University of Milano-Bicocca
Master's Degree in Data Science
Digital Signal and Image Management
Academic Year 2022-2023

Emotion Recognition on RAVDESS dataset

Authors:
Gianluca SCURI
Niccolò PUCCINELLI

Task: Emotion recognition

- **7 emotions** are human universally understood [1]
- **Applications**: Improve HRI, monitoring attention while driving, facial expressions codification for blind people, detect depression, ...
- Technology works better if it uses **multiple modalities** (video, audio, words used, body movement, gestures...)
- Chosen modalities:
  - Facial Emotion Recognition (**FER**)
  - Speech Emotion Recognition (**SER**)

[1] Matsumoto, D., Keltner, D., Shiota, M. N., O'Sullivan, M., & Frank, M. (2008). *Facial expressions of emotion*. In M. Lewis, J. M. Haviland-Jones, & L. F. Barrett (Eds.), Handbook of emotions.

Dataset: RAVDESS

- **Clips**:
  - 24 actors gender balanced
  - 60 x 24 = 1440 speech clips (audio + video)
  - ~4/5 sec
  - Standardized sentence and setup
- **Emotions**:
  - 1 emotion per clip
  - 8 classes: neutral, calm, happy, sad, angry, fearful, surprise, and disgust (7 used)
- **Benchmark** dataset [2]
- 16 actors **train**, 4 actors **validation**, 4 actors **test**

[2] Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions

Architecture: Two-stream model

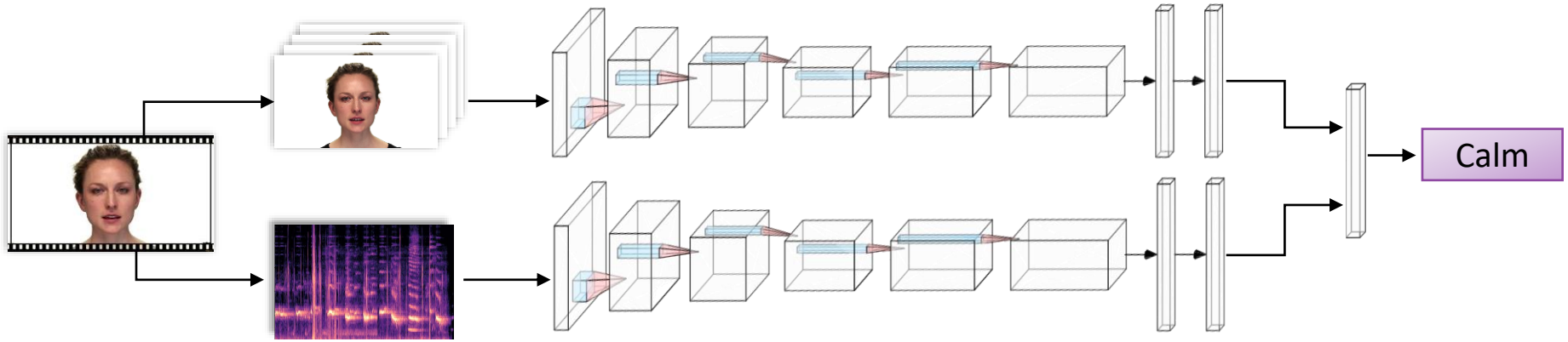
- **Training:**
  - Independent streams
  - 896 training clips
- **Inference:**
  - Frame scores fusion
  - Audio + video scores fusion

