社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

DCNN特徴を用いたWebからの質感画像の収集と分析

下田　　和†　　柳井　啓司†

† 電気通信大学 情報理工学部 総合情報学科

あらまし　近年、大規模物体認識のためのデータセットで学習した Deep Convolutional Neural Network (DCNN) の活性化信号を bag-of-features や Fisher Vector などの代わりに画像特徴ベクトルとして用いることが物体認識やシーン認識、属性認識など様々な認識に対して有効であることが示されている。本研究では、DCNN 特徴を利用した画像認識技術を用いて、質感を表現する言葉に対応した画像の認識可能性について、Web から収集した画像を用いて分析を行う。特に、物体の質感や状態に関する直感的な印象を表す擬音語 (オノマトペ) に対応した画像の認識可能性について分析を行う。実験では、オノマトペ単体と、オノマトペを含む形容詞と名詞の組み合わせについて対応する画像を Web から収集し、認識可能性について分析を行った。

キーワード　質感画像認識, DCNN 特徴, オノマトペ, Web 画像

Gathering and Analyzing Material Images on the Web with DCNN features

Wataru SHIMODA† and Keiji YANAI†

† Department of Informatics, The University of Electro-Communications, Tokyo

1. はじめに

画像を見たときに物体そのものよりも、ものの状態や質感や雰囲気など、直感的な印象のほうが目に入り、それに関係する言葉が頭に浮かぶことがある。人がそれを認識し、その印象を言葉にするのは簡単だが、コンピュータがそのような意味的に定義の曖昧なものの認識は苦手であることが知られている。しかしながら、近年画像認識の精度が向上し、動物画像、食事画像のような一般的な物体の認識だけでなく、物体の画像に写っている素材がなんであるかを認識する素材画像認識など、物体の形に依存しない画像の認識も行われるようになってきた。これまでは、物体の形に依存する局所特徴量による認識が主流であったが、近年は Deep Learning のような、より人間の脳に近い方式に基づく学習から特徴量を抽出する技術も注目されるようになった。このような最新の画像認識技術により、ものの状態や質感や雰囲気など、直感的な印象を表す言葉を画像から認識できる可能性がある。

日本には直感的にものの様子を表現する方法としてオノマトペがある。英語のオノマトペは、tic tac や quock など音源が主であるが、日本のオノマトペは物体の手触りや物体の形状や物体の食感などを表す際にも使われることがあり、他の言語に比べて豊富に存在する。

本研究ではこのオノマトペで画像を収集し、認識を行うことで、直感的な印象を表す言葉の画像からの認識可能性を確かめることを目的としている。また、オノマトペが表現するものは物体に限らず、物体の手触りや、食感など、多彩であるので、これらを識別することは、現在の画像認識手法がどのような画像に対して有効なのかを確かめることにも繋がる。

また、本研究ではオノマトペ画像の識別のみでなく、名詞とオノマトペのペアで画像を収集しその認識可能性を確かめた。1 つの名詞につき複数のオノマトペのペアで、名詞＋オノマトペの画像を収集する。オノマトペのみ用いて画像を収集すると、特定の物体の画像に偏って収集されることがあり、同じ名詞＋異なるオノマトペの組み合せに対応した画像を収集し、画像分類実験を行うことで、オノマトペ以外の物体カテゴリに関わる要素が認識に影響を与える可能性を抑えることができる。名詞＋オノマトペの認識結果は、オノマトペ単体の認識結果と比べて、よりそのオノマトペの視覚性を反映していると考えられる。

2. 関連研究

関連研究としては、素材画像の認識がある。オノマトペは、物体の形状、物体に触れたときの感覚、発生する音などを表現する際に用いられる。このようなオノマトペで集めた画像の認識は、一般的な物体認識よりも素材画像の認識との関係の方が大きいだろうと推測できる。

素材画像の研究としては、Liu らによる Flickr Material Database (FMD) [1] の認識が代表的である。FMD は繊維, ガラス、金属、プラスチック、水、葉、革、紙、石、木の 10 種類の素材カテゴリからなるデータセットである。FMD の分類にはどのような特徴量が有効であるかといったような研究が多くされてきたが、現在は Improved Fisher Vector (IFV) [2] と、

大規模物体認識のためのデータセットで学習した Deep Convolutional Neural Network (DCNN) の活性化信号を画像特徴ベクトルとして用いた Deep Convolutional Neural Network 特徴 (DCNN features) [3] が有効であることがわかっている。Cimpoi ら [4] は IFV と DCNN features を組み合わせることで 10 クラス素材画像分類で 67.1%を達成している。

また、Cimpoi ら FMD とは異なる Describable Textures Dataset(DTD) データセットを作った。DTD は 47 の画像の属性カテゴリからなる大規模なデータセットである。DTD の認識においても IFV と DCNN が有効であり、DTD の認識結果は FMD の認識の精度を向上させる助けにもなった。

