Department of Electrical, Computer, and Software Engineering

Part IV Research Project

Final Report

Project Number: 64

Talking Face Generation System using Deep Neural Networks

Yugyeong Hong

Gayeon Kim

Ho Seok An, Trevor Gee

13/10/2023

Declaration of Originality

This report is my own unaided work and was not copied from nor written in collaboration with any other person.

Name: Yugyeong Hong

Table of contents

ABSTRACT

Our project aims to create artificial humanoid avatars, specifically talking faces with dynamic emotions, using a single image and text input to enhance human-robot interaction. We emphasise the importance of precise lip motion, head movement, dynamic facial expressions, and video quality in making avatars more engaging to human users. Despite recent developments in talking face generation systems, their practical applications remain undefined. Moreover, the majority of research in talking face generation focuses on lip motion independently, often overlooking the integration of emotional facial expressions. These studies tend to heavily depend on audio or video inputs, which has resulted in a shortage of systems capable of generating facial videos with adjustable emotions from just a single image.

In this paper, we introduce EmoFaceGen, an innovative emotional talking face generation system. EmoFaceGen produces realistic talking face videos with emotions from a single facial image and text input, using Text-To-Speech to generate the audio. EmoFaceGen outperforms other open-source models, addressing current challenges in the field, particularly considering the memory and hardware limitations associated with conventional 3D graphics methods. Furthermore, this research explores the most effective techniques for generating realistic emotional talking faces, presenting two distinct approaches. Among these, the step-by-step approach has yielded the most promising results.

Acknowledgements

I would like to thank my project supervisors, Ho Seok Ahn and Trevor Gee for their guidance and support during this research project.

1. Introduction

In recent years, talking face generation has become a topic of significant interest, and numerous experiments have been conducted to investigate various strategies for generating talking faces. While the talking face generation system can be utilized for multiple purposes, including animation, virtual assistant development, enhancement of chatbot interaction, and in the healthcare field, the majority of current approaches [1],[2],[3],[4] primarily focus on generating realistic lip synchronization, identity preservation, incorporating realistic eye blinks, and simulating head movements in the generated talking face videos. Nevertheless, there are still many gaps in the talking face generation field, such as systems that can produce videos with controllable emotion, a single facial image, and an option for text input.

Emotions significantly influence our communication methods. Facial expressions, which might indicate emotions such as happiness, sadness, or anger, play a crucial role in our daily interactions. These emotional indicators are essential not only for genuine communication but also for utilizing talking face systems as therapeutic tools, especially for individuals diagnosed with autism spectrum disorder (ASD) [5].

Given the crucial role that emotions play in human communication, we aim to bridge this gap by creating talking faces that convey a variety of emotions, including happiness, sadness, and anger. Furthermore, many of the current models still fall short in terms of achieving high-quality realism, which can be noticeable to the human eye, particularly in models that are not specifically trained for a single individual.

Furthermore, by allowing the use of a single facial image as input, the range of potential faces available for video generation is expanded. This is significant because many users might have access only to facial images and not videos of the person they wish to recreate. This approach addresses a common limitation in several talking face generation systems, which often require a target video, as cited in references [6],[7],[8],[9]. Moreover, enabling text as an input mechanism improves both the accessibility and usability of the system. In many scenarios, using an audio file might be challenging or not an option. Given the current challenges, we introduce an innovative talking face generation system capable of taking an emotional condition, facial image, and target dialogue—either in text or audio form—to generate a talking face video. The primary emphasis of this report is the system's ability to generate emotional expressions, while my project partner has prioritized enhancing lip synchronization and head movement.

2. Literature Review

Our research integrates several areas of image augmentation, including facial expression changes, lip movements, head position dynamics, and visual quality improvement. In this section, it explores key literature that supports our approach in emotion generation.

2.1. Emotional Talking Face Video Generation

Currently, there is a limited number of research papers dedicated to emotional face generation. Due to the limited presence of annotated emotional audio-visual datasets, few of these methods can achieve realistic facial emotions.

