StackGAN を用いた感情を表すコミック画像生成に関する予備的検討

Preliminary Study on Generating Comic Images Expressing Emotion Using StackGAN

[1]陳 彦嘉　[2]柴田祐樹　[3]高間康史

Yen-Chia Chen[1], Hiroki Shibata[2], Yasufumi Takama[3]

東京都立大学大学院システムデザイン研究科
Graduate School of System Design, Tokyo Metropolitan University

Abstract: This paper reports the result of preliminary experiments of a method for generating comic images expressing emotions using StackGAN. The proposed method trains StackGAN using the dataset of comics of positive or negative facial images, which are associated with text describing emotions. Generated images are evaluated by applying facial expression recognition. Experiments are conducted changing the types of facial expressions in the training images to investigate the impact of diversity in the training data in terms of emotions on the quality of the generated images.

1　はじめに

本稿では，感情を表現するコミック画像を，StackGAN を用いて生成する手法について予備的検討を行った結果について報告する．企業において，サービスの向上は重要な課題であり，そのヒントとなるのはユーザからのフィードバックとなる．レビューやソーシャルメディアに投稿されたコメントは，企業のサービスに関するユーザの意見が含まれているため，これを活用することでアンケートやインタビューを実施せずにユーザからのフィードバックが得られることが期待できる．

レビューなどのコメントの特徴は，対象サービスなどに対する記述とともに，投稿者の感情が表現されていることである．サービスに満足しているコメントはポジティブな感情を示していることが想定される一方，商品に不満のある場合や，他者を攻撃するような悪意あるコメントの場合はネガティブな感情を示していると考える．従って，レビューコメントを分析する際には，このような感情の観点から分析する必要もあると考える．しかし，コメントを読んで感情の種類やその程度を読み取ることは分析者にとってコストのかかる作業である．また，sentiment analysis [1]などを適用して数値化しても，数値やグラフから感情を認識することは分析者にとって直感的ではないと考える．

本稿では，レビューコメントに含まれる感情を，表情を表す顔画像として可視化する手法を提案する．特に絵で物語を語るコミックは，登場人物の表情で様々な感情を読者に伝え，感情移入させることが可能であることから，コミック画像として生成することを試みる．提案手法の利用方法として，Charnoff face[2]の様に，多数のレビューコメントから生成したコミック画像を並べて提示することで，その概要把握を支援することが考えられる．また，レビュー投稿前にそれが表す感情をコミック画像として提示することで，その投稿が他者に与える印象を確認させる用途も期待できる．

前述の目的達成に必要な要素技術として，本稿では感情を表すコミック画像の生成手法について検討する．テキストから画像を生成する AI は近年急速に研究が進んでいる分野であり，代表的な生成手法の一つに GAN（Generative adversarial networks，敵対的生成ネットワーク）[3]がある．GAN は画像を生成する生成器（generator）と生成画像の真偽を判断する識別器（discriminator）から構成され，識別器の判断結果をフィードバックして，両者が反復的に学習を行い精度を向上させることで，生成器が高品質の画像を生成可能となる．GAN には様々なバリエーションが存在するが，本稿では感情を表すテキストから画像を生成することを目的とするため，StackGAN[4]を利用する．StackGAN は text-to-image タイプの GAN の中で代表的なものであり，多くの拡張手法が提案されている．画像を生成するタスクを，テキストを

用いて画像を生成するステージ１と，ステージ１の生成結果を用いて最終画像を生成するステージ２の２段階に分けて生成することで高解像の画像生成を実現しており，細かい線から構成されるコミック画像の生成にも有効であることが期待できる. しかし，StackGAN は鳥や花などに関する写真の生成に適用され，その有効性が示されているが，コミック工学においてコミック画像の生成に適用した研究は報告されていない.

提案手法では，ポジティブ、ネガティブな表情のコミック画像に，感情や表情を説明するテキストを付与して学習に用いる．生成した画像の評価において，客観性および再現性を重視するため，感情認識を適用する．学習に用いる画像が表す感情の種類を変更して行った実験結果を比較し，感情に関する学習データの多様性や画像の枚数が生成画像に与える影響を考察する.

２　関連研究

2.1 Stack GAN

テキストから画像を生成する研究のうち，Stack GAN は代表的な手法であり, 拡張手法も提案されている[5][6]．StackGAN はステージ１，２の二段階から構成される．ステージ１では，テキストをベクトル化し，画像を生成する．ステージ２ではこの画像を更に入力に用いて，より高解像度の画像を生成して最終的な出力とする．ステージ１は主にテキストの内容を正確に読み取ることを重視して，画像としての質に関してはステージ２で調整する.

