

AUTH - Multimodal facial expressions dataset

This is a dataset suitable for active and static single modality and multimodal (i) facial expression, (ii) identity and (iii) gender recognition methods. It contains:

1. Sequences of facial 3D models
2. Webots simulation environments
3. videos, along with synchronized audio (speech), captured from a grid of virtual cameras at 2 lighting levels

The dataset is available [here](#). Each part of the dataset can be utilized in a standalone manner, i.e., it can be downloaded and utilized separately from the rest.

Sequences of facial 3D Models: The facial 3D models (in OBJ format) were generated using DECA [1] method, applied on image sequences from the emotional speech part of RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) dataset[[link](#)]. RAVDESS contains videos (including audio) of 24 professional actors, vocalizing two statements in 8 different emotions/expressions in two intensity levels (normal, strong) and two repetitions per intensity. The neutral expression is also included in normal intensity only. Additionally, videos depicting the neutral expression (which were performed in normal" intensity only) were included in the data generation process. This part was generated following two strategies:

1. DECA generates sequences of accurate textured 3D facial models (by predicting the head pose, shape, detailed face geometry, expression parameters, and illumination) from a sequence of images.
2. DECA generates a 3D face from a given image, and then animates it by ``transferring" expressions from different faces/persons depicted in image sequences. This approach will be referred as expression transfer strategy.

Based on RAVDESS dataset, $24 \text{ (subjects)} \times 7 \text{ (expressions)} \times 2 \text{ (intensities)} \times 2 \text{ (statements)} \times 2 \text{ (repetitions)} + 96 \text{ (sequences of 3D models based on neutral expression)} = 1440$ sequences of facial 3D models were generated, depicting the actors posing their own expressions, i.e., without employing the expression transfer strategy.

The expression transfer strategy was employed, by transferring to each actor, expressions (1 repetition from one random intensity level from both statements for all emotions) extracted from 4 randomly selected same gender actors. Based on this strategy, $24 \text{ (subjects)} \times 8 \text{ (expressions)} \times 2 \text{ (statements)} \times 4 \text{ (subjects)} = 1536$ sequences of facial 3D models were additionally generated. In total, $1440 + 1536 = 2976$ sequences of 3D models were generated using both strategies.

Each sequence of 3D models adopts the following naming notation:

01_ExpressionID_Intensity_Statement_Repetition_Actor1_Actor2_Expression where:

- ExpressionID (01-08): neutral, calm, happy, sad, angry, fearful, disgust, surprised
- Intensity (01-02): Normal, Strong
- Statement (01-02): "Kids are talking by the door", "Dogs are sitting by the door"
- Repetition: (01-02): 1st or 2nd repetition
- Actor1 (01-24): 3D face of the corresponding actor. Odd numbered actors are male, even numbered actors are female.

- Actor2 (01-24): Expression performed from the corresponding actor. Odd numbered actors are male, even numbered actors are female).
- Expression: neutral, calm, happy, sad, angry, fearful, disgust, surprised

This part adopts the following structure:

```

3d_models
├── 01_01_01_01_01_01_neutral
│   ├── 000.json
│   ├── 000.mtl
│   ├── 000.obj
│   ├── 000.png
│   ├── 000_normals.png
│   ├── 001.json
│   ├── 001.mtl
│   ├── 001.obj
│   ├── 001.png
│   └── 001_normals.png
│   ...
├── 01_01_01_01_02_04_fearful
│   ├── 000.json
│   ├── 000.mtl
│   ├── 000.obj
│   ├── 000.png
│   └── 000_normals.png
│   ...
└── ...

```

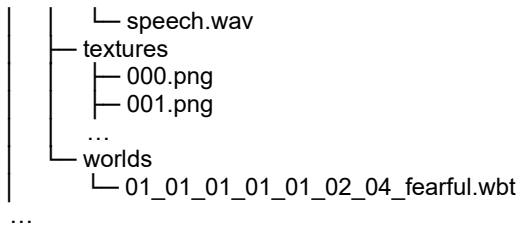
Webots simulation environments: The corresponding simulation environments are suitable for active multimodal facial expression recognition methods and other active perception tasks. A Webots controller is used to (i) animate the sequences of facial 3D models and (ii) enable the playback of the sound, which corresponds to the speech of the displayed actor. Finally, given two lighting levels (i.e. dark, bright), the user, through the controller, is also able to adjust the lighting conditions of the simulation scene. Overall, 2976 Webots simulation environments (equal to the 1440 + 1536 sequences of the generated facial 3D models) were created. Each simulation environment follows the same naming notation with the corresponding sequence of 3D models.