MEAD [10] has introduced an approach for generating emotional talking faces while offering explicit control over emotions. They have also made available the MEAD dataset [10], which encompasses a rich array of sentences and well-defined emotions spanning different intensities. However, this method [10] confines emotional expression primarily to the upper face, utilizing external emotion control via a one-hot emotion vector. Consequently, the lower part of the face is animated independently from audio, leading to inconsistencies in emotional expression across the entire face.

The video editing technique, EVP [11], emphasizes creating uniform emotions across the face by leveraging a distinct emotion latent feature extracted from the audio. Yet, these methods consistently depend on global landmarks or edge maps to directly produce textures infused with emotions, which limits their adaptability to unfamiliar target faces.

The Emotion-Aware Motion Model (EAMM) [12] introduced a technique for producing emotional talking faces using an emotion source video. It has two main parts: the Audio2Facial-Dynamics module (A2FD) which creates neutral-expression talking faces from one neutral frame, and the Implicit Emotion Displacement Learner, which pinpoints emotion-related facial dynamics. The A2FD employs three encoders to process source image, audio, and pose inputs, which are then combined and processed by an LSTM-based decoder. By leveraging a key-point detector and a flow estimator, it adds emotion patterns to motion representations, creating emotional talking face videos. However, EAMM [12] researchers highlighted issues such as facial distortions due to blending in unwanted features like identity and speech content. Our own tests found the method's emotional accuracy and visual quality lacking.

EmoTalkingFace [13] presented a unified model designed to produce emotional facial videos using just an input image, audio, and an emotional condition. This was accomplished by incorporating an emotional discriminator (essentially serving as a video emotion classifier) and integrating LSTM layers into the generator model, thereby learning distinct emotional traits. However, upon our evaluation, while the emotional condition was realized, the visual quality of the resulting video was less than desirable, with noticeable facial distortions appearing after 5-7 seconds depending on audio.

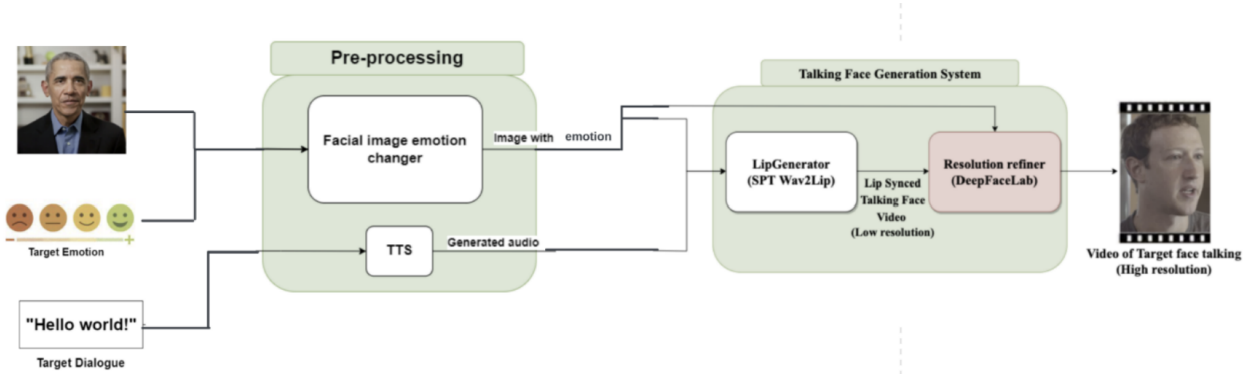

Fig. 1. Final pipeline of Talking face generation system in last year.

In last year's project, as shown in Fig1, Alan and Takahiro's approach to emotional talking face generation divides the creation of emotional expressions and lip synchronization into two separate phases, each powered by its own model. Using a facial image, emotion label, and text, this method produces a synchronized talking face video that conveys the chosen emotion. In the pre-processing phase, a facial image and desired emotion are processed through a "facial image emotion changer" which uses StyleGan [14] with NVIDIA's pre-trained weights. This adjusts the emotion in the image by encoding it to latent space and then altering this representation using a predefined emotional direction. The adjusted image is then integrated with lip synchronization to form a complete talking face video.