StackGAN の拡張手法として，StackGANv2[5]は三段階以上のステージにより構成されている．生成に用いる色の一貫性を保つため，正則化項を提案しており，この二点の改良によって，より高解像度の画像の生成が可能となっている．AttnGAN[6]は文単位の入力テキストを一つのベクトルに変換するのではなく，複数の単語に区切り，単語ごとにベクトル化する．これにより，画像は要素ごとに，関連性の高い単語から変換されたベクトルを用いて生成される.

2.2 Russell の感情円環モデル

Russell の感情円環モデルは，ヒューマン・ロボット・インタラクションや感情分析などの工学的研究において良く用いられており，快―不快と覚醒―非覚醒の二軸により構成される平面上に，感情を円環状に配置したものである．ポジティブな感情には元気，幸せや満足などがあり，快側に配置される．ネガティブな感情は不快側に配置され，緊張や心配，悲しみなどが存在する. 平面上に 16 種類の感情が配置されている．また，これら 16 感情の外周には，典型的感情エピソードである，驚き，幸せ，怒り，恐怖，嫌気と悲嘆が配置されている.

2.3 GAN によるコミック風肖像画生成

顔写真を入力として，コミック風の肖像画を生成する手法が提案されている[7]．この研究では，Pix2Pix[8]と CycleGAN[9]を用いている．入力された顔画像の特徴を残してコミック風の画像に変換するために，複数のタスクに分けて学習を行っている．第一段階では CycleGAN を利用して，コミック画像を用いて顔写真風の画像を生成している．第二段階では Pix2Pix を利用して，生成された画像を元のコミック画像に戻す. 最終的に, 学習した Pix2Pix を利用して，顔写真を入力に用いてコミック画像を生成する．また，作者の異なるコミック画像をデータセットとして用いることで，コミックの画風が学習に与える影響についても考察している.

３　提案手法

3.1 コミック画像データセットの構築

本稿では Manga109[1]から，研究用途に提供されているコミックデータセットを利用する．このデータセットには，109 タイトルのコミックが含まれており，それぞれ単行本の第一巻が収録されている．また, 本研究では使用しないが, 登場人物のセリフを，吹き出しの位置と発言者の情報とともに記録したデータも含まれている.

3.2 StackGAN による学習

顔写真とは違い，一枚のコミック画像に含まれている特徴量は少ない．色情報がグレースケールに限定されていること，奥行きに関する情報の欠落により，コミック画像の生成はアニメ風画像の生成よりも困難と考える. また, StackGAN をコミック画像の生成に用いた前例も著者らが調査した限り存在しない．そこで本稿では予備的検討として，段階的に画像データセットを増やすことで, StackGAN により生成されるコミック画像がどのように変化するかを検

[1] http://www.manga109.org/ja/index.html

証する．

Russell の感情円環モデルの快側の感情から serene（晴朗），contented（満足），happy（幸せ），不快側から stressed（重圧），および典型的感情エピソードから disgust（嫌気），anger（怒り）を選択し，それらに関する顔画像をコミックデータセットから選び出して学習に用いる．学習は表1に示すデータセットを用いて３回行う．表における数値はその感情に対応した画像の枚数である．

表 1：学習に用いるデータセットの枚数

データ	晴朗	満足	幸せ	重圧	嫌気	怒り
1	2			2		
2	20			20		
3	2	2		2	2	
4	20	20		20	20	
5	2	2	2	2	2	2
6	20	20	20	20	20	20

入力テキストは，感情の極性，感情の種類，表情に関する単語により構成される．感情の極性を表す単語は，Positive（快），Negative（不快）のどちらかである．感情の種類を表す単語は上述の通り，Positive の場合は serene，contented，happy，Negative の場合は stressed, disgust, anger のいずれかとする．表情を表す単語は，Positive の場合は laugh（笑い）と smile（微笑み）から，Negative の場合は bellow（怒鳴る）と grumpy（険しい表情）から一単語選ぶ．感情ごとに，2 種類の表情の画像を同数用意する．３回の実験において，学習率（learning rate）は両ステージで 0.0002，世代数はステージ 1 で 10000，ステージ 2 で 5000 とした．

