Human-Centered Design and User Experience, Vol. 114, 2023, 505–513
https://doi.org/10.54941/ahfe1004269

Exploring Anime Character Image Generation Based on User Preferences

Mitsuhiro Hayase

School of Culture-Information Studies, Sugiyama Jogakuen University, Japan

ABSTRACT

This paper presents a method for generating images of anime characters based on user preferences. A questionnaire is developed to measure user preferences, inspired by Rubin's "Love and Liking Scale". Specifically, user preferences are collected by presenting anime character images to participants through a crowd survey and collecting their responses. The model responsible for generating the anime character images is trained using deep learning techniques, using the survey data as a training set. However, attempts to generate images using the trained model did not produce the expected results. When analysing the survey data, it was found that there was limited variability in the "Love and Liking" scales for each anime character. This suggests that the trained models may not adequately reflect user preferences. Future work will focus on improving the model to accurately capture user preferences and developing a more appropriate model. This study provides fundamental knowledge and essential insights for the development and advancement of anime character image generation methods tailored to individual user preferences.

Keywords: Anime character image generation, User preferences, Deep learning

INTRODUCTION

In recent years, CG avatars have found application across a diverse array of media and content domains. Illustrative instances encompass digital signages, the YouTube platform, and the burgeoning Metaverse. Particularly in Japan, the proliferation of digital signages has disseminated to commercial establishments and public transit systems, among other venues. Within the realm of YouTube, Virtual YouTubers (VTubers), who leverage both two-dimensional (2D) and three-dimensional (3D) avatars for video dissemination and live streaming, have achieved conspicuous renown. And other Within the Metaverse, authentic commercial entities have transposed their physical presence into the virtual expanse. This phenomenon is concomitant with services that empower users to engage in shopping experiences through the agency of CG avatars.

These characters, often imbued with diverse attributes such as gender, age, visage, nomenclature, disposition, gesticulation, and vocal timbre, serve to sustain the congruity of the envisioned milieu. Conversely, in scenarios where users fabricate their own character (own avatar), an imperative for an idealized appearance commonly prevails. Nevertheless, the endeavour of characters origination necessitates a gamut of specialized knowledge,

equipment, and temporal investment, thereby incurring substantial fiscal and temporal outlays. Furthermore, the propensity of users to favour their self-constructed character (avatar) remains an indeterminate proposition.

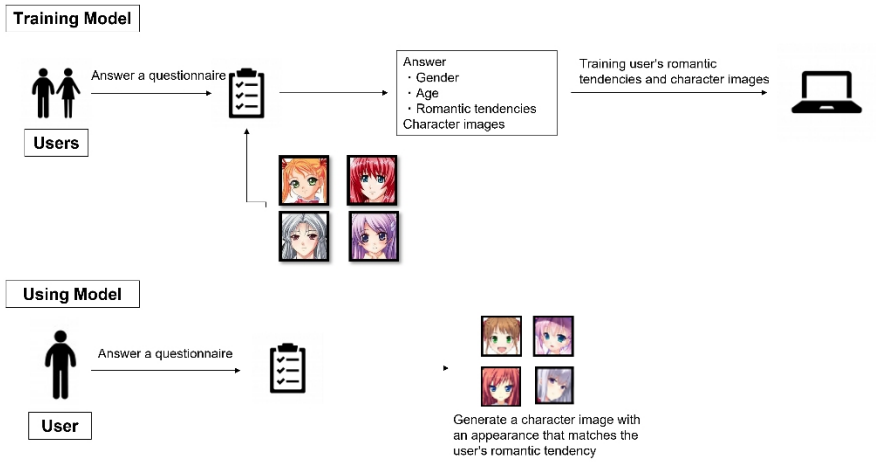

Figure 1: Overall overview (top: when training the model, bottom: when using the model).

In forthcoming times, as interactive agents attain maturation, CG characters entities could conceivably be furnished as the corporeal manifestations of artificial intelligence. In fact, AI-driven avatars and VTubers have appeared in the past couple of years. Therefore, AI-driven user-supported characters gain prominence in the impending era, efforts to create a character that is supported by all are likely to be accompanied by challenges.

