International Journal of Religion
2024
Volume: 5| Number 11 | pp. 9078 – 9094
ISSN: 2633-352X (Print) | ISSN: 2633-3538 (Online)
ijor.co.uk
DOI: https://doi.org/10.61707/n24d8d13

2D/3D Expression Generation Using Advanced Learning Techniques and the Emotion Wheel

Thi Chau Ma[1], Anh Duc Dam[2] and Quang Hieu Dao[3]

Abstract

Facial expression analysis is a critical component in numerous applications, ranging from human- computer interaction to digital character animation. Despite the availability of extensive datasets, most focus predominantly on basic emotions, limiting the expressiveness and applicability of models trained on them. This article introduces a novel approach to generating compound emotional expressions by leveraging the Emotion Wheel, a principle that captures the complex interrelations between basic emotions. Our method integrates the EMOCA (Emotion Driven Monocular Face Capture and Animation) [1] framework, which enhances 3D facial reconstruction by incorporating emotion recognition, stacking models [2, 54] with a sophisticated expression blending algorithm to synthesize nuanced 2D and 3D facial animations. Utilizing the VKIST dataset, which includes high-resolution facial images of Vietnamese individuals, we build a comprehensive database of emotion parameters. Through principal component analysis (PCA) and correlation-driven blending, our approach not only enhances the realism of generated facial expressions but also preserves the subtle nuances that are characteristic of human emotions. Experimental results demonstrate that our method consistently outperforms traditional linear interpolation techniques, producing more distinct and recognizable blended expressions. A user study further validates the naturalness and quality of the generated expressions, with average ratings indicating a strong preference for the proposed method over existing approaches. These findings suggest significant potential for improving emotional expressiveness in both static and dynamic digital character representations.

Keywords: *Expresion Retargeting, Expression Recognition, Expression Blending*

INTRODUCTION

The field of 3D facial expression reconstruction has witnessed significant advancements, driven by its side-ranging applications in healthcare [3, 4, 5] and entertainment [3, 6, 7]. The ability to generate diverse and nuanced emotional expressions for virtual characters is highly valued in the entertainment industry. This capability enhances storytelling and increases audience engagement by enabling the creation of more lifelike and relatable characters [6, 7]. The creation of emotionally expressive characters is crucial for delivering the immersive experience expected by audiences in modern digital media.

Similarly, emotion recognition technology is playing an increasingly significant role in healthcare, particularly in mental health diagnosis and treatment. By enabling remote emotion recognition and monitoring, this technology improves the accuracy of medical diagnoses and facilitates real- time assessment of patient emotions [4, 5]. Furthermore, emotion regulation is gaining recognition as a vital aspect of therapeutic settings. Research suggests that emotions typically last for approximately six seconds [8], and prolonged negative emotions can have detrimental effects on an individual's mood and daily life. Emotion recreation technologies offer a potential solution by transforming negative emotional states into more positive ones, thus promoting mental well-being [9].

However, most existing methods for facial expression analysis and generation are primarily focused on basic emotions [10-17]. This limitation restricts the expressiveness and applicability of models trained on these

[1] Faculty of Information Technology, VNU University of Engineering and Technology, No. 144 Xuan Thuy Street, Dich Vong Ward, Cau Giay District, Hanoi, Vietnam. Email: chaumt@vnu.edu.vn

[2] VNU University of Engineering and Technology, No. 144 Xuan Thuy Street, Dich Vong Ward, Cau Giay District, Hanoi, Vietnam Email: damanhduc140902@gmail.com

[3] Software Development Department, Flinters Vietnam, Duy Tan street, Dich Vong Ward, Cau Giay district, Hanoi, Vietnam. Email: dqhieu01@gmail.com

datasets. Consequently, there is a need for developing methods that can generate more complex and nuanced emotional expressions, encompassing a wider range of human emotions.

To overcome this limitation, our research aims to move beyond the constraints of basic emotions by creating more complex and nuanced emotional expressions for 3D faces. By achieving this goal, we seek to enhance both the realism and the therapeutic utility of 3D facial reconstructions. To this end, we propose a novel method using advanced learning techniques and the Emotion Wheel for combining detected emotional expression parameters with other carefully selected emotion expressions to generate compound emotional expressions. This approach enables the creation of more complex and nuanced emotional states in 3D facial reconstructions, resulting in significantly enhanced expressiveness and realism.

RELATED WORKS AND BACKGROUND

Expression Database Review

Numerous datasets have been compiled to advance the field of automatic facial expression analysis. However, a closer look at these datasets reveals a significant limitation: a predominant focus on basic emotions. The Radboud Faces Database (RaFD) [10], for instance, comprises images of 49 models expressing eight facial expressions, including the six basic emotions defined by Ekman, along with variations in gaze direction. Similarly, the Cohn-Kanade (CK+) database [11] consists of 593 image sequences from 123 subjects, focusing on posed transitions from neutral to peak expressions categorized into seven basic emotion categories. The Japanese Female Facial Expression (JAFFE) dataset [12], a smaller dataset with 219 images of 10 Japanese women, also employs posed expressions associated with the six basic emotions plus neutral. Other widely used datasets, such as the Static Facial Expressions in the Wild (SFEW) [13], the Denver Intensity of Spontaneous Facial Action (DISFA) [14], and the Facial Expression Recognition 2013 (FER) database [15] largely follow the same paradigm, classifying expressions into a limited set of basic emotion categories.

Even with the advent of larger-scale datasets like AffectNet [16] and Aff-Wild [17], which contain hundreds of thousands of images collected “in-the-wild”, this focus on basic emotions persists. While these datasets provide valuable resources for training and evaluating deep learning models, their reliance on a constrained set of emotion labels inherently limits the expressiveness and granularity of the learned representations. However, as evidenced by research in psychology and cognitive science, human emotion is far richer and more nuanced than can be encapsulated by these few categories. Real-world expressions are often complex, subtle, and blended, conveying a mix of emotions that defy simple categorization.

This need for more comprehensive and ecologically valid emotion representations has spurred the development of datasets like RAF-DB [18] and MAFW [19]. These datasets go beyond basic emotions by providing annotations for compound expressions, acknowledging the simultaneous experience of multiple emotions. However, the number and variety of compound emotion categories in these datasets remain limited, highlighting the need for continued efforts in this direction. The ability to accurately recognize, analyze, and generate compound emotions is critical for developing truly sophisticated and human-like facial expression analysis systems. The present work addresses this challenge by proposing a novel method for generating new, expressive emotion representations based on combining basic expressions.

Emotion Reconstructions

The task of 3D facial reconstruction from a single image is intricately linked to facial expression analysis. Traditional approaches, primarily based on 3D Morphable Models (3DMMs), have seen significant evolution, beginning with pioneering work like Blanz and Vetter's statistical model [20]. This foundational research has paved the way for more advanced frameworks such as the Basel Face Model (BFM) [21, 22] and the 3D Dense Face Alignment (3DDFA) framework [23]. However, despite these advancements, these methods often struggle to capture the subtle nuances that are crucial for rendering realistic and emotionally expressive faces.

Most existing techniques rely on differentiable rendering to compare predicted face meshes with input images, complicating the optimization process due to domain gaps. In contrast, SMIRK (Spatial Modeling for Image-based Reconstruction of Kinesics) [24] employs a neural rendering module that generates face images from the predicted mesh geometry and sparsely sampled pixels. This approach allows for a focus on geometry during supervision and enables the generation of varying expressions during training. By augmenting the training data with these generated images, SMIRK enhances generalization for diverse expressions. Evaluations demonstrate that SMIRK achieves state-of-the-art performance in accurate expression reconstruction.

Conversely, deep learning-based techniques provide a compelling alternative. Approaches such as volumetric CNNs [25], image-to-image translation networks [26], and end-to-end frameworks

[27] have emerged as powerful tools in this domain. Particularly noteworthy is the EMOCA model [1], which ingeniously integrates emotion recognition into a self-supervised learning framework for 3D facial reconstruction. By capturing subtle details of expressions while simultaneously predicting emotions, EMOCA exemplifies the potential of deep learning to synthesize 3D facial reconstructions that are not only accurate but also rich in emotional expressiveness.

