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

Bogdan Pikula, Steve Engels

Department of Computer Science, University of Toronto
pikula@cs.toronto.edu, sengels@cs.toronto.edu

Abstract

Facial expressions are a crucial but challenging aspect of animating in-game characters. They provide vital nonverbal communication cues, but given the high complexity and variability of human faces, the task of capturing the natural diversity and affective complexity of human faces can be a labour-intensive process for animators. This motivates the need for more accurate, realistic and lightweight methods for generating emotional expressions for in-game characters. In this work, we introduce FlexComb, a Facial Landmark-based Expression Combination model, designed to generate a real-time space of realistic facial expression combinations. FlexComb leverages the highly varied CelebV-HQ dataset containing emotions in the wild, and a transformer-based architecture. The central component of the FlexComb system is an emotion recognition model that is trained on the facial dataset, and used to generate a larger dataset of tagged faces. The resulting system generates in-game facial expressions by sampling from this tagged dataset, including expressions that combine emotions in specified amounts. This allows in-game characters to take on variety of realistic facial expressions for a single emotion, which addresses this primary challenge of facial emotion modeling. FlexComb shows potential for expressive facial emotion simulation with applications that include animation, video game development, virtual reality, and human-computer interaction.

Introduction

In animation and video games, facial expressions play an important role in conveying nonverbal information (Feldman and Rimé 1991) about a character's emotional state. While animators are experienced with modeling facial expressions for film and TV, it is more difficult to have video game characters dynamically change their facial expressions in-game, unless these expressions are modeled and programmed ahead of time. In cases where the character is feeling a combination of emotions, it isn't feasible to program these in advance, especially if the game is meant to support any combination of multiple emotions. If game characters intend to emulate the range of emotional facial expressions of which humans are capable, there is a pressing need to create more diverse and realistic techniques for generating such expressions.

Historically, facial expressions were created through a combination of manual work and software tools, usually referred to as Facial Rigs (Orvalho et al. 2012), allowing a user to manipulate a set of sliders, which in turn change corresponding facial attributes. This describes the so-called ”blend-shape rigging” or ”morph animation” pipeline. Another way to perform this task is by using the ”skeletal rigging” approach, where the facial mesh representing the skin is ”stretched” over a group of virtual ”bones” and ”muscles”. Handling the positioning of these features simulates the movement of facial features. This tends to generate a more physically accurate facial expression as it attempts to replicate the underlying facial structure. Despite having a lot of control over many intricate parameters, the process of using these approaches involves a great deal of trial and error with different permutations of these features. As a result, it leads to extensive iterative work. While it can express the space of facial expressions that humans can produce, there are also a higher number of permutations that result in anatomically incorrect faces.

Inspired by the concept of emotion combination and the incorporation of a large amount of facial expression data in the wild, we introduce FlexComb, a Facial Landmark-based (Wu and Ji 2019) Model for generating realistic facial expressions representing a variety of emotion mixes. The proposed approach leverages a vast CelebV-HQ (Zhu et al. 2022) dataset and an emotion-detection neural network, which has the training video clips trimmed to only contain emotional transitions, achieving a higher emotion detection accuracy during testing. The results show that we can build an emotional manifold (Chang, Hu, and Turk 2003), which can be sampled to yield a facial representation closest to the nearest emotion probability distribution.

Related Work

The related work can be described with regard to the following.

Motion Capture (MoCap) Systems

Motion Capture systems have always been a standard tool for capturing and reproducing human facial expressions. These systems employ actors to act out various emotions from given cues. The collected data is then used to animate digital characters, mimicking actors' original performance with a high degree of replication (Huang et al. 2011). This approach requires careful planning and usually has to be done well in advance before any animation work begins. MoCap provides a good solution but often faces the challenge of creating unusual and complicated emotions, since they are limited by the actors ability to express them on-demand. Furthermore, these systems are usually costly, since they require specialized equipment.

Emotion Recognition

Deep Learning techniques are a popular way to tackle emotion recognition tasks (Canal et al. 2022). The neural network models are primarily trained on various labeled emotional dataset, and leverage state-of-the-art Computer Vision techniques (Jain, Shamsolmoali, and Sehdev 2019), (Ko 2018). A few approaches rely on recognizing and embedding facial information into a low-dimensional space, such as FACS (Lien et al. 1998). These approaches can learn complex patterns of emerging emotions from relative positioning of facial attributes. We complement this approach by introducing a novel approach to preprocessing the training data by trimming the video sequences which contain strongly expressed emotions.