Research into preferences for individuals of the opposite gender encompasses the "law of similarity (Byrne 1961)," positing that individuals seek analogies with themselves, and the "matching hypothesis (Elaine 1966)," which advances the selection of partners who exhibit similarity and harmonious compatibility. Moreover, the " Mehrabian's Law" (Mehrabian 1972), which stipulates that visual input exerts a preeminent influence on individuals, supersedes the primacy of language or audition. Additionally, Rubin's Love and Liking Scale (Rubin 1970) is recognized as a gauge of predilection for individuals of the opposite gender.

In this study, our focus is anchored in Mehrabian's Law, with a concentrated inquiry into the implications of the visage of characters. Furthermore, we undertake an exploration of a methodology for crafting character appearances consonant with the romantic tendencies exhibited by each individual user, utilizing the framework of "Rubin's Love and Liking Scale."

METHOD FOR CHARACTER GENERATION

Overview

Figure 1 shows a conceptual diagram illustrating the method for character generation. First, training data is amassed through the administration of a

structured questionnaire. Next, the model is subjected to training via the utilization of a Conditional Deep Convolutional Generative Adversarial Network (cDCGAN), predicated upon the compiled dataset. When embarking upon the task of character generation, the identical questionnaire employed during the data collection phase is responded to. The trained model is then invoked to effectuate the generation of the character's appearance.

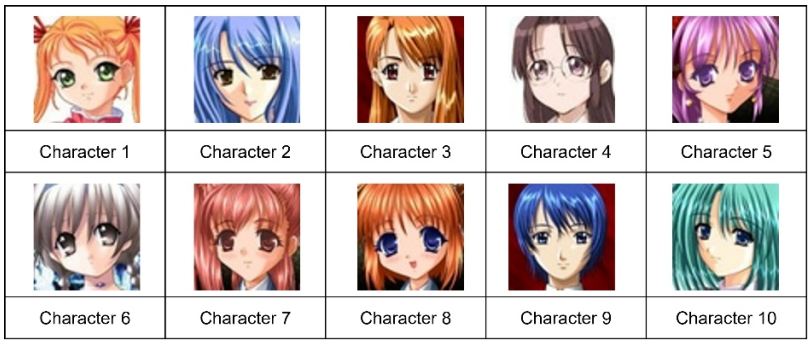

Figure 2: Examples of characters used in question the user's romantic tendency.

Data Collection

The questionnaire comprised dual inquiries concerning the user's domain particulars, namely gender and age, alongside an assemblage of 22 inquiries (culminating in a total of 24 queries) germane to the user's romantic tendency. These queries were grounded in the framework of Rubin's Love and Liking Scale (Rubin 1970). Elicitation of responses was conducted utilizing a 7-point Likert scale, with the rating "strongly agree" assigned to 7 and "mostly disagree" allocated to 1. Respondents furnished their responses contingent upon their appraisal of the appearance of character. The characters utilized were drawn from the "Anime Face Dataset" accessible on Kaggle.

The data acquisition was facilitated through the utilization of "Yahoo! Crowdsourcing," an initiative within the domain of Yahoo!Japan. The cohort of respondents numbered 2,200 individuals, comprising 1,408 males and 792 females. The average age of the respondents was 44.3 ± 0.4 years. Cumulatively, 550.0 ± 343.9 responses were garnered for each individual character.

Generative Model of Character Appearance

The generation of character appearance is accomplished through the utilization of the conditional Deep Convolutional Generative Adversarial Network (cDCGAN), an amalgamation of the foundational Deep Convolutional Generative Adversarial Network (Radford 2016), and the Conditional Generative Adversarial Network (Mirza 2014). The instantiation of the cDCGAN model is elucidated in Figure 3, while the architectural of the Generator (G) and Discriminator (D) are delineated in Figure 4 and Figure 5, respectively.

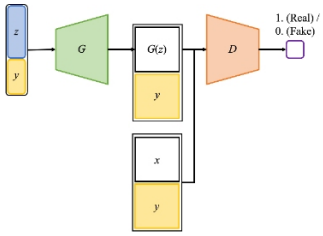

Figure 3: Definition of conditional DCGAN.

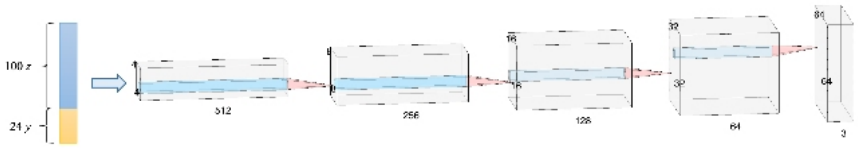

Figure 4: Generator.

