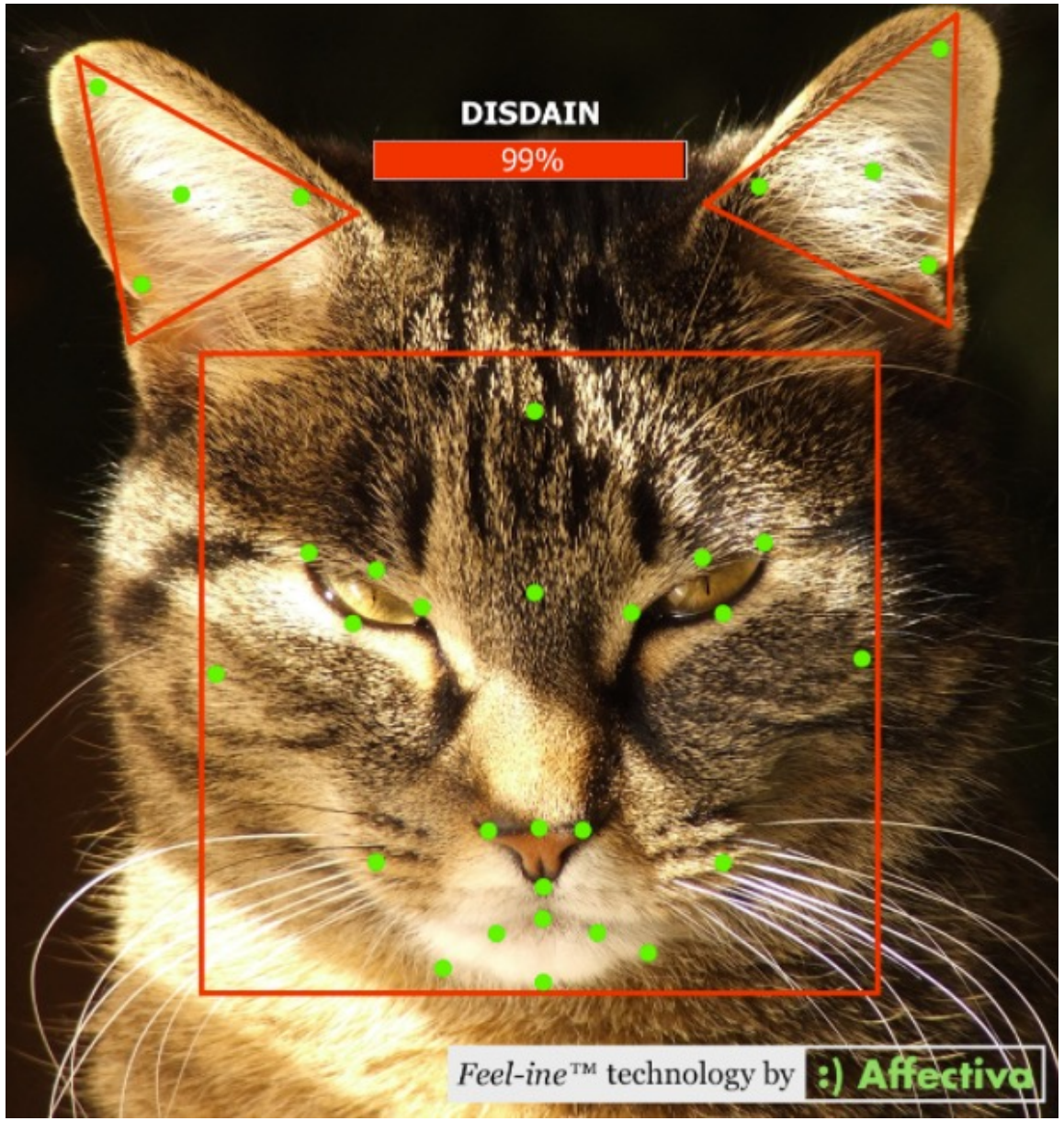
Comparing the Robustness of Humans and Deep Neural Networks on Facial Expression Recognition

