AffectNet

A Database for Facial Expression, Valence, and Arousal Computing in the Wild

Ali Mollahosseini, Behzad Hasani & Mohammad H. Mahoor

IEEE Transactions on Affective Computing, 10(1), 18–31 (2019)

Department of Electrical and Computer Engineering, University of Denver

Corpora II: naturalistic and in-the-wild data

Why this paper matters

It made dimensional facial affect a trainable problem

1M+

facial images with detected faces and 66 landmark points

450,000

images manually annotated by twelve trained experts

1,250

emotion-related queries across six languages

8 + 3

emotion categories, plus None, Uncertain and Non-face

The gap it fills

In-the-wild facial databases of the period were annotated with categories only, or with automatically inferred action units, or — where they carried valence and arousal — contained a few dozen subjects recorded in a laboratory. Nothing offered both models of affect, at scale, with manual annotation, outside controlled conditions.

The claim, and its most important qualifier

AffectNet is by far the largest manually annotated database of facial affect in the wild, and it made deep dimensional affect models trainable. It is also, by the authors' own measurement, a database whose two annotators agreed on the emotion category only 60.7% of the time.

1 Motivation

Three models of affect, and what the field was missing

Three ways to describe a face

The paper's framing, and the case for the dimensional model

Categorical

Choose from a list — Ekman's six basic emotions plus neutral, or an extended set.

Simple and interpretable. But mixed emotions cannot be transcribed into a limited set of words, and the model carries no notion of intensity.

Dimensional

Place the affect on continuous scales: valence for how positive or negative, arousal for how activating or calming.

Distinguishes subtly different displays and encodes intensity continuously. Expensive to annotate and needs trained annotators.

Facial Action Coding

Describe the visible muscle movements as action units, without naming an emotion at all.

Objective and fine-grained. But it describes movement, not affective state — mapping action units to emotion takes a further step.

Why the dimensional model needed a database

Categorical data existed at scale — FER-2013 had nearly 36,000 web images. Action-unit data existed at scale, though mostly annotated automatically. Dimensional data existed only in small laboratory corpora of a few dozen subjects. AffectNet exists to close that specific gap, and it annotates both models on the same images so they can be studied together.

The in-the-wild landscape before AffectNet

Every database covered one model, or few subjects, or few samples of hard classes

AFEW / SFEW	Video clips from 54 movies, 330 subjects, seven categories. The static subset has only 700 images and 95 subjects.
FER-2013	Roughly 36,000 web images at 48×48 pixels in greyscale — too low-resolution for landmark detection, only 547 disgust images, categories only.
FER-Wild	24,000 higher-resolution web images with landmarks, annotated by two labellers into seven categories. Still categorical only, still few disgust and fear samples.
EmotioNet	A million web images with action units; 100,000 manually coded, the rest annotated automatically at roughly 80% accuracy. Emotion categories are derived from action units, not labelled.
Aff-Wild	Over 500 YouTube videos annotated frame by frame for valence and arousal — the only dimensional in-the-wild database, but with only 500 subjects.

The recurring gap: dimensional annotation at scale, on still images, with many distinct subjects.

2

Collection

Querying the Internet for faces

Query design

The part of the method that determines what the dataset contains

Building the query set

Emotion words were combined with terms for gender, age and ethnicity, producing about 362 English strings — 'joyful girl', 'blissful Spanish man', 'furious young lady', 'astonished senior'.

These were then translated into Spanish, Portuguese, German, Arabic and Farsi, giving 1,250 queries in total.

Why translation was not enough

Direct translation did not reliably return the intended emotion, because each language and culture words emotion differently.

So native speakers proficient in English wrote their own query lists and inspected the results visually. The criterion for a good query: a high proportion of human faces showing the intended emotion, rather than drawings or objects.

- Negative terms were appended to suppress non-human results — 'drawing', 'cartoon', 'animation', 'birthday'.
- Stock photo sites were filtered out entirely, because their images are unnaturally posed and usually watermarked.
- No separate neutral query was run: enough neutral faces already came back from the emotion queries.
- Google, Bing and Yahoo were used. Baidu and Yandex were considered but either returned few suitable faces or lacked an API for bulk querying.

