ICML
International Conference
On Machine Learning

Facial Expression Recognition with Adaptive Frame Rate based on Multiple Testing Correction

Andrey V. Savchenko
[1]Scientific director at Sber AI Lab

[2]Sr. researcher, ISP RAS Research Center for Trusted Artificial Intelligence

[3]Full. Prof., Leading Researcher at HSE University

Email: andrey.v.savchenko@gmail.com
URL: www.hse.ru/en/staff/avsavchenko

2023

Problem statement

Facial expression recognition (FER) in video

Given the input facial video X = {X(t), t = 1, 2, ..., T } with T frames, it is necessary to associate it with one of C > 1 emotional classes. The classes are specified by the training set of N > 1 facial videos $X_n = \{X_n(t), t = 1,2,...,T_n\}, n = 1,2,...,N$ with known class label $y_n \in \{1, 2, ..., C\}$

Conventional approach

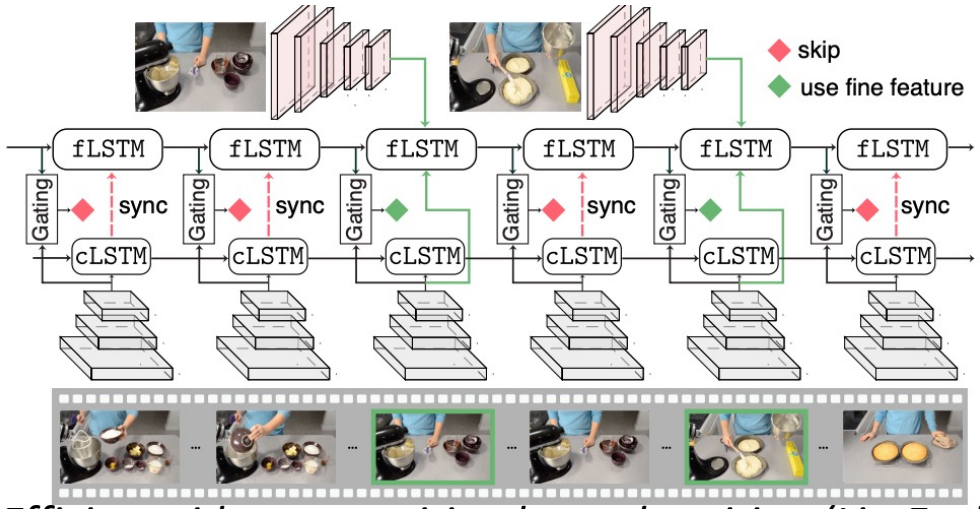
1. Detect/track faces and extract facial features (embeddings) **x**(t) in each frame using pre-trained DNN
2. Pool embeddings into a single video descriptor **x** (MaxPool/AvgPool, LSTM, attention, ...)
3. Feed x into a classifier C(l) (MLP, random forest, SVM,...).

Disadvantage: low speed due to *T* inferences in a DNN (plus slow face detection)

Solution: efficient video classification techniques from action recognition: AdaFrame, LiteEval, AR-Net, SCSampler, FrameExit,...

Their **disadvantages** for FER:

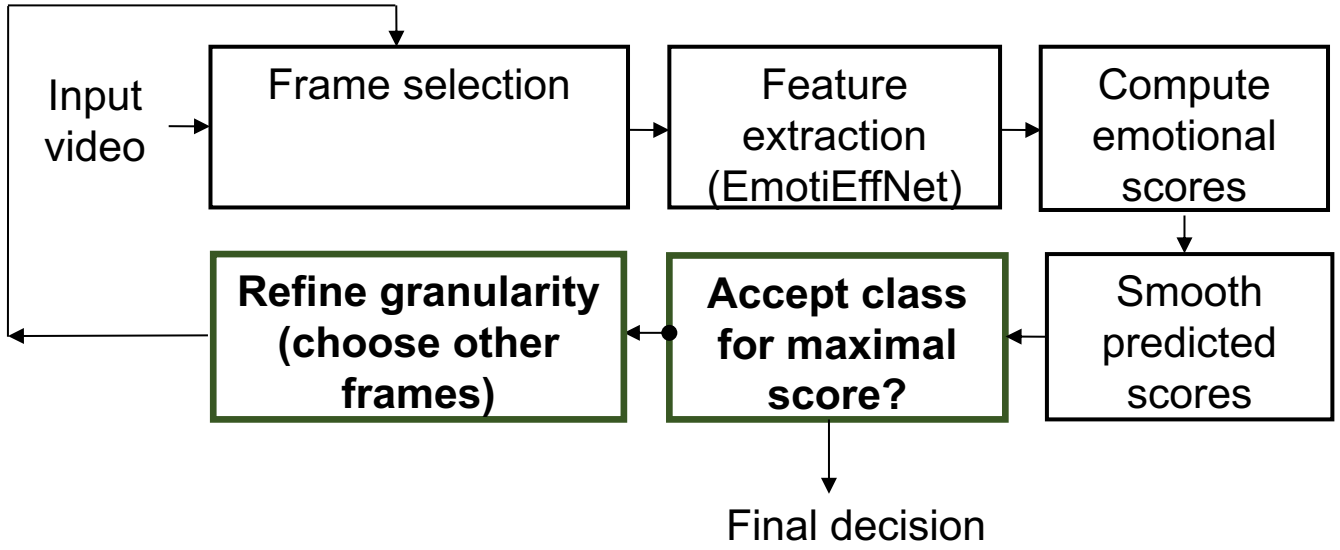
- rapid evolution of emotions (relatively short videos);
- presence of face detection/tracking step limits the widely-used RL-techniques with initial processing of all frames via lightweight models;
- small training sets with dirty and ambiguous labeling that limits the potential of deep models and forces the usage of lightweight models