Video stream: FER

- 2D CNN
- Frame-by-frame

Audio stream: SER

- 1D and 2D CNN
- Wav and Mel Spectrogram



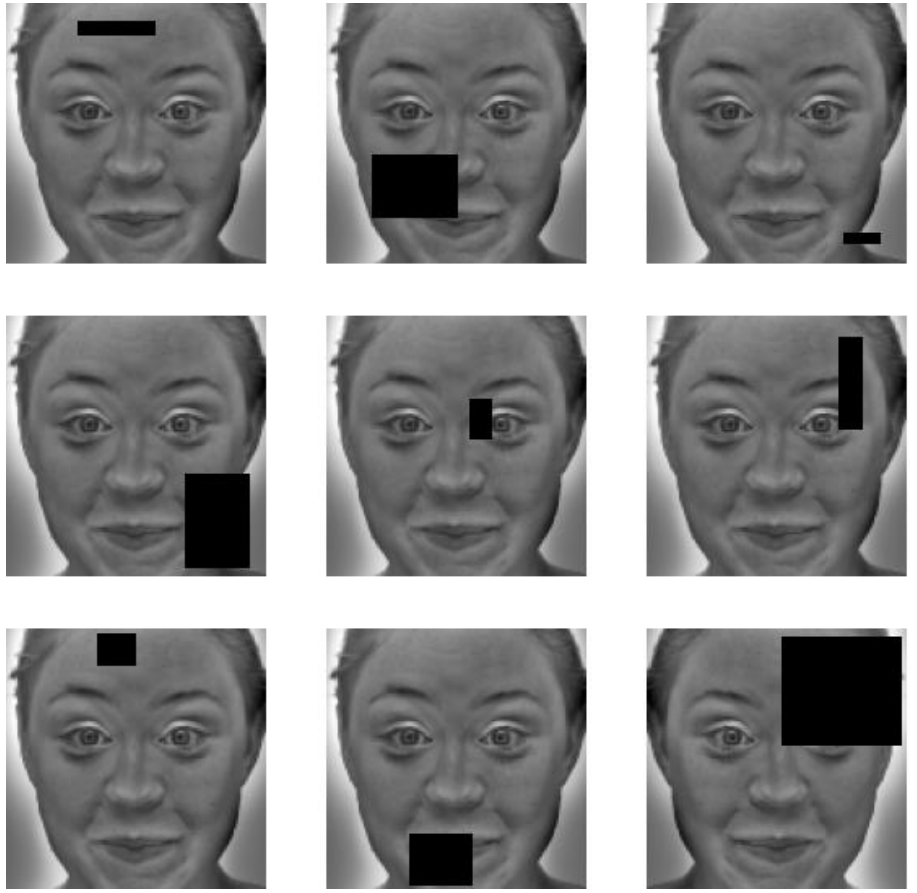
Video stream: Input processing

- **Frames extraction:**
  - 1 frame every 3 excluding the first 20 frames (~1 sec)
- **Training:**
  - 896 clips x 23 frames = 20608 frames (balanced classes)
- **Improvements:**
  - 224x224 → 112x112
  - RGB → GrayScale
  - Full frame → Face only (Haar Cascade)
  - Removed background → removed mean face
- **Key aspects:** mouth, eyes and eyebrows