From URLs to faces

The processing pipeline, and what it says about the images

~1,800,000

distinct URLs returned and stored across all queries

1,000,000+

images with at least one detected face and 66 landmark points

425 × 425

average face resolution (standard deviation 349 × 349)

The tooling

OpenCV face detection produced the bounding boxes. Facial landmarks came from a regression-based local binary feature alignment method, trained on the annotations from the 300-W competition.

Only images where a face and its landmarks were successfully extracted were kept.

One consequence worth noticing

Every filtering stage is also a selection stage. Faces that the detector missed — extreme poses, poor lighting, heavy occlusion, small faces, and any demographic the detector handles less well — never entered the database.

'In the wild' means as wild as a 2016 face detector could cope with.

The same structural point we made about MSP-Podcast, in a different modality: the automatic front end that makes harvesting possible also silently determines who is in your dataset. There the filter was signal-to-noise and diarisation; here it is a face detector.

Who is in the images

Estimated automatically on a 50,000-image sample

- Roughly half the faces are estimated to be men.
- Mean estimated age 33 years, with a standard deviation of about 17. Only 3.9% fall in the 10-to-20 range, while 30% are estimated at 20 to 30.
- Occlusion is rare: forehead 4.5%, mouth 1.1%, eyes 0.5%. Under 10% wear glasses.
- Roughly half the faces are recorded as wearing eye make-up and 41% lip make-up.
- Head pose is close to frontal on average — estimated mean pitch, yaw and roll all within about a degree of zero.

How to read these figures

They come from a commercial face API applied to a random sample, not from human annotation. They are estimates of perceived attributes, with all the error and bias that implies.

But the shape is informative

Near-frontal, unoccluded, well-lit, adult, frequently made up. This is what published photography looks like — not what a camera in a room looks like.

No ethnicity statistics are reported, despite ethnicity terms being used in the queries. That absence is worth noting for a dataset whose central claim is diversity in the wild.

3

Annotation

Twelve experts, and a deliberate rejection of crowdsourcing

Why not crowdsourcing

The opposite decision to the one MSP-Podcast made

The argument

Crowdsourcing is fast and cheap, but label quality varies considerably between annotators — and annotating valence and arousal requires a real understanding of the concepts.

So the team hired twelve full- and part-time annotators at their own university instead.

How they were trained

A written tutorial covering both models of affect with worked examples, then three training sessions in which each annotator labelled 200 images and had their results reviewed with feedback.

Annotators were supervised throughout, with guidance provided whenever they were uncertain.