In the subsequent phase, the emotionally changed facial image and text-to-speech (TTS) audio are input into the "LipGenerator", which is based on the Wav2Lip [15] model and its encoder-decoder structure. For this, Alan and Takahiro utilized the VoxCeleb2 [16] dataset, prioritizing it over the LRS2 [18] dataset due to its larger video content. They enhanced the Wav2Lip [15] model to form the SPT-Wav2Lip model by including a spatial transformer network [19], streamlining head-pose adjustments. Furthermore, to boost visual quality, they employed deep-fake technology, replacing lower-quality frames with higher-quality facial images.

2.2. Emotional Face Image Generation

2.2.1. Style-GAN

Style-GAN [14] originates from the "style-transfer" domain, which emphasizes realistic image-editing techniques. Diverging from traditional GAN models, Style-GAN [14] modifies the generator within the progressive GAN framework and alters the latent space input. These innovations led to the creation of a cutting-edge style editing model. It

demonstrated the capability to produce realistic results when combining different styles, facilitating expression transfers in images by leveraging latent spaces and pre-trained directional inputs.

2.2.2. Cycle-GAN

CycleGAN [17] introduced builds upon the well-known Pix-to-Pix model [20]. Its significant advancement is its ability to move beyond Pix-to-Pix's [20] dependency on "paired" datasets. In such datasets, paired images typically mirror attributes like positioning and colour style. In contrast, CycleGAN [17] can be trained using unpaired datasets, allowing for unique style transformations — for instance, converting a horse image to a zebra. This concept can be extended to expression transfer, where a CycleGAN [17] can be trained to transform images of neutral faces into any specified emotion.

2.2.3 StarGAN

StarGAN [21] builds on Conditional GANs to enable conditional image generation. Unlike traditional GANs, StarGAN [21] allows flexible image translation across multiple target domains by using conditional domain information. It uses a single generator to learn mappings between these domains. The setup includes a discriminator that identifies real vs. fake images and classifies real ones into their respective domains. The generator, on the other hand, accepts both an input image and a target domain label to produce a fake image. The aim during training is for the generator to recreate the original image from the fake, ensuring that the output is realistic and aligned with the target domain.

2.3. Visual Quality Improvement

2.3.1 Codebook lookup transformer

The Codebook lookup transformer, also referred to as the Codeformer [22], introduces a three-step technique to improve face restoration from subpar input data. The process begins with the training of an autoencoder using vector quantization, establishing a discrete codebook that aligns low-resolution inputs with high-resolution outputs. Subsequently, this is followed by the integration of a Transformer module and a Controllable Feature Transformation module, which together predict any corrupted or absent segments, refining the outcomes across different degradation extents.

3. Identified gaps and project goals.

Summarising the literature review on current systems focused on emotional facial video generation, we can see the need for a system capable of generating high-quality emotional facial videos is still required. Moreover, based on inputs alone, there is a stark lack of image-only input-driven models. Given the presented literature, it may seem that [13] has already solved what our system proposes. However, the generated videos' visual quality was very unrealistic based on initial

testing. Even following all the steps provided by [13]'s materials, no generated results could reproduce the same visual quality presented in their original paper. In last year approach, despite the introduction of the SPT-Wav2Lip model and the incorporation of Deep-fake technology to enhance the quality of the talking face videos, several issues remain unresolved. The system encounters challenges in producing accurate lip-synchronized frames during speaker pauses. Although DeepFaceLab [23] can address the visual quality issue, it is only effective when multiple images of the target speaker are provided and not when only a single image is available. Since the project requires a single image as input, an alternative solution should be explored to replace the DeepFaceLab [23] approach. In addition, the current state of image-to-video talking face technology is still far from realistic. One of the main reasons for this is that the generated videos lack head movement, resulting in a static image with only lip movement. Furthermore, replicating emotions using StyleGan's [14] latent spaces is challenging as it struggles to differentiate between negative emotions and neutral ones. In addition, observation indicates that our model based on Wav2Lip [15] tends to generate frames where the inside of the mouth is filled with black in most of the frames, leading to the loss of details such as teeth. As such, the main capabilities of the proposed systems are:

- Adequate visual quality of video frames
- The system must allow for controllable emotions
- The model is trainable using open-source datasets

With these capabilities proposed, the main research question of this paper is "How to generate realistic facial videos with emotion given a single image input with enhanced video quality?"

