CNNによるスタイル変換とWeb画像を用いた画像の任意質感生成

松尾　真†　柳井　啓司†

† 電気通信大学大学院 情報理工学研究科　〒 1828585 東京都調布市調布ヶ丘 1-5-1
E-mail: †matsuo-s@mm.inf.uec.ac.jp, ††yanai@cs.uec.ac.jp

あらまし　実世界画像の質感の任意変換は質感ベースの画像データの増量やデザイン，エンターテインメントなどの様々な分野での応用が期待できる。本研究では，2015 年に Gatys らが考案した，物体の形状を精密に維持して画像のスタイルを変換する Neural Style Transfer アルゴリズムを実世界画像の質感の転写に応用し，コンテンツ画像と質感単語による画像内物体の任意質感変換を目指す。単語概念からの自動スタイル生成では，手動で行っていたスタイル画像の選択を自動化し，質感単語で収集した Web 画像から DCNN 中間層の出力と画像のスタイル認識のための特徴量 Neural Style Vector を用いてスタイルクラスタを構築し，スタイル表現を自動生成するシステムを構築する。評価は生成画像へのアンケートを用いて行った。その結果，視覚的概念への連想が容易であり，類似スタイル表現を多く持つクラスタが生成される単語では人間に評価されやすい画像が生成され，またそのクオリティはコンテンツのカラーやスタイルとの共通構造によって，大きく影響を受けることが分かった。

キーワード

1. はじめに

近年，2012 年の ImageNet Large Scale Visual Recognition Challenge(ILSVRC) のクラス分類タスクにおいて最も高い精度を挙げた Deep Convolutional Neural Network（DCNN）[3] がコンピュータビジョンの分野で注目されており, 様々な分野へと応用する動きが進められている。

2015 年，Gatys ら [1] によって，Deep Neural Network(DNN) を用いた，画像のスタイルを絵画のものに変換するアルゴリズムが考案された。これにより，従来の画像合成技術よりも物体の形状を精密に維持して画像のスタイルを変換することが可能となった。

本研究では，スタイル変換アルゴリズムを Flickr Material Database(FMD) やオノマトペを用いて Web から収集した実世界画像の質感の転写に応用し，画像内物体の任意質感変換を目指す。

実世界画像の質感概念をスタイルとして，画像内物体の質感の任意変換が可能となれば，画像に対する心象を意図的に変化させることができ，画像データの増量やデザイン，エンターテインメントなどの様々な分野での応用が期待できる。

しかし，このアルゴリズムには目的のスタイルが画像の形式で存在する必要があるため，ある質感を与えたいときにはその質感を持つ画像をマニュアルで用意する必要がある。この問題の解決のため，画像のスタイル認識のために考案した特徴量 Neural Style Vector [2] を用いて，質感単語で収集した Web 画像から有用なものを自動選択し、スタイル表現を生成するシステムを構築する。

2. 関連研究

2.1 Neural Style Transfer

Deep Neural Network を用いたスタイル変換の関連研究は Gatys ら [1] の研究が挙げられる。この研究では，Imagenet1000 クラスを学習済みの DNN の中間層のフィルタ出力を使用することで，スタイル画像のスタイルをコンテンツ画像に転写しており，画像のクラス分類が中心であった DNN の研究分野に衝撃を与えた。

本研究では Gatys ら [1] の手法を用いて画像を合成することで，画像の質感の変換を行う。変換させる画像をコンテンツ画像 x_c，スタイル画像を x_s，合成結果画像を x_g とする。x_c，x_s，x_g のコンテンツ表現とスタイル表現を CNN の特定の layer の活性値から求め，x_g のコンテンツ表現が x_c に，スタイル表現が x_s に近くなるように反復的に合成する。

使用した CNN は VGG19 [3] であり，コンテンツ表現に使用する layer は conv4_2，スタイル表現に使用する layer は conv1_2，conv2_2，conv3_4，conv4_4，conv5_4 である。図 1 にスタイル変換アルゴリズムの概略を記す。

layer l におけるコンテンツ表現はパラメータ数 N_l の活性値行列 $F_l(x)$，その損失関数は x_c と x_g の差であり，式 1 で表される。

$$L_c(x_c, x_g) = \frac{1}{2}\sum_{i,\,j}(F_{l,\,i,\,j}(x_c) - F_{l,\,i,\,j}(x_g))^2 \tag{1}$$

layer l におけるスタイル表現は活性値行列の式 2 で表される相関行列 $G(x, l)$，その損失関数は x_s と x_g の差であり，式 3 で表される。使用する layer 全体の誤差は重み w_l を用いて式 4 で表される。

