自然言語による編集要求に対して効率的に確認を行う対話的画像編集システム

品川 政太朗[1,2,a)]　吉野 幸一郎[1,2]　Seyed Hossein Alavi[3]　Kallirroi Georgila[3]　David Traum[3]
Sakriani Sakti[1,2]　中村 哲[1,2]

概要

自然言語による編集指示文を用いた画像編集タスクでは，入力となる編集指示文が多様性なために編集結果の制御が難しい．これまでマスク機構を導入することで制御性を向上する手法が提案されてきたが，これは画像編集モデルに対して強い制約として働き，一部の編集を難しくしてしまう．本研究では，マスクありとなしのモデルによる編集結果をユーザに提示して確認する対話的画像編集システムを提案する．また，ユーザの労力を減らすために，マスクのエントロピーに基づいて，システムがマスクありのモデルで対応できない場合のみ確認するという確認戦略を提案する．

1. はじめに

画像編集システムは，ユーザとのインタラクションを介して所望の画像の作成を支援することを目的とする．このような，ユーザとインタラクションするシステムに自然言語インタフェースを導入することは，ユーザがシステム利用のための特別なスキルの習得をしなくて済むという利点がある．自然言語という人間に扱いやすい手段を用いることで，ユーザは他人に対して頼みごとをするかのように，システムに要求を入力することができる．近年では，人手では労力がかかる画像の意味的な編集を自然言語で手軽に実現するという趣旨の研究も盛んに行われてきている [1], [6].

このような自然言語による意味的な画像編集タスクの難しい点は，ユーザからの要求に多様な表現が許されるという点である．システムは，内部の画像編集モデルによってユーザの意図に沿った画像編集を行うが，編集指示文がモデルにとって苦手なものであったり，モデルが想定していないような指示文には，意図しない編集結果をユーザに提示してしまうという問題がある．

この問題を解決する最も簡単な方法は，ユーザの編集指示に対して複数の編集結果の候補を提示し，ユーザに選択させることであろう．例えば，特定の編集指示に特化した複数のモデルによる編集結果が用意できれば，その中から最もユーザの意図に合った結果を選んでもらえば良い．しかし，編集指示が入るたびに毎回ユーザに確認するのはユーザの労力を増大させてしまうという問題がある．効率的なシステムには，どの画像が適切か選べる自信がない場合のみユーザに確認するという対話戦略が必要である．

本研究では特に，マスク機構を持つモデルに利用できる確認手法を提案する．マスク機構は，一般的な画像変換タスクで広く使われており，編集箇所をマスクが明示することで生成画像の品質を向上させるのに有用であることが知られている [3], [4]. 自然言語による画像編集タスクは，自然言語に条件づけられた画像変換タスクであり，マスク機構が重要な役割を果たしている．マスク機構は，多様な表現がなされる自然言語の編集指示文が，画像のどの部分の編集に対応しているかを明確にし，編集指示文に沿わない編集の発生を抑制することができる [6]．しかし，このマスク機構はモデルへの強い制約として働くため，アバターの顔画像編集において目や鼻などの小さな領域の編集精度を向上させた一方で，髪の毛を長くするなど，大きな領域の編集が難しくなってしまうという問題がある．したがって，編集指示文によっては，マスクなしのモデルの編集結果がマスクありのモデルの編集結果よりユーザの意図に沿った画像である可能性があり，システムはどちらのモデルによる編集結果がユーザにとってより望ましいかを適切に選択する必要が出てくる．

マスクはどの領域を編集すべきかモデルが出力する尤度だと解釈できる．そこで，我々はマスクのエントロピーに基づき，ユーザの編集指示に対して，エントロピーが小さい編集—すなわち，マスクありモデルで対応できる可能性が高い編集—については，確認せず編集結果をユーザに提示し，エントロピーが大きい編集—すなわち，マスクありモデルで対応できる可能性が低い編集—については，マス

[1] 奈良先端科学技術大学院大学
[2] 理研 AIP
[3] Institute for Creative Technologies, University of Southern California
[a)] shinagawa.seitaro.si8@is.naist.jp

クあり・なし両方のモデルの編集結果をユーザに提示して確認する対話戦略をとる枠組みを提案する．これにより，システムがユーザに効率的な画像編集プロセス（対話的画像編集）を提供できることを示す．

2. 対話的画像編集タスク

対話的画像編集タスクの概要を図 1 に示す．このタスクは，人間のユーザとシステムから構成されている．システムの目的は，対話を通じて，ユーザの目標画像 X^g を生成することである．タスクを通して X^g はシステムからは不可視であり，ユーザは自然言語で現在の画像を目標に近づけるよう編集指示文を入力し，システムは元画像と編集指示文に基づいて新しい画像を生成する． 具体的には，以下のステップに基づいて対話が進行する．