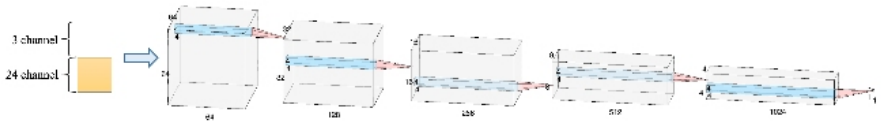

Figure 5: Discriminator.

The Generator is constituted by a series of layers, encompassing convolutional-transpose layers, batch normalization layers, and Rectified Linear Unit (ReLU) activation functions. Conversely, the Discriminator's composition encompasses convolutional layers, Batch Normalization layers, and Leaky Rectified Linear Unit (LeakyReLU) activation functions. Ultimately, the terminal output of discriminator is proffered as a probability value via the Sigmoid function. Compare the two methods of model training.

Model Training Method 1

The Generator receives a 124-dimensional input vector, forged through the fusion of a 100-dimensional latent vector (denoted as 'z') emanating from a standard normal distribution and a 24-dimensional vector (represented as 'y'), encapsulating attributes pertaining to romantic inclinations, age, and gender. Egressing from this process is a $3 \times 64 \times 64$ color image.

On the other hand, the input to the discriminator constitutes a 27-channel vector, comprised of 24 vectors, each commensurate with the dimensions of the input image, conjoined with the 3 channels of the image itself. The dimensions of the input image stand at 64×64. The resultant output is the ascertained probability signifying the authenticity of the image.

The data pertaining to the proclivity for romantic tendencies are subjected to normalization, thereby mapping the range from [1,7] to the interval [0,1]. Simultaneously, age undergoes normalization within the confines of [0,1],

operating under the premise that the upper age threshold spans 100 years. Gender is denoted as 1 for males and 0 for females.

Model Training Method 2

The inputs directed towards the Generator comprise One-Hot vectors, encompassing 154-dimensions for romantic tendencies (computed as $22 \times 7 = 154$) and an additional two dimensions for gender. The age component remains consistent with the specifications outlined in Method 1. Notably, the Generator operates on an input vector spanning 257-dimensions, achieved by the amalgamation of the 157-dimensional vector and the 100-dimensional latent vector.

Conversely, the input supplied to the discriminator entails an assemblage of 160 channels, each accommodating 157-dimensional vectors. These vectors maintain dimensions identical to those of the input image, and they are further conjoined with the 3 channels corresponding to the image.

Model Training

D and G endeavours to maximize the logarithmic probability $\log D(x)$, wherein D aptly categorizes both authentic and synthesized images, while concurrently minimizing the logarithmic probability $\log(1 - D(G(x)))$ wherein G generates images. Consequently, the loss function is formulated as follows.

$$\min_G \max_D V(D, G) = E_{x \sim pdata(x)}\left[\log D(x)\right] + E_{z \sim pz(z)}\left[\log(1 - D(G(x)))\right]$$

The optimization employs the Adam optimizer, with a learning rate (lr) of 0.0002 and a Beta1 value of 0.5. For the initiation of parameters, we leverage a model previously trained via DCGAN employing the Celeb-A Face Dataset. The DCGAN is trained subsequent to the random initialization of model weights from a normal distribution with a mean of 0 and a standard deviation of 0.02. Additionally, the training spans 100 epochs. The training approach adopts the mini-batch methodology, with a batch size defined as 64.

TRAINING RESULT

Model Training Method 1

Figure 6 shows the loss values of the Generator and Discriminator employing Method 1, while Figure 7 showcases the outcomes of image generation. The generated results are presented across distinct batch counts, with a consistent latent vector across all images, yet varied romantic tendencies, ages, and genders.

Upon scrutiny of Figure 6, it becomes evident that the separation between the Generator's loss and the Discriminator's loss remains proximate and co-linear, suggesting favourable progress in learning. However, given the proximity of Discriminator's loss to zero, it can be inferred that the Discriminator exhibits superior performance vis-à-vis the Generator. Examining Figure 7, visual noise manifests within the generated images. This outcome

arises due to the generator's loss hovering around 6, indicating that an increment in the number of learning iterations is unlikely to culminate in enhanced performance.