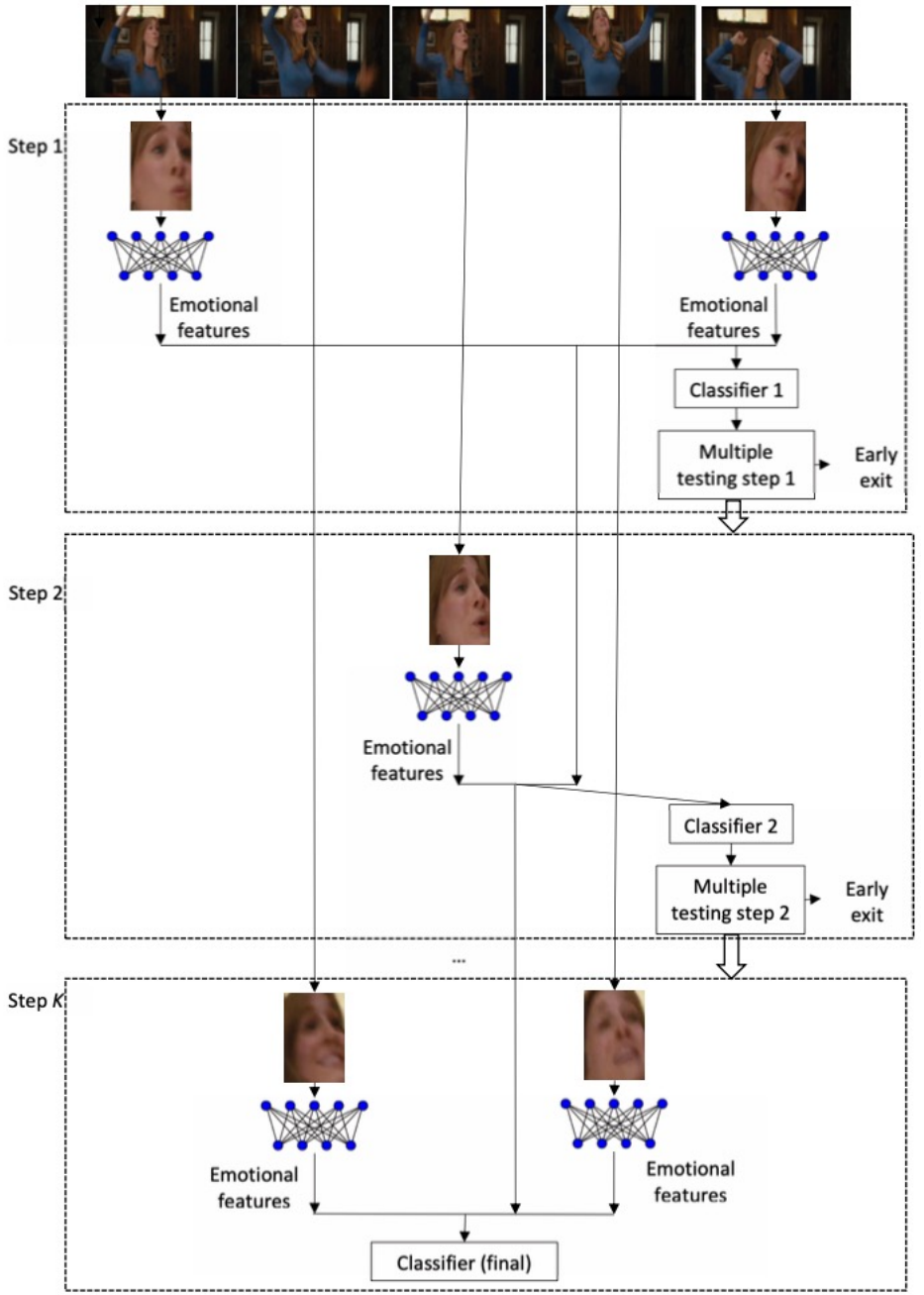


Efficient video recognition by early exiting (LiteEval)

Proposed approach (1)

Adaptive frame rate





- Video processing is inspired by sequential statistical analysis of A. Wald with at most L = $\lceil \log_k T \rceil$ steps.
- Deterministic frame sampling policy: the frame rate factor $FR^{(l)} = k^{L-l}$ is used at the *l*-th stage (*l* = 1,2,...,L). The sequence of frames at the *l*-th stage is a subset of frames from the (*l* + 1)-th stage.
- The reliability at the *l*-th level is verified by using the scores at the output of video classifier: $\max_{y \in \{1,\ldots,C\}} s_y^{(l)}(\mathbf{x}^{(l)}) > s^{(l)}$
- If we assume that every step has the same exit probability of 1/*L*, the average complexity: $\frac{1}{L}\sum_{l=1}^{L}(1+k^{l-1}) \approx \frac{T}{L}$.

Proposed approach (2)

Multiple testing correction

How to choose thresholds?

- Train classifier on part of training set, predict confidence scores on the remaining *M* training examples and fix the false acceptance rate (FAR) α_l.
- Threshold is chosen as the α_l–quantile of the maximal scores of other classes

$$\left\{ \max_{y \neq y_n} s_y^{(l)}(\mathbf{x}_{n_m}^{(l)}) \middle| \, m \in \{1, ..., M\} \right\}$$

How to choose FAR for every *l*-th step given the confidence level α of the whole procedure?

- It is a multiple testing problem of sequential analysis. We use the Benjamini-Hochberg correction:

$$\alpha_l = \frac{\alpha \cdot l}{L}$$

for each training example $n \in \{1, ..., N\}$ **do**
  for each frame $t \in \{1, ..., T_n\}$ **do**
    Extract facial region in $X_n(t)$ using an arbitrary face detector
    Feed the facial image into a neural network feature extractor and compute the embeddings $\mathbf{x}_n(t)$
  end for
  Compute video descriptor $\mathbf{x}_n = Pool(\{\mathbf{x}_n(t) | t \in \{1, 2, ..., T_n\}\})$
  for each step of adjusted frame rate $l \in \{1, ..., L-1\}$ **do**
    Compute $\mathbf{x}_n^{(l)} = Pool(\{\mathbf{x}_n(t) | t \in T^{(l)}\})$ (1)
  end for
end for
for each step of adjusted frame rate $l \in \{1, ..., L-1\}$ **do**
  Split N instances in a stratified fashion to get indices $\{n_1, ..., n_M\}$ of validation set
  Train the l-th classifier $\mathcal{C}$ using remaining training examples
  Initialize a list $S = []$
  for each validation instance $m \in \{1, ..., M\}$ **do**
    Append the maximal inter-class confidence score $\max_{y \neq y(n)} s_y^{(l)}(\mathbf{x}_{n_m}^{(l)})$ to S
  end for
  Assign the $\lfloor \alpha l / L \rfloor$-th largest element from S to the threshold $s^{(l)}$ using the Benjamini-Hochberg correction (4)
end for
Train an arbitrary classifier $\mathcal{C}$ using set of pairs $\{(\mathbf{x}_n, y_n)\}$.
return classifier $\mathcal{C}$ and thresholds $s^{(l)}, l = 1, 2, ..., L$

