

# Altering Facial Expression based on Textual Emotion

Mohammad Imrul Jubair<sup>a</sup>, Md. Masud Rana<sup>1</sup>, Md. Amir Hamza, Mohsena Ashraf,  
Fahim Ahsan Khan and Ahnaf Tahseen Prince

*Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Bangladesh*

**Keywords:** Facial Expression, Image to Image Translation, Emotion Detection.

**Abstract:** Faces and their expressions are one of the potent subjects for digital images. Detecting emotions from images is an ancient task in the field of computer vision; however, performing its reverse—synthesizing facial expressions from images—is quite new. Such operations of regenerating images with different facial expressions, or altering an existing expression in an image require the Generative Adversarial Network (GAN). In this paper, we aim to change the facial expression in an image using GAN, where the input image with an initial expression (i.e., happy) is altered to a different expression (i.e., disgusted) for the same person. We used StarGAN techniques on a modified version of the MUG dataset to accomplish this objective. Moreover, we extended our work further by remodeling facial expressions in an image indicated by the emotion from a given text. As a result, we applied a Long Short-Term Memory (LSTM) method to extract emotion from the text and forwarded it to our expression-altering module. As a demonstration of our working pipeline, we also create an application prototype of a blog that regenerates the profile picture with different expressions based on the user's textual emotion.

## 1 INTRODUCTION


As a result of the widespread usage of social media and blogs, people have been accustomed to expressing their feelings and thoughts digitally, whether by text, voice, or image. When it comes to these emotions, the facial expression plays an important role in our everyday modes of communication and connection, particularly when it comes to pictures, videos, online conferences, etc. While speaking of facial expressions, we commonly refer to happiness, sadness, anger, disgust, etc, which are very natural for humans (Barrett et al., 2019)(Frank, 2001)(Xu et al., 2017).

Photos and their expressions have an undoubtedly strong impact in the field of computer vision since researches have been going on for years to extract expressions. The recent progress of machine learning has brought enough accuracy in detecting and recognizing facial expressions. It is not simple to tackle the opposite difficulty of this task—imposing varied emotions on a pre-existing face in a photograph without the help of another human being—and this subject is still mostly studied. The recent breakthrough of the

Generative Adversarial Network (Goodfellow et al., 2014) has influenced the researchers to work with image-to-image translations and to develop different stunning tools, for example converting a horse into a zebra (Zhu et al., 2017). This type of application of GAN typically converts the image from a source domain  $X$  to a target domain  $Y$  by learning from an adequate amount of image data. Hence it also make it possible to create new images by learning from a large dataset, they are extremely appealing tools for creating images with a variety of expressions. When applied to an input image, it may be used to transform the facial expression into the intended expression; for instance, from a joyful face to an angry face. A dataset of faces of diverse people with a range of expressions is required for this type of modification.

### 1.1 Contribution

We found this research domain of regenerating different facial expressions very exciting and, in this paper, we experimented on different GAN-based methods on a variety of datasets as an attempt to determine the fittest combination. To make our research even more interesting, we broadened the scope of our domain to include an application that went beyond just modify-

<sup>a</sup>  <https://orcid.org/0000-0003-0112-7524>

<sup>1</sup>These authors contributed equally to this work.

ing the facial expression; we concentrated on textual emotion transmission to images. In our method, there are two phases: first, we extract emotion from a text, and then we transmit the outcome as an input to the facial expression generating module, which then modifies the emotion in a person’s photograph.

In real-world situations—such as blogs and social media—our method can be utilized. As an example, consider a blog where the facial expression of a user’s photo may be modified based on the content of his or her most recent post. For instances, in the case of a sorrowful statement like “I’m not feeling well today”, the expression on his already-existing profile photo—which had a happy face—would instantly change to one of sadness. Our approach requires two inputs: one is a photograph of the person’s face and the other is the text of the person’s post. The image is then sent through an artificial neural network module to be translated into yet another image of the same person but with a different expression based on the emotion collected from a text. In spite of this, our suggested text-to-image emotion transmission pipeline may be implemented into any instant messaging program, where it will identify emotions from the discussion and create expressions on thumbnail images of the people being spoken with. For example, if there is some talking that may contain something unpleasant or linked to harassment, one’s photo may be changed out of disgust or rage.