$$G_l(x) = F_l(x)F^T(x, l) \tag{2}$$

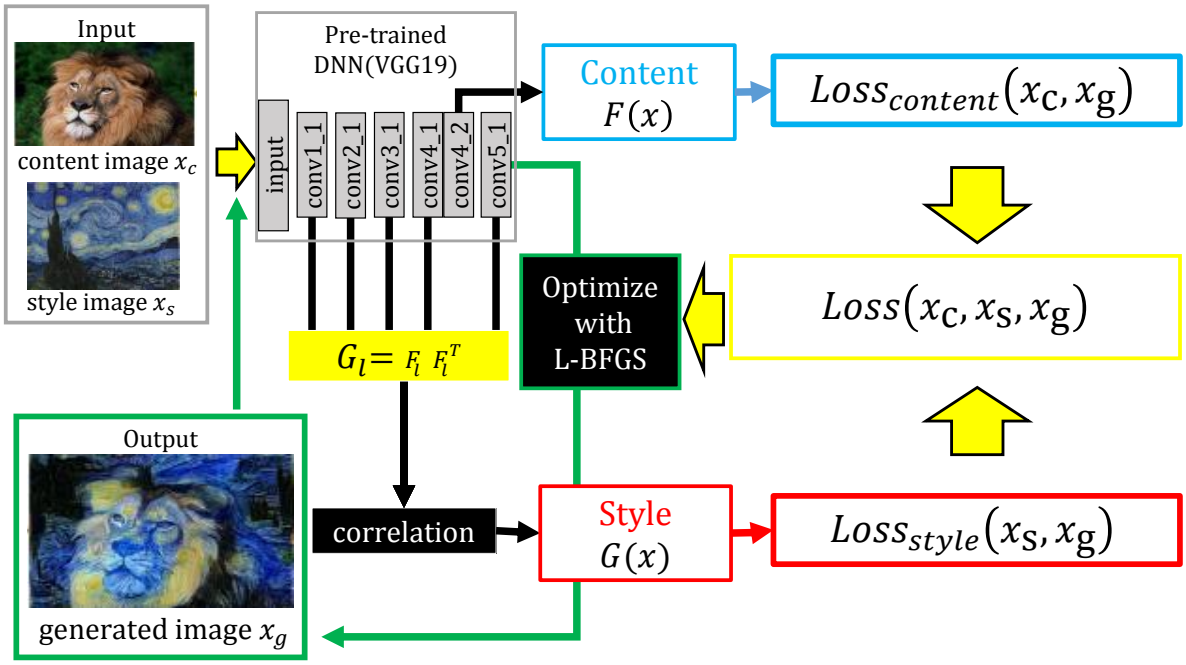

図 1　スタイル変換アルゴリズム

$$Loss_{s,\,l}(x_s,\,x_g) = \frac{1}{4N_l^2}\sum_{i,\,j}(G_{l,\,i,\,j}(x_s) - G_{l,\,i,\,j}(x_g))^2 \tag{3}$$

$$Loss_s(x_s,\,x_g) = \sum_l w_l Loss_{s,\,l}(x_s,\,x_g) \tag{4}$$

全体のエラー関数は重み w_c, w_s を用いて式 5 で表される。この式の値が最小となるように x_g を L-BFGS 法を用いて最適化する。

$$Loss(x_c,\,x_s,\,x_g) = w_c Loss_c + w_s Loss_s \tag{5}$$

現在このアルゴリズムには様々な改良が考案されている。Li ら [4] は Markov Random Fields(MRF) と CNN の組み合わせによって実世界画像の転写により適したスタイル変換を行った。Novak ら [5] は Gram Matrix のシフトや layer 間の Gram Matrix の操作によってスタイル変換の質を向上させた。Berger ら [6] は特徴マップの空間転移により，スタイルを広範囲に保持したスタイル変換や Inpainting を行った。

また，Feed forward network を用いてスタイルを学習することによる高速化 (Fast style transfer) が Johnson ら [7] や Ulyanov ら [8] により考案された。さらに単一の Feed forward network で複数のスタイルを学習したマルチスタイルの Fast style transfer が Dumoulin ら [9] によって考案された。

2.2　画像のスタイル認識

画像のスタイルを分析する研究は数は少ないが進められている。Datta ら [10] は画像の彩度, 浸透性, 被写界深度を元に aestheic rating を計測した。Marchesotti ら [11] は Aesthetic Visual Analysis (AVA) dataset [12] のユーザーのコメントを特徴量化し，定量化した美的評価を求めた。Keren ら [13] は絵画データに付与されたテキストを元に特徴を生成し, Pollock と Dali の絵画をクラス分類した。最新の技術を用いて画像のスタイルを分類した研究として, Karayev ら [14] の研究が挙げられる。Karayev らは Flickr Style, Wikipaintings, AVA Style から得られたスタイルメタデータ付き画像データセットに対して，それぞれのスタイルによるクラス分類を学習済みの DCNN 中間層から得られた特徴量を用いて行った。

2.3　bilinear CNN を用いたテクスチャ認識とクラスによるスタイル変換

Lin ら [15] は Neural Style Vector と同じく，Style Matrix をもとにした，bilinear CNN 特徴を用いて，FMD などのテクスチャデータセットにおけるテクスチャ認識を行い，Fisher Vector-CNN を上回る精度を挙げた。

また，bilinear CNN 特徴を用いて学習したクラス分類器の出力を用いて，新たな損失関数を設定し，スタイル変換アルゴリズムに組み込むことで学習済みの属性を持つテクスチャの外見にスタイル変換できるように改良した。