Model Training Method 2

Figure 8 shows the loss trends of the generator and discriminator employing Method 2, while Figure 9 showcases the outcomes of image generation. Analogous to Method 1, the generated results are presented across varying batch counts, maintaining a consistent latent vector across all images, albeit with varying romantic tendencies, ages, and genders.

Upon examining Figure 8, it is evident that the loss trajectories of the Generator and Discriminator align in a parallel fashion. In comparison to Figure 6, the elongated separation in Figure 8 signifies enhanced discriminator performance in Method 2. This augmentation is indicated by the greater magnitude of respective distances. Notably, upon inspection of Figure 9, an increased prevalence of noise is observable in contrast to Figure 7, resulting in images characterized by coarseness.

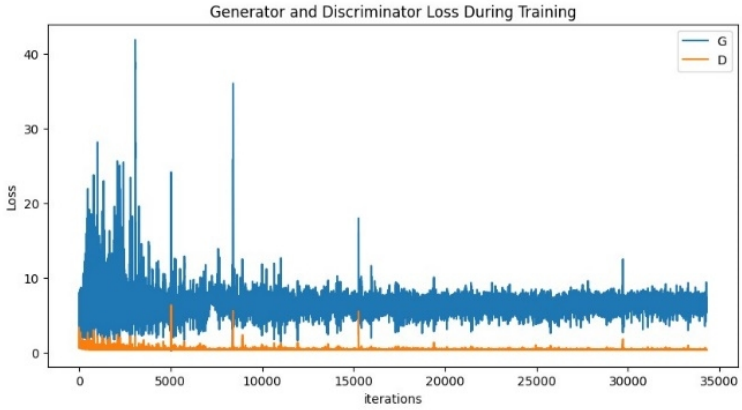

Figure 6: Transition of loss by method 1.

Figure 7: Generation character's appearance by method 1.

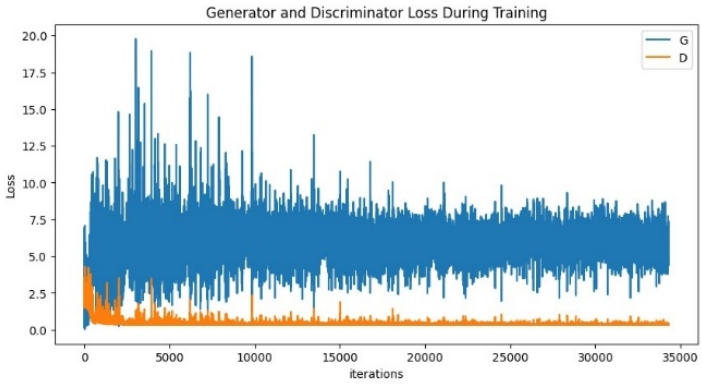

Figure 8: Transition of loss by method 2.

Figure 9: Generation character's appearance by method 2.

EXPERIMENT

A comparative analysis was conducted involving two scenarios: "When the latent vector remains constant while a vector composed of random data representing romantic tendency, age, and gender is introduced as input," and "When the vector comprising randomly generated romantic tendency, age, and gender data is held constant while the randomly generated latent vector is introduced as input". This comparison was performed separately for both Examination Method 1 and Examination Method 2.

Figure 10 shows the outcomes of generation with fixed latent vectors, whereas Figure 11 shows the outcomes of generation with fixed vectors representing romantic tendency, age, and gender.

Upon scrutiny of Figure 10, it is evident that subfigure 10(a) yields analogous characters, whereas subfigure 10(b) produces only a single distinct character. Turning attention to Figure 11, it becomes apparent that the same character is generated in subfigures 11(a) and 11(b). Nevertheless, in 11(b), although the character remains consistent, an image with a dissimilar hair color is yielded.

(a) Result of method 1 (b) Result of method 2

Figure 10: Result of character image generation when latent vector is fixed.

(a) Result of method 1 (b) Result of method 2

Figure 11: Result of character image generation when romantic tendency, age, and gender vectors are fixed.