3.3 生成画像の評価

生成した画像に対して，生成したい感情の極性を表す表情が生成されているかを評価する．生成した画像を人手で評価する方法も考えられるが，本稿では再現性，客観性を考慮して感情認識を適用して評価する．感情認識には Py-Feat[2]を利用する．Py-Feat が認識する感情のカテゴリは Russell 感情円環モデルとは異なり，快の感情は happiness，不快の感情は anger, disgust, fear, sadness, それ以外の感情は neutral と，surprise である．Py-Feat は各感情に該当する度合いを数値として出力するため，認識結果で，度合いが一番高い感情が属する極性を認識結果とする．

4 実験結果

4.1 データセット

画像に関しては，Manga109 のデータセットから，キャラクターの顔画像部分を 256x256 のサイズで切り出したものを利用する．コミック画像の特性上すべて正面の顔画像にするのは難しく，セリフ（吹き出し）に関しても学習に影響を及ぼす可能性があるため，データセットの拡張も兼ねて左右反転のコピーをする．すなわち，学習には表1に示した枚数の２倍の画像を用いる．

学習に用いる画像は必ずしもセリフを伴うものではなく，また感情的な単語を含んでいるとも限らない．そこで，画像に対応するテキスト（３種類の単語）については，著者の一人が画像から判断して付与する．

4.2 実験結果

4.2.1 実験 1

実験１では，表 1 に示したデータセット１を用いて学習する．ステージ１の学習では，世代数 8500 から生成画像がはっきりし，9000 辺りで崩壊する傾向が見られた．その後のステージ２では顔画像が生成されなかったため，学習に失敗したと考える．

4.2.2 実験 2

実験２では，表 1 に示したデータセット２を用いて学習する．ステージ１では，世代数 500 から生成画像がはっきりし，世代数 7400 と 8800 からそれぞれ約 200 世代崩壊する以外，安定して顔画像が生成された．ステージ２では，世代数 1000，2000 で顔の輪郭が生成されたが，生成する画像にすべてグレーのノイズが発生するため，感情認識ができなかった．

4.2.3 実験 3

実験３では，表 1 に示したデータセット３を用いて学習する．ステージ１では，世代数 750 から生成画像がはっきりし，世代数 3000 から 3700 の間は崩壊していた．ステージ２では，世代数 250 から生成画像がはっきりした．しかし，世代数 1100 から 3200 の間は崩壊し，その後 3600 までの 400 世代ではノイズがかかった顔画像が出力された．以降は再びグレーのノイズ画像しか出力されなかった．各感情と表情の組み合わせ 8 通りについて，学習世代数 1000 の生成画像に対する極性判定結果の混同行列を表 2 に

[2] https://py-feat.org/pages/intro.html

示す．行が認識結果に対応する．

表 2：実験 3 の分析結果

感情	晴朗	満足	重圧	嫌気
快				
不快	1		1	
失敗	1	2	1	2

表 2 より，感情が認識できた画像は 8 枚中 2 枚のみであった. 認識できた晴朗の画像は, 怒り（anger）のスコアが一番高く，快に関する感情は喜び（happiness）のスコアがわずかに出力された.

4.2.3 実験 4

実験 4 では，表 1 に示したデータセット 4 を用いて学習する．ステージ 1 では，世代数 1100 から生成画像がはっきりし，その後生成画像の崩壊は見られなかった．ステージ 2 では，世代数 2400 から，顔の輪郭部分とコマ枠部分にノイズが発生した．はっきりした顔画像は生成されなかったが，ノイズの分布は実験 1，2 と違い，画像すべてにかかるのではなく，外回りに集中していた.

4.2.3 実験 5

実験 5 では，表 1 に示したデータセット 5 を用いて学習する．ステージ 1 では，世代数 1000 から生成画像がはっきりし, 5400 から 5800 と 7000 から 7800 の間で崩壊した．ステージ 2 では，世代数 300 から生成画像がはっきりした．2100 から 2800 は生成画像にノイズがかかり，以降はノイズが強くなる傾向が観測された. 学習世代数 2000 の生成画像に対する極性判定結果の混同行列を表 3 に示す.

表 3：実験 5 の分析結果

感情	晴朗	満足	幸せ	重圧	嫌気	怒り
快						
不快			1			
驚き						1
中立	2					
失敗		2	1	2	2	1

表 3 より，実験 5 では認識できた 4 枚のうち 3 枚が快の感情を表す画像であった. 図 1, 2 に例を示す. Py-Feat ではノイズの有無にかかわらず，輪郭線が細い画像や，デフォルメされたキャラクターに対して正しく認識できない傾向が見られた.