本研究では，学習済みのテクスチャに限らず，任意の質感への変換を目指す。

3.　提案手法

スタイルの選択を自動化することで，コンテンツ画像と質感単語を入力とした任意質感生成を目指す。そのために，質感単語から収集された Web 画像から DCNN 特徴と Neural Style Vector を用いて有用なスタイル表現の抽出を行い，そのスタイル表現によるスタイル変換の結果を評価する。収集した画像から，以下の流れで N 個のスタイル表現を自動生成し，それぞれをスタイルとしてスタイル変換を行う。

質感単語をキーワードとした画像収集には Bing API を使用する。

（1）　コンテンツ画像と質感単語を入力

（2） 質感単語をキーワードとして Web 画像を収集

（3） 類似スタイルを持つ画像同士にクラスタリング (クラスタ数 $N = 10$)

（4） 各クラスタの上位 K 枚の画像を用いてスタイル変換

3.1 スタイル画像のクラスタリング

類似スタイルを持つ画像同士が集中したクラスタをスタイル表現として構成するために，以下の流れで Web 画像に対してクラスタリングを行い，スタイルの類似する画像同士を集め，スタイルクラスタを構築する。これらを通じて図 2 のように複数のスタイル表現が構成される。実験では $N = 10$ 個のスタイルクラスタを構成し，DCNN 特徴によるクラスタリング時に式 6 により各クラスタにスコアを設定し，上位 3 個のクラスタを実験では使用した。

（1） 画像を $W_r \times W_r$ にリサイズ

（2） 全ての画像から DCNN 特徴を抽出

（3） 不適切な画像を除去

（4） 残った画像から $W_p \times W_p$ のパッチを切り出し

（5） 全てのパッチから Neural Style Vector を抽出

（6） DCNN 特徴を用いて画像をクラスタリング (クラスタ数 $N = 10$)

（7） Neural Style Vector を用いて各クラスタ内の画像をリランキング

$$score(C) = \frac{N(C)}{\sum_i Dist(x_i, C) + \lambda} \tag{6}$$

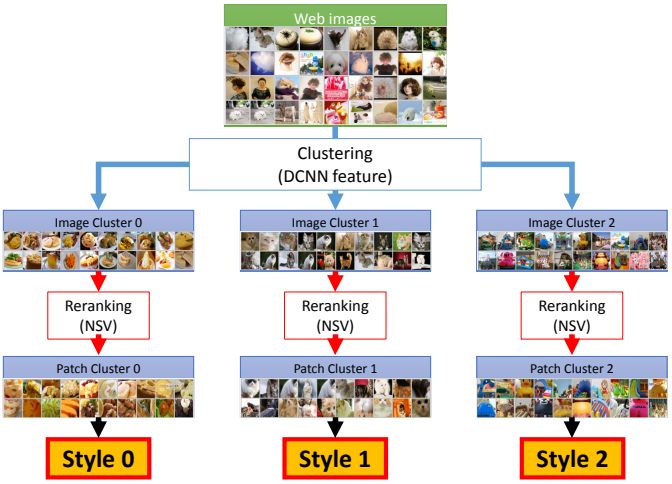

図 2　Web 画像からのスタイルクラスタの構築

収集された Web 画像は $W_r \times W_r$ にリサイズする。実験では $W_r = 1024$ とした。さらに Neural Style Vector を抽出する際は画像中央から $W_p \times W_p$ のパッチを切り出し，使用する。実験では $W_p = 512$ とした。

図 3 のように，収集した Web 画像には低解像度な画像や人間単体などの明らかにスタイル表現として不適切な画像が含まれているため，これらを除外した。低解像度画像はリサイズ前の長辺が $W_r/2$ 以下のものとし，人間単体の画像はノイズ画像を含んだ Web 画像をクラスタリングして得られた，1000 枚の顔画像をポジティブ，同数のランダム画像をネガティブとして学習した SVM を使用して除外した。図 3 から不要な画像を除外した結果が図 4 である。

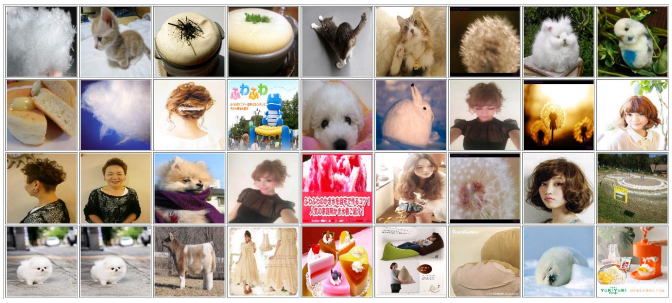

図 3　収集された Web 画像

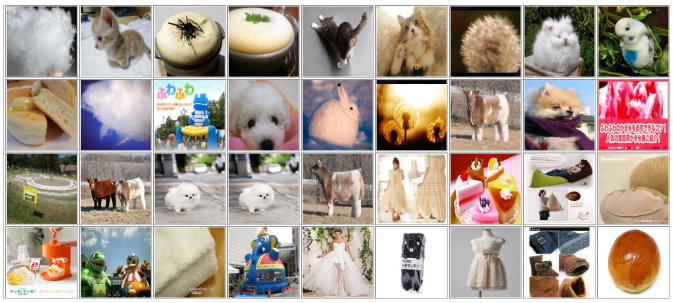

図 4　ノイズ画像の自動除去処理後の Web 画像

同じ質感単語の要素を持つ画像であっても，視覚的な表現が一致しているとは限らない。例えば，図 4 のように「ふわふわ」という質感単語で集まる画像には動物，風船，ケーキなどの様々な外見の物体カテゴリが含まれている。全く異なる物体間で類似したスタイルを持つ例は少ないため，まず DCNN 特徴を用いて k-means クラスタリングを行うことで，図 5 のように物体カテゴリ別のクラスタに分離させる。