Lucie Lévêque, Emmanuel Sampaio, François Villoteau, Matthieu Perreira Da Silva, and Patrick Le Callet

Nantes Université, France

Introduction – emotions

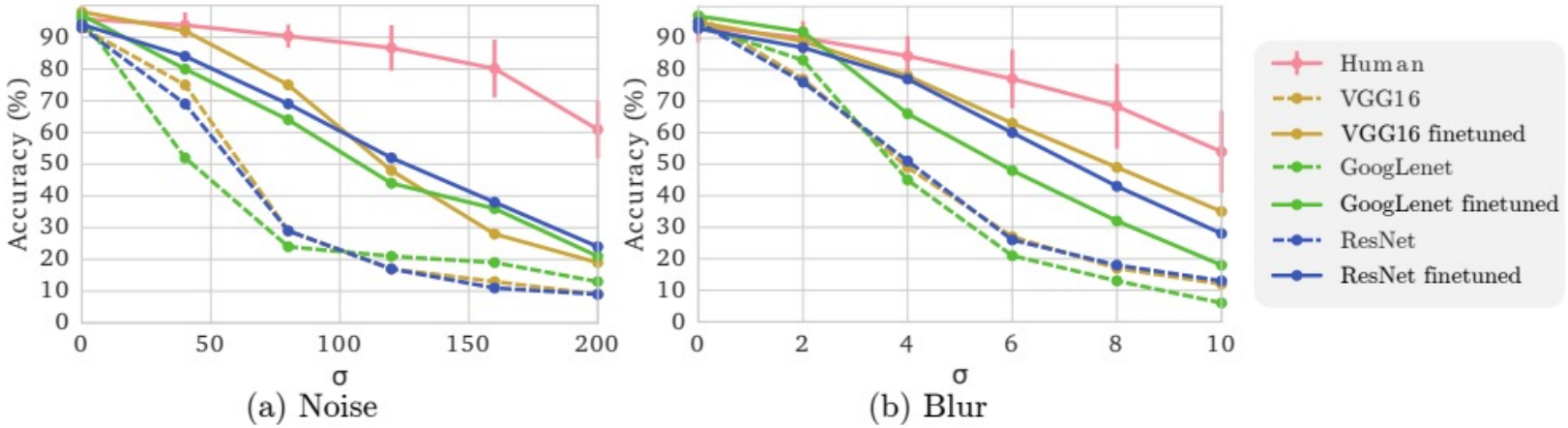
- **Emotions** play an important role in communication
- Emotions have been studied in various fields: psychology, HCI, and computer science (*affective computing*)
- Applications: education, mental health, marketing…
- **Facial expression recognition** (FER) made possible thanks to advances in computer vision, yet is remains a **challenging** task


DISDAIN
99%
Feel-ine™ technology by :) Affectiva

Introduction – humans *vs.* machines

- Comparing the performance of humans and machines can help diagnose cases where machines fail

- Dodge & Karam (2017) compared humans and DNNs on a **classification task** of dog images distorted with Gaussian noise and Gaussian blur at different levels → humans more robust to distortions



Introduction – FER

- Yang *et al*. (2021): 5 commercial systems on distorted images
  → limitations of studied models under certain **manipulations**

- Abate *et al*. (2022): how FER systems deal with face masks
  → **mouth** strongly contributes to emotion classification

- Krumhuber *et al*. (2021): comparison of FACET and humans
  → both perform better on **posed** expressions than spontaneous

- Dupré *et al*. (2020): comparison of 8 commercial models and humans
  → **accuracy** of 62% for best model (FACET), while of 75% for humans

Introduction – contributions

- **DisFER**: new FER dataset composed of distorted images of faces (released soon)
- Large-scale experiment conducted on a **crowdsourcing** platform
- **Comparison** of the performance of humans with pre-trained and fine-tuned open-source FER models
- Discussion on the definition of facial expression **ground truth**

Materials & methods – dataset

- **84 sources images** from FER-2013 dataset (Goodfellow *et al*.)
- 12 images per basic emotions
  → anger, disgust, fear, happiness, neutral, sadness, and surprise (Ekman *et al*., 1992)
- 3 types of **distortion** (GB, GN, SP) at 3 different levels
- Total of **840 stimuli** (including the original)

