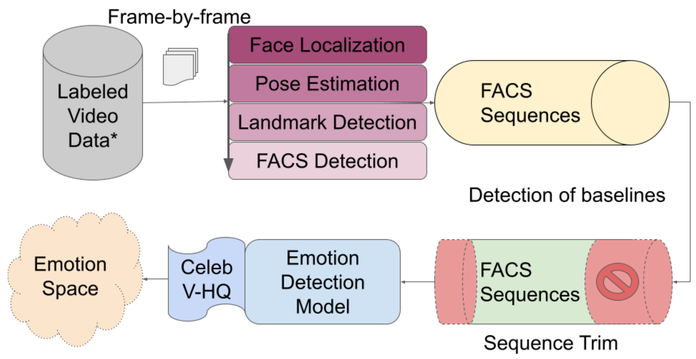
FlexComb: A Facial Landmark-Based Model for Expression Combination Generation

Bogdan Pikula, Dr. Steve Engels

Department of Computer Science, University of Toronto

Dynamic Graphics Project
University of Toronto
www.dgp.toronto.edu

Creating Facial Expressions



The general pipeline of FlexComb. We use a mixture of the following datasets containing facial expression video data: MMI, OULU-CASIA, DDCF. The video data is processed frame-by-frame to extract Facial Action Units (FACS). We then trim the resulting sequences of FACS by identifying the baselines and removing everything but the emotion shifts. This is fed into the emotion detection model which produces a space of distributions for emotion likelihoods.

Emotion Space

The centerpiece of our approach revolves around the CelebV-HQ dataset. Being one of the most extensive facial expression dataset available, it encompasses a great variety of emotions found "in the wild" captured from YouTube video clips.

1. To process the data, we extract the corresponding FACS from all the video clips found in the dataset. Upon gathering the sequences 20-valued vectors, we conduct a preminary exploratory analysis of the data.
2. Sequences are processed through an emotion detection model, yielding $1,290,000$ samples of emotional probability vectors for the following emotion labels: neutral, anger, disgust, fear, happiness, sadness, surprise.
3. Principal Component Analysis visualizes the emotion space, revealing three principal emotions: disgust, sadness, and anger.

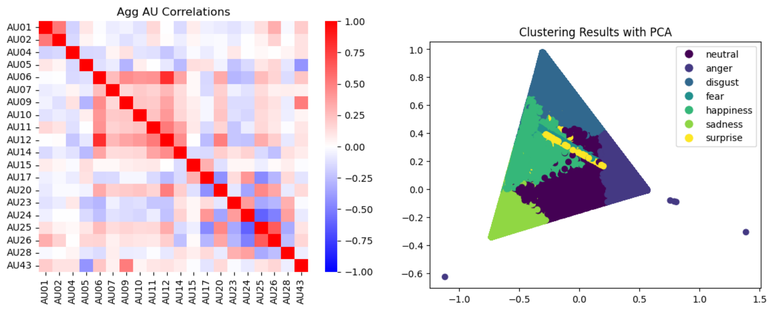


Figure 1. FACS correlation shows that despite a few exceptions, most facial activations occur independently, which further reinforces the complexity of human emotional expression. PCA Decomposition of Emotion Probability Vectors clustered by the emotion label.

Problem Formulation

The pipeline for FlexComb and its components.

- **Facial landmark analysis** and corresponding **FACS** in the video data on a frame-by-frame basis;
- **Detection of emotional shifts** in the video data by observing sudden FACS changes correlating with deviations from facial expression baselines.
- **Training an emotion detection model** that can accurately detect emotions outlined in the training dataset. [See below]
- **Generate a space of emotion likelihoods** by running the emotion Detection Model over the CelebV-HQ dataset.

Architecture

- **FACS Analysis:** The video data from several datasets is processed frame-by-frame to identify facial landmarks and extract FACS. Models utilized include:
  - *RetinaFace* for face localization
  - *Img2pose* for facial pose estimation and 3D head pose with head rotations.
  - *MobileFaceNet* to detect the 68 unique facial landmarks.
  - *XGBoost Classifier* is then employed to detect facial muscle activations.
- **Sequence Trim:** Place focus on frames where there's a notable shift in emotion, creating a dataset centered on emotional transitions. Emotional shifts are defined by notable changes in the FACS values between consecutive frames, with segments bound by shifts and an applied buffer size for capturing transitions.
- **Emotion Detection:** An emotion detection model trained on the trimmed FACS sequences identifies temporal dynamics of facial expressions and classifies them into seven unique emotions. The model boasts a 90% accuracy in emotion recognition when tested.

Evaluation

To evaluate the performance of FlexComb, we created different configurations for combinations of emotions, incrementing by 25%. We sampled the space for the closest probability vectors and determined the corresponding FACS values. For visualization, iClone8 FaceRigging software was employed, displaying muscle activations per an FACS-to-FaceKey mapping derived from the Action Unit reference descriptions. In essence, any emotion probability combination is feasible, given that the generated emotion space allows for an infinite array of configurations, supporting limitless combinations and variations.

- **Simplicity.** One significant advantage of FlexComb over traditional blendshapes is its ability to generate facial expressions without the need for manual modeling of key emotions. Traditional blendshapes require a dedicated facial model for each emotion.
- **Expression Realism and Coherence.** Another strength of FlexComb is in the ability to consistently create realistic facial expressions, which are inherently coherent and natural. This is thanks to the CelebV-HQ dataset being fundamentally rooted in the methodology.

All of the facial expressions sampled exist in the dataset one way or another. By analyzing the different emotion each combination of facial activation depicts, we are able to create a large space of emotional combinations. We use FACS to provide a simple and light facial representation that can be utilized to animate facial meshes in FaceRig.

Generation Examples

Figure 2. An example of a facial expression generated by FlexComb, representing equal parts Fear, Sadness and Surprise, as visualized on a FaceRig.

Figure 3. An example of facial expressions generated by FlexComb, representing different ratios of combining Happiness and Disgust emotions, as visualized on a FaceRig.

Figure 4. An example of facial expressions generated by FlexComb, representing different ratios of combining Neutral and Angry emotions, as visualized on a FaceRig.

References

[1] Rinat Abdrashitov, Fanny Chevalier, and Karan Singh.
Interactive exploration and refinement of facial expression using manifold learning.
In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20, page 778–790, New York, NY, USA, 2020. Association for Computing Machinery.

[2] Vitor Albiero, Xingyu Chen, Xi Yin, Guan Pang, and Tal Hassner.
img2pose: Face alignment and detection via 6dof, face pose estimation, 2021.

[3] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han.
Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices, 2018.

[4] Jiankang Deng, Jia Guo, Yuxiang Zhou, Jinke Yu, Irene Kotsia, and Stefanos Zafeiriou.
Retinaface: Single-stage dense face localisation in the wild, 2019.

[5] Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy.
Celebv-hq: A large-scale video facial attributes dataset.
In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part VII*, pages 650–667. Springer, 2022.

Artificial Intelligence and Interactive Digital Entertainment

AAAI
Association for the Advancement of Artificial Intelligence