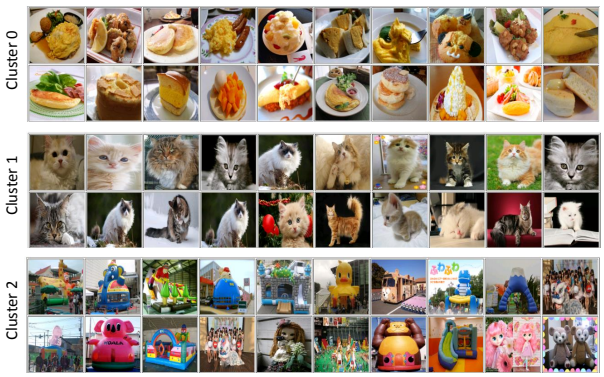

図 5　DCNN 特徴によりクラスタリング

各クラスタからのスタイル表現の導出のため，Neural Style Vector [2] によるクラスタ内のリランキングを行い，各クラスタの上位 K 枚を選出する。リランキングにはまずクラスタ内の画像をさらにクラスタリングし，小クラスタ C に含まれる画像数 $Num(C)$，小クラスタ内の画像のクラスタ中心との距離 $Dist(x_i, C)$ の総和から，式 6 のようにスコアを与え小クラスタをリランキングし，さらに小クラスタ内の画像をクラスタ中心との距離が小さい順にリランキングする。

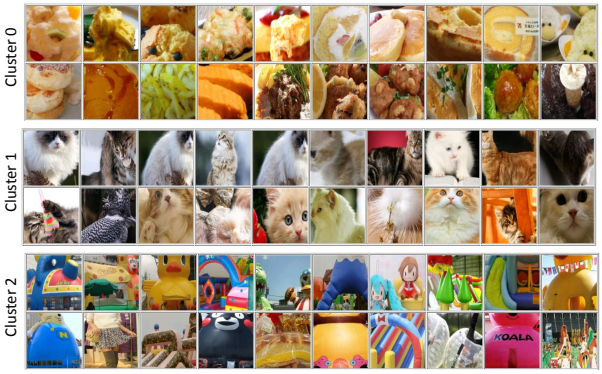

図 6　Neural Style Vector によるリランキング

3.2　スタイル表現の統合

図 7 は K 枚の異なるスタイル画像を用いる場合のスタイル表現の統合方法として，各画像の Gram Matrix の値の最大値 (max)，平均値 (mean)，幾多ら [16] の用いたコンテンツ画像に基づく最適化 (minimize) およびスタイル画像の結合 (connect) を使用した結果である。max，mean ではそれぞれ Gram Matrix を式 7，8 の G' に置き換えて使用した。

$$G'_{l,i,j} = \arg\max_{1 \leqq k \leqq K} G_{l,i,j}(x_k) \tag{7}$$

$$G'_{l,i,j} = \frac{1}{K} \sum_{k=1}^{K} G_{l,i,j}(x_k) \tag{8}$$

minimize では式 9 で表されるコンテンツ画像の Gram Matrix との 2 乗差を最小化する各画像の重みベクトル r を定義し，Gram Matrix を式 10 の G' に置き換えて使用した。

$$\arg\min_{r}(G_{l,i,j}(x_c) - \sum_{k=1}^{K} r_k G_{l,i,j}(x_{s,k}))^2 \tag{9}$$

