Emotion Recognition in the Wild

from Videos using Images

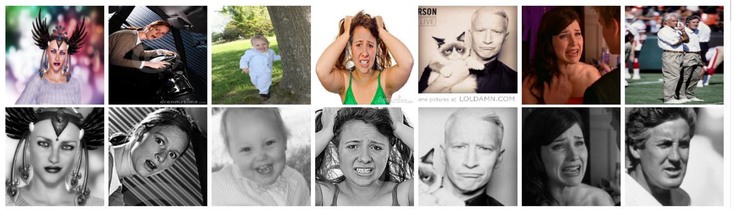
Sarah Adel Bargal, Emad Barsoum, Cristian Canton Ferrer, Cha Zhang
[1] Boston University, [2] Microsoft Research

Main Goal

Emotion Recognition from Videos for the classes: Anger/ Disgust/Fear/Happy/Neutral/Sad/Surprise.

Emotion Images Dataset

- MSR dataset collected with the help of the Bing team
- 148K images crawled from the web
- Each image was annotated by 12-15 crowd workers for one of the basic emotions



Emotion Videos Dataset

- AFEW 6.0 Dataset
- 1.5K movie videos
  - –748 train, 382 Valid videos, 593 Test Videos

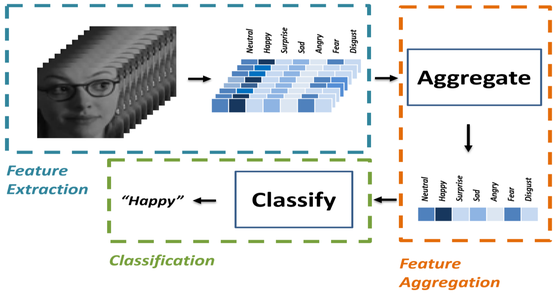
Spontaneous — Acted

Training our Models

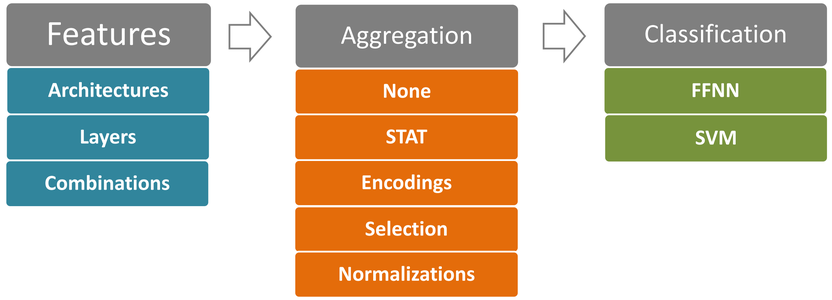
- We train three networks on the emotion images dataset:
  - a modified VGG (13 layers) [Barsoum *et al.* ICMI'16]
  - a second VGG (16 layers) [Simonyan and Zisserman arXiv'14]
  - RESNET (91 layers) [He *et al.* arXiv'15]
- Probabilistic Label Drawing training [Barsoum *et al.* ICMI'16]
  - A random emotion tag is drawn from the crowd-sourced label distribution of an image and used as the ground truth for that image in a certain epoch.

Recognition Pipeline

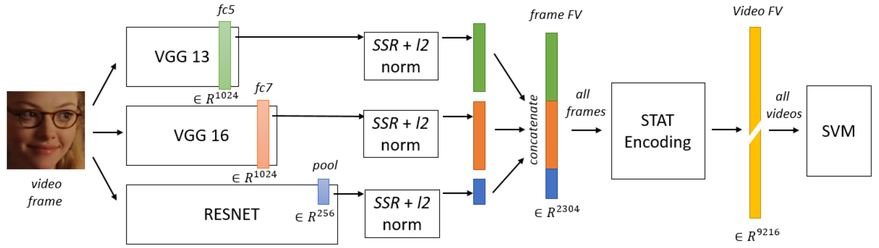
- Spatial approach: Compute frame features, then aggregate these features into a single feature representing the entire video.



How can we do better?



Best-performing Pipeline



Feature Comparison for STAT Encoding

We perform statistical (STAT) encoding for various:

- **Architectures**
  VGG13/VGG 16/RESNET
- **Layers**
  Fully Connected/Last Convolutional/Output
- **Layer Combinations**
  Same architecture/ Different architectures

Approach	Validation Acc (%)
challenge baseline	38.81
op VGG13	57.07
op VGG16	55.24
op RESNET	53.66
op VGG13+*op* VGG16+*op* RESNET	57.33
fc5 VGG13	58.9
fc7 VGG16	56.02
pool RESNET	52.62
fc5 VGG13 + *fc7* VGG16 + *pool* RESNET	**59.42**
fc5 VGG13 + *fc7* VGG16 + *pool* RESNET **	59.16

Best results are obtained using layer combinations of different architectures: *fc5* of VGG13 + *fc7* of VGG16 + *pool* RESNET.
*** means using a selection of the middle 90% of frames.*

State-of-the-art performance on AFEW 6.0

- Using this pipeline on 593 Test Videos (including reality shows)

Approach	Validation Acc (%)	Test Acc (%)
challenge baseline	38.81	40.47
fc5 VGG13 + *fc7* VGG16 + *pool* RESNET	**59.42**	**56.66**