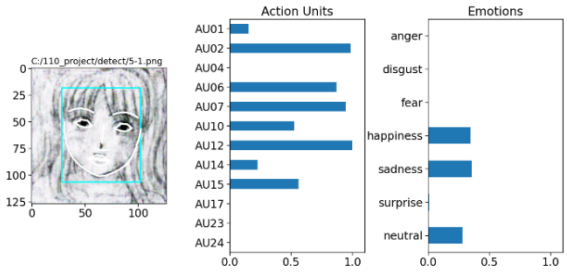

図 1：認識に成功した画像例（実験 5）：
Positive, happy, smile.

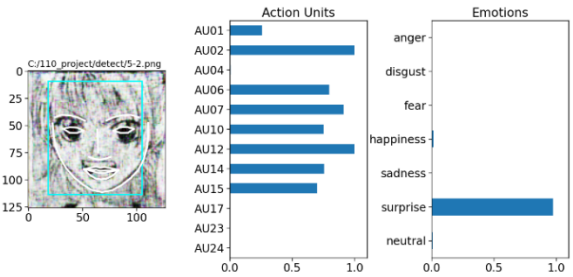

図 2：異なる感情として認識された画像例
（実験 5）：Negative, anger, bellow.

実験 3 と 5 において，同じセリフから生成された同一キャラクターの画像を図 3，4 に示す．表 2，3 に示した通り，実験 3 では不快の感情が，実験 5 では快の感情が生成しやすい傾向が見られたが，今後生成枚数を増やして確認する必要があると考える.

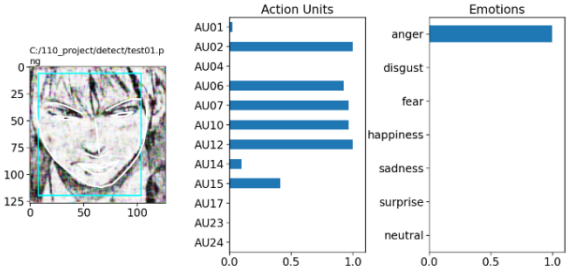

図 3：生成画像例（実験 3）：
Negative, stressed, grumpy.

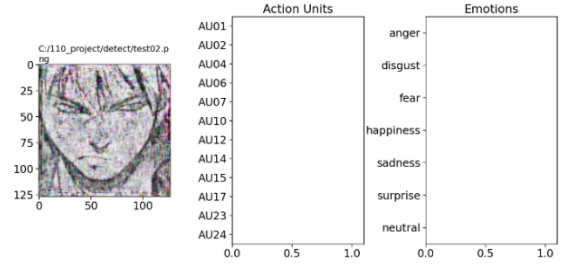

図 4：生成画像例（実験 5）：
Negative, stressed, grumpy.

実験３と５で生成された，感情を認識できた画像のうち，同じキャラクターの画像が一種類存在した（図 5，6）．これより，画像として同様の表情に見えるが，感情認識の結果は大きく異なっており，感情認識の頑健性は低いといえる．

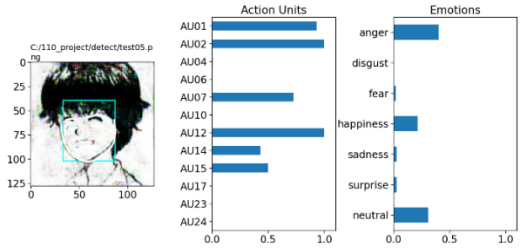

図 5：生成画像例（実験３）：
Positive, serene, laugh.

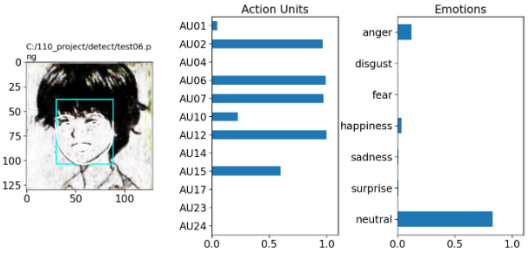

図 6：生成画像例（実験５）：
Positive, serene, laugh.