This integration of emotional context into the reconstruction process represents a significant leap forward, allowing for more nuanced and lifelike representations of human faces. As the field continues to advance, the combination of traditional modeling techniques with deep learning innovations could lead to even more sophisticated methods for understanding and generating 3D facial expressions, ultimately enhancing applications in areas such as virtual reality, animation, and human-computer interaction.

Emotion Blending

The ability to synthesize and manipulate facial expressions, including the creation of compound emotions, is crucial for a wide range of applications, from animation and gaming to virtual reality and mental health assessments. Traditional techniques, such as Active Appearance Models (AAMs) [28-31], 3D Morphable Models (3DMMs) [31-35], and Blendshape Models [36-39], primarily rely on parametric representations and pre-defined deformations to generate facial expressions. While effective in certain contexts, these methods often fall short in capturing the full complexity of human emotions.

In contrast, image-based approaches, such as the Expression Ratio Image (ERI) technique [32, 40- 42] and various image warping methods [29, 40, 43-46], manipulate image data directly, offering greater flexibility and adaptability in expression synthesis. These methods allow for more dynamic changes in facial features, but they can still struggle with the subtleties of nuanced emotional expression.

The emergence of conditional Generative Adversarial Networks (cGANs) has significantly transformed the landscape of facial expression generation. By leveraging continuous emotion representations [47-51], cGANs excel at producing realistic and varied facial expressions. Frameworks like GANmut [51] exemplify this advancement, enabling the synthesis of complex emotions from basic categorical labels. This capability allows for a richer emotional portrayal, making digital characters more relatable and lifelike.

Furthermore, ongoing research into techniques that preserve individual identity during expression manipulation enhances the realism of synthesized faces. Methods that maintain unique facial features while allowing for emotional variability are essential for creating authentic virtual interactions. These advancements collectively point toward a future where synthesizing realistic and emotionally expressive faces is not only achievable but also practical for various applications. The integration of these sophisticated techniques reflects a significant leap forward in the quest for lifelike digital interactions, paving the way for innovations in entertainment, therapy, and human-computer interaction.

Emotion Wheel

The "Emotion Wheel" [52] (Figure 1) is a visualized principle used to help people identify, understand, and communicate about their emotions. It typically consists of a circular diagram with different emotions arranged in a way that represents their relationships to one another.

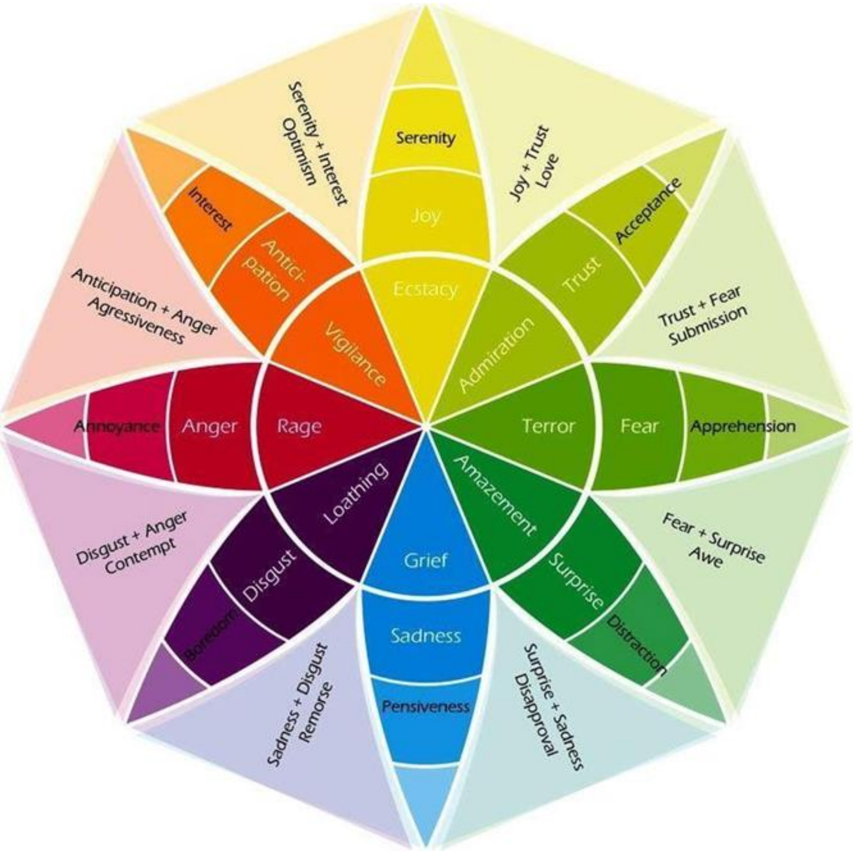

Figure 1. Emotion Wheel [52]

The Emotion Wheel is divided into different segments, each representing a basic emotion such as joy, anger, fear, sadness, or disgust. These basic emotions are then further broken down into more specific emotions related to or derived from the basic emotions. For example, the joy segment may include emotions like happiness, contentment, pride, and excitement. The anger segment may include disappointment, irritation, rage, and resentment. The fear segment may include worry, anxiety, terror, and dread.

The Emotion Wheel is designed to help people: (i) Identify their emotions: By looking at the different emotions on the wheel, people can pinpoint the specific emotion they are experiencing;

(ii) Understand emotional relationships: The way the emotions are arranged on the wheel shows how different motions are connected and can influence each other; (iii)Communicate about emotions: The wheel provides a common language and framework for discussing and expressing emotions. By leveraging the emotional relationships depicted on the Emotion Wheel, individuals can gain insights into how different emotions are connected and potentially generate new emotional experiences. The Emotion Wheel can be a useful tool in various contexts, such as: Character development and building in applications related to healthcare, filmmaking, or video games; Personal reflection, therapy, and emotional intelligence training; and Team-building activities.

Proposal for Expression Blending Techniques in Character Animation

Reconstructing realistic and expressive facial animations is crucial for creating convincing 2D and 3D characters. However, capturing the necessary 2D and 3D data required to achieve this level of facial expressiveness can be a complex and resource-intensive process. The data capture equipment and workflows needed to generate high-quality 3D facial expression models are often expensive and time-consuming. Additionally, many of the existing 2D and 3D facial expression databases available to developers tend to be limited in scope, typically containing only around 6 to 8 basic, prototypical expressions, as outlined in Section 2.1. This limitation in available data can make it challenging for creators to imbue their characters with the full range of nuanced and dynamic facial expressions that contribute to naturalistic and engaging character performances. Overcoming these technical and resource barriers remains an ongoing challenge in the field of digital character creation and animation.

In this work, we present a novel approach to generate a wider range of character facial expressions by leveraging the structure and relationships defined within the Emotion Wheel principle. The proposed process, including 3 main steps (i) Emotion Prediction; (ii) Expression Blending; and (iii) Target reconstruction, is outlined in Figure 2.

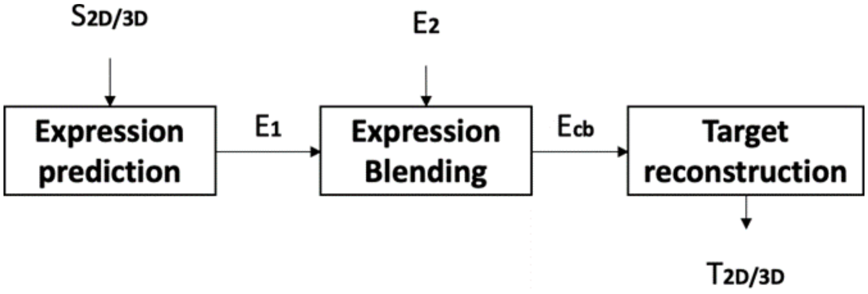

Figure 2. Process of Expression Blending for Character Faces

We refer to source $S2D/3D$ as the 2D image or 3D model of the character whose emotional expression we want to transform. To create a new emotional expression for the character based on the Emotion Wheel, we first predict the initial expression $E1$ of the $S2D/3D$. In this context, $E1$ represents the parameters that define the initial emotional expression of the S2D/3D character.