4. Research method

4.1. System Approach

Given a target face, emotion & dialogue (text or audio), our emotional talking face generation system aims to produce a video of the target face speaking the dialogue with the specified emotion. Two distinct pipelines have been developed for this process. The End-to-End pipeline integrates emotion generation and lip-syncing within one unified model. Conversely, the Step-by-Step pipeline spits these tasks into individual stages and models.

5. Pipeline 1: End-to-End Approach

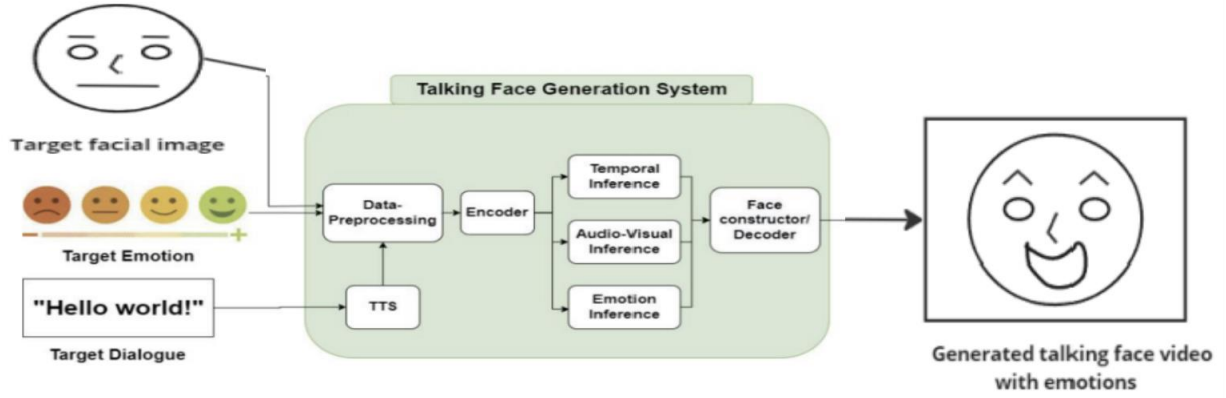

Fig. 2. End-to-End pipeline.

In the initial approach, the goal was to adjust emotion and synchronize lip and head movements using a single model. As illustrated in Figure 2, the process starts by taking a facial image, a chosen emotion, and a dialogue. The dialogue is then transformed into an audio file through a text-to-speech generator. This information is processed by a unified generator model, resulting in a video where the selected face communicates with the given emotion. The foundational work of EmoTalkingFace [13] served as a starting point due to its capabilities for emotional facial generation. However, evaluations showed that the videos lacked visual realism. The prolonged training durations, sometimes spanning over a month, were impractical, especially when aiming to achieve integrated lip synchronization, head movement, and improved video quality. These findings prompted a shift towards a step-by-step approach.

