

15th IEEE International Conference on Automatic Face and Gesture Recognition

Real-time Facial Expression Recognition "In The Wild" by Disentangling 3D Expression from Identity

Mohammad Rami Koujan, Giorgos Giannakakis, Nicolas Pugeault, Anastasios Roussos

16-20 November 2020

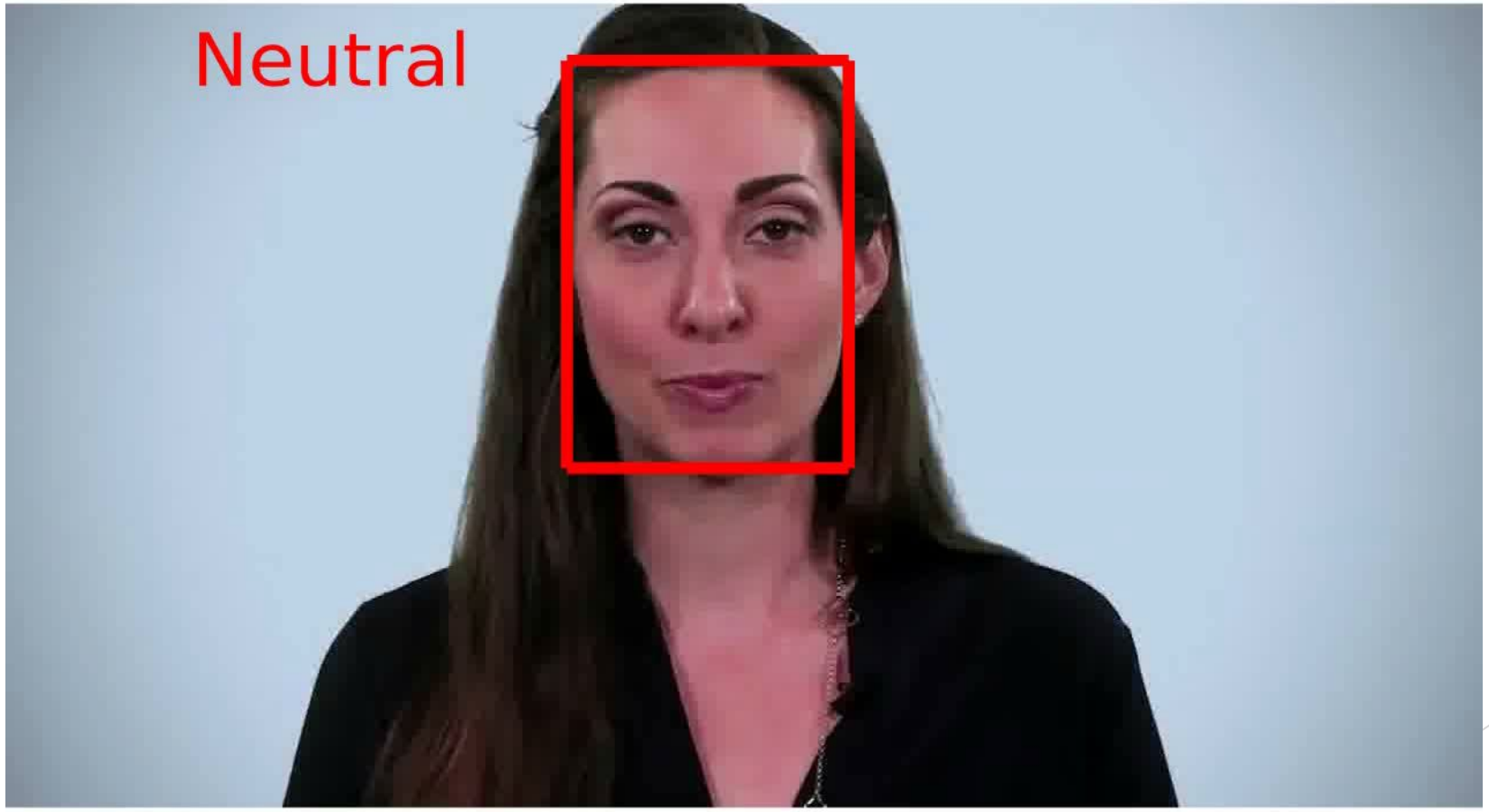
UNIVERSITY OF EXETER