$E2$ represents the parameters that define one of the expressions on the adjacent circle r in the

Emotion Wheel. In this phase, we propose utilizing advanced learning techniques, such as machine learning or deep learning, given the availability of datasets that are highly conducive to training, to effectively extract expressions from the input data.

We have built a database DBEX containing the parameter sets representing the different basic emotions that appear in the Emotion Wheel, including: Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, and Anticipation (section 3.1). We propose a formula to combine $E1$ and $E2$ to create a new composite expression, Ecb(section 3.2). For example, if $E1$ is Sadness, then $E2$ could be Surprise or Disgust. When combining Sadness and Surprise, we expect the resulting Ecb to be Disapproval, while combining Sadness and Disgust, we expect the resulting Ecb to be Remorse. The new emotional parameter Ecb will be applied to the 2D or 3D character through the target reconstruction phase to obtain the target $T2D/3D$.

DBEX Building

To build the DBEX database containing the parameters of the basic expressions: Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, and Anticipation, we utilized the VKIST facial expression image dataset. The VKIST dataset consists of 441 sets of facial expression images from 441 individuals. Each set of images for each individual has 7 facial expressions: Neutral, Joy, Fear, Surprise, Sadness, Disgust, and Anger. Firstly, we generated 2 additional sets of Trust and Anticipation expression images using the formula mentioned in section 3.2. We used the 8 expression image sets including Joy, Trust, Fear, Surprise, Sadness, Disgust, Anger, and Anticipation to generate the expression parameter set for each individual expression. Next, we calculated the average of these parameter sets, which we refer to as $B1$.

Recognizing the inherent variability in facial expressions and the potential for ambiguous emotional portrayals, we adopted a refined approach to enhance the clarity and accuracy of our database. Instead of averaging all available expression parameters, we meticulously curated a subset of images that exemplify prototypical expressions for each emotion category, guided by the Facial Action Coding System (FACS) [53], which we refer to as $B2$.

The parameter set $B2$represents clearer and more accurate emotional expressions (Figure 3) through evaluation and comparison with the FACS expression definitions. Therefore, we used $B2$ as the DBEX parameter database.

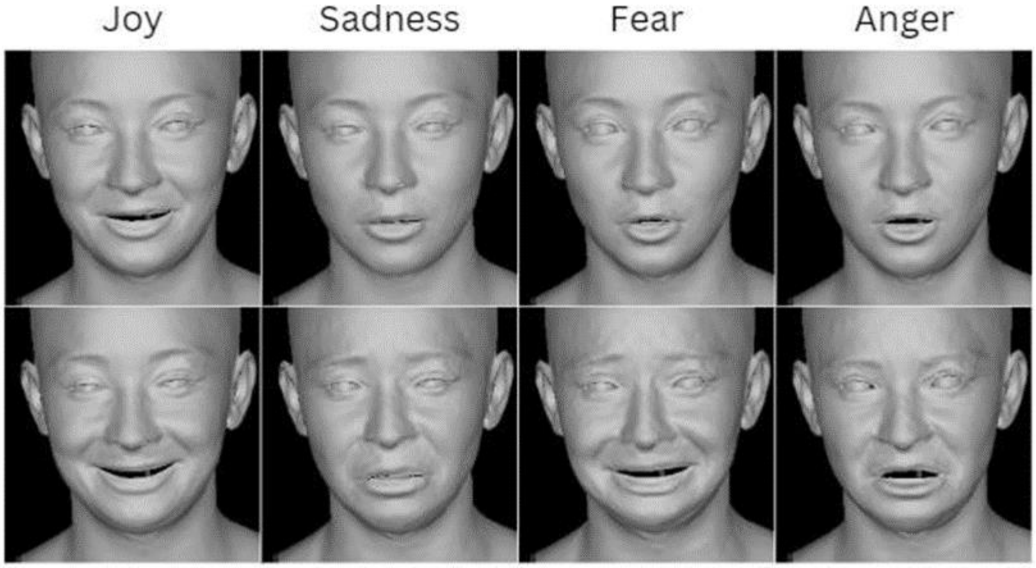

Figure 3. DBEX parameter database building (top) $B1$ and (bottom) $B2$

Expression Blending

To combine two sets of expression parameters, associated with expressions $E1$ (predicted expression) and $E2$ (expression selected to combine) respectively, we introduce a formula that leverages the correlations between these parameters.

Ecb = $w1$ $E1$+$w2E2$ (1)

where, $w1$+$w2$=1.

The expression parameters $E1$ and $E2$ are typically represented as vectors or matrices. Within the components of these vector/matrix representations, there are certain key elements that have a significant influence on the visual characteristics of the two emotional expressions. Let

$IE1$ and $IE2$ represent the sets of indices corresponding to the most correlated components for expressions $E1$ and $E2$, respectively. We define the overlapping indices as:

Io = $IE1$ ∩ $IE2$ (2)

The specific formula will be provided as follow:

$$E\ \ [i] = \begin{matrix} w_1 E_1[i] + w_2 E_2[i]\ \ if\ i \in I_{E_1} \backslash Io \\ w_2 E_1[i] + w_1 E_2[i]\ if\ i \in I_{E_2} \backslash Io \end{matrix} \quad (3)$$

$$\frac{E_1[i] + E_2[i]}{2}$$

The formula above shows that the combined expression will focus the weighting on the key components within the parameters $E1$ and $E2$.

Executing the Proposal through Expression Recognition with Stacking Model and Face Reconstruction Using EMOCA

Dataset

The dataset used in our research is the VKIST dataset, a novel and valuable compilation obtained from the Vietnam-Korea Institute of Science and Technology. The VKIST dataset comprises high- quality facial images from 441 Vietnamese individuals, each exhibiting seven distinct basic emotions: Neutral, Joy, Sadness, Fear, Anger, Surprise, and Disgust. This results in a total of 3087 diverse facial expression images in the dataset. The VKIST dataset represents a unique and comprehensive resource for studying human facial expressions. The breadth of emotions captured, the high image quality, and the consistent framing and preprocessing of the data make it an invaluable asset for our research. All images were captured from a frontal view with an impressively high resolution of 2987×1984 pixels. This exceptional level of detail ensures that the subtle facial features and nuances of each emotional expression are clearly visible and can be accurately analyzed. To further prepare the VKIST dataset for the Expression Blending phase, the facial regions were carefully extracted from the full images and resized to a standardized 224×224 pixel format. This standardization step preserves the critical facial details while also enabling efficient processing and analysis of the data.

Experiment Scenario

We conducted experiments to validate the proposal presented in Section 3. When conducting experiments and evaluating the results of expression blending, we utilized the EMOCA framework [1] and Stacked Machine Learning Models [54]. With a source input $2DImageS$ of a character with expression $E1$, we aimed to reconstruct a target $2DImageT$ and a 3D model of a character with a new expression Ecb by blending a queried expression $E2$ from the VKIST dataset with the character's expression $E1$. We followed the experimental process outlined below (Figure 4):