Contributions of this paper are summarized below.

- We provide brief explanations of the image datasets of face expression for facial emotion creation, which can serve as a useful reference for future studies. We experimented with various GAN models on picture datasets in order to create faces with the required emotion, and in this paper, we highlight the outcomes for further investigation. Furthermore, we make required adjustments to the datasets in order to get better results. We used Long Short-Term Memory (LSTM) model (Hochreiter and Schmidhuber, 1997) to train for the task of extracting emotion from text. The retrieved emotions are fed into our facial expression creation algorithm, which then generates facial expressions.
- We provide the findings of our suggested pipeline as well as a prototype application to illustrate our point of view. We developed a blog where the user can upload a post and our model first detects the emotion of the post and then apply the emotion over his/her face, and generates the expression corresponding to the emotion.

**Paper Organization.** The following describes the structure of this paper. Section 2 discusses related studies on face expression creation as well as relevant datasets in more detail. In Section 3, we describe our suggested pipeline and methodology for our image emotion transfer from text, and in Section 4, we describe the outcomes of our experiment. Section 5 concludes our discussion by outlining the limits of our research as well as possible future directions.

## 2 RELATED WORKS

Various relevant studies on GAN for facial expression generation are discussed in this section. We also explore several picture datasets of face expressions, which are subsequently followed by a number of text-based datasets for the purpose of emotion recognition.

### 2.1 Facial Expression Generation

Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) were used in our research to generate pictures with a variety of face emotions. GAN is an adversarial method that is comprised of two neural network models: the generator and the discriminator. The generator model attempts to learn the data distribution, while the discriminator model attempts to differentiate between samples taken from the generator and samples taken from the original data distribution. During the training process, these two models are trained in parallel, with the generator learning to create more and more realistic examples while the discriminator learns to become more and more accurate at differentiating produced data from actual data. As a continuous game, both networks strive to make the produced samples seem as indistinguishable from actual data as possible. The loss function of GAN can be represented by the following equation where the generator tries to minimize it and the discriminator tries to maximize it. The loss function of a GAN model is shown in Eq. 1.

$$L(G, D) = E_x[\log(D(x))] + E_y[\log(1 - D(G(y)))] \quad (1)$$

Here,  $x$  is the real data sample where  $E_x$  is the expected value over all  $x$ .  $D(x)$  is the discriminator’s probability estimation of  $x$  being real, and  $D(G(y))$  is the discriminator’s probability estimation that a fake instance  $y$  is real.

#### 2.1.1 Study on Different Methods

In recent years, a large number of variants of the Generative Adversarial Network have been developed.

Various kinds of GANs have been employed to produce face pictures with specific expressions by a variety of researchers, including several types of Convolutional Neural Networks (Albawi et al., 2017). We've compiled a list of some of the more noteworthy ones.