AI-based methods

Generative AI has been on the rise recently, with many developments in Generative Adversarial Networks (Siddiqui 2022a), (Siddiqui 2022b) and Diffusion Models (Zou et al. 2023) being used for facial expression generation. These can be quite robust, able to learn complex facial characteristics, although they run into the risk of creating unrealistic facial expressions, that wouldn't exist in nature. Another approach utilizes Manifold Learning (Abdrashitov, Chevalier, and Singh 2020), which guarantees that the generated facial expression stays inside the space of realistic facial expressions.

Method

This section elaborates on the pipeline for FlexComb and its components. We break down the emotional detection part of the problem, and discuss the application for utilizing it in a diverse dataset. Finally, we explore techniques for the facial emotion generation process in creative applications.

Problem Formulation

Our pipeline, described in Figure1, works up the problem into following components:

- Facial landmark analysis and corresponding Facial Action Coding System units (FACS) (Hjortsjö 1970) 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.
- Running the emotion Detection Model over the CelebV-HQ (Zhu et al. 2022) dataset to produce an space of emotion likelihoods.

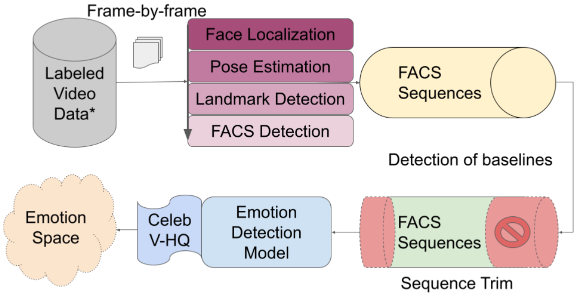

Figure 1: The general pipeline of FlexComb. We use a mixture of the following datasets: MMI (Pantic et al. 2005), OULU-CASIA (Zhao et al. 2011), DDCF (Dalrymple, Gomez, and Duchaine 2013). The video data is processed frame-by-frame to extract Facial Action Units (Hjortsjö 1970). We then trim the resulting 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.

Pipeline Architecture

FACS Analysis. We begin by breaking the video data from MMI (Pantic et al. 2005), OULU-CASIA (Zhao et al. 2011), DDCF (Dalrymple, Gomez, and Duchaine 2013) datasets into frames, which get processed individually. A series of models are employed to analyze the frames, locate the faces, find corresponding landmarks and extract Facial Activation Units for each. **RetinaFace** (Deng et al. 2019) is used for face localization, isolating the faces in each frame. **Img2pose** (Albiero et al. 2021) is used for facial pose estimation. It detects the faces and outputs a 3D head pose estimation with account for head rotations. **MobileFaceNet** (Chen et al. 2018) is used for detecting Facial Landmarks. These are 68 pairs of points (x, y) for the unique location on the face. The relative positioning and alignment of these points help with identification of Action Units. For this task, we utilize an **XGBoost Classifier** model, resulting in a vector of AU values representing facial muscle activations or ”movements”. These values serve as a facial representation for emotions, as they can be used to manipulate blended shapes to reconstruct an underlying emotion.

Sequence Trim. Next, we take the resulting FACS sequences and perform a trimming procedure. We look at the changes of FACS values from frame to frame, identifying points where a shift in emotion occurs. By eliminating long and unchanging segments from each video, we place the focus exclusively on the emotional shifts. This creates a more concise dataset, which is focused on emotional transitions. Let n be the number of frames, for which we have a sequence of FACS vectors for a given video clip: $V = (v_1, v_2, \dots, v_n)$. We define an emotional shift as a substantial change in the FACS representation. The difference of FACS values between two consecutive frames is denoted as $\Delta v_i = v_{i-1} - v_i$. We set a threshold parameter t, such that

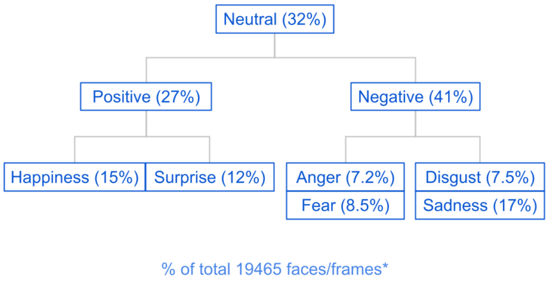

Figure 2: Distribution of emotions in the training dataset. This diagram illustrates the proportion of each of the seven unique emotions - anger, disgust, fear, happiness, sadness, surprise and neutral.