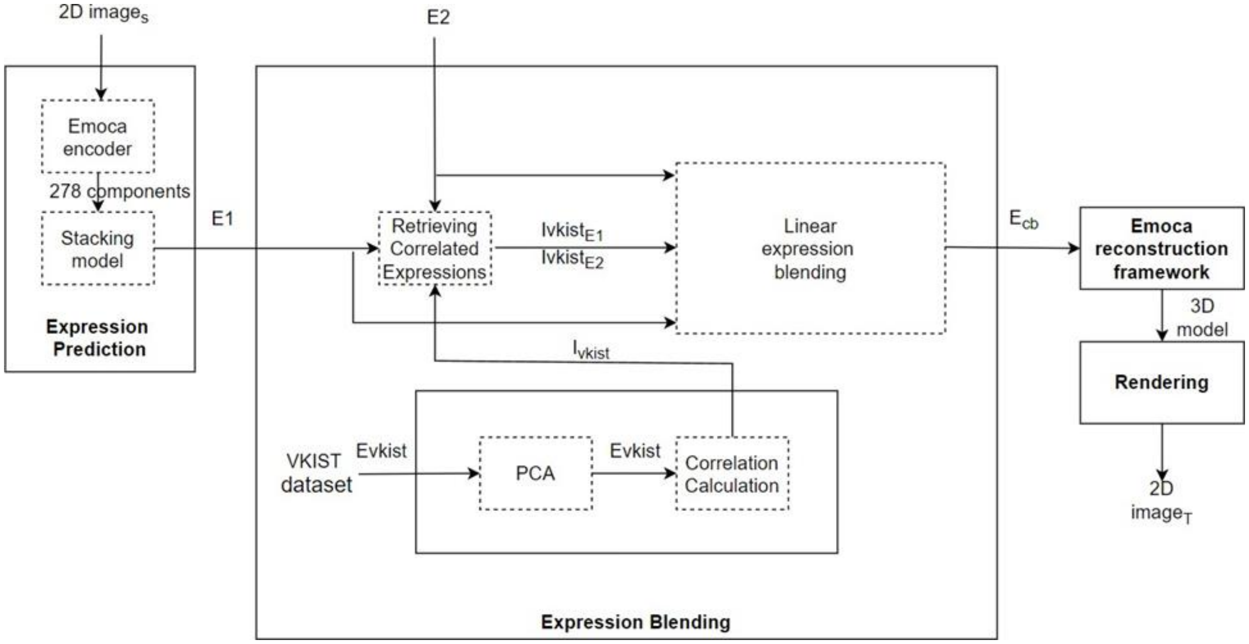

Figure 4. Experiments of 8 new expression generation

The EMOCA encoder within the EMOCA framework architecture extracts 278 components from shape, expression and detail parameters (100 from shape, 50 from expression and 128 from detail) from the $2DImageS$. These parameters are then fed into a stacked model [54] to predict the corresponding emotion label. Both the predicted emotion label and the expression parameters $E1$ will be used as input for the subsequent expression blending phase, where they will be combined with the expression parameters $E2$ from the VKIST dataset.

To perform expression blending, we first utilize the entire VKIST dataset to create a correlation- driven blending method. We apply Principal Component Analysis (PCA) to the 50 parameters extracted from all images in the VKIST dataset. Our analysis, as illustrated in Figure 5, reveals that the first 16 principal components (PCs) capture approximately 95% of the variance in the expression parameters. This suggests that these 16 PCs contain the most significant information about the variations in facial expressions, allowing us to effectively represent the high- dimensional expression space in a more compact and efficient manner.

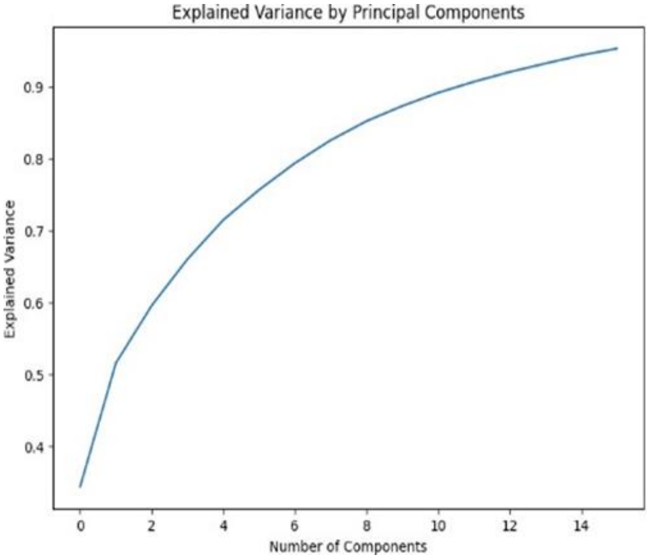

Figure 5. Explained Variance by principal component.

Building upon this insight from PCA, we delve deeper into the relationship between individual expression parameters and specific emotions.

We then calculate the Pearson correlation coefficients between each of the 50 expression components and each of the seven emotion labels. This analysis allows us to pinpoint the 16 parameters that exhibit the strongest correlation with each emotion, effectively identifying the most influential parameters in conveying specific emotional states. These correlations serve as the foundation for our expression blending algorithm, as they provide a quantitative measure of the relationship between individual expression parameters and specific emotions.

Appling formula (3), the expressions were blended linearly as formula (4) in detail and $w1$ and

$w2$ are the weights assigned to the primary and secondary emotions.

$$E_{cb}[i] = \begin{cases} w_1 E_{1PCA}[i] + w_2 E_{2PCA}[i] \ \ if \ i \in I_{E_{1PCA}} \backslash I \\ w_2 E_{1PCA}[i] + w_1 E_{2PCA}[i] \ if \ i \in I_{E_{2PCA}} \backslash Io \\ \dfrac{E\ [i] + E\ [i]}{2} \end{cases} \quad (1)\,(4)$$

Where, $IvkistE1$ and $IvkistE2$ represent the sets of indices corresponding to 16 the most correlated components for expressions $E1$ and $E2$, respectively, and $Io = IvkistE1 \cap IvkistE2$. By adjusting these weights, we can control the relative contribution of each emotion to the final blended expression. For example, setting $w1 = w1 = 0.5$ would result in a simple average of the two expressions.

To ensure the combined parameters stay within the valid range of expression values, we apply a clipping operation:

$Ebounded = clip(Ecb, Emin, Emax)$ (5)

where $Emin$ and $Emax$ are the minimum and the maximum values observed in the VKIST dataset for each expression parameter. This clipping step prevents the generation of unrealistic or exaggerated expressions.

Evaluation

Figure 6 and 7 illustrate the results of blending the expressions on the 2nd ring of the Emotion Wheel. On the ring, the Expression E1 appears before proceeding counterclockwise.

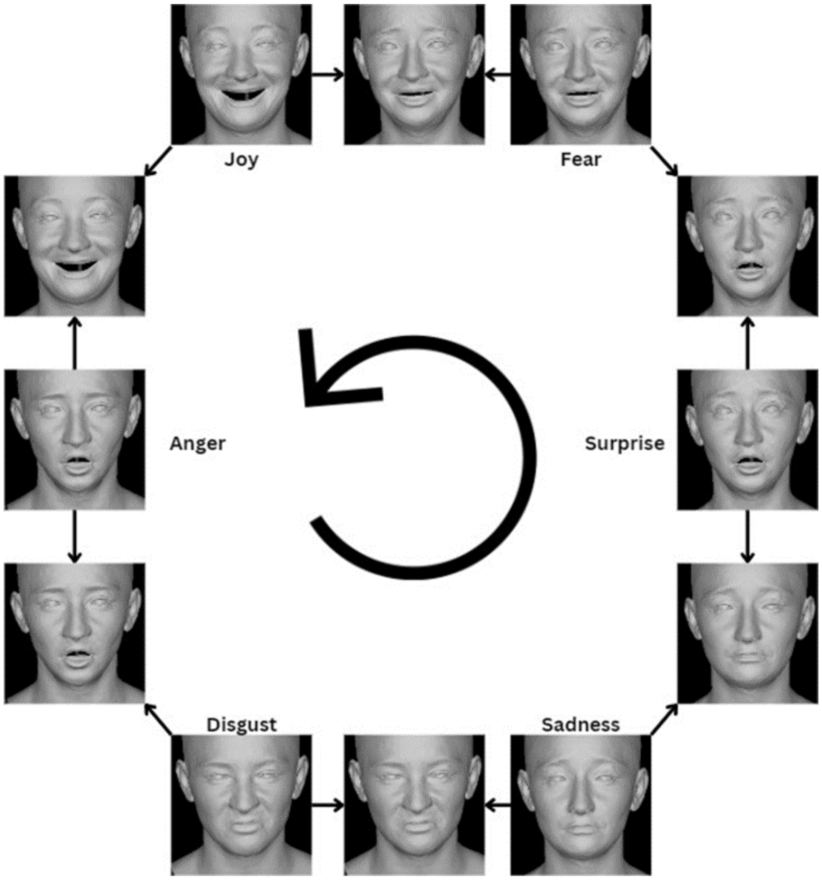

Figure 6. Linear expression blending (counterclockwise)