FORTH INSTITUTE OF COMPUTER SCIENCE

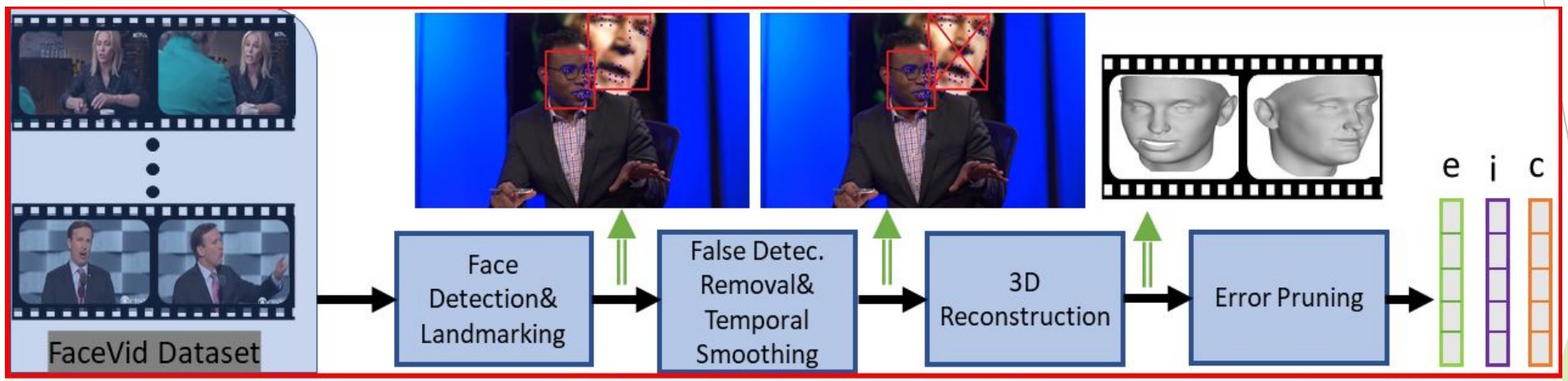
Motivation: Key Non-Verbal Signals



Seven Basic Emotions- Results (ours)



FaceVid Annotation



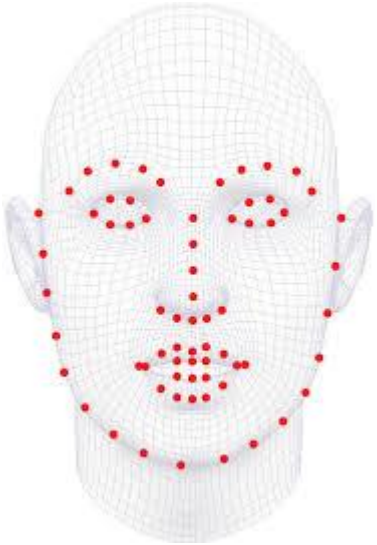
- FaceVid dataset: 6K YouTube videos (total of 12M frames)
- Pre-processed and annotated with 3D reconstruction information in four subsequent stages