$$s.t. \sum_{k=1}^{K} r_k = 1, 0 \leqq r_k \leqq 1$$

$$G'_{l,i,j} = \sum_{k=1}^{K} r_k G_{l,i,j}(x_k) \tag{10}$$

connect では図 8 のようにスタイル画像を空間的に結合した上で，境界線をアルファブレンドした一枚の画像をスタイル画像として使用した。

実験では最もスタイルの内容が色濃く反映された connect の手法でスタイル画像の統合を行った。

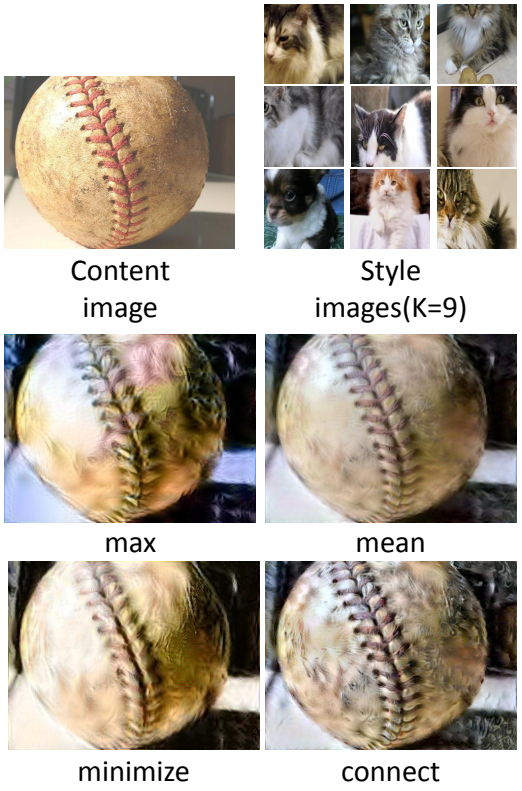

図 7　スタイル画像の統合手法（max:各画像の最大値，mean：各画像の平均値，minimize:コンテンツ画像に基づく最適化 [16]，connect：スタイル画像の結合）

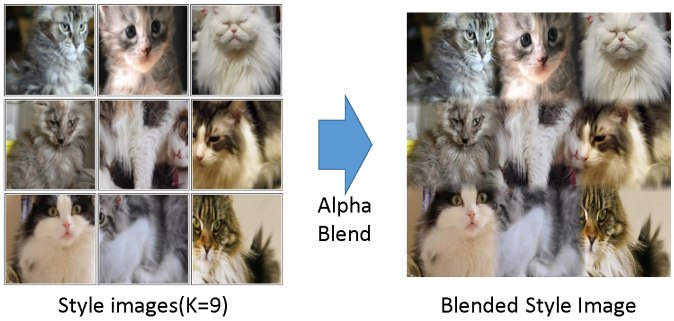

図 8　スタイル画像の統合

4.　実験

実験では FMD および質感単語で構築したスタイルクラスタを用いてスタイル変換を行い，ユーザー評価を行う。使用するスタイルクラスタは各単語の上位 3 つを選出し，クラスタ上位 9 枚を使用してスタイル変換画像を生成する。

4.1　実験データ

コンテンツ画像は図 9 の 3 種類の画像を使用した。質感単語はオノマトペ 23 単語を使用した。

4.2　ユーザー評価

ユーザー評価では図 10 のように，質感単語と正解の画像 3 枚，ランダムに選出した同コンテンツの不正解の画像 12 枚をブラウザに表示し，ユーザーが最もその単語に当てはまると感じた画像を選択し，その正解の可否で評価する。問題数はコン

図 9　コンテンツ画像（左から content0，1，2 とする）

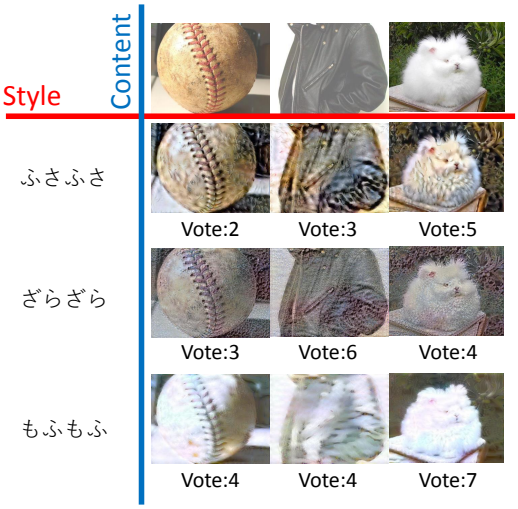

図 11　正解率の高い質感変換結果例（vote：正解者の数）

テンツ数 3，単語数 23 語の 69 問，回答したユーザー数は 9 人だった。

4.3　評価結果

各単語のユーザー評価の結果は図 13 のようになった。正解率が高かった単語は「ふさふさ」,「ざらざら」,「もふもふ」だった。また，コンテンツ別の評価結果はそれぞれ図 14，図 15，図 16 のようになった。

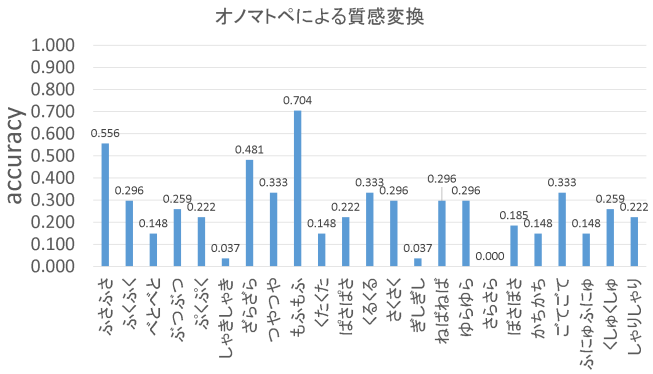

図 13　ユーザー評価の結果

各単語のユーザー評価の結果は図 13 のようになった。正解率が高かった単語は「ふさふさ」,「ざらざら」,「もふもふ」だった。正解率に繋がる要因として考えられるのは,「類似するスタイルクラスタの構成の成功」,「スタイル変換のクオリティ」,「質感単語の視覚的イメージのしやすさ」であると考えられる。