Step 1 ユーザに元画像 X_0^s と目標画像 X^g が与えられる．元画像はシステムと共有される．

Step 2 i ターン目において，ユーザは編集指示文 I_i を入力する．

Step 3 システムは，編集指示文 I_i と元画像 X_{i-1}^s に基づいて，新しい画像 X_i を生成する．

Step 4 システムは，X_i を新しい元画像 X_i^s として再設定し，ユーザは対話を継続するかどうかを選択する．継続する場合は，新しい編集指示文を入力し，次のターンに進む（$i \mathrel{+}= 1$ とし，Step 2 に進む）．対話を終了する場合，対話は終了し，元画像 X_i^s を目標画像 X^g と比較し評価する．

確認戦略を持つシステムは，Step 3 で { 確認，確認なし } の二つの行動の内一つを選択する．システムが確認を選択した場合は，Step 4 の前に，以下の確認手順のサブステップを挿入する．

Step 3-c1 システムは，ユーザに画像候補を提示し，どの画像がユーザの意図に沿うかを確認する．

Step 3-c2 ユーザは，目標画像を達成するために適切な画像を選択する．システムは，選択された画像をシステムの編集結果 X_i とする．

3. 条件付き DCGAN による画像編集モデル

Deep Convolutional Generative Adversarial Network (DCGAN) [5] は，画像生成に広く用いられているモデルであり，Generator (G) と Discriminator (D) を交互に学習する敵対的学習 [2] によってモデルを訓練することができる．これを自然言語による画像編集タスクに合わせて拡張した手法 [6] では，G の入力として元画像と編集指示文による条件付けを行うことで，目標の画像を生成することができる．このとき，元画像は CNN によるエンコーダ E_{im}，編集指示文は LSTM によるエンコーダ E_i により特徴抽出され，全結合層 FC によって統合された出力が G の条件付けとなる．また，マスクありのモデルでは，新たにマスク生成器 G_m とマスクされた元画像のエンコーダ E_{imm} が追加され，元画像と編集指示文を基に生成したマスクを元画像に重畳して特徴量抽出を行い、この出力を新たな G の生成条件として追加する．マスク機構は，編集すべき領域と元画像を保持すべき領域を明示的に分ける制約として働き，画像編集の性能向上に貢献する．しかし，このような強い制約は「髪を長くして」といった，大きな変化を要求する編集操作に対して悪影響を与えてしまう問題がある．マスクあり・なしのモデルはそれぞれ得意とする編集が異なることから，現実的には，ユーザの要求に合わせて適切に選択を行うという動機が生じる．本研究では，この選択をマスクのエントロピーに基づいて行う．

4. システムの確認戦略

システムが編集結果として複数の候補画像を持つとき，システムにとって最も安全な戦略は，毎回ユーザにどの画像が一番目的に適うか確認することである．しかし，この戦略はユーザに選択させる追加の行動を要求してしまうため，ユーザに負担のかかる戦略である．効率的な対話を行うには，システムが候補画像に対して，どの画像がユーザの意図に沿うか判断できない場合のみユーザに確認すれば良い．本研究では，マスクありのモデルによる画像編集時に生成されるマスクのエントロピーに基づいて確認を行う．

$$entropy = -\frac{1}{WH}\Sigma_i^W\Sigma_j^H\{m_{ij}\log(m_{ij}) + (1-m_{ij})\log(1-m_{ij})\} \leq -\log 0.5. \quad (1)$$

マスクの横，縦のサイズをそれぞれ W，H とする．ここで，m_{ij} はマスクにおける (i, j) 位置の画素値である．実験では確認の閾値を $-\alpha\log 0.5$ $(0 \leq \alpha \leq 1)$ とし，$entropy \geq -\alpha\log 0.5$ のとき，システムは確認を選択する．

5. 実験設定

本研究の対話的画像編集タスクにおけるモデルの学習およびシステムの評価には，Avatar Image Manipulation with an Instruction (AIMI) dataset [6] を用いた．AIMI データセットは，アバター作成サイトから収集したアバターの顔画像とクラウドソーシングで集めた人手による編集指示文のデータセットである．各サンプルは (元画像，目標画像，編集指示文) の組で構成される．編集指示文は 22 種類あり，髭や眉毛，髪の毛の変化などがある．編集指示文の語彙サイズは，単語分割する場合で 1,892 となる．訓練：開発：テストセットは先行研究 [6] と同様に，それぞれ 4,296：230：230 サンプルで分割して用いた．また，画像生成器を学習するにはサンプルサイズが十分でないため，組となってない単一の画像 161,065 サンプルを教師なし学習用のデータとして学習に利用した．