6. Pipeline 2: Step-by-step approach

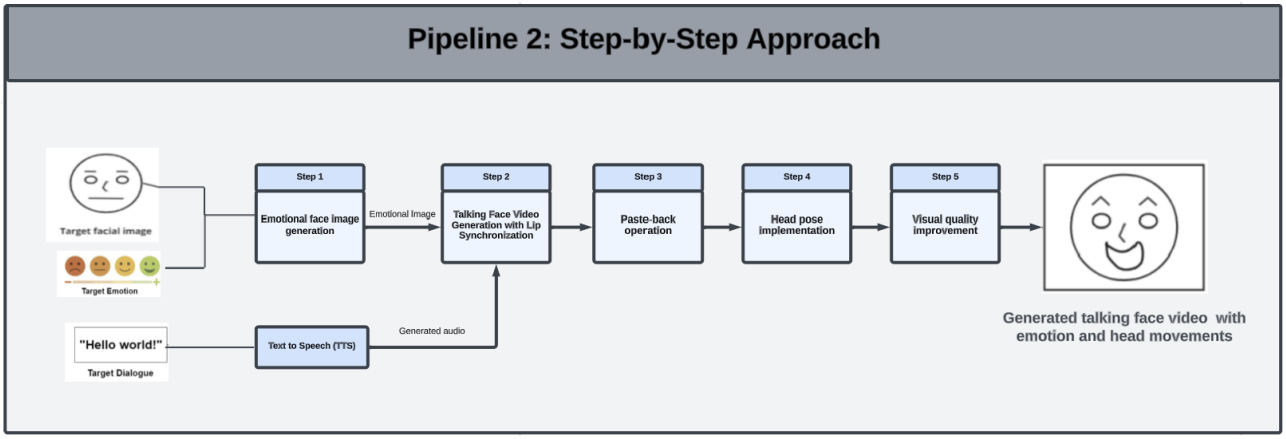

Fig. 3. Step-by-step pipeline.

Given the extended training duration and the unrealistic visual and emotional quality, a shift was made from the initial approach. Unlike the initial pipeline where integrated processes such as lip synchronization, emotion generation, and head movements were merged into one model, our revised pipeline breaks these tasks into five separate stages, each with its own dedicated model. The procedure is depicted in Figure 3, with a particular focus on steps 1 and 5 in this report.

6.1. Step 1: Emotional Face Image Generation

As mentioned previously for the motivation of facial image expression modification, we considered integrating three systems into our framework: a pre-trained StyleGAN model [14], a CycleGAN model [17], and a StarGAN model [21] that we trained independently. However, given the constraints of these models, we decided to build upon EmoTalkingFace [13].

6.1.1. Style-GAN

Utilising NVIDIA's original pre-trained models and weights, combined with pre-established expression directions, modifying image expressions was attempted. However, using latent spaces and image encoding, distinguishing between negative emotions and neutral ones proved to be a challenge.

6.1.2. Cycle-GAN

For the image expression model, CycleGAN's potential to train on unpaired datasets was attempted. This method made use of open-source facial datasets such as FFHQ [14], instructing the model with two different emotional face sets. Initially, the goal was to translate neutral faces into happy ones. However, CycleGan’s design limits it to understanding relationships between just two domains simultaneously, which limits its output to transitioning from a neutral face to a happy one.

6.1.3. StarGAN

To address the limitations of the CycleGAN [17] method, the StarGAN [21] framework was employed. StarGAN offers flexible image translation across multiple target domains using conditional domain information, enabling a single generator (G) to learn mappings between various domains. For training the model, we initially planned to utilise the RaFD [24] dataset, which provides eight labels for facial expressions, such as 'happy,' 'angry,' and 'sad.' However, while awaiting approval for the RaFD [24] dataset, we used the RAF [25] dataset, which contains seven labels for facial

expressions, and the AffectNet [26] dataset, which offers eight labels for facial expressions. However, StarGAN's outputs produced emotional images that were challenging to identify.

6.1.4. Emotion image generator with step 5

To generate emotional images, EmoTalkingFace [13], which already facilitates emotional face generation, was trialled. Drawing from the concept of EmoTalkingFace's emotional discriminator [13], which they defined as a video emotional classifier, and inspired by the success of invertible frowns' [27] ability to convey target emotions through their emotional loss function driven by an "off-the-shelf" emotional classifier, adopting a similar strategy seemed feasible. The challenge with this approach was the suboptimal visual quality of the resulting video, marked by evident facial distortions emerging after 5-7 seconds depending on the audio. To address this, the Codeformer [22] was integrated to enhance visual quality and produce emotional images rather than videos. Consequently, we successfully generated enhanced-quality emotional images, as depicted in Figure 4.