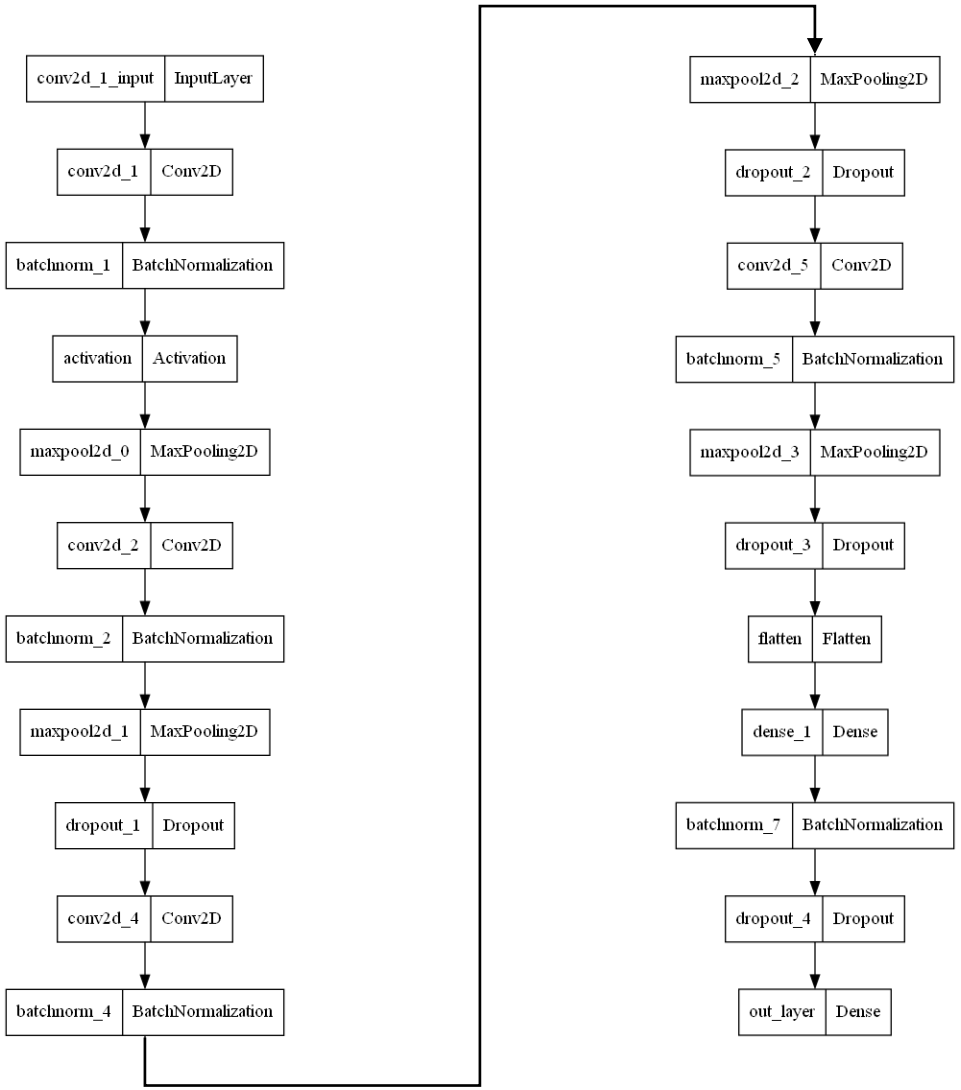



Video stream: Models

- **Before:**
  - **Transfer learning**: ResNet50
- **After:**
  - Training **from scratch**
  - **Grid search 5-Folds CV fine-tuning**: leaving out 4 actors for each fold
  - Data **augmentation**
    - Random Flip
    - Random Cutout
  - Dropout