Figure 6 illustrates the linear combination of all components of E1 and E2, while Figure 7 illustrates the linear combination using Formula (4) when using PCA, focusing on the important components of the expression parameter. Our method using PCA consistently produces more distinct and recognizable blended expressions compared to linear interpolation. This difference is more pronounced when using SEP based on Action Units (AUs), as our method effectively incorporates characteristic AUs of the secondary emotion into the blended expression. Our correlation-driven method also excels in generating subtle emotional blends that remain distinguishable from the neutral state. For instance, the Contempt expression (Anger + Disgust) (Fig 7) exhibits a lowered lip corner (AU15) and a wrinkled nose (AU9). These nuances are often lost in linearly interpolated expressions. This is particularly important for the VKIST dataset, which comprises facial images of Vietnamese individuals, as East Asian individuals tend to express emotions more subtly, often relying on eye movements. Our method effectively captures these subtle changes in the eyes, reflecting the nuances of the combined emotions.

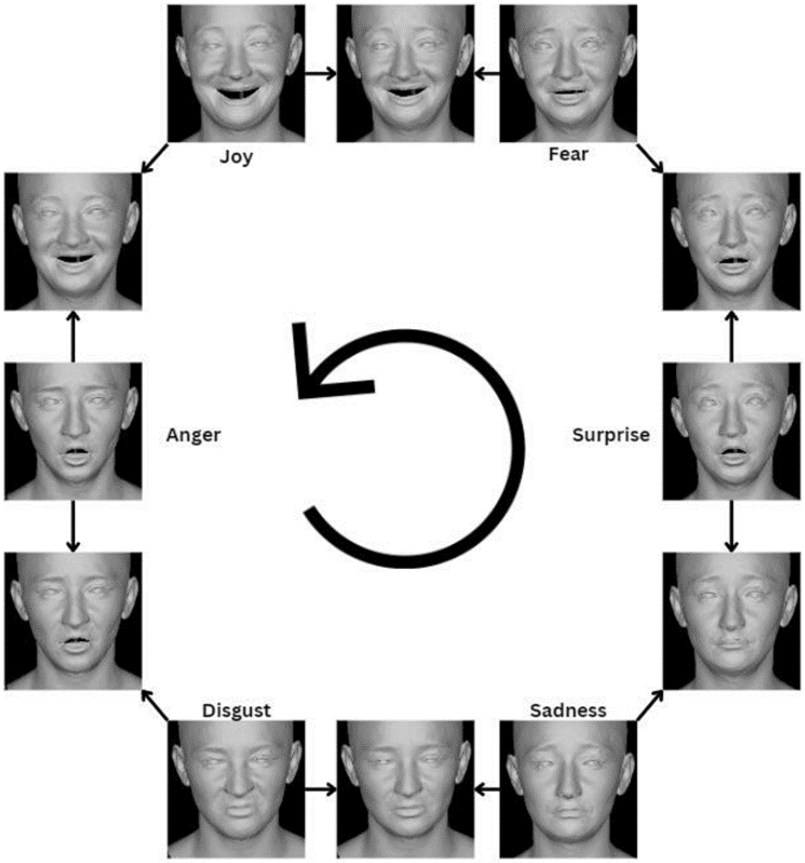

Figure 7. PCA expression blending (counterclockwise)

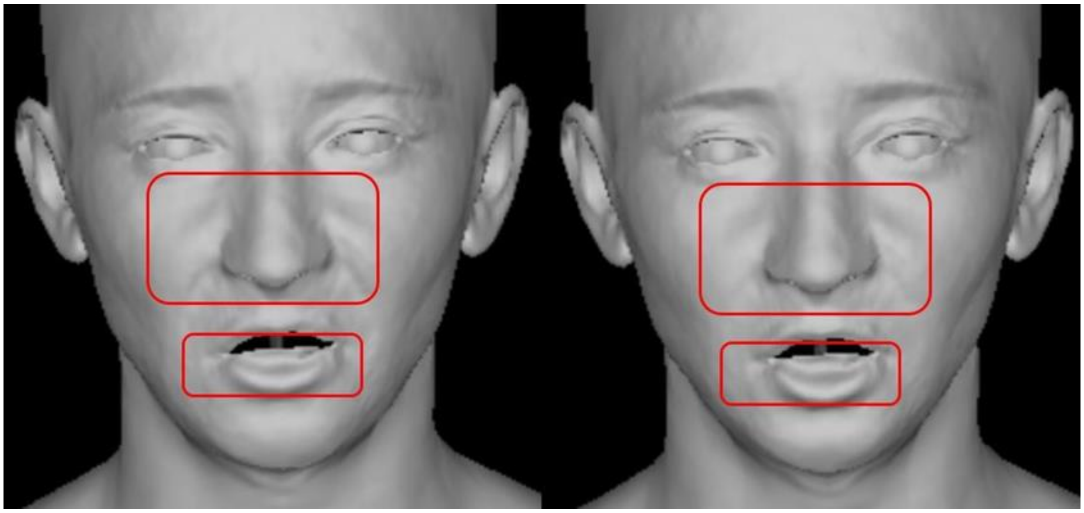

Figure 8. Contempt expression: PCA (left), linear (right).

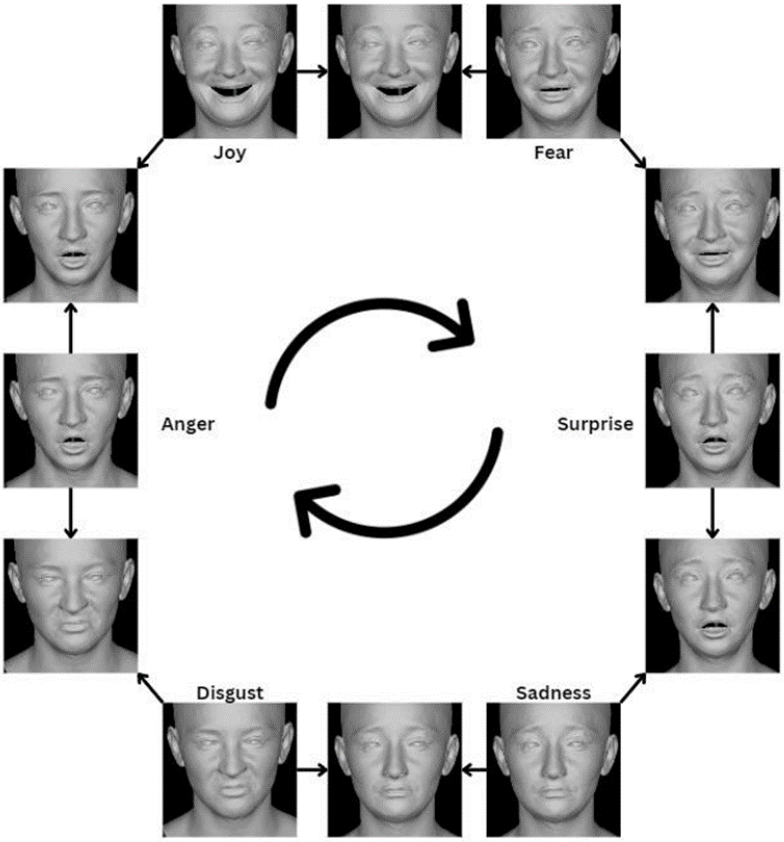

Figure 9. Linear expression blending (clockwise)