これらの単語のクラスタの内，最も選択されることが多かったものが図 11 であり，これらのスタイルクラスタは図 12 のよ

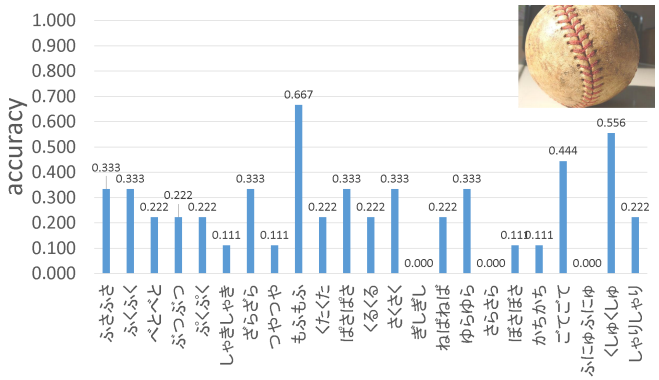

図 14　ユーザー評価の結果 (content0)

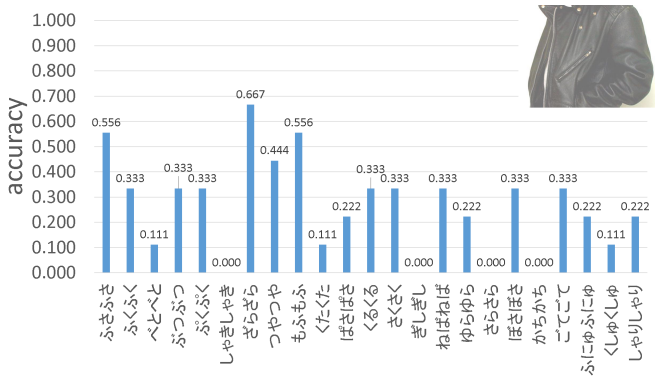

図 15　ユーザー評価の結果 (content1)

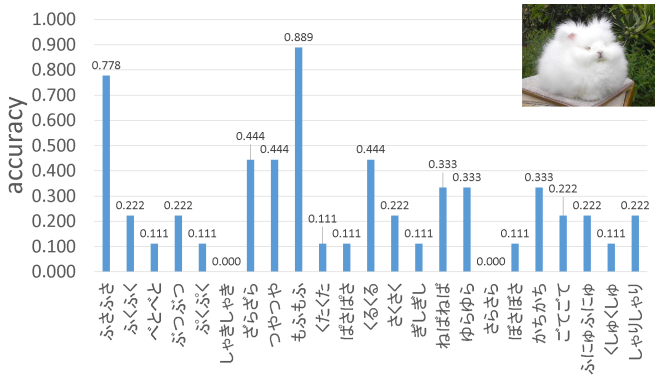

図 16　ユーザー評価の結果 (content2)

うに，全て類似したスタイルが上位に集中しており，スタイル画像の収集が成功した例だと言える。

また，コンテンツ別の評価結果はそれぞれ図 14，図 15，図 16 のようになった。これらを比較すると質感単語の正解率はコンテンツによって大きく異なることが分かる。

図 17，18，図 19，20，図 21，22，図 23，24 はそれぞれ「ふさふさ」,「ざらざら」,「もふもふ」,「くしゅくしゅ」から生成されたスタイルクラスタおよび変換画像とアンケートにおいての正解者の数である。

「ふさふさ」,「くしゅくしゅ」のスタイルクラスタでは全てのクラスタで似たスタイルのものが集まり，正解数にクラスタ毎の差はあまり見られなかった。それに対し,「ざらざら」,「もふもふ」のスタイルクラスタは 3 つのクラスタの内，1 つのク