This part adopts the following structure:

```

webots_sim
├── 01_01_01_01_01_01_neutral
│   ├── controllers
│   │   └── supervisor_controller
│   │       └── supervisor_controller.py
│   ├── sound
│   │   └── speech.wav
│   ├── textures
│   │   ├── 000.png
│   │   ├── 001.png
│   │   └── ...
│   └── worlds
│       └── 01_01_01_01_01_01_neutral.wbt
├── 01_01_01_01_02_04_fearful
│   ├── controllers
│   │   └── supervisor_controller
│   │       └── supervisor_controller.py
│   └── sound

```



Five keyframes of a video depicting a RAVDESS actor pronouncing a statement in “calm” emotion (top). The same keyframes of the synthetic video captured from the generated Webots environment (bottom).

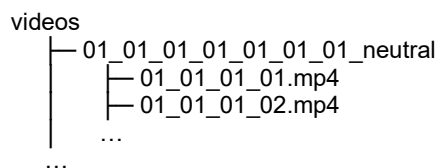
Video dataset: This part contains videos depicting the various sequences of 3D models within Webots. The footage was collected from a grid of virtual cameras, placed in a set of 25 angles ($-60^{\circ} \dots +60^{\circ}$ in pan with 30° increments and $-30^{\circ} \dots +30^{\circ}$ in tilt with 15° increments) and 3 distances (0.5, 0.75 and 1 meters). In addition, aiming to simulate various illumination conditions, the video-footage was collected at “dark” and “bright” lighting levels. Overall, for each sequence of 3D models, we collected 2 (lighting conditions) \times 3 (camera-to-face distances) \times 5 (tilt angles) \times 5 (pan angles) = 150 videos. The dataset is provided as a set of videos in MP4 format. The resolution of the videos was set to 300 \times 300 pixels. The videos also contain audio, which corresponds to the speech of the displayed actor.

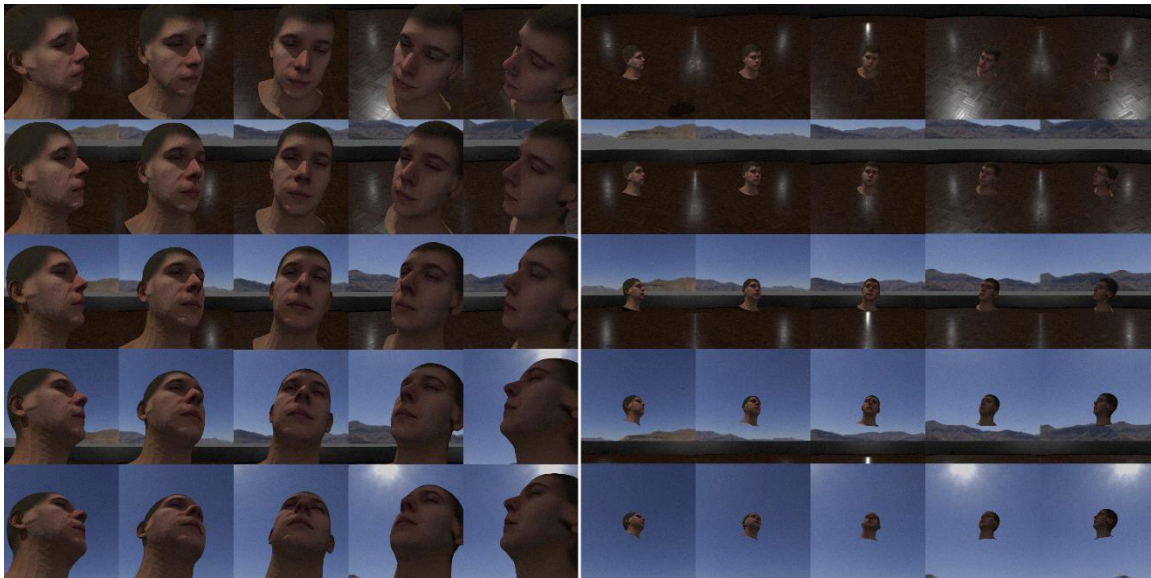
Each video adopts the following naming notation:

Lighting_Distance_Tilt_Pan.mp4 where:

- Lighting (01-02): Dark, Light
- Distance (01-03): 01 \rightarrow 0.5m, 02 \rightarrow 0.75m, 03 \rightarrow 1.0m
- Tilt (01-05): 01 \rightarrow -30° , 02 \rightarrow -15° , 03 \rightarrow 0° , 04 \rightarrow 15° , 05 \rightarrow 30°
- Pan: (01-05): 01 \rightarrow -60° , 02 \rightarrow -30° , 03 \rightarrow 0° , 04 \rightarrow 30° , 05 \rightarrow 60°

This part adopts the following structure:





Depiction of the 25 camera angles in 0.5 meters (left) and 1 meter (right).

[1] Feng, Y., Feng, H., Black, M. J., & Bolkart, T. (2021). Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (ToG)*, 40(4), 1-13.