- *ExprGAN (Ding et al., 2018)*: The Expression Generative Adversarial Network (ExprGAN) is a neural network that generates random and low-resolution pictures of faces using an input face image and a labeled expression. With the face picture and their emotion label, it trains the encoder to produce fresh photos of the same person's face with a different expression. With the assistance of this model, the strength of the produced facial expressions may be adjusted from high to low.
- *StarGAN (Choi et al., 2018)*: An image-to-image translation method known as Star Generative Adversarial Network (StarGAN) produces fixed input facial expressions for various domains based on a single input face. It is capable of learning mappings across domains using just a single generator and a discriminator, which makes it very efficient. The majority of the work on this model was done using CycleGAN, which is a tool for transferring pictures from one domain to another. As previously stated, StarGAN is comprised of two convolutional layers, with the generator using instance normalization and the discriminator employing no normalization. In addition to the PatchGAN (Isola et al., 2016) discriminator network, which determines whether local image patches are genuine or false, the StarGAN discriminator network is based on the Discriminator network of StarGAN. However, it will not be able to create a face emotion that is not already included in the training set.
- *G2GAN (Song et al., 2018)*: In the training phase of the Geometry-Guided Generative Adversarial Network (G2GAN), a pair of GANs is used to execute two opposing tasks, which are performed by the G2GAN. One method is to eliminate the expressions from face pictures, while another is to create synthesized expressions from facial photographs. In conjunction with one another, these two networks form a mapping cycle between the neutral facial expression and the random facial expressions on the face. The face geometry is used to regulate the synthesis of facial expressions in this method of control. Additionally, it maintains the individuality of the expressions when synthesizing them.
- *CDAAE (Zhou and Shi, 2017)*: The Conditional Difference Adversarial Autoencoder (CDAAE) produces synthetic facial pictures of a previously unknown individual with a desired expression based on the conditional difference between the two images. While learning high-level facial expressions, CDAAE uses a long-range feedforward connection that runs from the encoder layer to the decoder layer, and it only takes into consideration low-level face characteristics while learning high-level facial emotions. Instead of using the same pictures as input and output, the network is trained using pairs of photographs of the same person with different expressions rather than using the same images as input and output. This method maintains the identity of the data and is appropriate for use with even smaller datasets.
- *A Text-Based Chat System Embodied with an Expressive Agent (Alam and Hoque, 2017)*: Here the author proposes a framework for a text-based chat system with a life-like virtual agent that seeks to facilitate natural user interaction. They created an agent that can generate nonverbal communications like facial expressions and movements by studying users' text messages. This agent can generate facial expressions for six fundamental emotions: happy, sad, fear, furious, surprised, and disgust, plus two more: irony and determination. To depict expressiveness, the authors used the software programs—*MakeHuman* and *Blender*—to build two 3D human characters, a male and a female and to create realistic face expressions for these agents. Instead than modifying the user's picture, the writers utilized an animated figure to convey emotions.

There are, however, different kinds of GAN models that may be used for the creation of face expressions. In the papers (Deng et al., 2019) and (Liu et al., 2021), the authors used conditional GAN (cGAN) for the generation of 7 expressions (anger, disgust, fear, happy, sad, surprise, and neutral) and 6 expressions (anger, disgust, fear, surprise, sadness, and happiness), respectively. (Chen et al., 2018) used Double Encoder Conditional GAN (DECGAN) to generate seven different expressions in a single run. A further development is the Geometry—Contrastive Adversarial Network (GC-GAN), which was developed by (Qiao et al., 2018) for the generation of face pictures with target expressions. But none of these methods took into account the possibility of creating face pictures with emotions from text. As a result, in order to achieve our goal of emotion transmission to image from text, we concentrated on extracting emotions from the text and then creating face expressions

by combining those feelings with others.

### 2.1.2 Study on Facial Expression Datasets

Several kinds of datasets were explored for generating facial expressions and generating emotion from text. The datasets that were utilized for the creation of face expressions are listed below.

- *CelebFaces Attributes Dataset (CelebA)* (Liu et al., 2015): It is a large-scale face attributes dataset including more than 200K celebrity pictures, each of which has 40 attribute annotations. It is available for download here. The pictures in this collection depict a wide range of posture variations as well as a cluttered backdrop. There are over 10,000 identities, over 202 thousand facial pictures, and five landmark locations with 40 binary characteristics annotations each image.
- *Multimedia Understanding Group (MUG)* (Aifanti et al., 2010): In order to address some of the constraints of previous comparable databases, such as high resolution, consistent lighting, a large number of subjects and several takes per subject, the MUG database was developed. It is made up of picture sequences of 86 people expressing themselves via facial expressions. Each image was recorded in the jpg format with a resolution of  $896 \times 896$  pixels. There were 35 women and 51 men who took part in the database creation. The participants were divided into two groups: women and men. With 1462 sequences accessible, each including more than 1 thousand pictures and seven different face expressions, the possibilities are endless. The reactions range from surprise to delight to fear to rage to neutrality to sorrow to contempt.
- *Facial Expression Research Group Database (FERG)* (Aneja et al., 2016): It is mostly a 2D animation dataset that contains pictures of six stylised characters' facial expressions, most of which are animated (3 males and 3 females). Annotated facial expressions are used to create a database of stylised characters in the game. It includes approximately 55K annotated face pictures of six stylised characters, which are organized into categories. The characters were created using the MAYA program and have six distinct face emotions: furious, disgusted, fear, delight, surprise, neutral, and sad.
- *Oulu-CASIA NIR-VIS Database (Oulu CASIA)* (Zhao et al., 2011): Approximately 74% of the participants in the Oulu-CASIA database are male, with a total of 80 subjects ranging in age