Materials & methods – dataset

- Example of stimuli:

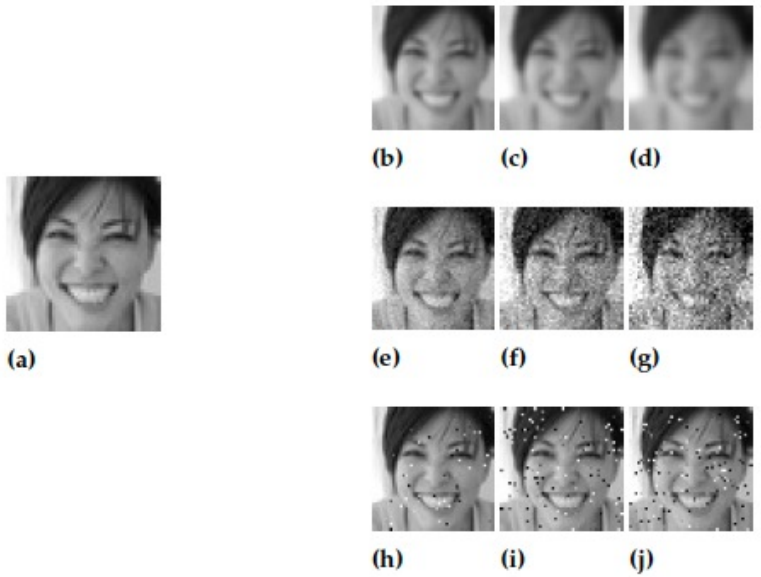


Figure 1. Illustration of sample distorted images used in our study: (a) represents an original image from the FER-2013 dataset; (b), (c), and (d) are its GB versions; (e), (f), and (g) the GN versions; and (h), (i), and (j) the SP versions (at low, medium, and high levels, respectively).

Materials & methods – crowdsourcing experiment

- Rating 840 images is time-consuming and tiring
  → **crowdsourcing** experiment

- DisFER carefully split into **21 playlists** of 40 images
  (same numbers of images of a given consideration)

- **1051 participants** (50% females) were recruited using Prolific
  → 20 playlists watched and rated by 50 distinct participants,
  1 playlist by 51 participants)

Materials & methods – DNNs

- Preliminary benchmark (Sampaio *et al*., 2022):
  2 **open-source pre-trained models**
  → Residual Masking Network (RMN): 51% accuracy on FER-2013
  → DeepFace: 55% accuracy on FER-2013

- Both models **fine-tuned** for each distortion type (GB, GN, and SP)
  → Training phase: sparse categorical cross entropy as loss function
  → 27 batches of 42 images
  → each batch homogeneously divided among the 7 emotions
  (half of a batch: original images, other half: distorted)