マスクあり・なしのモデルの学習時には，画像生成を学

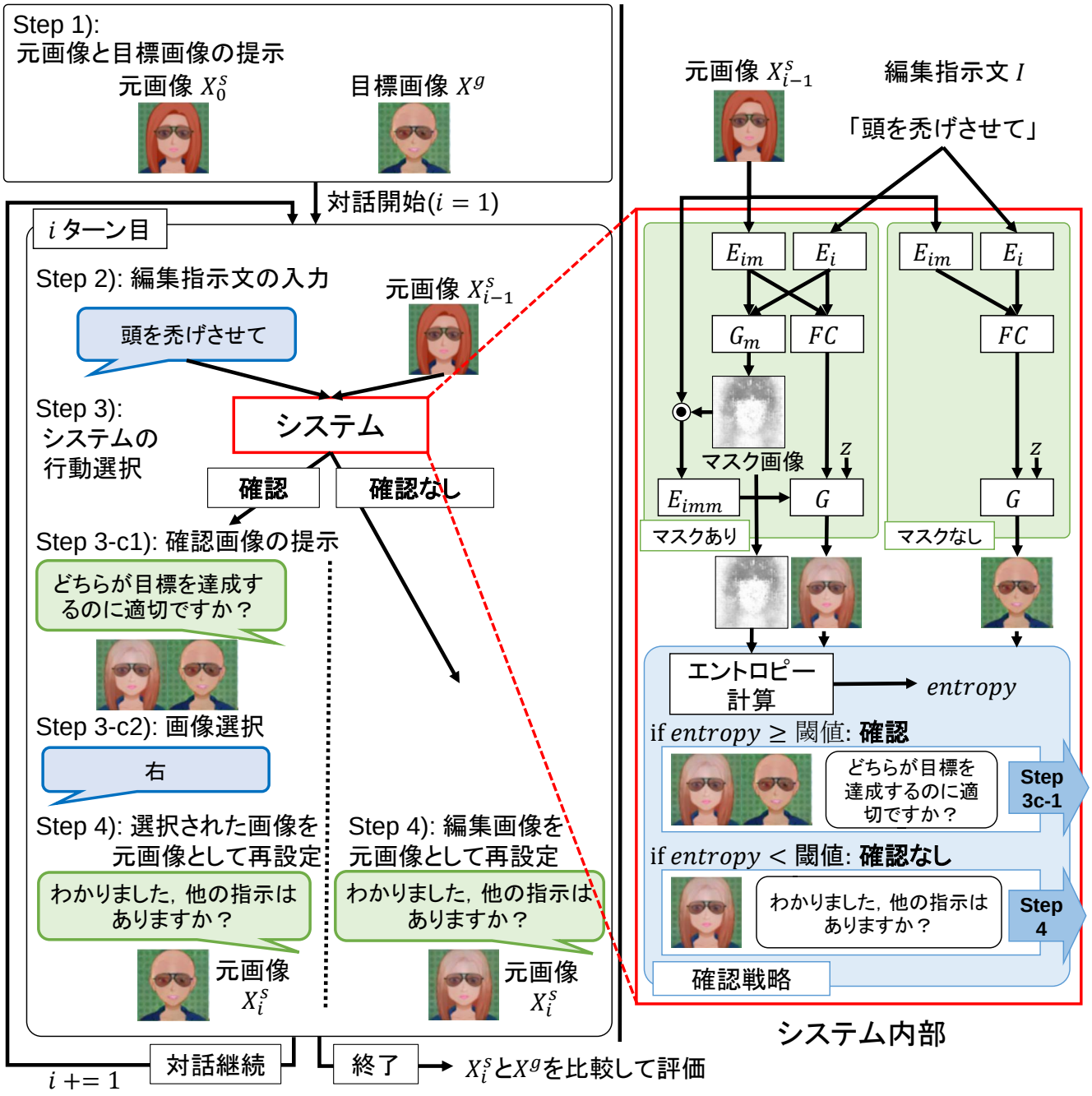

図 1　対話システムの全体像

習させるフェイズと画像編集を学習させる二つのフェイズを 2,200 サンプルごとに交互に行った．ここで前者のフェイズでは，同一の教師なし学習用の画像サンプルを元画像と目標画像両方に用いて，編集指示文の特徴ベクトルはゼロベクトルで固定した．目標画像は元画像と同一であることから，E_i，FC，G のみ学習した．後者のフェイズでは，(元画像，目標画像，編集指示文) の組のサンプルによって全てのモジュールの学習を行った．このとき，マスクありモデルではマスク生成器 G_m の学習を促進するため，マスクの教師データを元画像と目標画像の差分から作成した．具体的には，同じ位置のピクセル値が異なる場合マスクの値を 0，同じ場合 1 とした．マスクの損失関数には，生成されたマスクと教師のマスクとの間の mean squared error (MSE) 関数を用いた．