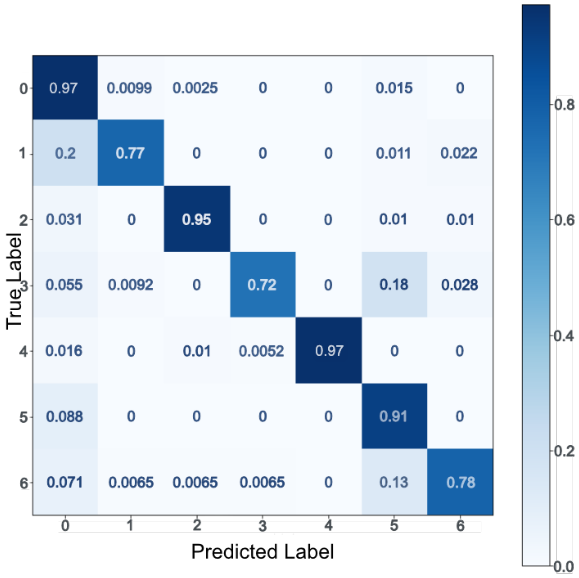

Figure 3: Confusion Matrix for the Emotion Detection Model. The matrix provides a detailed breakdown of the model's performance for classifying each of the seven emotions: neutral, anger, disgust, fear, happiness, sadness, surprise (from left to right).

if the absolute value of Δv_i exceeds it, the change is labeled as substantial, and the frame i depicts an emotional shift. We denote the index of frames where the emotional shift is detected as $s_j \in \mathcal{S}$. We then slice the sequence into segments, bound by $[s_j : s_{j+1}]$. Given the nature of the data, we remove the first segment, while defining a constant variable d, serving as a buffer size for capturing the approximate start and end of the transition. The slicing process is illustrated in 1.

$$\hat{Seq} = \{V[s_j - d : s_{j+1} - d], s_j \in S, j \neq 0\} \quad (1)$$

Emotion Detection. In this step, we introduce an emotion detection model, trained on the previously acquired data comprised of FACS sequences. This approach allows the model to learn the temporal dynamics of facial expressions, providing more context for emotion recognition. The model is trained to classify a FACS sequence into seven unique emotions, shown on Figure 2. The model is capable of detecting the emergence of these emotional states over time as it captures patterns in the sequences of FACS. The trim data was split into training/test data using an $80/20$ ratio. When assessing the performance of our sequence neural network model on the test set, it achieved an accuracy of 90%, indicating a high level of overall precision. For a more detailed look, we show the confusion matrix in Figure 3.

Method

In this section, we elaborate on the structure of FlexComb and its components. We start with the dataset, which includes the processing of the video clips from the dataset to extract AU sequences. We then break down the idea of the emotion space, and explore ways of interracting with it to generate facial expressions for emotion combinations. Additionally, we investigate different practices using this method in creative applications.

Dataset

The centerpiece of our approach revolves around the CelebV-HQ (Zhu et al. 2022) 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. Thanks to this, the data provides valuable insight into the complex, diverse and often subtle nature of human facial expressions, which occur naturally, without any acting involved.

The randomness of these expressions, paired with a combination of various environmental conditions, makes for a more organic dataset when compared to actor-based ones. To aid with diversity and complexity, CelebV-HQ (Zhu et al. 2022) also has the longest combined duration of video clips, along with manually labeled sets of facial attributes, and corresponding actions.

Data Preprocessing

In order to use the data, we extract the corresponding AU's 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. As the FACS extraction provides a high-dimensional understanding of the facial data, we are provided with a interpretable representation, which can be used to understand the relationships for different facial actions. The correlation of AU's found in the dataset is depicted on Figure 4. This visualization shows that despite a few exceptions, most facial activations occur independently, which further reinforces the complexity of human emotional expression.