from 23 to 58 years old in the database. Fifty topics are Finnish, and thirty subjects are Chinese, according to the course description. Images are captured at a rate of 25 frames per second and at a resolution of 320 by 240 pixels by using imaging gear.

In the database, there are six different face expressions to choose from: surprise, happiness, sorrow, rage, fear and contempt.

- *AffectNet Database* (Mollahosseini et al., 2017): Developed via the collection and annotation of face pictures, AffectNet is a library of naturalistic facial emotions. There are almost one million face pictures in this collection, which was compiled from the Internet by searching three major search engines with 1,250 emotion-related keywords in six different languages. Manual labeling was performed on about half of the recovered pictures (440K), which included seven different facial expressions and their respective intensities in terms of valence, arousal, and agitation. It includes seven different face expressions: anger, contempt, fear, joy, neutrality, and sorrow.

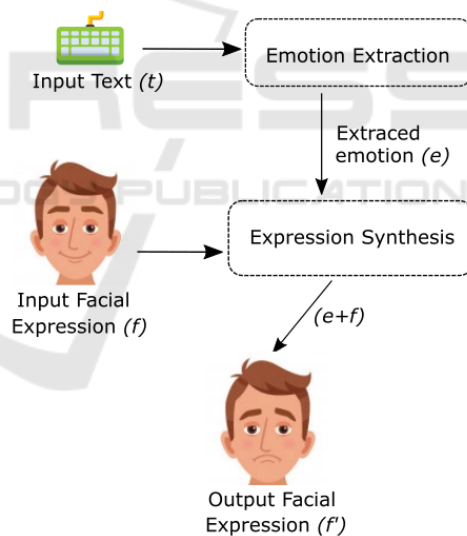


Figure 1: The proposed pipeline of our approach.

## 2.2 Dataset for Emotion Detection from Text

In order to extract the emotion from the text, we utilized the EmoBank Dataset (Buechel and Hahn, 2017). A text corpus of emotion is available, including texts gathered from different social media platforms and the internet as a whole. Emotion categories are manually assigned to each text corpus in a process called manual labeling. Seven different kinds of

Table 1: A statistical overview of EmoBank dataset (Buechel and Hahn, 2017).

Expressions	# of entries
happy	1092
sadness	1082
anger	1079
fear	1076
shame	1071
disgust	1066
surprise	1050
<b>Total</b>	<b>7516</b>

emotions are represented by a total of 7,516 items, including joy, sorrow, rage, fear, humiliation, disgust, and feeling guilty. Table 1 shows the total number of entries for each emotion in this dataset.

We gathered and analyzed a variety of datasets and used a variety of techniques to these datasets while maintaining the same experimental setup in order to determine the most appropriate combination. The (LSTM+EmoBank) combination was chosen in this study to identify emotional content in text since it fulfills our goal in a simple way. As a side note, we discovered that StarGAN on a tweaked version of the MUG dataset performs the best for us. After describing our pipeline, which is based on the previously stated blended approach, we will go through the experimental setups and comparisons in the next part.