Datasets

1) AffWild: Affective Behavior Analysis in-the-wild (ABAW) challenge

Frame-level video-based FER: assign each frame $X(t)$, =1,2,…,T to emotional category $c \in [1, 2, \ldots, C_{EXPR}]$, C_{EXPR} =8 классов (anger, disgust, fear, happiness, sadness, surprise, neutral, other)

- Official training set: 585,317 frames
- Official validation set: 280,532 frames
- https://ibug.doc.ic.ac.uk/resources/cvpr-2023-5th-abaw/

2) AFEW (Acted Facial Expression In The Wild): EmotiW 2013-2019 challenges

Audio-video emotion recognition: assign the whole video with T frames to emotional category $c \in [1, 2, \ldots, C_{EXPR}]$, C_{EXPR} =7 классов (Anger, Disgust, Fear, Happiness, Sad, Surprise, and Neutral)

- Official training set provided by organizers: 773 clips (1-5 seconds)
- Official validation sets: 383 videos.
- https://sites.google.com/view/emotiw2019/challenge-details

Emotional feature extraction

EmotiEffNets from HSEmotion library

- EfficientNet-B0 from repository

https://github.com/HSE-asavchenko/face-emotion-recognition/

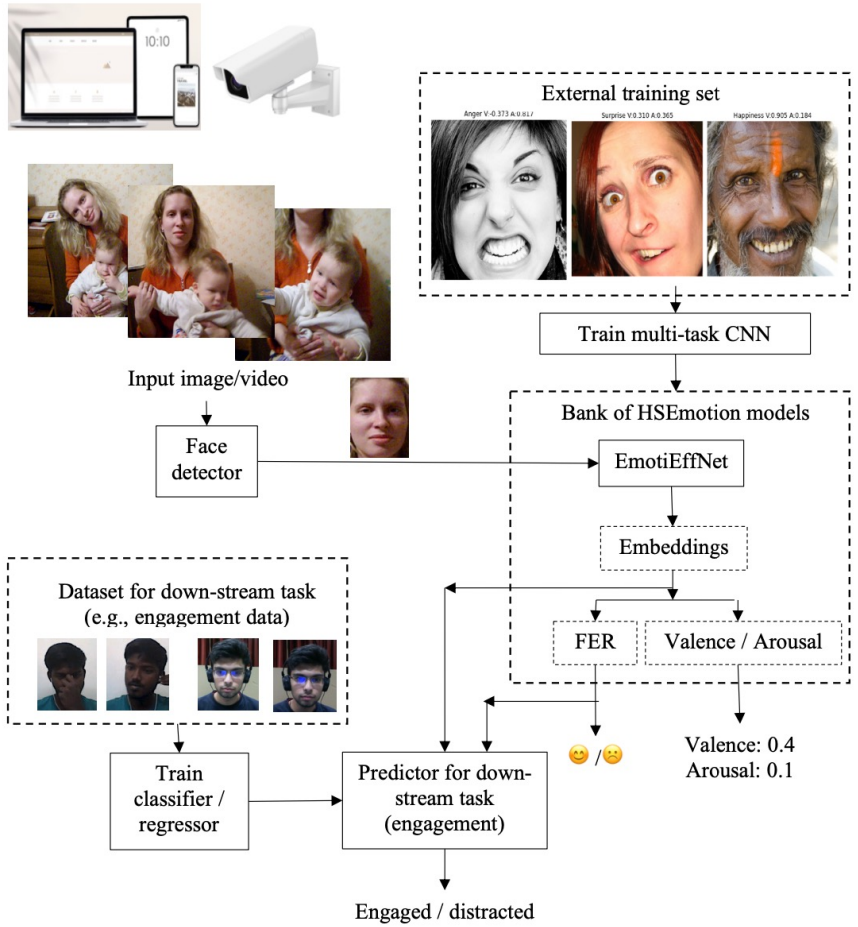
- Python packages hsemotion, hsemotion-onnx:

https://github.com/HSE-asavchenko/hsemotion

pip install hsemotion

Rank	Model	Accuracy (8 emotion)	Accuracy (7 emotion)	Extra Training Data	Paper
1	**Multi-task EfficientNet-B2**	63.03	66.29	×	Classifying emotions and engagement in online learning based on a single facial expression recognition neural network
2	**MT-ArcRes**	63		×	Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace
3	**DAN**	62.09	65.69	×	Distract Your Attention: Multi-head Cross Attention Network for Facial Expression Recognition
4	**SL + SSL in-panting-pl** (B0)	61.72		×	Using Self-Supervised Auxiliary Tasks to Improve Fine-Grained Facial Representation
5	**Distilled student**	61.60	65.4	✓	Leveraging Recent Advances in Deep Learning for Audio-Visual Emotion Recognition
6	**Multi-task EfficientNet-B0**	61.32	65.74	✓	Facial expression and attributes recognition based on multi-task learning of lightweight neural networks
7	**SL + SSL puzzling** (B2)	61.32		×	Using Self-Supervised Auxiliary Tasks to Improve Fine-Grained Facial Representation

https://paperswithcode.com/sota/facial-expression-recognition-on-affectnet



- Savchenko IEEE SISY 2021;
- Savchenko et al., IEEE Trans. on Affective Computing 2022;
- Savchenko CVPRW 2022, 2023;
- Savchenko ECCVW 2022

ABAW

Experimental results

Efficient video classifiers, EmotiEffNet-B0 features

Method	F1-score	Time $\bar{t}$
Smoothing (all frames)	**0.4262**	55.94±0.25
AdaFrame	0.4205	42.32±0.30
LiteEval	0.4220	50.71±0.26
AR-Net	0.4051	22.39±0.25
OCSampler	0.3928	4.85±0.22
FrameExit	0.4177	5.97±0.37
Proposed approach	0.4217	**3.70±0.20**

Fixed FAR vs proposed multiple testing correction

Sequence of frame rates	Thresholds estimator	F1-score	Time $\bar{t}$
(200–>100–>50–>10–>1)	Fixed FAR	0.4190	20.15±0.35
	Proposed	0.4217	3.70 ±0.20
(100–>50–>10–>1)	Fixed FAR	0.4205	23.82±0.29
	Proposed	0.4221	11.03 ±0.32
(50–>25–>1)	Fixed FAR	0.4257	26.58±0.31
	Proposed	0.4253	17.12 ±0.23
(50–>10–>1)	Fixed FAR	0.4258	25.03±0.30
	Proposed	0.4258	15.51 ±0.19
(200–>50–>1)	Fixed FAR	0.4203	29.39±0.28
	Proposed	0.4207	20.41 ±0.20
(100–>50–>1)	Fixed FAR	0.4225	27.26±0.25
	Proposed	0.4230	20.31 ±0.21