Emotion Space

The sequences from the preprocessing are passed through the emotion detection model, effectively processing and generating 1290000 samples of emotional probability vectors. As a result, each processed sample yielded a probability

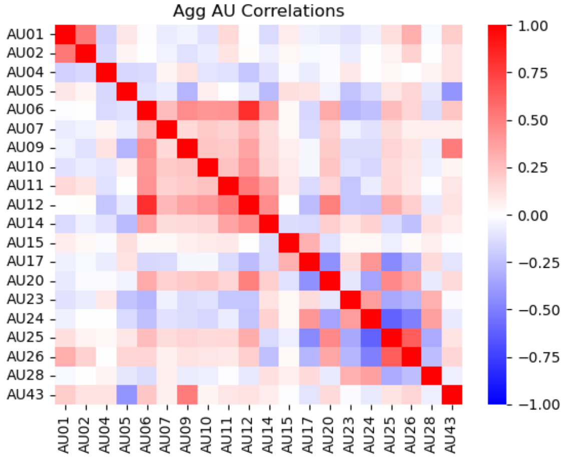

Figure 4: Correlation matrix of Action Units (AUs) in the CelebV-HQ (Zhu et al. 2022) dataset. Each cell represents the correlation coefficient between pairs of AUs, providing a generatlized visual depiction of their co-occurence patterns. Red colors indicate higher correlation, while blue - the opposite.

vector indicative of any emotions identified, and their corresponding intensity. To visualize and understand the structure of this high-dimensional emotional space, we utilize Principal Component Analysis (Abdi and Williams 2010). The resulting plot, taking the shape of a triangle, indicates of there being three principal emotions: disgust, sadness and anger. Other emotions fall along edges of the triangle, except for fear and surprise which intersect towards the middle, suggesting there exists a strong correlation between them and indicating that they can co-occur or be confused.

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 AU-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.

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. In contrast, FlexComb dynamically creates diverse facial expressions by leveraging an extensive dataset and bypassing manual emotion modeling. Additionally, the computational efficiency of FlexComb ensures faster generation of facial expressions without compromising on quality.

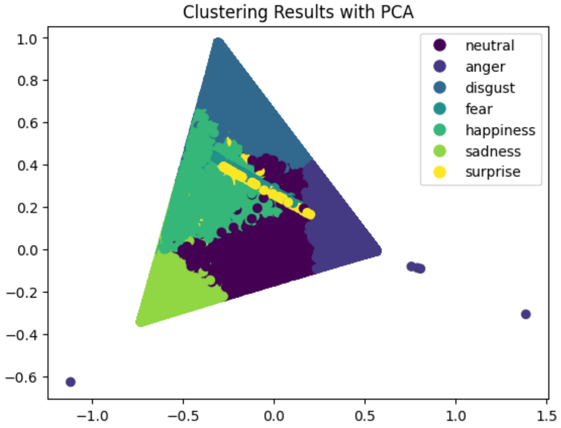

Figure 5: PCA Decomposition of Emotion Probability Vectors clustered by the emotion label.

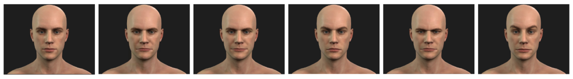

Figure 6: An example of five generated by FlexComb facial expressions, representing different ratios of combining Neutral and Angry emotions, as visualized on a FaceRig.

Generation Examples

Figure 6 showcases the spectrum of neutral to angry expressions. It highlights progressive blending, with anger's intensity gradually amplifying, juxtaposed with a baseline state with no activated action units for reference.

Figure 7 offers an insight into the blending of happiness and disgust. This serves as evidence of FlexComb's ability in managing intricate emotional combinations, emphasizing its ability to seamlessly blend contrasting emotions.

To show that the model is capable of combining more than two emotions, we generate a combination of equal parts fear, sadness and surprise, shown in Figure 8. The resulting facial expression convincingly conveys a subtle combination of these emotions.

Expression Realism and Coherence

One of the big strengths 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 (Zhu et al. 2022) 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 (Hjortsjö 1970) to provide a simple and light facial representation, that can be utilized to animate facial meshes in Facial Rigs (Orvalho et al. 2012).

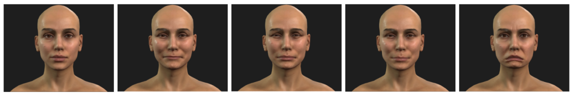

Figure 7: An example of facial expressions generated by FlexComb, representing different rations of combining Happiness and Disgust emotions, as visualized on a FaceRig.