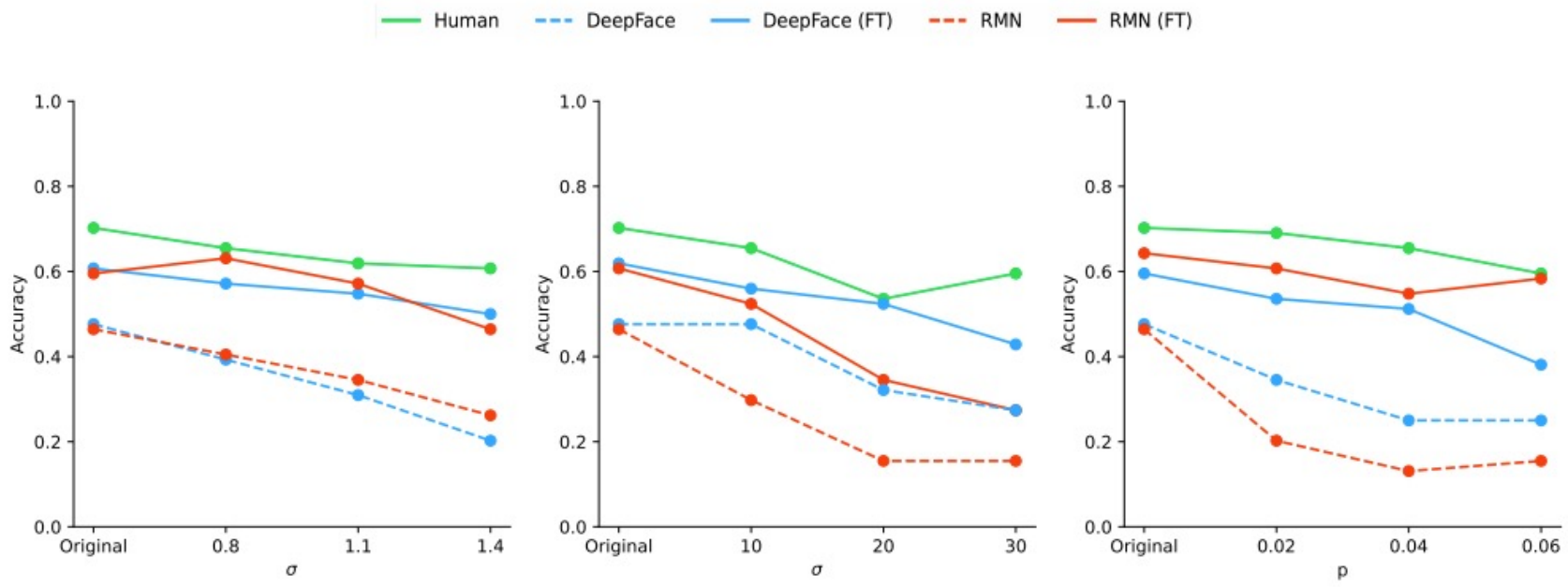
Results – overall

- **Human baseline**: expression selected by the highest number of participants for a given image

- Overall human accuracy using original FER-2013 labels: **63%** (close to human accuracy on FER-2013)

	RMN	RMN FT	DeepFace	DeepFace FT
Accuracy	0.26	0.52	0.33	0.52
Cohen's kappa	0.13	0.43	0.22	0.44

Results – accuracy



(a) Gaussian blur **(b)** Gaussian noise **(c)** Salt-and-pepper

Figure 2. Illustration of the accuracy achieved by humans, pre-trained (i.e., DeepFace and RMN) and fine-tuned (FT) models, for each distortion type (i.e., GB, GN, and SP) and level.