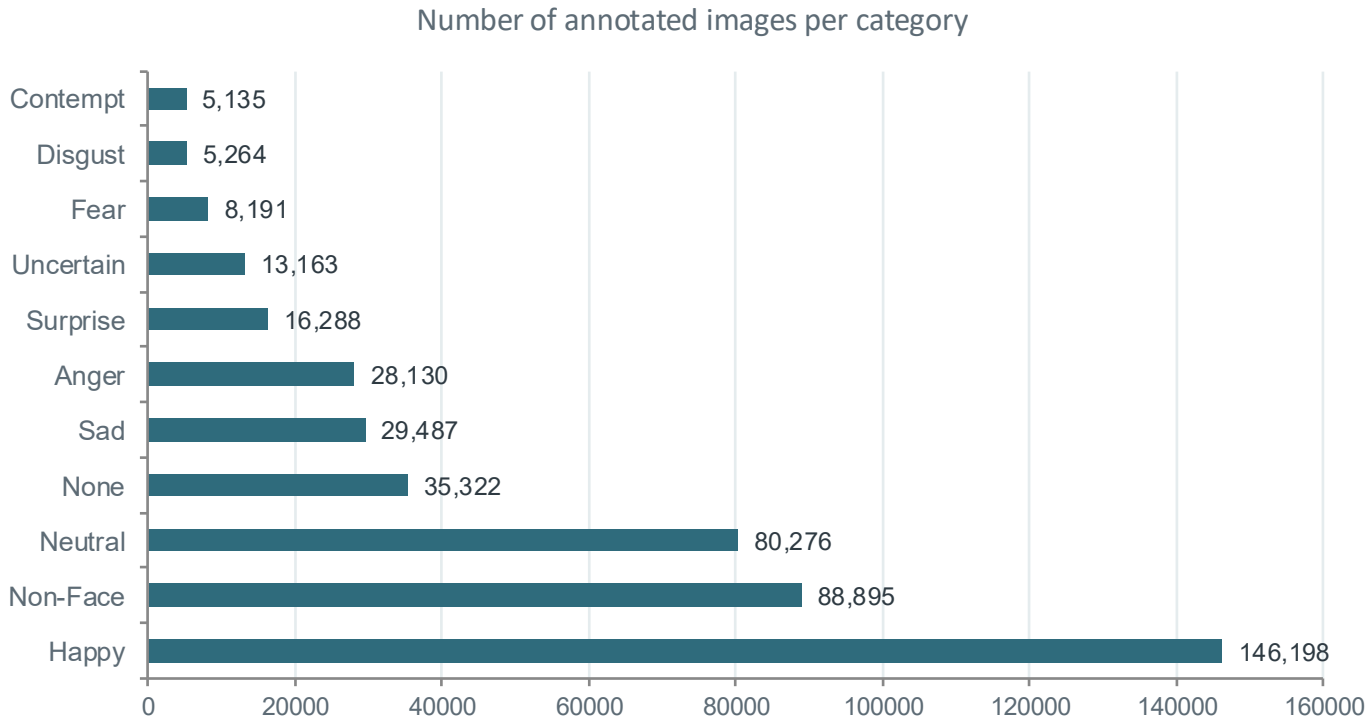
And the price paid for it

Due to time and budget constraints, each image was annotated by one annotator. Only 36,000 of the 450,000 — eight percent — were seen by two.

So for 92% of the database, the label is one trained person's judgement of one still image, with no second opinion and no way to detect a mistake. Compare MSP-Podcast: at least five raters per segment, with the disagreement itself recorded.

Eleven categories

Three of which are not emotions



- None means an expression that is real but fits no listed category — sleepy, bored, focused, ashamed. Valence and arousal are still assigned.
- Uncertain means the annotator could not decide. No valence or arousal is assigned.
- Non-Face covers five distinct problems: no face present, a watermark over the face, a failed bounding box, a drawing or painting, or a distorted face.
- Happy alone is 146,198 images — more than sad, anger, surprise, fear, disgust and contempt combined.
- Contempt and disgust have roughly 5,000 images each, a twenty-eight-fold imbalance against happy.

Nearly a fifth of everything annotated is Non-Face. That is the true cost of harvesting images from search engines, and it is paid in annotator time.

Annotating valence and arousal

A circumplex interface, with one questionable guardrail

The interface

Annotators click a point inside a circumplex; the software will not accept points outside it, so every annotation has a radius between 0 and 1.

The numeric valence and arousal values are displayed as the annotator moves, which the paper says helped them place points precisely and with more confidence.

The guardrail

Each emotion category has a predefined expected region in the plane — for happy, valence in (0, 1] and arousal in [-0.2, 0.5].

If an annotator places a point outside their chosen category's region, the software warns them. They may proceed, and the image is flagged for later review by the authors.