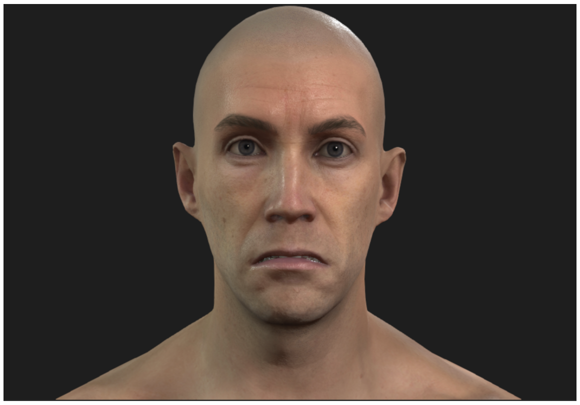

Figure 8: An example of facial expressions generated by FlexComb, representing equal parts Fear, Sadness and Surprise, as visualized on a FaceRig.

Discussion

Model Performance Analysis

The model is able to generate a facial expression, randomly sampled from an emotional space, for one of specified emotion(neutral, anger, disgust, fear, happiness, sadness, surprise), as well as arrangement of any of them using a user-defined combination ratio, representing the intensity for each. The emotional distributions for each expression have a high degree of confidence due to the emotion detection model's test accuracy of 90% on a dataset of video clips containing sequences of video clips.

Since neutral facial expressions are the most common, they have a higher percentage of occurrence, allowing for a more detailed emotional combination. Combinations of other emotions with neutral fundamentally allow the user to model that particular emotion more accurately, increasing the intensity by decreasing the ratio for the ”neutral” emotion.

Visual Interpretation

The model generates sequences of facial activation units, which need to be interpreted by another model or a piece of software, such as any Face Rig. We show that it's easily possible to visualize the expression by using an out-of-the-box Face Rig solution that gives direct control over the textured blended facial mesh. We show that facial expressions can get fascinating by tweaking the emotion combination ratio, even displaying unusual and potentially strange combinations of polarizing emotion labels.

Applications

One of the main idea for using FlexComb is in video game development. Modern video games have a significant amount of Non-playable characters (NPC's), each requiring separate input from the animator. This approach could speed up the development, while also creating a more realistic and expressive facial expression for a more immersive experience. The area of animation and film could also see benefit from using FlexComb to set the starting point for an emotional look of a character. Another idea for using FlexComb is in the field of Virtual Reality. Recent developments of virtual reality environments, such as ”Metaverse” (Mystakidis 2022), call for animated characters representing the user wearing the VR headset. These characters mimic the user in many ways, including facial expressions. This along with a deeper understanding of emotions could help model virtual interactions more closely.

Limitations and Future Work

Currently, we only focus on generating static facial expressions. While having them can serve as a good starting point for an animation, it's still a major challenge to animate facial expressions for emotional transitions. Another limitation would be the emotional labels used for emotion detection. Different interpretation of the affective domain can dictate a different approach for classifying emotions, meaning that there is a much bigger range of emotions humans can express. In some cases, it is limiting to reduce that range to just seven emotions. While being lightweight and natural, the proposed method only produces the expressions that exist in the dataset, relying heavily on it. This creates limitations for the space of expressions represented. In the current state, it cannot generate novel and unseen expressions that fall outside of the dataset.

In future work, we would like to address these limitation, and work on the generation of sequences of facial expression representing facial expressions stemming from one (or more) static ones.

Conclusion

In this work, we address the need for a lightweight framework for generating natural facial expressions for in-game characters, ones that reflect the character's dynamic emotional state. We introduce FlexComb, a Facial Landmark-based Expression Combination model that leverages a transformer-based architecture, trained on the extensive labelled CelebV-HQ dataset. The result is a system that consistently generates a diverse set of natural facial expressions for any combination of emotions, instead of being limited to a set of single (dominant) emotions.

Contrary to traditional blendshapes or other prevalent methods which produce fixed, often linear, facial movements, FlexComb offers a nuanced approach to facial animation by creating expressions that can blend multiple emotions. It capitalizes on real-world nuances captured in the CelebV-HQ dataset, ensuring that facial expressions are not just a replication of standard, dominant emotions but instead reflect the complex combinations of emotions humans often exhibit.

We show that a simple sequence trimming technique for video clip training data contributes to a better performance in predicting the emerging emotions, demonstrating its ability to capture emerging emotional cues. FlexComb is capable of generating realistic facial expressions for any emotion combination, paving the way for animating more realistic in-game characters by drawing from real-world facial expressions. The result is a natural static expression, which can be generated in sequence to create a series of facial expressions. Future work will aim to address these limitations by developing a model that can generate this continuous sequences of facial expressions. In addition, FlexComb will integrate an interface to help fine-tune the visualization of the sequences, thereby improving the application potential in the video game development and animation industries.