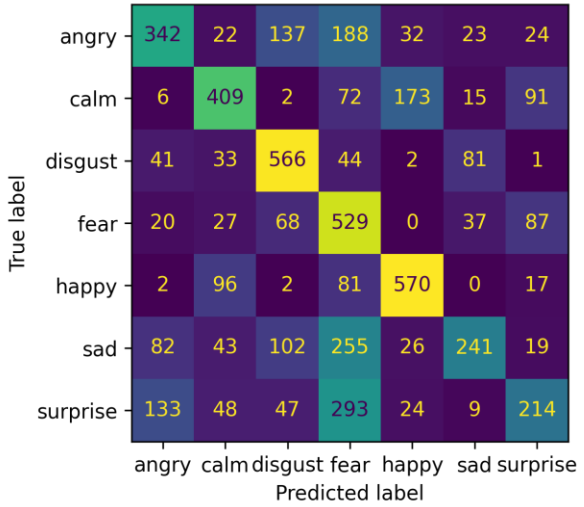


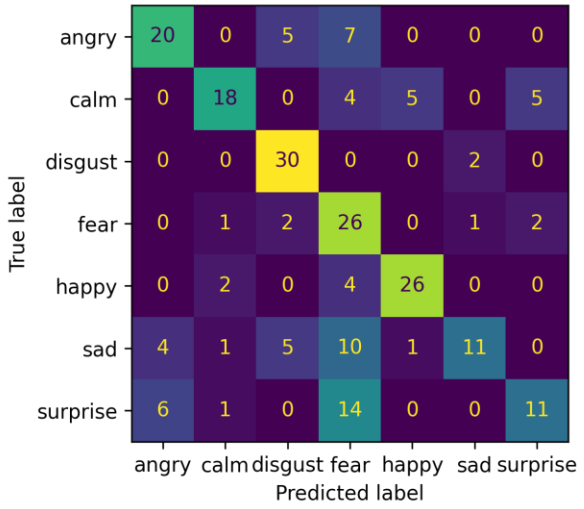
Video stream: Architecture



Video stream: Results

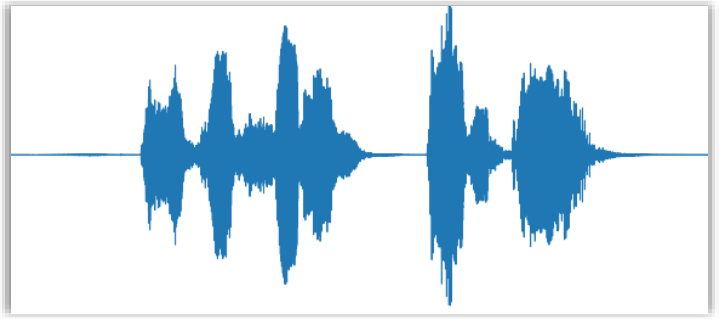
- **Single frame:**
  - Top-1 accuracy: **53.8%**
  - Top-3 accuracy: **83.3%**
- **Full video:**
  - **Mean**, mode, median, weighted mean (highest softmax value)
  - Top-1 accuracy: **62.5%**
  - Top-3 accuracy: **90.1%**

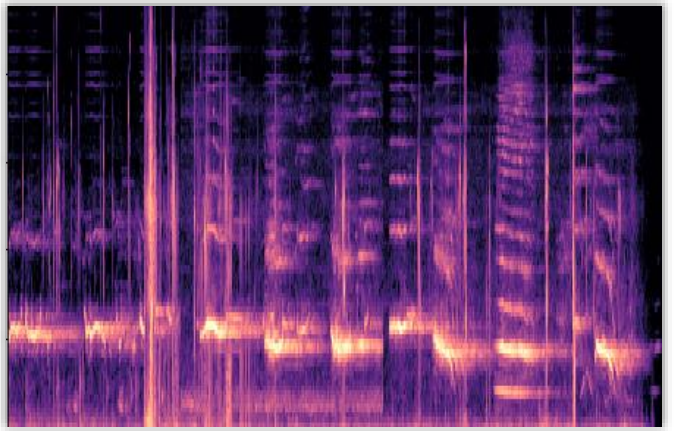




Audio stream: Input processing

- **Audio extraction:**
  - Mel spectrogram **signal** (sampling rate = 48000 Hz) - **1D CNN**
  - Mel spectrogram **image** (dimensions = 128x282) - **2D CNN**
- **Training:**
  - 896 samples (balanced classes)
- **Improvements:**
  - Cut 3 middle seconds
  - Standardization
  - Limit to min frequency (50Hz)
  - No improvements with other features (e.g. mfcc)





Audio stream: Models

1D CNN

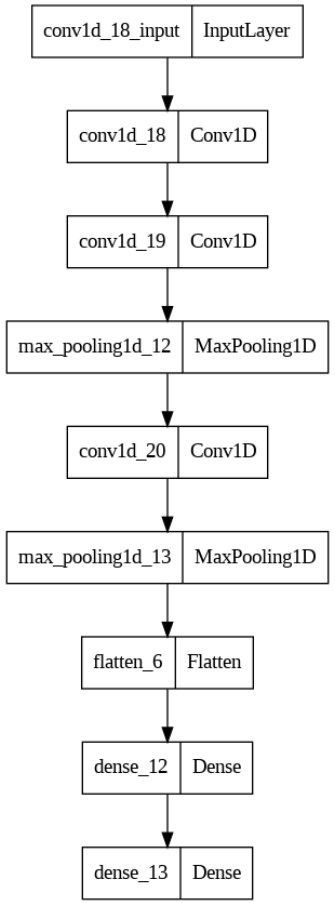
- Training from scratch
- Augmentation on original audio:
  - Noise
  - Pitch
  - Stretch
  - Shift
- **Grid search 5-Folds CV fine-tuning**: leaving out 4 actors for each fold