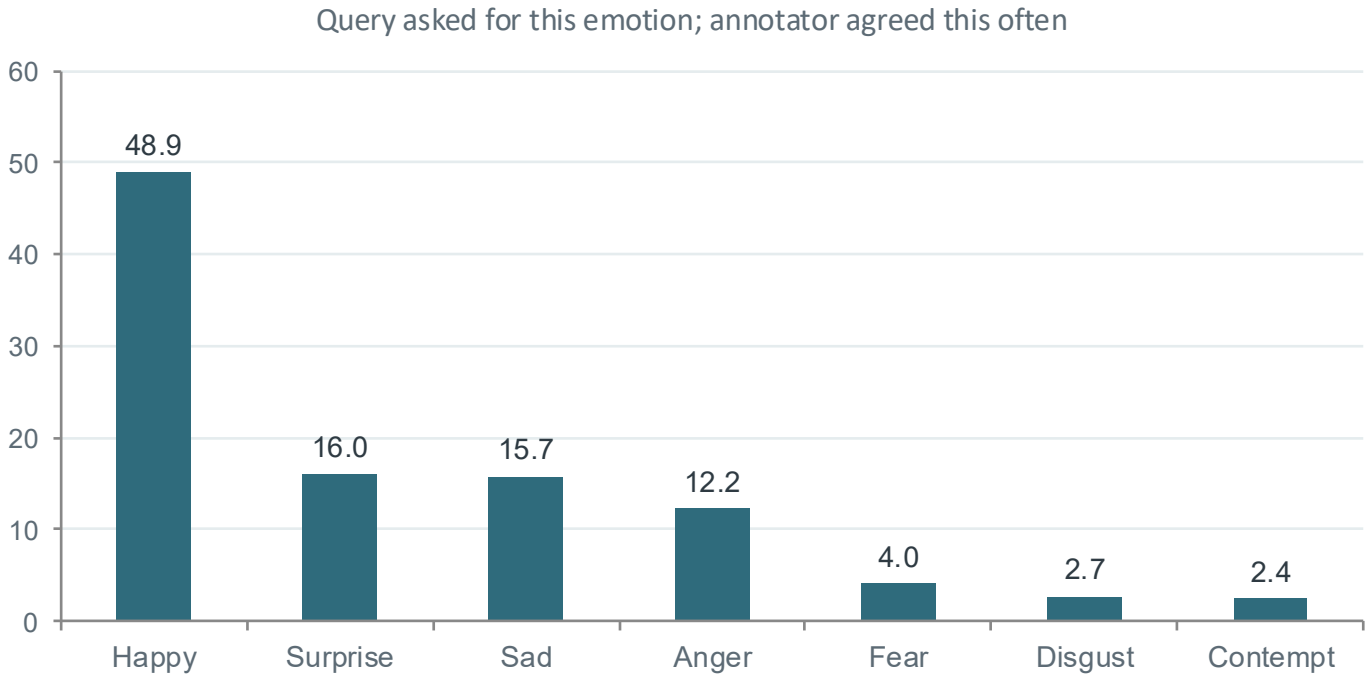
Why the guardrail is worth arguing about

It is presented as an error-prevention device, and it plainly catches genuine mistakes. But it also softly enforces a prior relationship between the categorical and dimensional labels: an annotator who sees a low-arousal happy face is nudged toward a more typical rating.

Any study using AffectNet to investigate how categories map onto the valence–arousal plane should know that the mapping was partly assumed by the annotation tool.

Search queries do not return what you ask for

Percentage of images annotated as the emotion that was queried



- Only happy exceeds 20%. Contempt, disgust and fear queries return the intended emotion under 5% of the time.
- About 15% of every query's results are Non-Face, and roughly 15 to 20% come back as neutral regardless of what was asked.
- Roughly 30% of results for every emotion query are annotated happy — including the queries for sadness, fear and anger.
- The authors' explanation: people publish photographs of themselves looking positive, not looking disgusted.

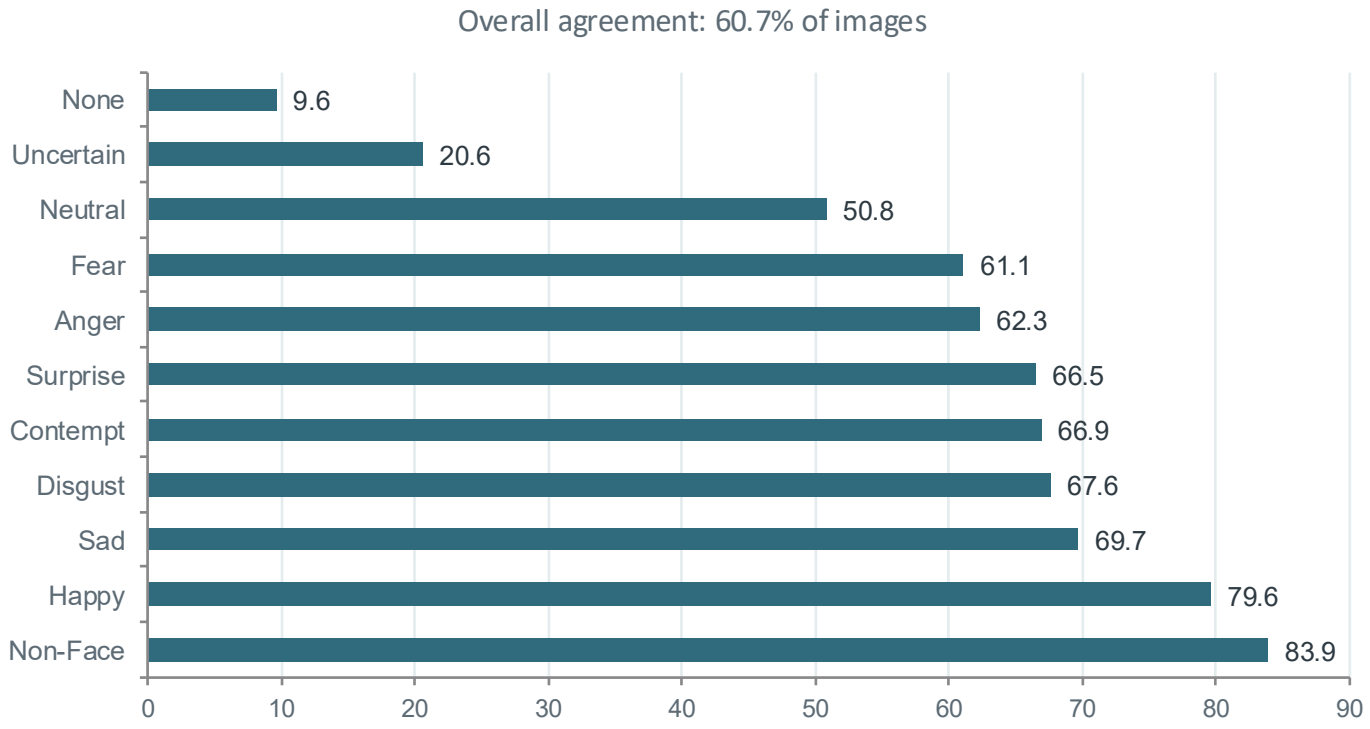
This table is the single best piece of evidence in the paper that manual annotation was necessary. A dataset built by trusting query terms as labels would be wrong more than half the time — and catastrophically wrong for the rare classes.