Annotation Stages

1- Face Detection and Landmarking



Input frame



3DMM Mean Face

2- False Detections Removal and Temporal Smoothing

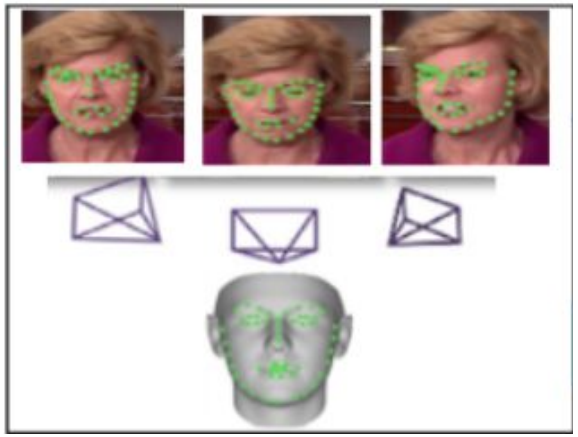
- Track detected face in the first frame
- Check if the BBs between consecutive frames stay within a reasonable margin
- Stop tracking when BBs go out of margin for K consecutive frames
- Temporally smooth extracted and tracked landmarks using cubic splines

Annotation Stages

3- 3D Reconstruction

- 3DMM and sparse landmarks-based reconstruction
- Large scale Morphable Model built from 10K scans was employed for this purpose

$$\mathbf{x}(\mathbf{i},\mathbf{e}) = \bar{\mathbf{x}} + \mathbf{U}_{id}\mathbf{i} + \mathbf{U}_{exp}\mathbf{e}$$



Fitting Procedure:

- Minimize a cost function that consists of three terms
  1. sum of squared 2D landmark reprojection errors over all frames
  2. Shape priors imposing a quadratic prior on the identity and expression parameters
  3. Temporal smoothness enforcing smooth facial expression parameters

Annotation Stages

4- Error Pruning

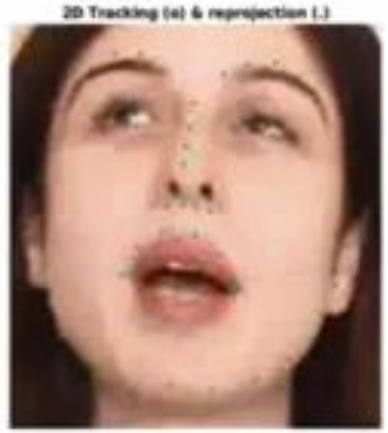
- Remove videos corresponding to failed 3D reconstruction, in two stages:

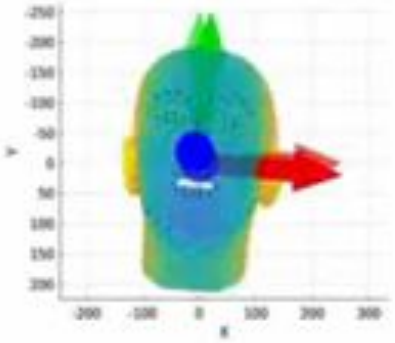
1. Automatic:
   - Estimated identity parameters should be independent and identically distributed normal random variables
   - Remove videos with estimated ||i|| above a threshold
   - 300 videos were eliminated at the end of this stage

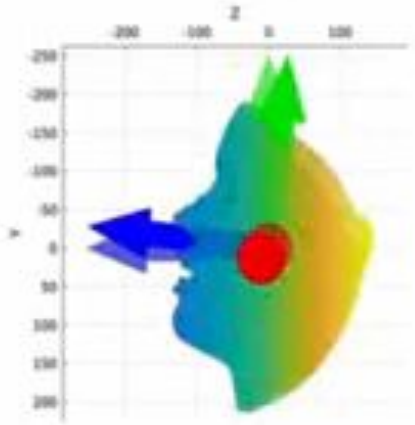
2. Manual:
   - inspect the reconstructions of all remaining videos and manually flag and prune videos where it is evident that the 3D face reconstruction has failed
   - 250 videos were removed at the end of this stage

