box coordinate or score was 0.000. HTTP statuses: 200: 275, 423: 25. From a client in Bangkok the median end-to-end time was 0.58 s (90th percentile 1.08 s). The service's own processing time had a median of 0.37 s and a 90th percentile of 1.12 s per image. `/multi` costs 0.3 IC per image and `/single` 0.2 IC.

7 Limitations

- **Benchmark versus use.** WIDER FACE is dominated by crowds and small faces; most identity and selfie images contain one large face. The results by face height are the better guide for those images.
- **Fixed operating point.** The service does not return low-confidence faces, so AP cannot credit detections it would make at a lower threshold. Systems evaluated on WIDER FACE usually return every candidate box.
- **Box convention.** A face counts as found only if the box overlaps the annotation by at least 0.5 IoU; a correct detection drawn tighter or looser than the annotation style can count as a miss.
- **No Thai data.** WIDER FACE images come from the web; Thai faces and Thai documents are not specifically represented.

8 Reproducibility

The scripts `get_data.sh`, `bench.py` and `gen_results.py`, and the returned boxes and scores for every image (no images, no face crops), are available from iApp Technology on request: sale@iapp.co.th.

References

- [1] S. Yang, P. Luo, C. C. Loy, X. Tang. *WIDER FACE: A Face Detection Benchmark*. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.

About iApp Technology. iApp Technology Co., Ltd. is a Thai artificial-intelligence company providing OCR, document understanding, speech, identity verification and large-language-model services for the Thai language.
Web: iapp.co.th · E-mail: sale@iapp.co.th · Tel: +66 2 124 4041