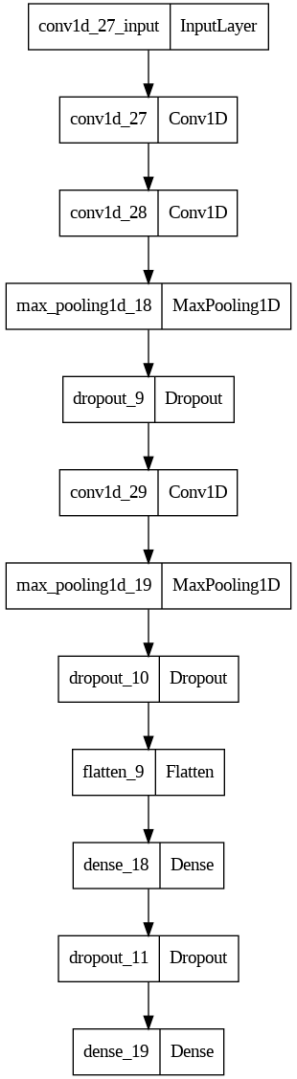
2D CNN

- Training from scratch
- Augmentation on original audio:
  - Noise
  - Pitch
  - Stretch
  - Shift
  - No improvements with image augmentation (e.g. masking on Mel spectrogram)
- **Grid search 5-Folds CV fine-tuning**: leaving out 4 actors for each fold
- Doubled the number of filters
- **Rectangular** kernels