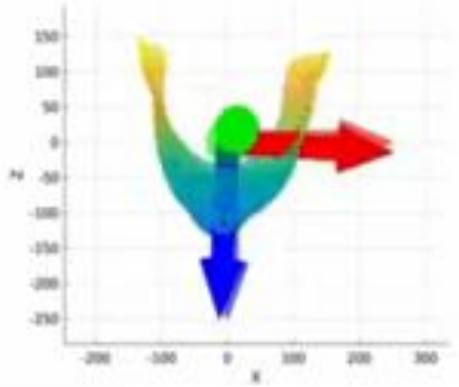
FaceVid Dataset



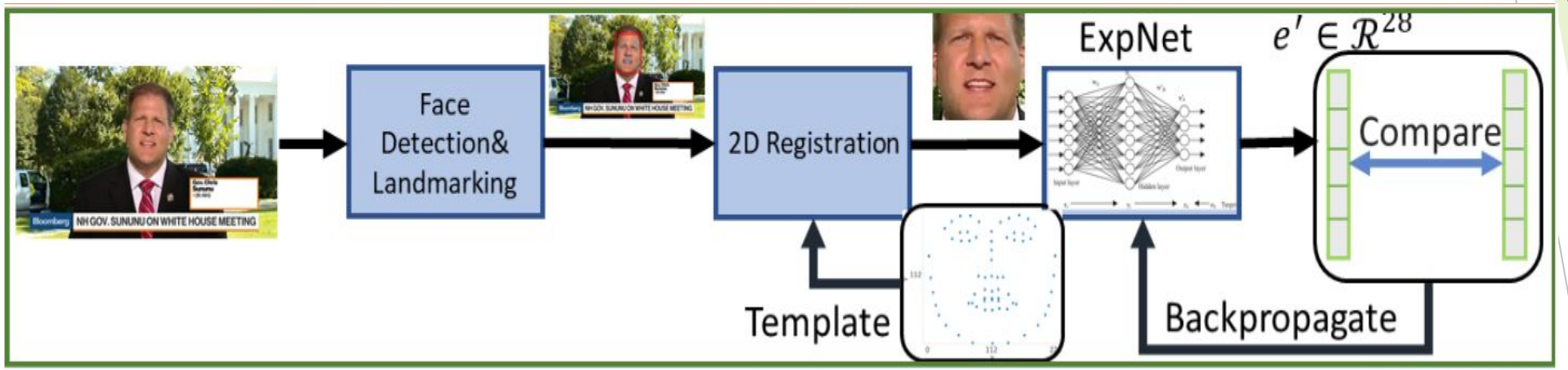






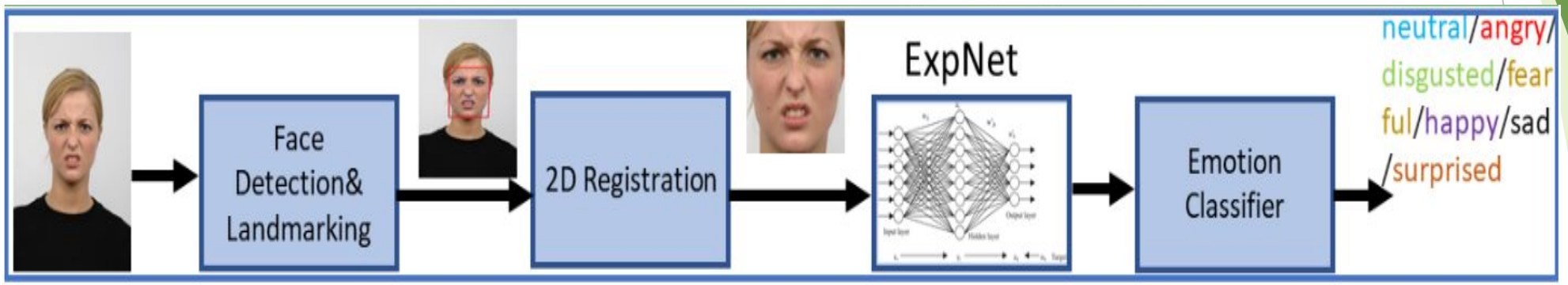


DeepExp3D Training



- Given the FaceVid dataset, a Convolutional Neural Network (CNN) N : I→e, aiming at regressing the 3DMM facial expression coefficients from a given RGB image (I) was trained in a supervised fashion
- As a pre-processing step, input frames were registered to a 2D template of facial landmarks
- The trained network was fine-tuned on a dataset of real 3D facial scans (4DFAB)