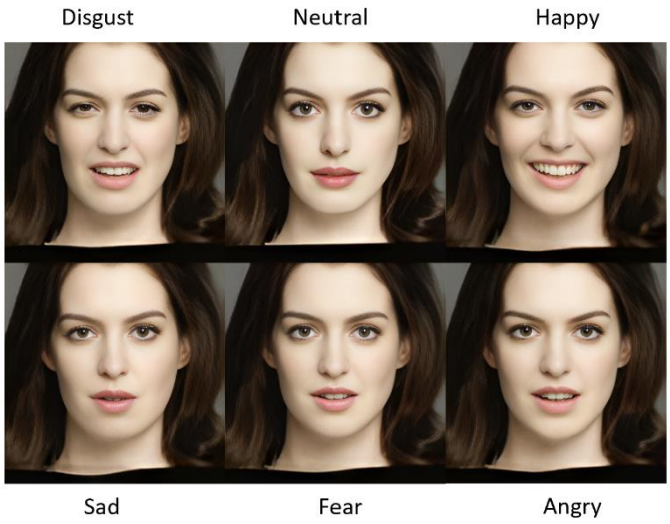

Fig. *4.* Generated emotional images.

7. Evaluation

Given the emphasis of this report on the emotional dynamics of the system, evaluation is largely qualitative. The system is tailored to human use and emotion recognition inherently demands a human-centric approach. While emotional classifiers might exhibit high accuracy, their biases toward their training datasets can introduce inaccuracies or detect nuances typically unnoticed by individuals. Therefore, this report will exclusively showcase qualitative emotional

outcomes. Quantitative evaluations, especially concerning lip synchronization and head movement, will be addressed by the project partner.

8. Experiments

8.1. Software and hardware setup

All development was tested and developed on university-provided hardware, such as the Robotic lab computers and the university lab computer. CARES provided a 15 GeForce RTX 3090 GPU, 256GB RAM and AMD Thread Ripper 3970X CPU, whilst the lab computer provided an RTX 3080, 32 GB RAM and Intel i9-10900 CPU. Both computers provided a native Linux environment. Git and GitHub were used for version control of created code and code bases. Conda [28] was used to create and manage project-specific dependencies and environments.

8.2. Emotional face image

Regarding the Emotional Face Image generator, an anonymous survey was conducted in the lab with 10 participants. To determine the recognizability of emotions for the model, users were asked to classify the emotions of a series of generated images from the model. The findings from this survey are showcased in six pie charts, referenced as Figure 6, each representing emotions: Angry, Sad, Disgusted, Neutral, Happy, and Fear. Key insights from the results are:

- In the ‘Disgusted’ category, 70% of participants correctly identified the emotion, whereas 30% of participants believed it showcased ‘Angry’
- Both ‘Happy’ and ‘Neutral’ emotional images received unanimous correct responses
- Images labelled as ‘Sad’, and ‘Fear’ predominantly received ‘Neutral’ as the response, with only a single correct identification for each
- The ‘Angry’ category revealed diverse responses, with answers spanning across all emotions, except for ‘Happy’

Consequently, while this model effectively generates emotions such as Neutral, Happy, and Disgusted, it struggles to accurately represent Sad, Angry, and Fear emotions.

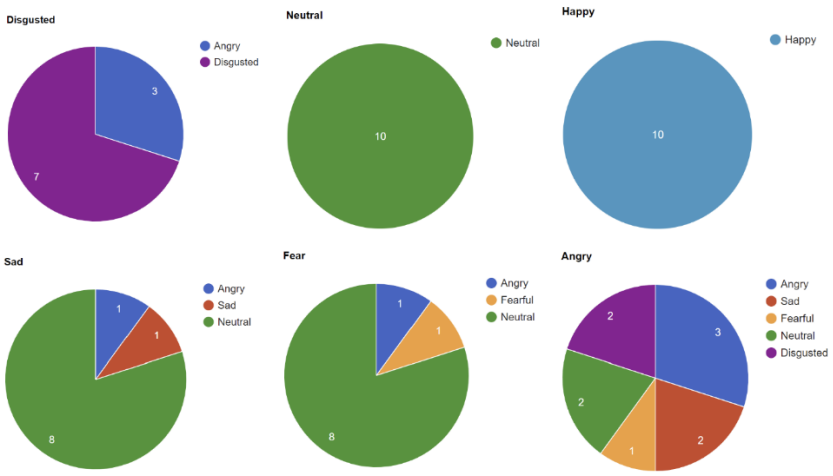

Fig. 5. Survey results for emotional image generator.