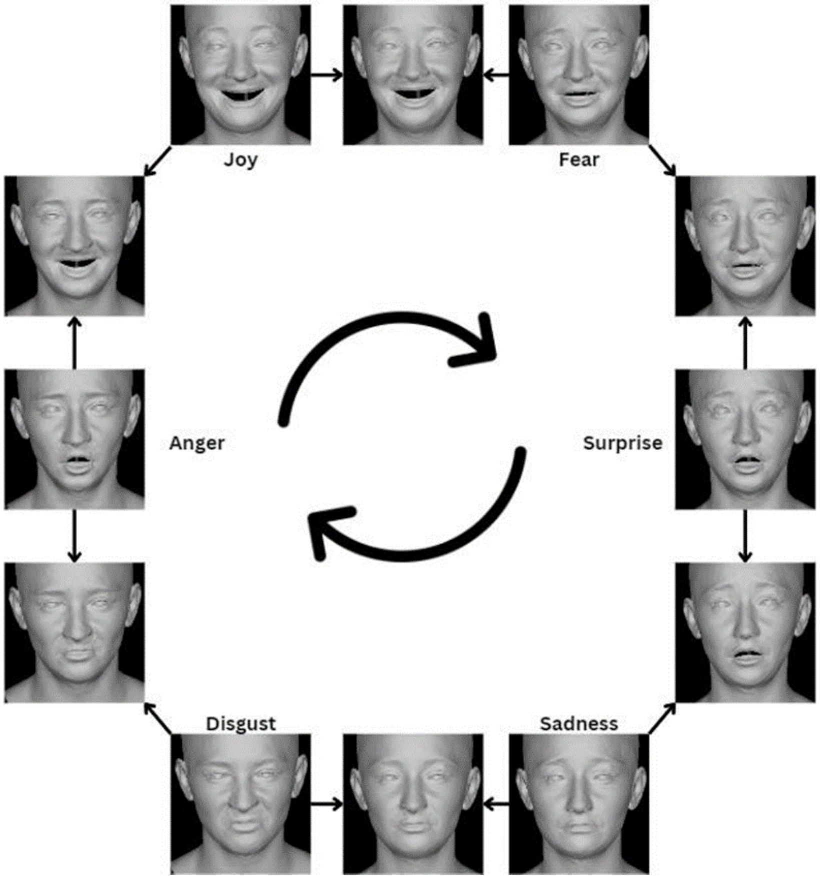

Figure 10. PCA expression blending (clockwise)

Figure 9 and 10 illustrate the results of blending the expressions on the 2nd ring of the Emotion Wheel, while the Expression E1 appears before proceeding clockwise. From the experiment, we can see that the result of the new expression Exp will be different when we change the roles of the first two input expressions E1 and E2. With the combination method presented in formula 4, E1 plays the primary role, so the output expression is more biased towards E1 than E2. Figure 11 illustrates the generation of the expression of remorse when blending disgust and sadness. With E1 being Disgust, the mouth region of the new expression exhibits a more pronounced effect, while with E1 being Sadness, the eyebrows of the new expression are more heavily influenced. This is consistent with the greater degree of influence of E1.

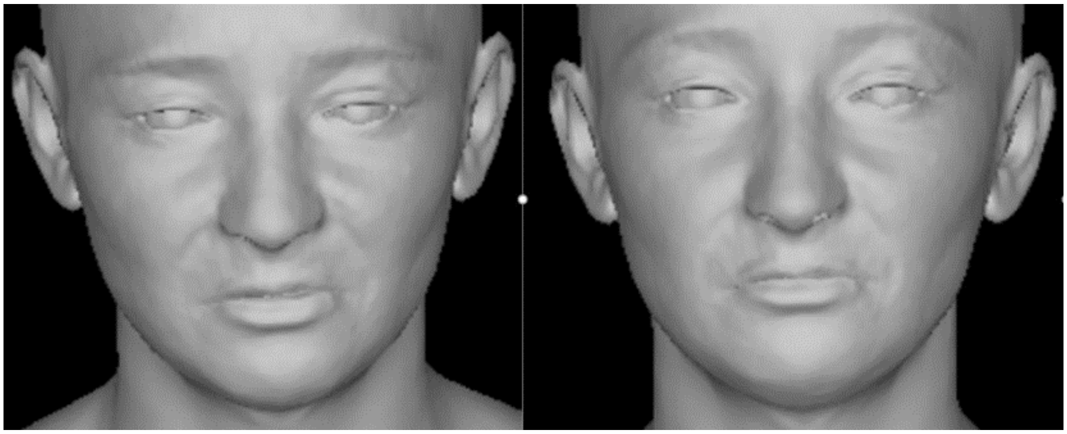

Figure 11. Expression remorse with E1- Disgust (left) and E1-Sadness (right)

Furthermore, to evaluate the quality of naturalness of the blended expressions, we conducted a user study involving 20 participants. The study focused on the special case where the "Neutral" emotion is blended with other emotions. We generated a set of 3D face reconstructions by combining the "Neutral" expression with each of the six remaining basic emotions, and we added 2 new emotions generated using our method based on the emotion wheel which are "Trust" and "Anticipation".

With the six basic emotions, the participants were presented with these reconstructions and asked to label the emotion expressed in each. This evaluation aims to assess the effectiveness of our method in generating subtle emotional expressions that are still recognizable and distinguishable from the neutral state. Additionally, participants rated the naturalness and quality of the 3D face reconstructions on a scale of 1 to 5, where 1 represents the lowest quality and 5 represents the highest quality.

In contrast to the basic emotions, recognizing "Trust" and "Anticipation" poses a greater challenge due to their inherent complexity and nuanced expression. These emotions are not explicitly defined within the FACS, making it difficult for participants to accurately label them. Therefore, for these two emotions, we focused solely on assessing the perceived naturalness and quality of the generated expressions.

The results of the user study are summarized in Table I. The average naturalness rating across all blended expressions was 3.53 out of 5, indicating that the generated expressions were perceived as reasonably natural.

Table I. User Evaluation of 3D Facial Expressions Generated with Correlation-Driven Blending

Emotion Label	Accuracy (%)	Average NaturalnessRating
Surprise	85	3.50
Joy	90	3.60
Anger	90	3.40
Disgust	90	3.90
Fear	70	3.05
Sad	85	3.60
Trust	NA	3.60
Anticipation	NA	3.55

With this approach, we enhance facial reconstruction by blending emotions from our custom DBEX database using the Emotion Wheel, enabling the generation of complex emotional expressions beyond basic reconstructions. As shown in Figure 12 (bottom row), our method synthesizes more nuanced emotions compared to SMIRK and EMOCA, which focus on individual expression reconstruction. Our approach captures a broader range of emotional depth and subtle facial details.

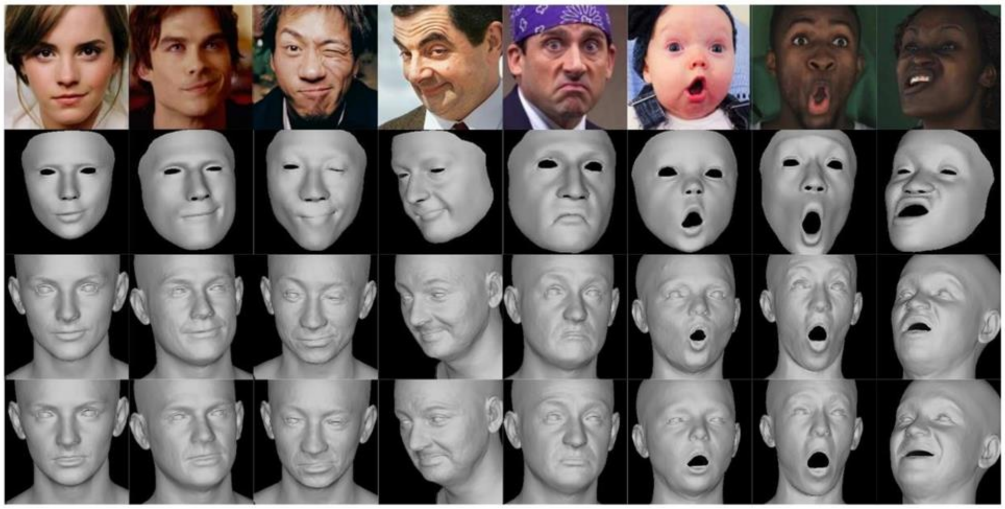

Figure 12. Expression generation from monocular images: monocular images (top), SMIRK [24] (second row), EMOCA (third row), Ours (bottom).