モデルの学習には最適化手法に $Adam$ ($\alpha = 2.0 \times 10^{-4}, \beta = 0.5$) を用いた．画像は 64×64 に大きさを変更し，ミニバッチサイズは 64 とし，5000 phase 学習させた．

6. 実験結果

効果的な対話戦略は，対話によって目標画像に近い画像を生成することだけでなく，ユーザの労力が小さい効率的な戦

略であることが求められる．これら 2 つの点を同時に評価するため，対話戦略の評価指標として $\Delta SSIM/\#user\ turn$ を用いた．ここで，$\Delta SSIM$ は目標画像と元画像の間の SSIM について，対話終了時の SSIM から対話開始時の SSIM を引いたものであり，$\#user\ turn$ は対話終了までのユーザの労力（編集指示文の入力回数と画像選択を行った回数の和）を表す．対話戦略を比較するため，Mann–Whitney U test を行い，$\alpha = 1.0$ またはランダムと，$\alpha = 0.0, 0.25, 0.50, 0.75$ を比較した結果，$p < 0.001$ の有意差が $\alpha = 0.0$ と $\alpha = 1.0$，$\alpha = 0.25$ と $\alpha = 1.0$, $\alpha = 0.50$ と $\alpha = 1.0$，そして $\alpha = 0.0$ と $random$ の間に認められた．この結果から，$\alpha = 0.0$，$\alpha = 0.25$ がより短い対話でより良い SSIM を達成していることが分かった．

しかし，$\alpha = 0.0$，$\alpha = 0.25$ はほぼ全ての場合に確認を行う戦略となっていた．ここで，$\alpha = 0.0$，$\alpha = 0.25$ と $\alpha = 0.50$ の間の $\Delta SSIM/\#user\ turn$ の有意差を調べたところ，有意差は確認できなかったが，$\#user\ turn$ のみを比較した場合，$p < 0.001$ で $\alpha = 0.50$ が $\alpha = 0.0$ よりも有意に短いことが分かった．このことから，SSIM では多少劣るものの，効率的な対話としては $\alpha = 0.50$ が有用であることが示唆された．

効率的な対話の成功例を次に示す（図 2）．最初ユーザは元画像の髪の毛をウェーブがかった髪に変更するように要求した．これは大きな変化を伴う編集であることから，マスクありのモデルが苦手とする操作である．システムは，エントロピーに基づき，マスクありモデルによる編集結果ではユーザの意図に沿わない可能性が高いと判断し，確認を選択することができた．ターン $i = 2$ では，システムが目標画像に近いポニーテール画像をマスクなしモデルから生成することに成功し，ユーザがこれを選択したが，このマスクなしモデルの編集結果は目の色が青から緑に変化してしまった．そこでユーザはこれを修正するため「目の色を青にして」と追加の編集指示文を入力し，システムはこれをマスクありモデルで対応できると判断して確認なしを選択した．この選択によって，ユーザに確認する手間を削減することに成功していることがわかる．

7. まとめ

本研究では，マスクのエントロピーに基づく確認対話戦略を持つ対話的画像編集システムを提案した．実験では，DCGAN を拡張したマスクあり画像編集モデルから得られるマスクのエントロピーを用いた．システムはユーザの編集指示に対して，マスクありモデルで対応できる可能性が高いと判断したときにはマスクありモデルで編集した画像をそのまま提示し，可能性が低いときのみマスクあり・なしのモデルによる編集画像を両方提示してユーザに確認することで，冗長な対話を抑制できた．今後の展望としては，ユーザとの対話データを収集して確認タイミングを強化学習し，より適応的な対話戦略を学習することが考えられる．

図 2　$\alpha = 0.50$ (閾値 $= 0.35$) と設定したときの対話例

参考文献

[1] Dong, H. et al, Semantic Image Synthesis via Adversarial Learning, Proc. ICCV, 2017.
[2] Goodfellow, I. et al, Proc. NIPS, pp. 2672–2680, 2014.
[3] Mejjati, Y. et al, Unsupervised attention-guided image-to-image translation, Proc. NeurIPS, pp. 3693–3703, 2018.
[4] Mo, S. et al,Instagan: Instance-aware image-to-image translation, Proc. ICLR, 2019.
[5] Radford, A. et al, Unsupervisedrepresentation learning with deep convolutional genera-tive adversarial networks, Proc. ICLR, 2016.
[6] Shinagawa, S. et al, Image manipulation system with naturallanguage instruction, IEICE Trans., Vol.J102-D, No.8, pp.514–529, 2019.