9. Discussion

As seen by the results in figure 5, images expressing happiness and disgust are easily distinguishable; however, differentiating between anger, sadness, and fear proves challenging. However, there remains substantial debate and uncertainty in the academic community regarding which emotion model—categorical or dimensional—most accurately represents the true nature of emotions. Our model is trained based on the categorical emotion model, which assumes there are a small number of emotions hard-wired into the human brain. Using the CREMA-D dataset [29], a resource featuring 91 actors (48 male, 43 female) expressing all of Ekman's categorical emotions, six basic emotion categories are suggested: anger, disgust, fear, happiness, neutral, and sadness. However, the dimensional model argues that emotions are correlated, and each emotion category can be represented with a combination of values from emotional dimensions. These dimensions determine if the emotion is active or passive, with a valence dimension that ascertains if the emotion is positive or negative. Thus, our approach can be formulated using the dimensional emotion mode, focusing on the intensity levels of the emotion.

10. Conclusion

Our project initially aimed to create a realistic talking face generation system using a single image and text-converted audio. The initial single-model approach faced challenges related to visual quality and training duration, leading us to

adopt a more successful two-step approach. This revised system effectively captures emotions such as Neutral, Happy, and Disgusted, enhancing user interaction.

However, our study has difficulties in generating emotional images, especially in distinguishing between certain emotions. The ongoing debate between categorical and dimensional emotion models further underscores the complexities of emotional representation. Our future work will involve adapting our emotional image generator based on the dimensional model, focusing on emotion intensity to enhance emotional expression in avatars.

Moreover, our current system has limitations when dealing with non-frontal faces, affecting emotion generation. To overcome this, we propose integrating alignment tools as a pre-processing step to improve accuracy while minimizing unintended facial alterations.

Despite room for improvement, we believe our system has practical applications, such as an emotion teaching tool for individuals who are diagnosed with autism spectrum disorder, creating realistic conversational experiences with deceased individuals, and enhancing human-robot interactions with social robots.

References

[1] Z. Sibo, Y.Jiahong, L.Miao, Z.Liangjun, “Text2Video: Text-Driven Talking-Head Video Synthesis with Personalized Phoneme-Pose Dictionary,” Jan.2022, [Online]. Available: https://arxiv.org/abs/2104.14631

[2] Y. Zhou, X. Han, E. Shechtman, J. Echevarria, E. Kalogerakis, and D. Li, “MakeItTalk: Speaker-Aware Talking Head Animation,” Feb. 2021, [Online]. Available: https://arxiv.org/abs/2004.12992

[3] Z.Hang, S.Yasheng, W.Wayne, L.Chen, W.Xiaogang, L.Ziwei Liu, “Pose-Controllable Talking Face Generation by Implicitly ModularizedAudio-VisualRepresentation,”Apr.2021,[Online].Available:https://arxiv.org/abs/2104.11116

[4] W.Suzhen, L.Lincheng, D.Yu, Y.Xin, “One-shot Talking Face Generation from Single-speaker Audio-Visual Correlation Learning,” Dec.2021, [Online]. Available: https://arxiv.org/abs/2112.02749

[5] S. A. Cassidy et al., “Expressive visual text-to-speech as an assistive technology for individuals with autism spectrum conditions,” Comput Vis Image Underst, vol. 148, pp. 193–200, Jul. 2016, doi: 10.1016/j.cviu.2015.08.011

[6] R. Kumar, J. Sotelo, K. Kumar, A. de Brebisson, and Y. Bengio, “ObamaNet: Photo-realistic lip-sync from text,” Dec. 2017, [Online]. Available: http://arxiv.org/abs/1801.01442

[7] L. Li et al., “Write-a-speaker: Text-based Emotional and Rhythmic Talking-head Generation,” Apr. 2021, doi: 10.48550/arxiv.2104.07995.