4 Agreement

How hard is this task, really?

Categorical agreement

36,000 images annotated twice, fully blind and independently



- Neutral is strikingly weak at 50.8% — it leaks into None, Sad and Contempt, because a face with no strong expression is genuinely ambiguous.
- None collapses at 9.6%. The images the authors inspected contain very subtle emotions that are easily confused with other categories.
- Fear and surprise are heavily mutually confused, at 15.3% and 14.0% respectively — the same pair that CREMA-D's perception study found hard.
- Disgust and anger are the other confusable pair, at 13.1%.

The structure of the disagreement is interpretable, not random — which means it reflects genuine perceptual ambiguity rather than careless annotation.

Dimensional agreement

How closely do two annotators place the same face?

	RMSE	Correlation	Sign agreement	CCC
Valence, same category	0.190	0.951	0.906	0.951
Arousal, same category	0.261	0.766	0.709	0.746
Valence, all images	0.340	0.823	0.815	0.821
Arousal, all images	0.362	0.567	0.667	0.551

Valence is easier than arousal

Consistently, on every metric. Judging how positive a face is turns out to be less subjective than judging how activated it is.

Agreeing on the category matters

With the same category chosen, valence correlation is 0.951; across all images it falls to 0.823, and arousal from 0.766 to 0.567.

Note a discrepancy: section 4.3 of the paper quotes annotator RMSE as 0.367 and 0.481, while the table and the conclusion give 0.340 and 0.362. Cite the table.

What the agreement number means for everything else

Human performance is the ceiling, and it is low

For the labels

92% of the database has a single annotator's label. If two experts agree only 60.7% of the time, then a substantial fraction of those single labels would have been different with a different annotator.

For the test set

The test set is the doubly annotated subset. Where the two disagreed, one annotation was picked at random — except that in 29.5% of disagreements, preference was given to the original search query.

For the benchmark

A model reported at 65% accuracy on AffectNet is not obviously worse than a human. The paper says as much: even human annotations carry more error here than automated methods achieve on easier databases.

The middle card deserves particular attention

Resolving annotator disagreement in favour of the search query means part of the test set's ground truth is the query term — the very signal the hit-rate table showed to be unreliable, and the reason manual annotation was needed in the first place.

It affects a minority of a minority: 29.5% of the disagreements within the 36,000-image test set. But it is a systematic tie-break in a known-biased direction, and anyone reporting AffectNet test accuracy should know it is there.

5

Baselines and appraisal

Class imbalance, and what 'in the wild' actually means

The imbalance problem

Four strategies, and a lesson about evaluation metrics

Imbalanced

Train on the data as it comes. Fails badly on contempt and disgust — too few examples to learn from.