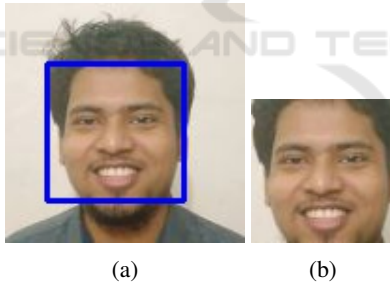


Figure 2: Results of applying Haar-cascade. (a) main input image, and (b) shows final cropped and resized image.

### 3 PROPOSED PIPELINE

The goal of our work is to change the facial expression of a photograph depending on the emotion derived from a particular text. The pipeline for our system is shown in Fig. 1. There are two stages to the pipeline's operation. In the beginning, it accepts the text  $t$  and the first face picture that is entered. To identify the emotion  $e$ , the text is delivered to the emotion extraction module, and the picture of the facial expression  $f$  is provided to the expression synthesis module. An image of the person's face  $f'$  is produced

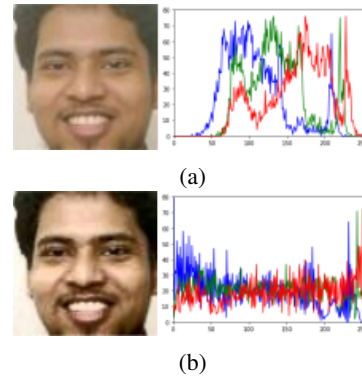


Figure 3: Results of applying histogram equalization. (a) input image of the face (left) with its histogram (right), and (b) output image of the face (left) with its histogram (right) after the equalization.

using an expression based on the data in the synthesis phase ( $f + e$ ) during the synthesis phase. Following that, we will go through each of these stages in more detail.

#### 3.1 Emotion Extraction from Text (LSTM+EmoBank)

The EmoBank dataset (Buechel and Hahn, 2017) was used in conjunction with Long Short Term Memory to aid in the emotion identification process (Sherstinsky, 2018). In order to improve adaption, we do preprocessing on the dataset, which includes case conversion, removal of white space and punctuation marks, spell correction, and handling of numerical symbols and the unknown term, among other things. Afterwards, we embed the text using the GloVe (Pennington et al., ) representation method and train our model to recognize emotions in the text provided by the user.

#### 3.2 Facial Expression Synthesis (StarGAN + tunedMUG)

We use the StarGAN(Choi et al., 2018) to change the facial expression of a person depending on the emotion expressed in a text message. In order to test the technique, we used a modified version of the MUG dataset (Aifanti et al., 2010), which we refer to as the *tunedMUG* dataset. The actions that were taken in order to acquire this version are detailed below.

##### 3.2.1 Face Extraction

Generalization is required in order to run a model across any kind of data. Face expressions need data from a variety of backgrounds, individuals, and situations to be accurate. In order to construct a general-

ized model, we first applied the Haar-cascade (Padilla et al., 2012) to the MUG dataset, which allowed us to concentrate on the faces as much as possible throughout the training process. Thousands of positive pictures (such as photos of faces) and thousands of negative images are used to train the Haar-cascade (images without faces). It provided us the location of the faces and we stored this face as an image of  $128 \times 128 \times 3$  size. Fig. 2 shows the output of applying the Haar-cascade.

### 3.2.2 Histogram Equalization

The Histogram Equalization method (Cheng and Shi, 2004) is used to ensure that all of the picture data has the same distribution. The samples in the dataset are given a generic attribute as a result of this technique. The outcome of the procedure is shown in Fig. 3.