CONCLUSION

This paper presents a novel method for generating new and expressive emotional representations by combining basic emotions using the Emotion Wheel principle. Our approach addresses the limitations of existing facial expression datasets, which predominantly focus on a limited set of basic emotions. By leveraging the structure of the Emotion Wheel, we successfully created a more nuanced and diverse set of compound emotions, which can be applied to both 2D and 3D facial reconstructions.

The proposed method, which includes steps for emotion prediction, expression blending, and target reconstruction, was validated using the VKIST dataset. Our results demonstrate that the combination of basic emotions yields new, composite emotional expressions that are both visually distinct and emotionally meaningful. The use of correlation-driven blending techniques allows for the precise adjustment of expression parameters, ensuring that the resulting expressions capture the subtle nuances of human emotion.

Furthermore, the user study conducted as part of this research confirmed the effectiveness of our method, with participants recognizing and rating the naturalness of the newly generated expressions highly. This outcome suggests significant potential for applications in fields such as digital animation, virtual reality, and human-computer interaction, where the ability to convey complex emotions is crucial.

In conclusion, our research provides a robust framework for advancing the field of facial expression analysis, offering a pathway to more authentic and emotionally rich digital characters.

Future work will explore the expansion of our method to incorporate real-time emotion generation and the integration of additional emotional dimensions.

Acknowledgement

This work has been supported by VNU University of Engineering and Technology under project number QG25.03 “The research of reconstructing facial expressions and movements on 3D face model in acting using, serving the purpose of teaching in traditional arts training institutions”.

REFERENCES

Danecek, Radek and Black, Michael J. and Bolkart, Timo. EMOCA: Emotion Driven Monocular Face Capture and Animation. (2022). Conference on Computer Vision and Pattern Recognition (CVPR).

Wolpert, David. Stacked Generalization. Neural Networks. (1992). 5. Pp. 241-259. 10.1016/S0893- 6080(05)80023-1.

Sharma, S., Kumar, V. 3D Face Reconstruction in Deep Learning Era: A Survey. (2022) Arch Computat Methods Eng 29, pp. 3475–3507. https://doi.org/10.1007/s11831-021-09705-4.

Nelson LA, Michael SD. The application of volume deformation to three-dimensional facial reconstruction: a comparison with previous techniques. (1998). Forensic Sci Int.;94(3): pp. 167-181. doi:10.1016/s0379-0738(98)00066-8.

Nguyen, D. P., Nguyen, T. N., Dakpé, S., Ho Ba Tho, M. C., & Dao, T. T. Fast 3D Face Reconstruction from a Single Image Using Different Deep Learning Approaches for Facial Palsy Patients. (2022). Bioengineering (Basel, Switzerland), 9(11), 619. https://doi.org/10.3390/bioengineering9110619.

Woodward, A., Delmas, P., Chan, Y.H., Gastelum-Strozzi, A., Gimel'farb, G.L., & Flores, J.M. An interactive 3D video system for human facial reconstruction and expression modeling. (2012). J. Vis. Commun. Image Represent., 23, pp. 1113-1127. https://doi.org/10.1016/j.jvcir.2012.07.005.

Shi, T., Yuan, Y., Fan, C., Zou, Z., Shi, Z.X., & Liu, Y. Face-to-Parameter Translation for Game Character Auto-Creation. (2019). 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 161-170.

Prudential Vietnam. 7 Surprising Facts About Human Emotions. (2024). Available: https://www.prudential.com.vn/vi/blog-nhip-song-khoe/7-su-that-bat-ngo-ve-cam-xuc-con-nguoi/.

Dai-ichi Life Vietnam. Balancing Emotions: 4 Ways to Escape Negative Emotions. (2024). Available: https://kh.dai-ichi-life.com.vn/song-vui-khoe/bi-quyet/thoi-quen-song-khoe/suc-khoe-tinh-than/can- bang-cam-xuc-4-cach-de-thoat-khoi-cam-xuc-tieu-cuc.

Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., & van Knippenberg, A. Presentation and validation of the Radboud Faces Database. (2010). Cognition and Emotion, 24(8), pp. 1377–1388. https://doi.org/10.1080/02699930903485076.

Lucey, P., Cohn, J.F., Kanade, T., Saragih, J.M., Ambadar, Z., & Matthews, I. The Extended Cohn- Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. (2010). 2010

IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, pp. 94-101, doi: 10.1109/CVPRW.2010.5543262.

Lyons, M.J. "Excavating AI" Re-excavated: Debunking a Fallacious Account of the JAFFE Dataset. (2021). PsyArXiv. DOI: 10.31234/osf.io/bvf2s.

A. Dhall, R. Goecke, S. Lucey and T. Gedeon. Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark. (2011). 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), Barcelona, Spain, pp. 2106-2112, doi: 10.1109/ICCVW.2011.6130508.

S. M. Mavadati, M. H. Mahoor, K. Bartlett, P. Trinh, and J. F. Cohn. DISFA: A Spontaneous Facial Action Intensity Database. (2013). In IEEE Transactions on Affective Computing, vol. 4, no. 2, pp. 151-160, doi: 10.1109/T-AFFC.2013.4.

Goodfellow, I.J. et al. Challenges in Representation Learning: A Report on Three Machine Learning Contests. (2013). In: Lee, M., Hirose, A., Hou, ZG., Kil, R.M. (eds) Neural Information Processing. ICONIP 2013. Lecture Notes in Computer Science, vol 8228. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-42051-1_16.

A. Mollahosseini, B. Hasani, and M. H. Mahoor. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. (2019). In IEEE Transactions on Affective Computing, vol. 10, no. 1, pp. 18-31, doi: 10.1109/TAFFC.2017.2740923.

Kollias, D., Tzirakis, P., Nicolaou, M.A. et al. Deep Affect Prediction in-the-Wild: Aff-Wild Database and Challenge, Deep Architectures, and Beyond. (2019). Int J Comput Vis 127, pp. 907–929. https://doi.org/10.1007/s11263-019-01158-4.

S. Li, W. Deng and J. Du. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild. (2017). IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 2584-2593, doi: 10.1109/CVPR.2017.277.

Yuanyuan Liu, Wei Dai, Chuanxu Feng, Wenbin Wang, Guanghao Yin, Jiabei Zeng, and Shiguang Shan. MAFW: A Large-scale, Multi-modal, Compound Affective Database for Dynamic Facial Expression Recognition in the Wild. (2022). In Proceedings of the 30th ACM International Conference on Multimedia (MM '22). Association for Computing Machinery, New York, NY, USA, pp. 24–32. https://doi.org/10.1145/3503161.3548190.

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. (1999). In Proceedings of the 26th annual conference on Computer graphics and interactive techniques (SIGGRAPH '99). ACM Press/Addison-Wesley Publishing Co., USA, pp. 187–194. https://doi.org/10.1145/311535.311556.

Pascal Paysan, Marcel Lüthi, Thomas Albrecht, Anita Lerch, Brian Amberg, Francesco Santini, and Thomas Vetter. Face Reconstruction from Skull Shapes and Physical Attributes. (2009). In Proceedings of the 31st DAGM Symposium on Pattern Recognition - Volume 5748. Springer-Verlag, Berlin, Heidelberg, pp. 232–241. https://doi.org/10.1007/978-3-642-03798-6_24.

James Booth, Anastasios Roussos, Stefanos Zafeiriou, Allan Ponniahy, and David Dunaway. A 3D morphable model learnt from 10,000 faces. (2016). In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'16), pp. 5543– 5552. Doi: 10.1109/CVPR.2016.598.

X. Zhu, Z. Lei, X. Liu, H. Shi, and S. Z. Li. Face Alignment Across Large Poses: A 3D Solution. (2016). IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 146-155, doi: 10.1109/CVPR.2016.23.

Retsinas, G., Filntisis, P.P., Daněček, R., Abrevaya, V., Roussos, A., Bolkart, T., & Maragos, P. 3D Facial Expressions through Analysis-by-Neural-Synthesis. (2024). ArXiv, abs/2404.04104.