Down-sampling

Cap each class at 15,000 images. Some classes have fewer, so the result is only semi-balanced.

Up-sampling

Replicate minority samples until every class matches happy. Contempt is replicated roughly twenty-fold and the network overfits those copies.

Weighted loss

Penalise misclassification in proportion to how rare the class is. Best skew-normalised performance, and much better on the rare classes.

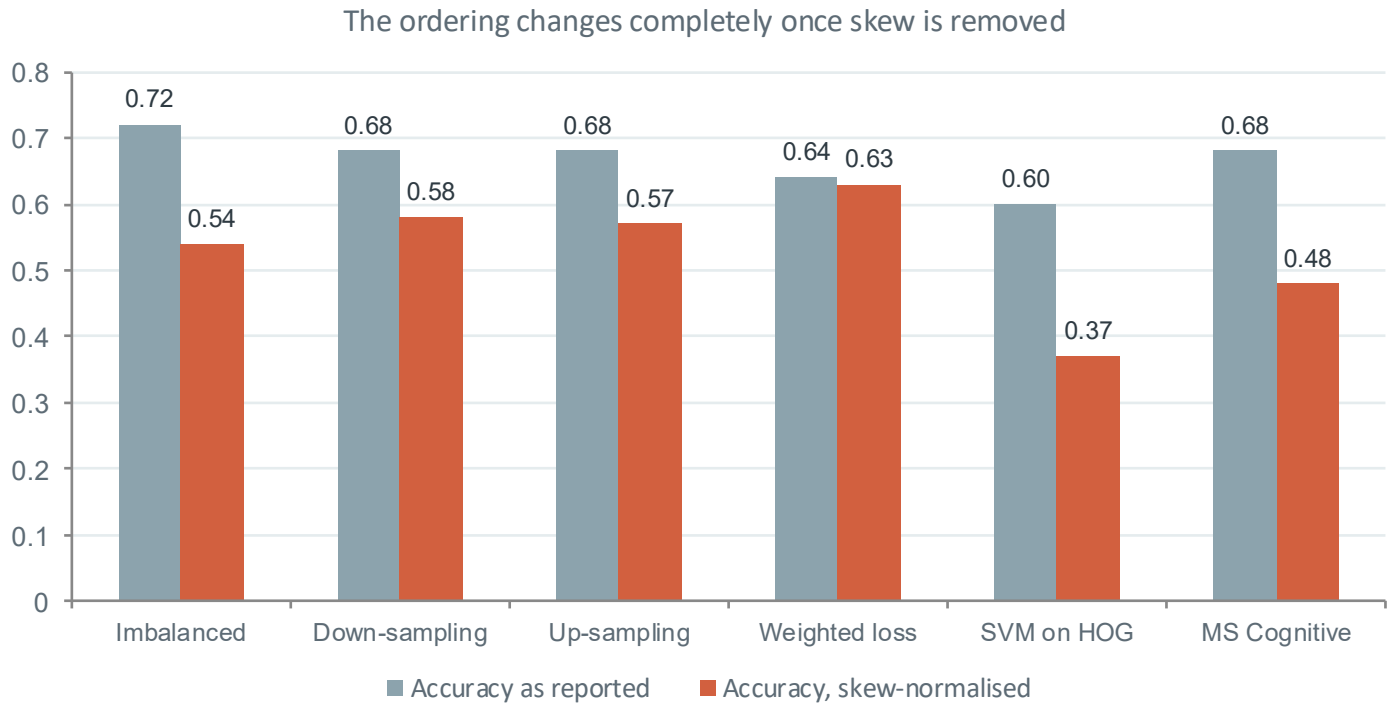
The methodological point behind all of this

The test set is imbalanced too, so the metric is affected as much as the model. Every measure except area under the ROC curve is distorted by skew.

The paper therefore reports both original and skew-normalised scores throughout, normalising by random under-sampling averaged over 200 trials. That practice is worth copying — an accuracy figure on an imbalanced test set tells you as much about the test set as about the model.