図 10　ユーザー評価実験（赤枠が正解）

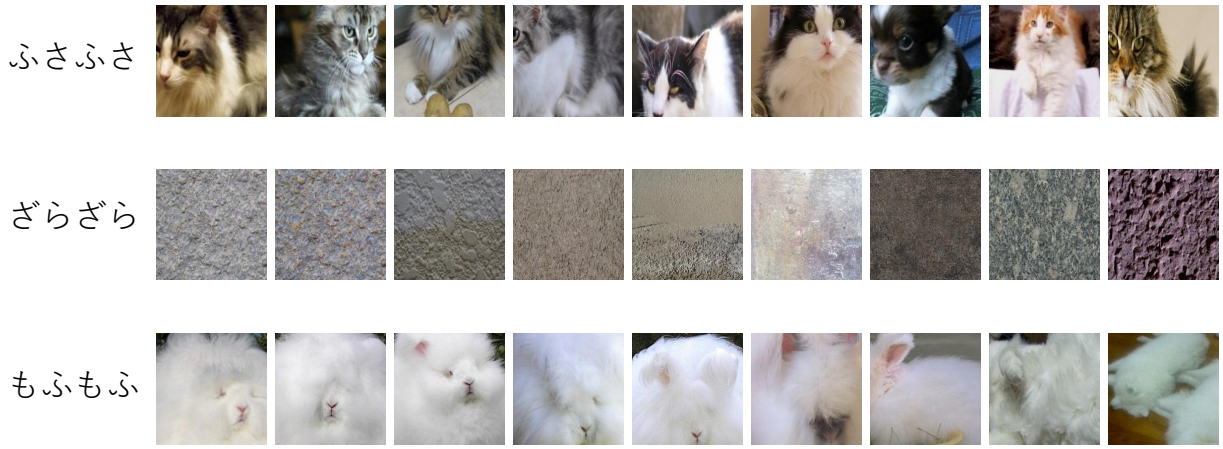

図 12　正解率の高いスタイルクラスタ

ラスタの正解率だけが非常に高くなっていた。「ざらざら」のスタイルクラスタは Cluster0 はパッチ状のスタイルとして有用性の高い画像があつまっており，それ以外のクラスタは類似した画像があまり集まっていなかったためだと思われる。「もふもふ」の Cluster0 は類似したスタイルが上位に集まってはいるが，アニメの画像など，スタイルとして好ましくない画像が多く，Cluster1 は「ふさふさ」などの他のオノマトペで収集された画像に似ていることから，選択されることが少なかったのではないかと思われる。このクラスタの順序と正解率の違いからクラスタのスコア（式 6）は改良の余地があると考えられる。

また，これらの結果から，スタイル変換のクオリティについてはコンテンツ画像とスタイルの相性に強く依存していると考えられる。コンテンツが基調とするカラーによって背景と物体のどちらが大きく変化するかが変わっており，後者のコンテンツの正解率が高くなっている。例えば，図 12 は図 11 の元となったスタイルクラスタであり，「ざらざら」は黒を，「もふもふ」は白を基調とした画像が集まっており，それぞれ同じ色を基調としたコンテンツを変換した画像（「ざらざら」と content1，「もふもふ」と content2）の正解率が高くなっている。

また，コンテンツとスタイルの間の共通構造（しわ，毛並みなど）の有無によって生じるスタイル変換のクオリティの違いも影響していると思われる。例えば，同じく白を基調としたコンテンツであるにも関わらず，content0 と content2 の「ふさふさ」と「もふもふ」の正解率には大きな差がある。これらの単語では content0 では新たに毛の質感を追加しているのに対し，content2 では元より持っている質感を強調される形で変換されており，自然な変換となっている。

したがって，コンテンツの持つ特徴を考慮したスタイルクラスタ構成により効果的なスタイル生成が可能になると思われる。

図 17 「ふさふさ」スタイルクラスタ

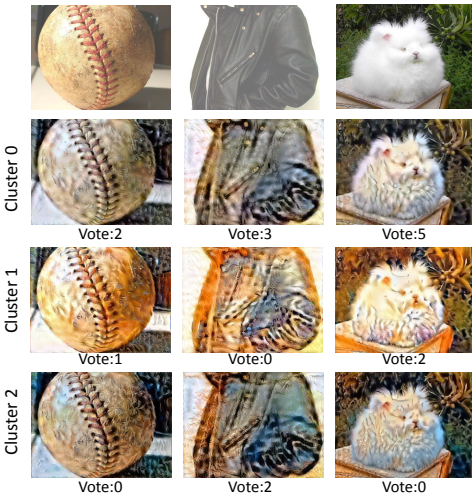

図 18 「ふさふさ」による質感変換結果

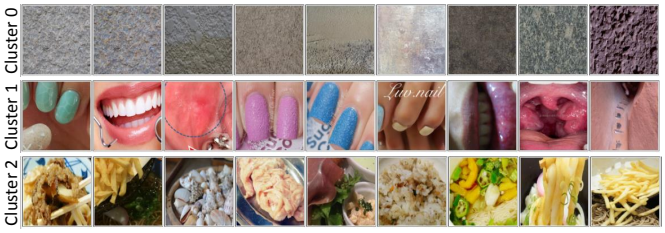

図 19 「ざらざら」スタイルクラスタ

5. おわりに

Neural Style Transfer アルゴリズムを用いた画像の任意質感生成を行うために，単語概念からスタイル表現を自動生成するシステムを構築し，複数の生成画像の評価をアンケートを用いて行った。その結果，視覚的概念への連想が容易であり，類似スタイル表現を多く持つクラスタが生成される単語では人間に評価されやすい画像が生成され，またそのクオリティはコンテンツのカラーやスタイルとの共通構造によって，大きく影響を受けることが分かった。

スタイル変換アルゴリズムの傾向から，変換に使用するスタイル画像はきめ細かな構造が多く，コンテンツ画像に類似する構造が有り，変換したい対象の領域のカラー傾向が共通していることが望ましいことが分かった。そこで，コンテンツのスタイル表現に依存したスタイルクラスタの構築や領域分割とスタイル画像の色変換を用いて対象領域のカラーをコンテンツに近づけることでより自然な質感変換が可能になると考えられる。