Emotion Classifier



- To classify the generated facial expression vectors e ∈R^28 produced by the DeepExp3D network N, Error Correcting Output Codes (ECOC) method was utilised to solve this 7-class (neutral + six basic emotions) classification problem
- The binary learner of choice is Support-Vector Machines (SVM)
- 10-fold cross validation with the one-versus-all coding scheme were implemented to train/test the emotion classifie

Quantitative Results (Runtime)

COMPARISON OF REQUIRED RUN-TIME TO PRODUCE ESTIMATIONS OF FACIAL EXPRESSION FROM A SINGLE IMAGE.

Method	ITW [5]	SfM3DMM [36]	3DDFA [68]	ExpNet [7]	**Ours**
Time (sec)	6.4	3.0	0.6	0.088	**0.02**

[5]: J.Booth et al. T-PAMI 2018
[36]: M.R.Koujan et al. CVMP 2018
[68]: X. Zhu et al. CVPR 2016
[7]: J. Chang et al. FG 2018

Quantitative Results (7-Basic Emotions)

COMPARISON OF FACIAL EXPRESSION RECOGNITION ACCURACIES ON 5 WIDELY-USED BENCHMARKS.

Dataset	Approach	Acc.(%)
RadFD	ExpNet [7]	75.00
	Ali et al. [1]	85.00
	Zavarez et al. [67]	85.97
	Jiang& Jia [33]	94.52
	Mavani et al. [48]	95.71
	Wu& Lin [63]	95.78
	Sun et al. [59]	96.93
	Yaddaden et al. [65]	97.57±1.33
	Ours	**97.65±1.00**
CFEE	ExpNet [7]	62.50
	Ours	96.43±1.1
	Du et al.[15]	**96.84±9.73**

Dataset	Approach	Acc.(%)
RAF-DB	ExpNet [7]	55.20
	Li& Deng [40]	74.20
	Lin et al. [43]	75.73
	Fan et al. [18]	76.73
	Gosh et al. [22]	77.48
	Shen et al. [57]	78.60
	Vielzeuf et al. [61]	80.00
	Deng et al. [10]	81.83
	Ours	82.06±0.73
	Li et al. [42]	**83.27**

Dataset	Approach	Acc.(%)
KDEF	ExpNet [7]	71.00
	Zavarez et al. [67]	72.55
	Ali et al. [1]	78.00
	Ruiz-Garcia et al. [55]	86.73
	Yaddaden et al. [65]	90.62
	Ours	**92.24±0.70**
CK+	ExpNet [7]	61.17
	Wang et al. [62]	86.3
	Jung et al. [34]	92.35
	Ours	**96.45±0.8**