References

Abdi, H.; and Williams, L. J. 2010. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4): 433–459.

Abdrashitov, R.; Chevalier, F.; and Singh, K. 2020. 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, 778–790. New York, NY, USA: Association for Computing Machinery. ISBN 9781450375146.

Albiero, V.; Chen, X.; Yin, X.; Pang, G.; and Hassner, T. 2021. img2pose: Face Alignment and Detection via 6DoF, Face Pose Estimation. arXiv:2012.07791.

Canal, F. Z.; Müller, T. R.; Matias, J. C.; Scotton, G. G.; de Sa Junior, A. R.; Pozzebon, E.; and Sobieranski, A. C. 2022. A survey on facial emotion recognition techniques: A state-of-the-art literature review. *Information Sciences*, 582: 593–617.

Chang, Y.; Hu, C.; and Turk, M. A. 2003. Manifold of facial expression. In *AMFG*, 28–35.

Chen, S.; Liu, Y.; Gao, X.; and Han, Z. 2018. MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices. arXiv:1804.07573.

Dalrymple, K. A.; Gomez, J.; and Duchaine, B. 2013. The Dartmouth Database of Children's Faces: Acquisition and Validation of a New Face Stimulus Set. *PLoS ONE*, 8(11): e79131.

Deng, J.; Guo, J.; Zhou, Y.; Yu, J.; Kotsia, I.; and Zafeiriou, S. 2019. RetinaFace: Single-stage Dense Face Localisation in the Wild. arXiv:1905.00641.

Feldman, R. S.; and Rimé, B. 1991. *Fundamentals of nonverbal behavior*. Cambridge University Press.

Hjortsjö, C.-H. 1970. *Man's face and mimic language*. Studentlitteratur Lund, Sweden.

Huang, H.; Chai, J.; Tong, X.; and Wu, H.-T. 2011. Leveraging Motion Capture and 3D Scanning for High-Fidelity Facial Performance Acquisition. In *ACM SIGGRAPH 2011 Papers*, SIGGRAPH '11. New York, NY, USA: Association for Computing Machinery. ISBN 9781450309431.

Jain, D. K.; Shamsolmoali, P.; and Sehdev, P. 2019. Extended deep neural network for facial emotion recognition. *Pattern Recognition Letters*, 120: 69–74.

Ko, B. C. 2018. A Brief Review of Facial Emotion Recognition Based on Visual Information. *Sensors*, 18(2): 401.

Lien, J.; Kanade, T.; Cohn, J.; and Li, C.-C. 1998. Automated facial expression recognition based on FACS action units. In *Proceedings Third IEEE International Conference on Automatic Face and Gesture Recognition*, 390–395.

Mystakidis, S. 2022. Metaverse. *Encyclopedia*, 2(1): 486–497.

Orvalho, V.; Bastos, P.; Parke, F. I.; Oliveira, B.; and Alvarez, X. 2012. A Facial Rigging Survey. *Eurographics (State of the Art Reports)*, 183–204.

Pantic, M.; Valstar, M.; Rademaker, R.; and Maat, L. 2005. Web-based database for facial expression analysis. In *2005 IEEE International Conference on Multimedia and Expo*, 5 pp.–.

Siddiqui, J. R. 2022a. Explore the Expression: Facial Expression Generation using Auxiliary Classifier Generative Adversarial Network. arXiv:2201.09061.

Siddiqui, J. R. 2022b. FExGAN-Meta: Facial Expression Generation with Meta Humans. arXiv:2203.05975.

Wu, Y.; and Ji, Q. 2019. Facial landmark detection: A literature survey. *International Journal of Computer Vision*, 127: 115–142.

Zhao, G.; Huang, X.; Taini, M.; Li, S. Z.; and Pietikäinen, M. 2011. Facial expression recognition from near-infrared videos. *Image and Vision Computing*, 29(9): 607–619.

Zhu, H.; Wu, W.; Zhu, W.; Jiang, L.; Tang, S.; Zhang, L.; Liu, Z.; and Loy, C. C. 2022. 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*, 650–667. Springer.

Zou, K.; Faisan, S.; Yu, B.; Valette, S.; and Seo, H. 2023. 4D Facial Expression Diffusion Model. arXiv:2303.16611.