Categorical baseline results

Eight-way classification on the test set

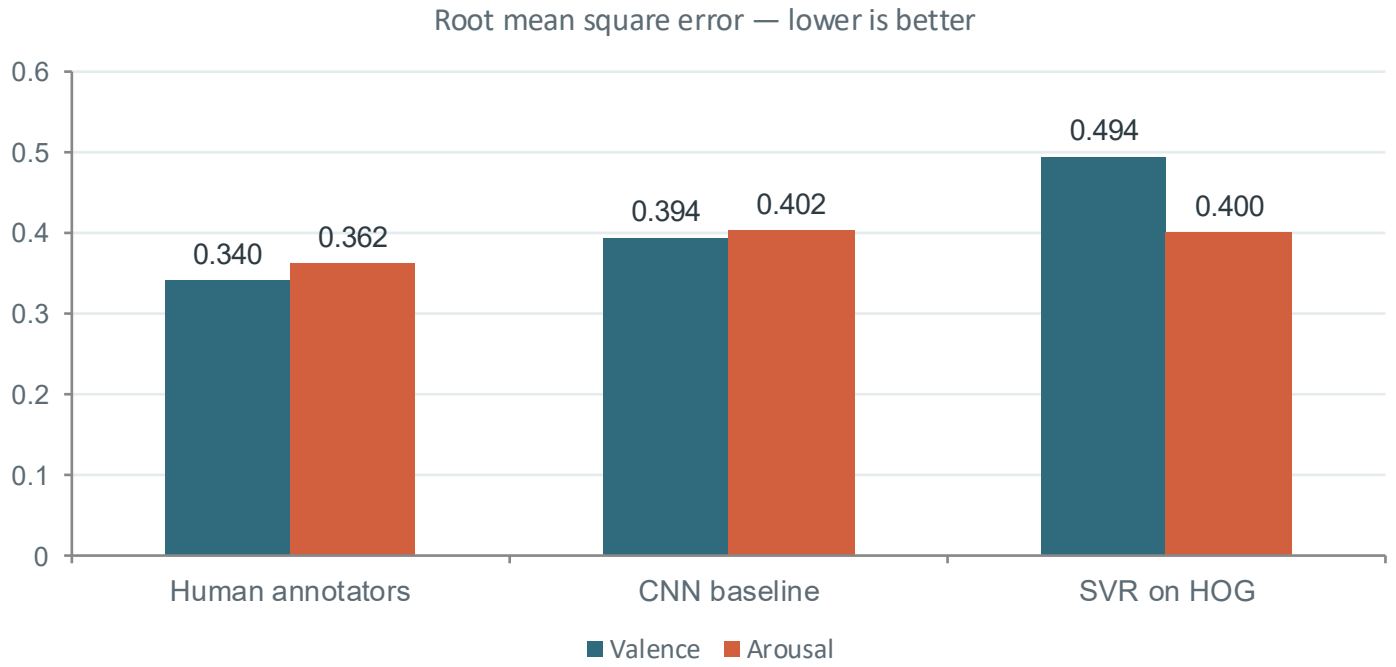


- Imbalanced training looks best on the raw metric and is fourth once skew is removed. That reversal is the whole argument for normalised reporting.
- The weighted-loss network is the only approach whose two bars are close together — a sign it is not exploiting the skew.
- A commercial system of the day scored 0.94 on neutral and 0.85 on happy, but 0.25 on fear, 0.27 on disgust and 0.04 on contempt.
- The CNN baselines beat both HOG features with an SVM and the off-the-shelf commercial system on every metric reported.

The contempt figure for the commercial system — four percent — is the clearest demonstration that in-the-wild affect was an unsolved problem in 2017, not a shipped product.

Dimensional baseline

And how close it comes to the human ceiling



- A plain AlexNet with a Euclidean loss and a single output neuron lands within about 0.05 RMSE of the disagreement between two trained humans.
- Concordance correlation is 0.541 for valence and 0.450 for arousal, against 0.821 and 0.551 between annotators.
- So the model is close to human on error magnitude and further away on agreement — it is roughly right, less often exactly right.
- Error is lowest near the centre of the circumplex. The hardest regions are low-valence mid-arousal and low-arousal mid-valence — contempt, boredom, sleepiness.

Read carefully: the model is not as good as a human. It is as close to one human as another human is. On a task this subjective, those are different claims, and only the second is supported.

Critical appraisal

What the design buys, and what it costs

Strengths

Both models of affect annotated manually on the same images, at a scale that made deep dimensional models trainable for the first time.