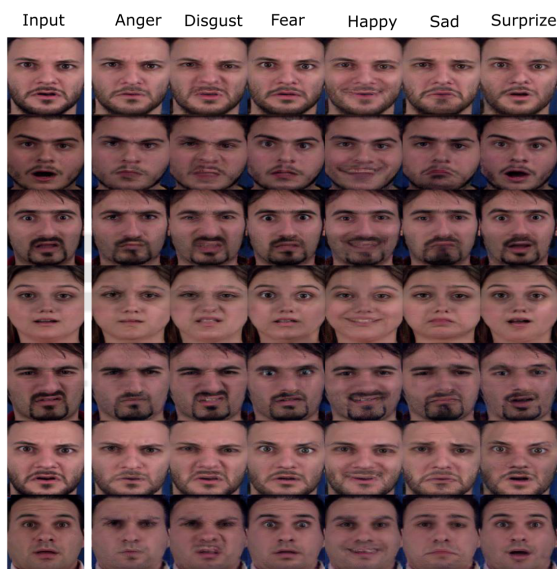


Figure 4: Results of facial expression synthesis for test images. Here the, images in the left column are the input images, and the other columns represent different expressions.



Figure 5: Additional results of facial expression synthesis of our pipeline for test images. Here the faces are of different sources than the original MUG dataset.

## 4 EXPERIMENTS

As part of our pipeline development, we use the (StarGAN + *tunedMUG*) model and the (LSTM + EmoBank) model. In this part, we describe our findings in suitable depth and with relevant comparisons.

### 4.1 Experimental Setup

With 11.17GB GPU support, we were able to put the StarGAN model into action on a Google Colab. It took more than 48 hours to complete our whole training process from start to finish. We utilized pictures from our *tunedMUG* dataset worth about 20k for training purposes. We utilized the Adam optimizer (Kingma and Ba, 2015) with decay rates of  $\beta_1 = 0.5$  and  $\beta_2 = 0.999$  for the 1<sup>st</sup> and 2<sup>nd</sup> moments of the gradient, respectively, with  $\beta_1 = 0.5$  and  $\beta_2 = 0.999$  for the 1<sup>st</sup> and 2<sup>nd</sup> moments of the gradient. By flipping data horizontally, we can also add data augmentation into the equation. Furthermore, we choose batch sizes of 16 and 0.0001 as the learning rates for the generator and discriminator, respectively, to get the best results. In addition, we base our LSTM

Table 2: Comparison of accuracy of RNN and LSTM based emotion extraction from text on EmoBank dataset.

Model+Dataset	Train Acc.	Test Acc.
RNN + EmoBank	44%	33%
LSTM + EmoBank	71%	59%

model on the Google Colab. To train our model, we created an embedding matrix using our dataset and used the `glove.6B.50d` (Lal, 2018), which improved the consistency of our model by identifying similar words in our dataset and included them in our embedding matrix. This embedding matrix assists the model in dealing with the user's input term that was not in our dataset, and the model identified a comparable word to this unknown word in order to accurately generate the outcome. LSTM was used with  $\beta_1 = 0.9$  and  $\beta_2 = 0.999$  for Adam optimizer and categorical cross-entropy (Zhang and Sabuncu, 2018) as the loss function with a batch size of 32.

### 4.2 Experimental Results

With the EmoBank dataset, we ran tests on it with a Recurrent Neural Network (RNN) (Sherstinsky, 2020) and a Long Short-Term Memory (LSTM) (Sherstinsky, 2018), and discovered that LSTM provided acceptable results in terms of test and train accuracy (see Table 2).

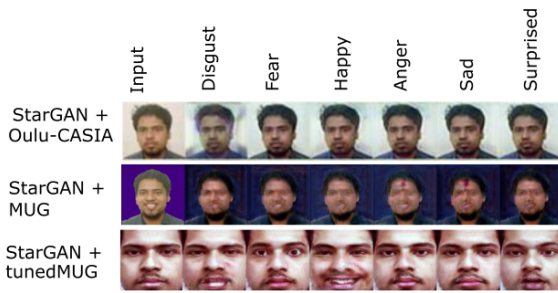


Figure 6: Comparison among Oulu-CASIA (*top*), MUG (*middle*) and our tunedMUG dataset (*bottom*) for applying StarGAN. Here, the leftmost column holds the input faces.