4.2.3 実験 6

実験 6 では，表 1 に示したデータセット 6 を用いて学習する．ステージ 1 では，世代数 200 から生成画像がはっきりし，4700 から 5200，5700 から 8300 の間と，9400 以降で崩壊した．ステージ 2 では，世代数 3000 から，実験 4 と同様のノイズ画像が生成され，表情認識が適用できなかった．

4.2.4. 学習データの多様性による影響の考察

感情の種類や，学習に用いる画像の枚数と認識結果の関係について考察する．感情の種類と生成画像の関係については，2 種類，4 種類と 6 種類の感情に対して，それぞれ 2 回の実験を行った結果，2 種類の感情を用いた学習では，2 回の実験どちらも画像が上手く生成されなかった．これより，感情の種類はある程度用意する必要があると考える．

学習に用いる画像の枚数と認識結果の関係については，感情認識ができた実験 3, 5 は各感情 2 枚ずつ学習に用いた場合であり，20 枚ずつ用いた場合は一度も感情認識が成功しなかった．この結果より，画像を増やしても生成画像の質は向上しない可能性が示されたが，今後さらに枚数を増やしての検証が必要と考える．また，学習過程は単調ではなく，一度表情が読み取れる画像が生成されるようになった後，世代を重ねると崩壊する場合が観測された．この現象についても，今後さらに検証が必要と考える．

Py-Feat の認識結果については，快の感情に分類された画像はなく，不快の感情が 3 枚，中立（neutral）が 2 枚，驚きが 1 枚という結果になった．図 5 と図 6 では顔の領域として検出された位置がずれている．これは，図 5 で眉毛の部分が上手く生成できてなく，目の部分が眉毛に誤認されたためと考える．また，図 1，2，3，4 を見る限り，多少のノイズがあっても表情が認識できる場合があるため，キャラクターの画風，輪郭線，描かれている角度の影響が認識精度に与える影響が大きいと考える

５　まとめ

本稿では，感情を表現するコミック画像を，StackGAN を用いて生成する手法について予備的検討を行った．実験を行った結果，4 種類以上の感情に対してそれぞれ 2 枚の画像を与えた場合に感情認識が可能であったことを示した．また，学習過程は単調ではなく，後半に生成画像が崩壊する現象が多くみられたことから，安定した画像生成にはまだ至っていないと考える．Py-Feat による感情認識結果は，快の感情を認識できない傾向にあるが，顔として検出された領域を見る限り，キャラクターの顔輪郭と顔パーツの位置はある程度正確に把握可能であることを確認した．

本稿で示した実験結果より，ステージ 2 での学習が失敗しやすい傾向がみられたため，安定した画像生成のために今後は，学習率の調整，及びモデル構造の変更を検討する予定である．また，テキスト入力の多様性向上や，ディフュージョンモデルを用いた同様の実験に取り組む必要があると考える．

参考文献

[1] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. “Sentiment analysis algorithms and applications: A survey.” Ain Shams Engineering Journal, Vol. 5, No. 4, pp. 1093-1113 (2014).

[2] Christopher J. Morris, David S. Ebert, and Penny L. Rheingans. “Experimental analysis of the effectiveness of features in Chernoff faces.” 28th AIPR Workshop: 3D Visualization for Data Exploration and Decision Making (2000).

[３] Antonia Creswell, Tom White, Vincent Dumoulin, Kai Arulkumaran, Biswa Sengupta, and Anil A. Bharath. “Generative Adversarial Networks: An Overview.” IEEE Signal Processing Magazine, Vol. 35, No. 1, pp. 53-65 (2018).

[４] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. “StackGAN: Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks.” ICCV2017, pp. 5907-5915 (2017).

[５] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N. Metaxas. “StackGAN++: Realistic Image Synthesis with Stacked Generative Adversarial Networks.” IEEE Trans, on Pattern Analysis and Machine Intelligence, Vol. 41, No. 8, pp. 1947-1962 (2018).

[６] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. “AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks.” CVPR2018, pp. 1316-1324 (2018).

[７] Yen-Chia Chen, Lieu-Hen Chen, Hiroki Shibata, and Yasufumi Takama. “Styled Comic Portrait Synthesis Based on GAN.” JSAI2021: Advances in Artificial Intelligence, Springer, pp. 69-80 (2022).

[８] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. “Image-to-Image Translation with Conditional Adversarial Networks.” CVPR2017, pp. 1125-1134 (2017).

[９] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. “Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks.” ICCV2017, pp. 2223-2232 (2017).