Genuine variation in pose, lighting, resolution, background and age — far beyond any laboratory corpus.

Trained, supervised expert annotators with a written tutorial and reviewed practice sessions.

Non-face, uncertain and none categories, so that failure and ambiguity are recorded rather than forced into an emotion label.

Agreement measured and published, including the uncomfortable 60.7% figure, in the abstract's own conclusion.

Skew-normalised metrics reported alongside raw ones throughout.

Limitations

One annotator for 92% of the database, so most labels have no second opinion and no measurable reliability.

The query tie-break imports search-term bias into the test set's ground truth on exactly the hard cases.

The circumplex guardrail softly enforces an assumed mapping between categories and dimensions.

Severe class imbalance — twenty-eight to one — driven by what people publish, not by what people feel.

No subject identity, so the same person can appear in both training and test data. There is no subject-independent split.

No consent from the people depicted, and no ethnicity composition reported despite ethnicity terms in the queries.

What 'in the wild' actually means

The claim both of today's papers rest on

What it does mean

Uncontrolled capture conditions. Variable lighting, pose, resolution, background, camera and subject. No laboratory, no instructions, no director.

What it does not mean

A representative sample of human emotional expression. What it samples is the set of images people chose to publish, that a search engine chose to return, and that a face detector could parse.

The same holds for podcasts

Public, performative speech by people who chose to record and release it, filtered for clean single-speaker wideband audio, and selected by models trained on actors.

The honest formulation

Both corpora replace the experimenter's control with a different, invisible set of controls: publication norms, search ranking, licence terms, and the failure modes of the automatic tools used to filter the raw material.

That is a real improvement over a recording studio — the behaviour is not acted. But 'in the wild' names where the data was captured, not how representative it is.

The two papers side by side

Two answers to the same question

	MSP-Podcast	AffectNet
Modality	Speech	Still images
Source	403 Creative Commons podcasts	Three search engines, 1,250 queries
Scale in the paper	2,317 annotated turns	450,000 annotated images
Selection	Machine learning retrieval for target arousal and valence	None — annotate what the queries return
Annotators per item	At least five crowd workers	One trained expert (two for 8%)
Quality strategy	Prevent bad labels with real-time monitoring	Train annotators, then trust them
Balance achieved	Yes for attributes, no for categories	No — 28:1 between largest and smallest class
Agreement reported	$\alpha \approx 0.46$ valence, 0.23 primary emotion	60.7% categorical, RMSE 0.34 valence
Redistribution	Permissive licence, commercial use allowed	Cropped faces and labels released for research
Official splits	Added in later releases	Test and validation defined; no subject independence

Both harvest existing material. They differ on almost every decision that follows from that.

Take-aways

Four things to carry out of this session

1	**Harvesting moves the bias, it does not remove it**	Publication norms, search ranking, licence terms and automatic filters replace the experimenter — and unlike the experimenter, they are undocumented.
2	**Balance is relative to a label space**	MSP-Podcast balanced arousal and valence and left categories skewed. Always ask: balanced with respect to what?
3	**Retrieval is a use for imperfect models**	A cross-corpus model too weak to deploy is strong enough to sort a list — and a human checks the shortlist afterwards. That idea transfers well beyond speech.
4	**Report your agreement, and your skew**	Both papers do, and both look worse for it in the short term. It is what makes their numbers interpretable at all.

Questions for discussion

1. Fixed annotation budget: 2,317 items with five raters each, or 450,000 items with one. Which do you buy, and does the answer change depending on whether you are building a benchmark or a training set?

2. Two trained experts agree on the emotion category 60.7% of the time. Is a single-annotator label on the remaining 92% of AffectNet a ground truth at all, and what should a paper reporting accuracy on it claim?

3. Both corpora sample what people chose to publish. Name a kind of emotional expression that neither source can contain — and say what it would take to collect it ethically.

4. The circumplex guardrail nudges annotators toward category-consistent valence and arousal. Where else in these two pipelines does the tooling quietly encode an assumption as data?

5. These are photographs and recordings of identifiable people who never consented to being in a training set. What would a defensible version of this data collection look like in 2026?

Mollahosseini, Hasani & Mahoor (2019), IEEE Transactions on Affective Computing 10(1), 18–31 · Lotfian & Busso (2019), same journal, 10(4), 471–483