The outputs of the expression synthesis module generated from our *StarGAN+tunedMUG* technique for the testing samples are shown in Fig. 5, which includes the results of the tests. We also present the findings for faces other than those from the MUG dataset, which are more diverse (Fig. 4). The findings show that our system is capable of producing acceptable outcomes in terms of expression synthesis, which is encouraging.

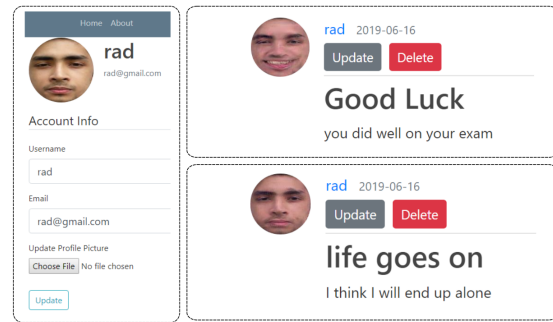
We also show in Fig. 6 a qualitative comparison between the Oulu-CASIA, MUG, and our *tunedMUG* datasets for the purpose of using StarGAN. The figure shows that our *StarGAN+tunedMUG* combination produces much better outcomes than the others.

### 4.3 Application

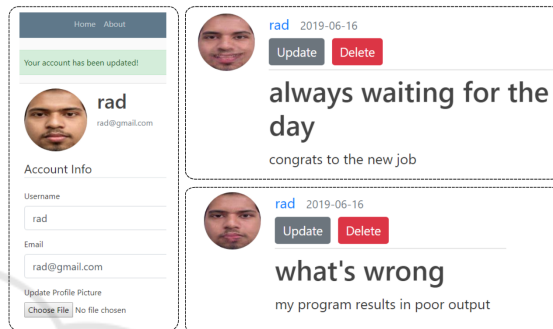
We developed a web application in order to demonstrate the overall performance of our image emotion transmission from text method. In this prototype of a social networking site, users can login, add a profile photo, and publish content to the site. Fig. 7 depicts a scenario in which our program is used, in which the user's expression in the image is changed depending on the written post that he has submitted.

## 5 CONCLUSION

In this article, we suggested a pipeline for the transmission of emotion from text to picture. Our system receives textual input from the user, extracts emotion from it, and then synthesizes an appropriate facial expression depending on the emotion derived from the text. In order to do this, we divided our system into two phases: one for emotion recognition from text, and another for picture creation using GAN. EmoBank dataset with minimal preparation was utilized for emotion processing, and LSTM was employed to get this result. Based on the MUG dataset,



(a) case - I.



(b) case - II.

Figure 7: Two example cases (a & b) of using our application. In both cases: *left*: user's profile page with a image of his face. *top-right*: user shares a post having a happy emotion and the expression in the profile picture is changed. The similar case occurs in the *bottom-right* but for sadness.

we developed a custom expression synthesis module that can be used in any environment. On this adjusted MUG dataset, we used the StarGAN technique to change the facial expression of the participants. In order to show our working pipeline, we have also created an application that reproduces the profile image with different expressions depending on the mood of the user's post in order to demonstrate our functioning process.

There are many possibilities for future endeavors in our system. Currently, our system is focused on the facial area, but we have plans to expand its capabilities to include pictures of the whole body in a variety of positions. In order to achieve better fusion, we want to do ablation studies on more datasets and GAN techniques in the future. We also plan to run user experiment evaluation of our system (Tarkkanen et al., 2015).

## REFERENCES

Aifanti, N., Papachristou, C., and Delopoulos, A. (2010). The mug facial expression database. *11th Interna-*