Results – confusion matrices

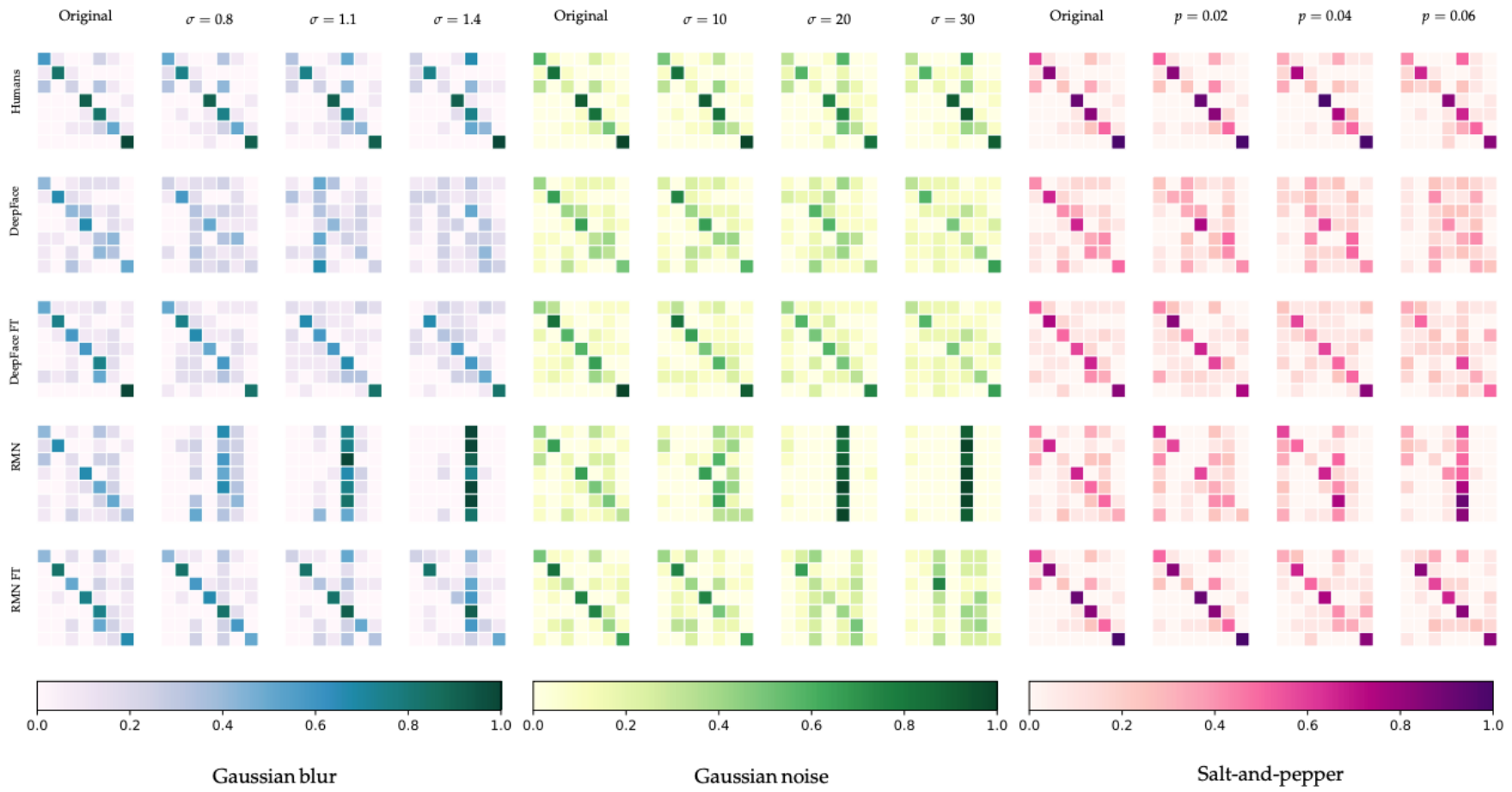


Figure 3. Illustration of the confusion matrices generated for humans, pre-trained (i.e., DeepFace and RMN) and fine-tuned (FT) models, for each distortion type (i.e., GB, GN, and SP) and level. Lines correspond to ground truth labels, while columns represent predicted labels. Emotions are presented in alphabetical order, i.e.: anger, disgust, fear, happiness, neutral, sadness, and surprise.

Discussion – on obtained results

- Human visual system more robust to signal distortions than DNN: **experience** may play an important role in human performance
- When considering individual performance: average accuracy of 55%, lowest of 30% (3 participants), and highest of 80% (1 participant)
- **Poor accuracy**!?
  → questions on labels

Figure 6. Illustration of sample images misclassified by human participants under all of their 10 versions; (a) was classified as anger, but labelled as fear; (b) classified neutral, labelled sadness; (c) classified anger, labelled disgust; and (d) classified fear, labelled disgust.

Discussion – on original labels

- FER-2013 original dataset labelled using Google image search API…
- Question raised: what is **ground truth** for facial expressions?
- Inter-observer **agreement**:
  → strong agreement on 10% of our dataset (mostly happy images), moderate agreement on 60%, and very poor agreement on 10%
  → Fleiss' kappa = 0.6 on our dataset (moderate agreement)
- New ways to define **ground truth**?
  → based on humans' classifications rather than on search engines
  → when humans tend to disagree; 2 (or more) labels?

ACM IMX 2023 held in Nantes next June!

- **Workshop proposals: 15th December 2022**
- **Call for associate chairs: 2nd January 2023**
- **Call for technical papers: 3rd February 2023**
- More information: https://imx.acm.org/2023

Thank you for your attention ☺

- Contact: lucie.leveque@univ-nantes.fr