Jackson, A.S., Bulat, A., Argyriou, V., & Tzimiropoulos, G. Large Pose 3D Face Reconstruction from a Single Image via Direct Volumetric CNN Regression. (2017). IEEE International Conference on Computer Vision (ICCV), pp. 1031-1039.

M. Sela, E. Richardson, and R. Kimmel. Unrestricted Facial Geometry Reconstruction Using Image-to- image Translation. (2017). IEEE International Conference on Computer Vision (ICCV), Venice, Italy, pp. 1585-1594, doi: 10.1109/ICCV.2017.175.

A. T. Tran, T. Hassner, I. Masi and G. Medioni. Regressing Robust and Discriminative 3D Morphable Models with a Very Deep Neural Network. (2017). IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, pp. 1493-1502, doi: 10.1109/CVPR.2017.163.

M. Song, Z. Dong, C. Theobalt, H. Wang, Z. Liu, and H.-P. Seidel. A Generic Framework for Efficient 2- D and 3-D Facial Expression Analogy. (2007). IEEE Transactions on Multimedia, vol. 9, no. 7, pp. 1384– 1395, doi: 10.1109/TMM.2007.906591.

A. Asthana, M. de la Hunty, A. Dhall, and R. Goecke. Facial performance transfer via deformable models and parametric correspondence. (2012). IEEE Transactions on Visualization and Computer Graphics, vol. 18, no. 9, pp. 1511–1519, doi: 10.1109/TVCG.2011.157.

K. Li, F. Xu, J. Wang, Q. Dai, and Y. Liu. A data-driven approach for facial expression synthesis in video. (2012). In Conference on Computer Vision and Pattern Recognition. IEEE, pp. 57–64, doi: 10.1109/CVPR.2012.6247658.

K. Li, Q. Dai, R. Wang, Y. Liu, F. Xu, and J. Wang. A data-driven approach for facial expression retargeting in video. (2014). IEEE Transactions on Multimedia, vol. 16, no. 2, pp. 299–310, doi: 10.1109/TMM.2013.2293064.

L. Xiong, N. Zheng, Q. You, and J. Liu. Facial expression sequence synthesis based on shape and texture fusion model. (2007). In IEEE International Conference on Image Processing, vol. 4. IEEE, pp. IV – 473–IV 476, doi: 10.1109/ICIP.2007.4380057.

Simon Haykin. Neural Networks: A Comprehensive Foundation (1st. ed.). (1994). Prentice Hall PTR, USA.

Q. Zhang, Z. Liu, B. Guo, D. Terzopoulos and H. -Y. Shum. Geometry-driven photorealistic facial expression synthesis. (2006). In IEEE Transactions on Visualization and Computer Graphics, vol. 12, no. 1, pp. 48-60, doi: 10.1109/TVCG.2006.9.

T. Kanade, J. Cohn, and Y.-L. Tian. Comprehensive database for facial expression analysis. (2000). In International Conference on Automatic Face and Gesture Recognition. IEEE, doi. 10.1109/AFGR.2000.840611, pp. 46 – 53.

Susskind, J.M., Hinton, G.E., Movellan, J.R., & Anderson, A.K. Generating Facial Expressions with Deep Belief Nets. (2008).

Kyle Olszewski, Joseph J. Lim, Shunsuke Saito, and Hao Li. High-fidelity facial and speech animation for VR HMDs. (2016). ACM Trans. Graph. 35, 6, Article 221, 14 pp. https://doi.org/10.1145/2980179.2980252.

Saito, S., Li, T., Li, H. Real-Time Facial Segmentation and Performance Capture from RGB Input. (2016). In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds) Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science(), vol 9912. Springer, Cham. https://doi.org/10.1007/978-3-319-46484-8_15.

T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. (2001). In IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 23, no. 6, pp. 681-685, doi: 10.1109/34.927467.

Huang, D., De la Torre, F. Bilinear Kernel Reduced Rank Regression for Facial Expression Synthesis. (2010). In: Daniilidis, K., Maragos, P., Paragios, N. (eds) Computer Vision – ECCV 2010. ECCV 2010. Lecture Notes in Computer Science, vol 6312. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642- 15552-9_27.

W. -f. Liu, J. -l. Lu, Z. -f. Wang and H. -j. Song. An Expression Space Model for Facial Expression Analysis. (2008). Congress on Image and Signal Processing, Sanya, China, pp. 680-684, doi: 10.1109/CISP.2008.216.

Liu, Z., Shan, Y., & Zhang, Z. Expressive expression mapping with ratio images. (2001). Proceedings of the 28th annual conference on Computer graphics and interactive techniques.

S. Agarwal, M. Chatterjee, and D. P. Mukherjee. Synthesis of emotional expressions specific to facial structure. (2012). In Indian Conference on Vision, Graphics and Image Processing. ACM, doi: 10.1145/2425333.2425361.

J. -C. Wang, Y. -H. Yang, H. -M. Wang and S. -K. Jeng. Modeling the Affective Content of Music with a Gaussian Mixture Model. (2015). In IEEE Transactions on Affective Computing, vol. 6, no. 1, pp. 56-68, doi: 10.1109/TAFFC.2015.2397457.

M. D. Zeiler, G. W. Taylor, L. Sigal, I. Matthews, and R. Fergus. Facial expression transfers with input- output Temporal Restricted Boltzmann Machines. (2011). In Advances in Neural Information Processing Systems, pp. 1629–1637, doi: 10.5555/2986459.2986641.

Bhaskar, H., Torre, D.L., Al-Mualla, M. Exaggeration Quantified: An Intensity-Based Analysis of Posed Facial Expressions. (2016). In: Kawulok, M., Celebi, M., Smolka, B. (eds) Advances in Face Detection and Facial Image Analysis. Springer, Cham. https://doi.org/10.1007/978-3-319-25958-1_5.

Y. Choi, M. Choi, M. Kim, J. -W. Ha, S. Kim and J. Choo. StarGAN: Unified Generative Adversarial Networks for Multi-domain Image-to-Image Translation. (2018). IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 2018, pp. 8789-8797, doi: 10.1109/CVPR.2018.00916.

Dimitrios Kollias and Stefanos Zafeiriou. VA-StarGAN: Continuous affect generation. (2020). In Advanced Concepts for Intelligent Vision Systems: 20th International Conference, ACIVS 2020, Auckland, New Zealand, February 10–14, 2020, Proceedings 20, pp. 227–238. Springer, doi: 10.1007/978-3-030- 40605-9_20.

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. 2022. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. (2022). IEEE Trans. Pattern Anal. Mach. Intell. 44, 10_Part_1, pp. 5962–5979. https://doi.org/10.1109/TPAMI.2021.3087709.

Pumarola A, Agudo A, Martinez AM, Sanfeliu A, Moreno-Noguer F. GANimation: Anatomically-aware Facial Animation from a Single Image. (2018). Comput Vis ECCV; 11214:835-851. doi:10.1007/978-3-030- 01249-6_50.

S. d'Apolito, D. P. Paudel, Z. Huang, A. Romero and L. V. Gool. GANmut: Learning Interpretable Conditional Space for Gamut of Emotions. (2021). IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp. 568-577, doi: 10.1109/CVPR46437.2021.00063.

Plutchik, Robert. Emotion: Theory, research, and experience: Vol. 1. Theories of emotion, vol. 1. (1980). New York: Academic, https://doi.org/10.1016/C2013-0-11313-X.

Ekman P, Friesen W . Facial Action Coding System: A Technique for the Measurement of Facial Movement. (1978). Palo Alto: Consulting Psychologists Press, https://doi.org/10.1037/t27734-000.

Anh Duc Dam, Thi Chau Ma. Enhancing Emotion Recognition with Stacked Machine Learning Models: Insights from a Novel Vietnamese Facial Expression Dataset. (2024). Conference: ICIIT 2024: 2024 9th International Conference on Intelligent Information Technology. DOI:10.1145/3654522.3654523.