- tional Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*, pages 1–4.
- Alam, L. and Hoque, M. M. (2017). A text-based chat system embodied with an expressive agent. *Advances in Human-Computer Interaction*, 2017:1–14.
- Albawi, S., Mohammed, T. A., and Al-Zawi, S. (2017). Understanding of a convolutional neural network. In *2017 International Conference on Engineering and Technology (ICET)*, pages 1–6.
- Aneja, D., Colburn, A., Faigin, G., Shapiro, L., and Mones, B. (2016). Modeling stylized character expressions via deep learning. In *Asian Conference on Computer Vision*, pages 136–153. Springer.
- Barrett, L. F., Adolphs, R., Marsella, S., Martinez, A. M., and Pollak, S. D. (2019). Emotional expressions reconsidered: challenges to inferring emotion from human facial movements. *Psychological Science in the Public Interest*, 20(1):1–68.
- Buechel, S. and Hahn, U. (2017). EmoBank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 578–585, Valencia, Spain. Association for Computational Linguistics.
- Chen, M., Li, C., Li, K., Zhang, H., and He, X. (2018). Double encoder conditional gan for facial expression synthesis. In *2018 37th Chinese Control Conference (CCC)*, pages 9286–9291. IEEE.
- Cheng, H. and Shi, X. (2004). A simple and effective histogram equalization approach to image enhancement. *Digital Signal Processing*, 14(2):158–170.
- Choi, Y., Choi, M., Kim, M., Ha, J.-W., Kim, S., and Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8789–8797.
- Deng, J., Pang, G., Zhang, Z., Pang, Z., Yang, H., and Yang, G. (2019). cgan based facial expression recognition for human-robot interaction. *IEEE Access*, 7:9848–9859.
- Ding, H., Sricharan, K., and Chellappa, R. (2018). Exprgan: Facial expression editing with controllable expression intensity. In *Thirty-Second AAAI Conference on Artificial Intelligence*.
- Frank, M. (2001). Facial expressions. In Smelser, N. J. and Baltes, P. B., editors, *International Encyclopedia of the Social & Behavioral Sciences*, pages 5230–5234. Pergamon, Oxford.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680.
- Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A. (2016). Image-to-image translation with conditional adversarial networks.
- Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*.
- Lal, A. (2018). glove.6b.50d.txt. <https://www.kaggle.com/watts2/glove6b50d.txt>. Last accessed: 22 Nov 2021.
- Liu, L., Jiang, R., Huo, J., and Chen, J. (2021). Self-difference convolutional neural network for facial expression recognition. *Sensors*, 21(6):2250.
- Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.
- Mollahosseini, A., Hasani, B., and Mahoor, M. H. (2017). Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*.
- Padilla, R., Filho, C., and Costa, M. (2012). Evaluation of haar cascade classifiers for face detection.
- Pennington, J., Socher, R., and Manning, C. D. <https://nlp.stanford.edu/projects/glove/>. Last accessed: 22 Nov 2021.
- Qiao, F., Yao, N., Jiao, Z., Li, Z., Chen, H., and Wang, H. (2018). Geometry-contrastive gan for facial expression transfer. *arXiv preprint arXiv:1802.01822*.
- Sherstinsky, A. (2018). Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *arXiv preprint arXiv:1808.03314*.
- Sherstinsky, A. (2020). Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network.
- Song, L., Lu, Z., He, R., Sun, Z., and Tan, T. (2018). Geometry guided adversarial facial expression synthesis. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 627–635.
- Tarakanen, K., Harkke, V., and Reijonen, P. (2015). Are we testing utility? analysis of usability problem types. In Marcus, A., editor, *Design, User Experience, and Usability: Design Discourse*, pages 269–280, Cham. Springer International Publishing.
- Xu, Q., Yang, Y., Tan, Q., and Zhang, L. (2017). Facial expressions in context: Electrophysiological correlates of the emotional congruency of facial expressions and background scenes. *Frontiers in Psychology*, 8:2175.
- Zhang, Z. and Sabuncu, M. R. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels.
- Zhao, G., Huang, X., Taini, M., Li, S. Z., and Pietikäinen, M. (2011). Facial expression recognition from near-infrared videos.
- Zhou, Y. and Shi, B. E. (2017). Photorealistic facial expression synthesis by the conditional difference adversarial autoencoder. In *2017 seventh international conference on affective computing and intelligent interaction (ACII)*, pages 370–376. IEEE.
- Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Computer Vision (ICCV), 2017 IEEE International Conference on*.