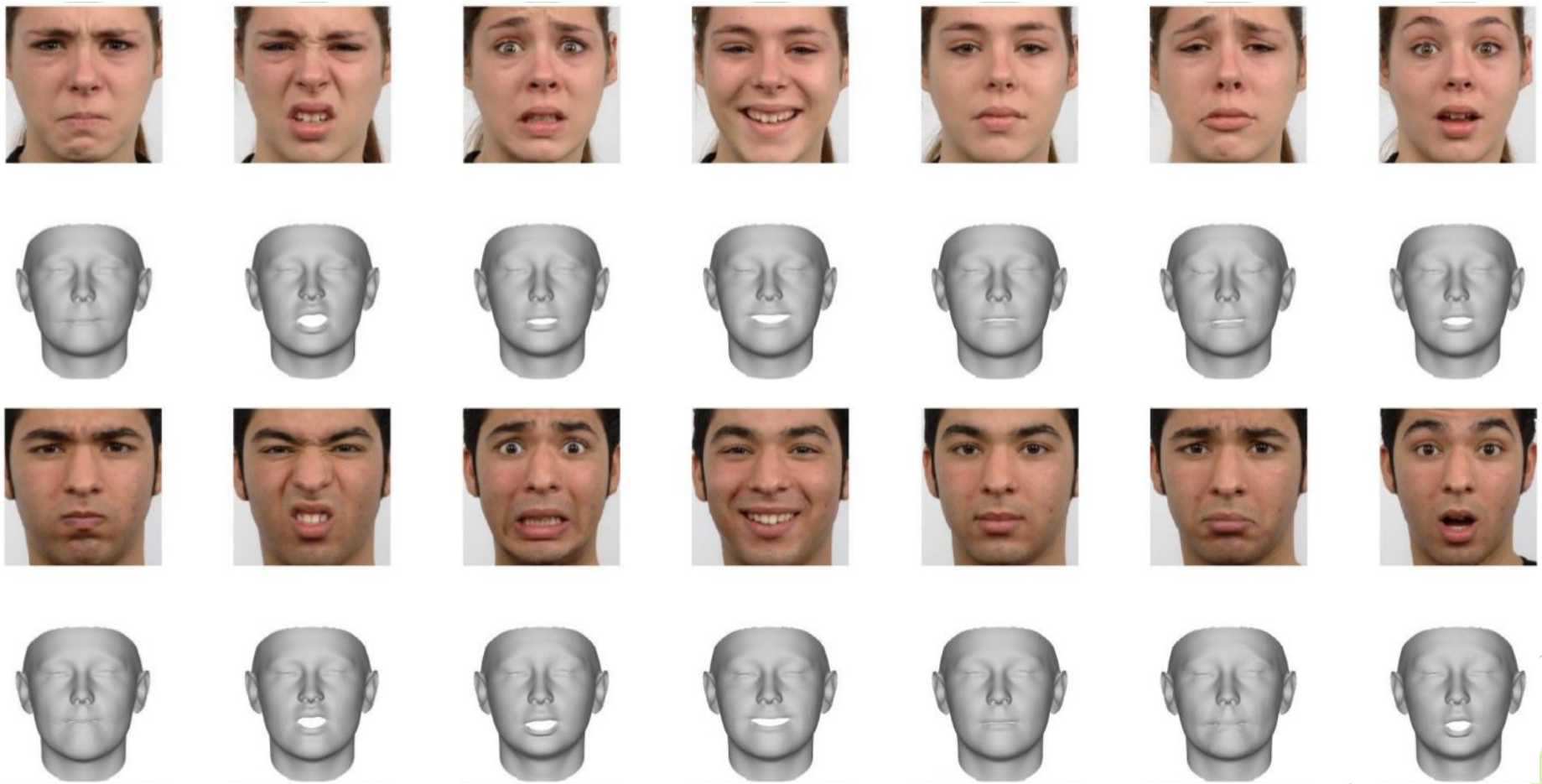
Quantitative Results (Stress Detection)

STRESS DETECTION ACCURACY COMPARISON ON DATASET USED IN [24]