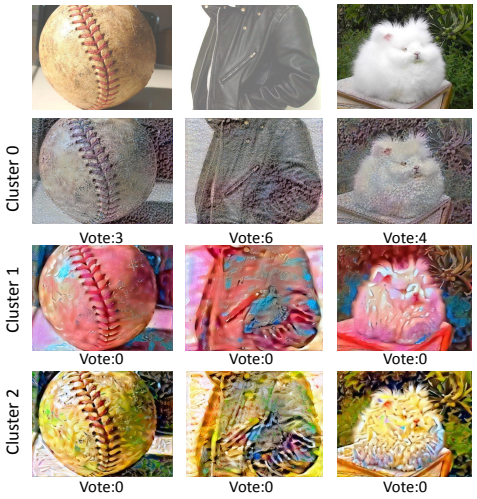

図 20 「ざらざら」による質感変換結果

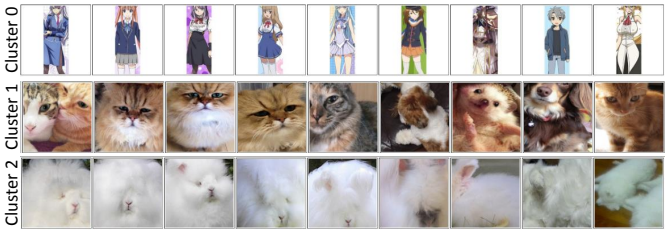

図 21 「もふもふ」スタイルクラスタ

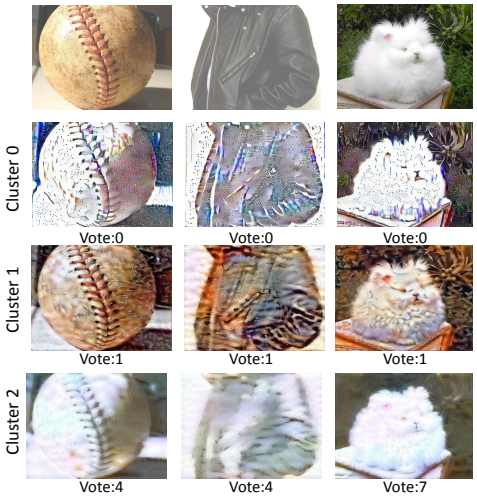

図 22 「もふもふ」による質感変換結果

文　献

[1] L. A. Gatys, A. S. Ecker, and M. Bethge. Image style transfer using convolutional neural networks. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2016.

[2] S. Matsuo and K. Yanai. Cnn-based style vector for style image retrieval,. In *Proc. of ACM International Conference on Multimedia Retrieval*, 2016.

[3] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *Proc. of arXiv:1409.1556*, 2014.

[4] C. Li and M. Wand. Combining markov random fields and convolutional neural networks for image synthesis. In *Proc. of arXiv:1601.04589*, 2015.

[5] R. Novak and Y. Nikulin. Improving the neural algorithm

図 23 「くしゅくしゅ」スタイルクラスタ

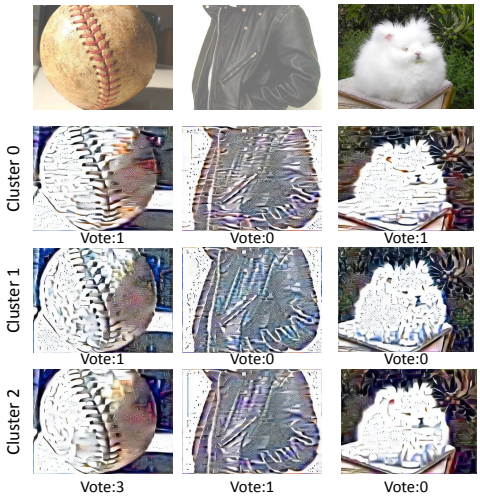

図 24 「くしゅくしゅ」による質感変換結果

of artistic style. In *Proc. of arXiv:1605.04603v1*, 2016.

[6] G. Berger and R. Memisevic. Incorporating long-range consistency in CNN-based texture generation. In *Proc. of arXiv:1606.01286v1*, 2016.

[7] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. of arXiv:1603.08155*, 2016.

[8] D. Ulyanov, V. Lebedev, A. Vedaldi, and V. Lempitsky. Texture networks: Feed-forward synthesis of textures and stylized images. In *Proc. of arXiv:1603.03417v1*, 2016.

[9] V. Dumoulin, J. Shlens, and M. Kudlur. A learned representation for artistic style. In *Proc. of arXiv:1610.07629v1*, 2016.

[10] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Studying aesthetics in photographic images using a computational approach. In *Proc. of European Conference on Computer Vision*, 2006.

[11] L. Marchesotti and F. Perronnin. Learning beautiful (and ugly) attributes. In *Proc. of British Machine Vision Conference*, 2013.

[12] Naila. Murray, De. Barcelona, L. Marchesotti, and F. Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2012.

[13] Keren. D. Painter identification using local features and naive bayes. In *Proc. of International Conference on Pattern Recognition*, 2012.

[14] S. Karayev, M. Trentacoste, H. Han, A. Agarwala, A. Darrell, T. Hertzmann, and H. Winnemoeller. Recognizing image style. In *Proc. of British Machine Vision Conference*, 2013.

[15] T. Y. Lin and S. Maji. Visualizing and understanding deep texture representations. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2016.

[16] 幾田光, 大垣慶介, 小田桐優理. 畳み込みニューラルネットワークを用いたテクスチャ特徴量の混合に基づく自然なテクスチャ転写. 第 19 回 画像の認識・理解シンポジウム (MIRU 2016), 2016.