Various neural networks and sequences of frame rate factors

Sequence of frame rates	EmotiEffNet-B0		MT-EmotiEffNet-B0		EmotiEffNet-B2	
	F1-score	Time $\bar{t}$	F1-score	Time $\bar{t}$	F1-score	Time $\bar{t}$
(200)	0.3624	0.55±0.05	0.3323	0.56±0.04	0.3062	1.15 ±0.12
(1)	0.4262	55.94±0.25	0.3913	56.68±0.25	0.3532	116.04 ±0.30
(200–>100–>50–>10–>1)	0.4217	**3.70±0.20**	0.3820	**1.34±0.08**	0.3503	**4.63 ±0.13**
(50–>25–>1)	0.4253	17.12±0.23	0.3861	1.81±0.06	0.3518	19.09 ±0.19
(50–>10–>1)	0.4258	15.51±0.19	**0.3898**	3.02±0.12	0.3521	14.58 ±0.13
(200–>50–>1)	0.4207	20.41±0.20	0.3771	1.15±0.09	0.3488	24.82 ±0.22
(100–>50–>1)	0.4230	20.31±0.21	0.3787	1.07±0.05	0.3503	24.57 ±0.23
(200–>1)	0.4205	48.27±0.37	0.3832	31.37±0.28	0.3477	74.01 ±0.27
(100–>1)	0.4228	36.48±0.19	0.3840	14.43±0.07	0.3505	47.49 ±0.18
(50–>1)	**0.4258**	33.01±0.21	0.3885	12.65±0.09	**0.3528**	43.73 ±0.14

AFEW

Experimental results

Efficient video classifiers

METHOD	F1-SCORE	TIME $\bar{t}$
FAN (RESNET-18)	0.5118	35.18±0.08
DENSENET-161	0.5144	170.61±0.31
IR-50	0.5378	92.64±0.24
VGG-FACE + BLSTM	0.5391	165.90±0.45
NOISY STUDENT	0.5517	29.26±0.06
FBP FUSION	0.6550	232.02±0.33
EmotiEffNet-B0		
ALL FRAMES	0.5927	55.94±0.19
ADAFRAME	0.5906	49.95±0.25
LITEEVAL	0.5927	52.20±0.31
AR-NET	0.5526	32.43±0.23
OCSAMPLER	0.5530	30.27±0.18
FRAMEEXIT	0.5726	31.89±0.34
PROPOSED APPROACH	0.5910	29.75±0.15

Various neural networks and sequences of frame rate factors

SEQUENCE OF FRAME RATES	EMOTIEFFNET-B0		MT-EMOTIEFFNET-B0		EMOTIEFFNET-B2	
	ACCURACY	TIME $\bar{t}$	ACCURACY	TIME $\bar{t}$	ACCURACY	TIME $\bar{t}$
(18)	0.5085	3.60±0.03	0.5013	3.65±0.03	0.5040	7.48 ±0.06
(1)	0.5927	55.94±0.19	0.5699	56.68±0.20	0.5937	116.04 ±0.29
(18–>9–>1)	0.5850	**29.75±0.15**	0.5515	**27.55±0.14**	0.5778	**53.74 ±0.21**
(18–>6–>1)	**0.5927**	32.79±0.17	0.5515	30.09±0.15	0.5831	54.00 ±0.20
(9–>3–>1)	0.5903	38.70±0.18	0.5831	37.41±0.17	0.5989	73.03 ±0.23
(6–>3–>1)	0.5903	40.01±0.17	0.5726	38.93±0.17	0.5937	76.06 ±0.22
(18–>1)	0.5824	31.02±0.17	0.5541	30.30±0.16	0.5778	58.00 ±0.20
(9–>1)	0.5903	34.53±0.17	**0.5752**	33.01±0.17	0.5910	63.15 ±0.23
(6–>1)	0.5877	34.31±0.16	0.5726	34.04±0.16	0.5910	61.30 ±0.22
(3–>1)	0.5903	40.38±0.19	0.5726	39.45±0.19	**0.5937**	72.75 ±0.28

Conclusion

We present the novel framework for efficient video-based FER using sequential analysis of various frames:

1. The most remarkable feature is the multiple testing correction that makes it possible to automatically reach a balance between efficiency and accuracy.

2. The recognition trustworthiness is improved by maintaining only one hyper-parameter, FAR

3. It can be applied with an arbitrary emotional feature extractor, frame pooling strategy, and video classifier

Source code

Disadvantage

Need to know the number of frames *T* to predict facial expression in the whole video fragment.

Thank you!

This work was supported by a grant for research centers in the field of artificial intelligence, provided by the Analytical Center for the Government of the RF in accordance with the subsidy agreement (agreement identifier 000000D730321P5Q0002) and the agreement with the Ivannikov Institute for System Programming of the RAS dated November 2, 2021 No. 70-2021-00142.