[8] S. Suwajanakorn, S. M. Seitz, and I. Kemelmacher-Shlizerman, “Synthesising obama: Learning lip sync from audio,” in ACM Transactions on Graphics, 2017, vol. 36, no. 4. doi: 10.1145/3072959.3073640

[9] P. K. R, R. Mukhopadhyay, J. Philip, A. Jha, V. Namboodiri, and C. v. Jawahar, “Towards Automatic Face-to-Face Translation,” Mar. 2020, doi: 10.1145/3343031.3351066

[10] K. Wang, Q. Wu, L. Song, Z. Yang, W. Wu, C. Qian, R. He, Y. Qiao, and C. C. Loy, “MEAD: A large-scale audio-visual dataset for emotional talking-face generation,” in ECCV (21), vol. 12366, 2020, pp. 700–717.

[11] X. Ji, H. Zhou, K. Wang, W. Wu, C. C. Loy, X. Cao, and F. Xu, "Audio-driven emotional video portraits," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14080–14089, 2021.

[12] Xinya Ji, Hang Zhou, Kaisiyuan Wang, Qianyi Wu, Wayne Wu, Feng Xu, Xun Cao,” EAMM: One-Shot Emotional Talking Face via Audio-Based Emotion-Aware Motion Model,” Sep. 2022. [Online]. Available: https://arxiv.org/abs/2205.15278

[13] S. E. Eskimez, Y. Zhang, and Z. Duan, “Speech Driven Talking Face Generation from a Single Image and an Emotion Condition,” Aug. 2020, [Online]. Available: http://arxiv.org/abs/2008.03592

[14] T. Karras, S. Laine, and T. Aila, “A Style-Based Generator Architecture for Generative Adversarial Networks,” Dec. 2018, doi: 10.48550/arxiv.1812.04948.

[15] K. R. Prajwal et al, “A Lip Sync Expert Is All You Need for Speech to Lip Generation in The Wild,” 2020. [Online]. Available: arXiv:2008.10010.

[16] J. S. Chung, A. Nagrani, and A. Zisserman, “VoxCeleb2: Deep Speaker Recognition,” Jun. 2018, doi: 10.21437/interspeech.2018-1929.

[17] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks,” Mar. 2017, doi: 10.48550/arxiv.1703.10593

[18] T. Afouras, J. S. Chung, A Senior and A. Zisserman, “Deep Audio-Visual Speech Recognition,” 2018. [Online]. Available: arXiv:1809.02108.

[19] M. Jaderberg, K. Simonyan, A. Zisserman and K. Kavukcuoglu, “Spatial Transformer Networks,” 2015. [Online]. Available: arXiv:1506.02025

[20] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-Image Translation with Conditional Adversarial Networks," 2016.

[21] 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," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 8789-8797.

[22] S. Zhou, K. C.K. Chan, C. Li, and C.C. Loy, "Towards Robust Blind Face Restoration with Codebook Lookup Transformer," 2022.

[23] I. Perov et al, “DeepFaceLab: Integrated, flexible and extensible face-swapping framework,” 2020. [Online]. Available: arXiv:2005.05535

[24] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. Van Knippenberg. Presentation and validation of the radboud faces database. Cognition and Emotion, 24(8):1377–1388, 2010. 2

[25] L.Shan, D.Weihong and D.JunPing “Reliable Crowdsourcing and Deep Locality-Preserving Learning for Expression Recognition in the Wild,” 2018, doi:http://www.whdeng.cn/raf/model1.html#dataset

[26] M.Ali, H.Behzad, Mohammad H. Mahoor. “AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild,” Oct.2017, doi:https://arxiv.org/abs/1708.03985

[27] I. Magnusson, A. Sankaranarayanan, and A. Lippman, "Invertible Frowns: Video-to-Video Facial Emotion Translation," 2021.

[28] Anaconda Inc., “Anaconda Software Distribution,” Anaconda Documentation, 2020.
[29] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, “CREMA-D: Crowd-sourced Emotional Multimodal Actors Dataset,” IEEE Trans Affect Comput, vol. 5, no.4, pp. 377–390, 2014, doi: 10.1109/TAFFC.2014.2336244.