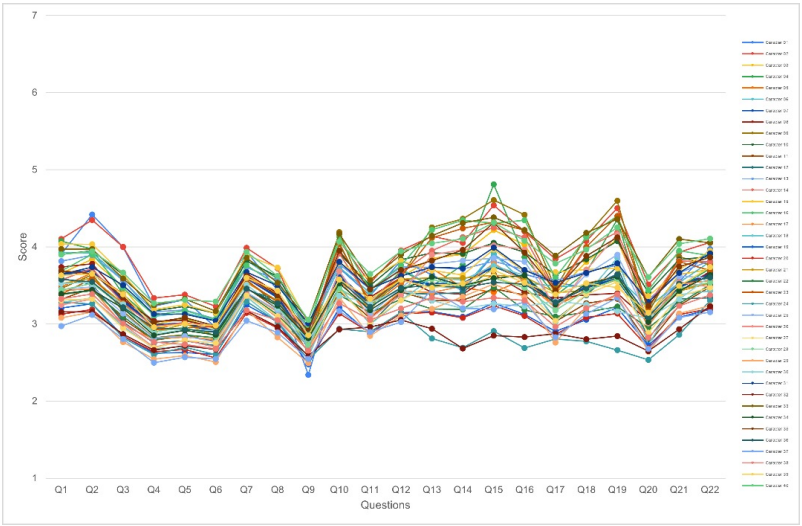

Figure 12: Average value of question item for each character image.

ANALYSIS OF COLLECTED DATA

The randomly generated vectors representing inclinations romantic tendencies displayed marginal variance in character outcomes. Consequently, we undertook an analytical examination of the compiled dataset. Figure 12 presents the computed average values for each character across the spectrum of inquiry items. Upon perusal of Figure 12, it becomes evident that the average values pertaining to individual question items evince proximity. Noteworthy differences are discernible in question items 15 and 19. Specifically, Question 15 queries, "Do you exude seriousness?" while Question 19 inquires, "Do you appear intelligent?" Notably, these inquiries bear in Japanese a semblance in meaning. Accordingly, an unpaired t-test was executed between these two items, yielding a significant difference with a p-value of 3.02×10^{-43}.

Subsequently, modifications were affected to the values of these two items, with an attempt to render the remaining values as the computed average (rounded to the nearest integer), and gender designated as 1. The outcomes are depicted in Figure 13. In each figure, the character's score for each question item shifts progressively from 1 to 7 along the horizontal axis. Despite the conspicuous difference between the two aforementioned items, manipulation of a solitary value did not engender discernible variations in character appearance.

CONCLUSION

Within this study, we embarked on an investigation concerning the formulation of a methodology for character appearance generation that aligns with the user's romantic disposition. Employing the cDCGAN, we endeavoured to establish it as the mode of character generation. A comparative assessment was conducted between the original romantic tendency vector and its transformation into a one-hot encoded vector. The findings revealed that the accuracy of the generated outcomes was notably superior for the non-one-hot encoded vectors. Nonetheless, both methodologies failed to engender distinct characters solely predicated upon the spectrum of romantic tendencies. It can be inferred that an effective learning of the relationship between character images and questionnaire-derived data remains elusive.

In forthcoming endeavour, we envisage delving into the exploration of the nexus between images and questionnaire-derived data, leveraging methodologies such as StyleGAN (Karras 2019), CLIP (Radford 2021), and similar approaches. Furthermore, we contemplate the integration of user attributes like the Big5 index, as we navigate the course ahead.

REFERENCES

Albert Mehrabian (1972) "Silent Messages: Implicit Communication of Emotions and Attitudes", Wadsworth Publishing Company.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever (2021) "Learning Transferable Visual Models From Natural Language Supervision", ICML 2021.

Alec Radford, Luke Metz, Soumith Chintala (2016) "Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks", ICLR 2016.

Byrne. D (1961) "Interpersonal attraction and attitude similarity", The Journal of Abnormal and Social Psychology, Vol. 62, No. 3, pp. 713–715.

Mehdi Mirza, Simon Osindero (2014) "Conditional Generative Adversarial Nets", arXiv:1411.1784.

Rubin, Z. (1970) "Measurement of romantic love", Journal of Personality and Social Psychology, Vol. 16, No. 2, pp. 265–273.

Tero Karras, Samuli Laine, Timo Aila (2019) "A Style-Based Generator Architecture for Generative Adversarial Networks", CVPR2019.

Walster Elaine, Aronson Vera, Abrahams D arcy, Rottman Leon (1966) "Importance of physical attractiveness in dating behavior", Journal of Personality and Social Psychology, Vol. 4, No. 5, pp. 508–516.