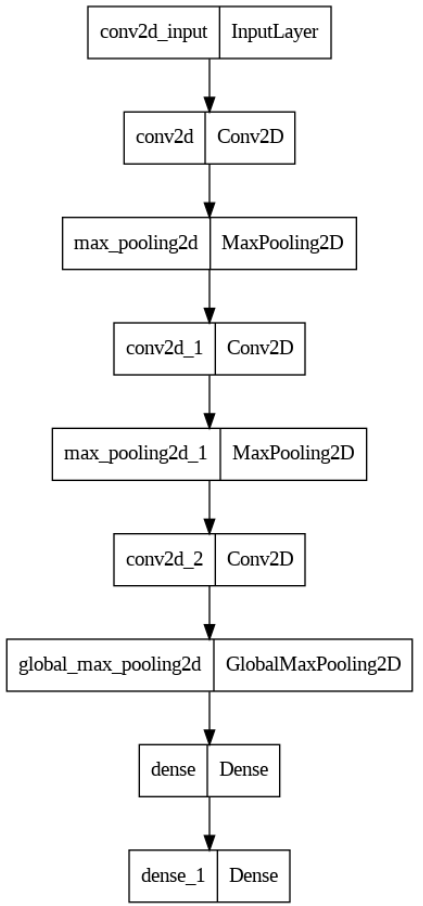
Audio stream: Architecture

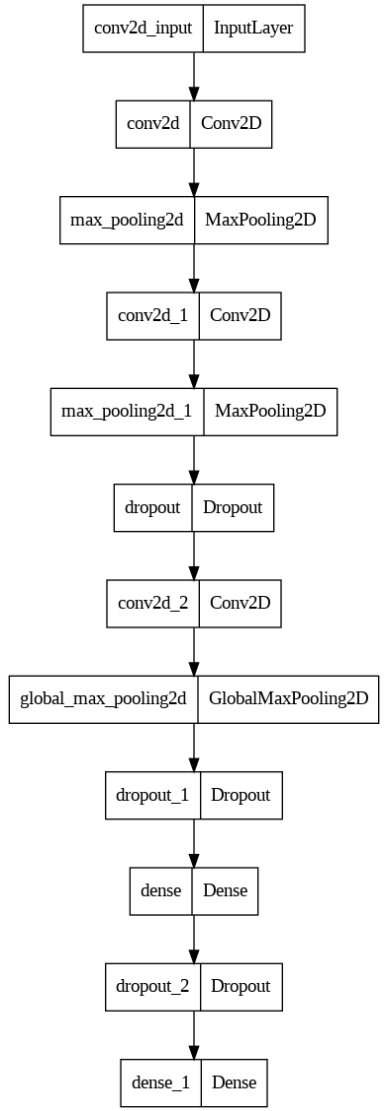
1D CNN





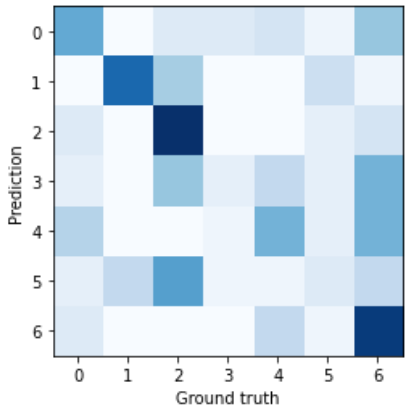
2D CNN



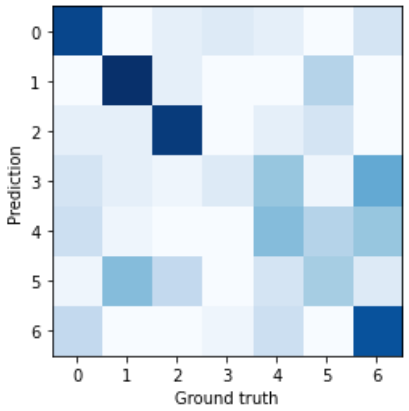


Audio stream: Results

1D CNN

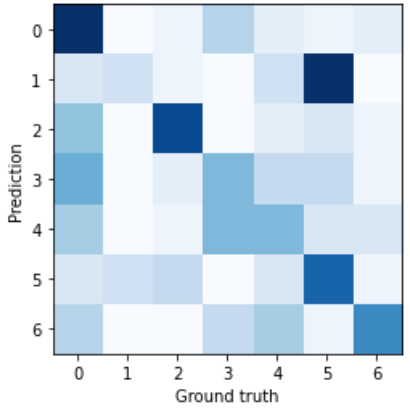


- Vanilla model
- Accuracy:
  - Top-1: 41%
  - Top-3: 79%

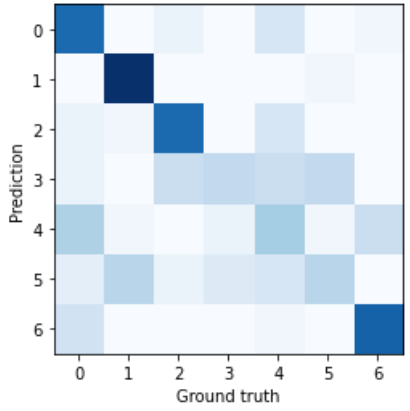


- **Tuned** model
- Accuracy:
  - Top-1: **48%**
  - Top-3: **80%**

2D CNN



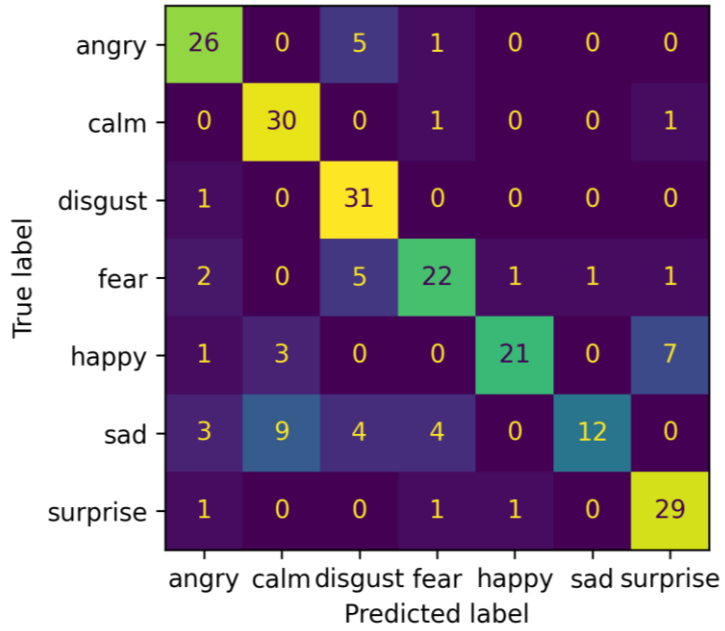
- Vanilla model
- Accuracy:
  - Top-1: 40%
  - Top-3: 78%



- **Tuned** model
- Accuracy:
  - Top-1: **59%**
  - Top-3: **88%**

Full Clip: Late fusion results

- **Mean**, median, weighted mean
- **Accuracy:**
  - Top-1: **76.3%**
  - Top-3: **95.7%**
- Possible further **improvements:**
  - Better **fusion** technique
  - Integrate another **dataset** (more subjects)



Emotion Recognition: Demo