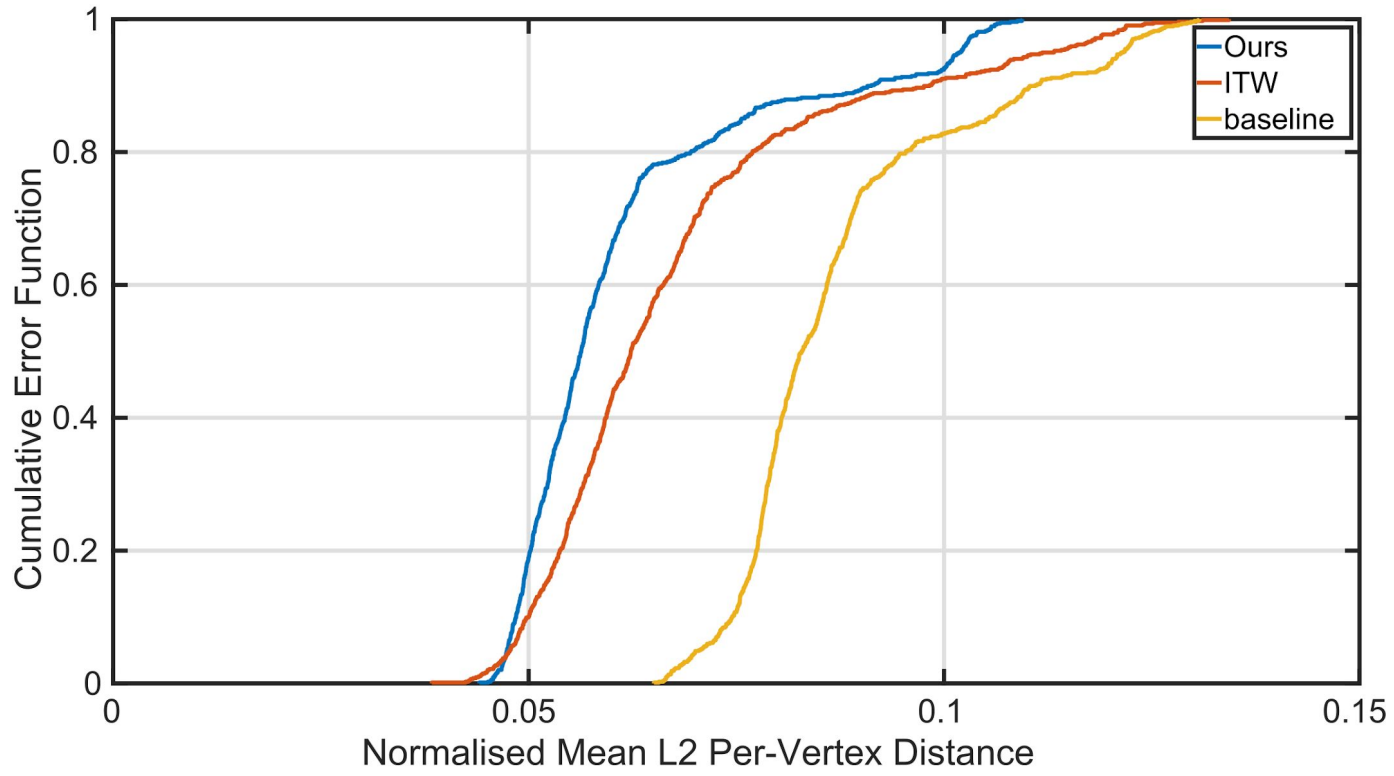
Phase	**Head motility** [24]	**DWNet1D** [27]	**Ours**
Emotional Recall	82.99 %	83.50 %	**86.70 %**
Stressful images	85.42 %	**92.60 %**	88.42 %
Stressful videos	85.83 %	85.90 %	**88.83 %**
Average	84.75 %	87.33 %	**87.98 %**

[24]: G. Giannakakis et al. IEEE BHI 2018
[27]: G. Giannakakis et al. ACIIW 2019

Qualitative Results



Ablation Study



Mean (over all vertices of every mesh) vertex-to-vertex error between the 3D reconstructions produced by our method, ITW and a baseline (linear 3DMM fitting) and the ground-truth 3D meshes of the 4DFAB dataset

Conclusion

- Large-scale dataset of facial monocular videos was collected and annotated
- Framework for the automatic recognition of human emotions and stress detection from monocular in-the-wild images is proposed
- During inference, our framework is runnable at 0.02 seconds per frame
- Proposed framework capitalizes on a well-trained CNN for facial expression regression that outperforms other SOTA methods
- Competitive results on five benchmarks for emotion recognition and one for stress detection

Thank You!