素材画像の認識以外に、属性 (attribute) の認識も本研究と関連が深い。尾関らは属性を用いた認識のための AwA (Animals with Attributes) データセットにおける DCNN 特徴の有効性を示している [5]。しかしながら、AwA は対象が動物に限定されており、様々な対象の属性に関して DCNN の有効性を確認しているわけではない。

本研究ではこれらの先行研究を参考にして、IFV と DCNN を用いて質感画像の認識を行う。オノマトペ単体の認識可能性だけでなく、特定の物体に限定してオノマトペが分類可能であるかについても分析を行う。

なお、本研究では、主としてオノマトペ画像の認識を行うが、現在著名なオノマトペ画像データセットは存在しない。そこで、独自にオノマトペデータセットを用意する必要がある。FMD、DTD においては、データセットを構築するうえで、カメラでの撮影などは行わず、Web 画像をマイニングする方法が用いられている。Web 画像のマイニングによるデータセットを構築する際には、ノイズ画像の除去方法が常に問題になるが、これらの研究ではクラウドソーシングによってノイズ画像を除去している。クラウドソーシングはネットを利用して複数人の人間の手によって画像をアノテーションする方法であり、Amazon Mechanikal Talk が一般に使われている。クラウドソーシングは人手によるものなので、高精度なデータセットを構築できる。しかし、クラウドソーシングは、利用にコストがかかる。また、ワーカーの多くは日本人でないため日本のオノマトペ画像のアノテーションは困難であることが想定され、本研究には向いていない。そこで、本研究では、ノイズ除去を機械的に行う。画像認識を利用したリランキングによって、人手を介さずにデータセットを構築する。

3. 質感画像の収集と認識可能性の評価

本節では、オノマトペ単体に対応する画像の収集・分析について、手順を説明する。まず、クエリとしてオノマトペを用いて Web 画像を収集し、自動でノイズ除去を行うことで、オノマトペ画像データセットを構築する。次に、このオノマトペ画像データセットについて、その認識率を評価する。

3.1 オノマトペ画像の収集

オノマトペ画像データセットの構築には Web 画像検索エンジンの WebAPI を用いる。Web 画像はオノマトペをクエリとして、Bing API を利用して収集する。ネットを利用する多くの人がオノマトペに相当していると考える画像が検索結果の上位に表れるはずである。しかし、意図していない画像がノイズとして混じる可能性が考えられるのでノイズ除去を行う必要がある。上位の画像はよりクエリとの関係が強いと考え、これを疑似ポジティブ画像として自動でノイズ除去を行う。

3.2 画像のリランキング

本研究では、ノイズ除去や、認識可能性の評価に画像認識によるリランキングを利用している。このリランキングは人手を用いずにすべて自動で行う。以下の手順でリランキングを行う。

まず、各画像から特徴量を取り出し、ランキングの上位の画像を疑似ポジティブ画像として SVM を学習する。このモデルをランキングの各画像に適用すると、SVM の出力値が得られる。この出力値の昇順に各画像をランキングし、リランキング結果とする。

3.3 データセット構築手順

データセットは以下の手順で構築する。リランキングは一度では不十分なので二度行う。

（1） Web 画像を BingAPI により 1000 枚収集する
（2） 収集した Web 画像 1000 枚の特徴量を抽出する
（3） 収集画像の上位 10 枚で SVM を学習し、学習した SVM で収集した 1000 枚の画像をリランキング
（4） リランキング結果の上位 20 枚で SVM を学習し、学習した SVM で収集した 1000 枚の画像を再びリランキング
（5） リランキング結果の上位 50 枚を質感画像データセットとする

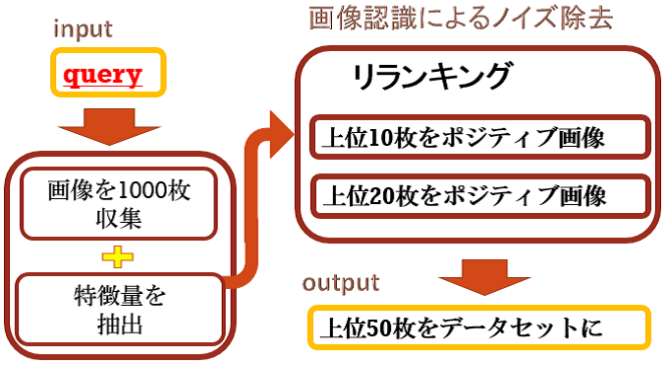

図 1　データセット構築の流れ

図 2、図 3、図 4 に、リランキングによるノイズ除去の過程を示す。図 2 は BingAPI の検索結果、図 3 は検索結果上位 10 枚を使った SVM モデルによるリランキング結果、図 4 はリランキング結果上位 20 枚を使った SVM モデルによるリランキング結果となっている。

3.4 認識可能性の評価

正例（質感）画像を 50 枚、負例（ランダム）画像を 5000 枚として、ランダム画像 5000 枚を質感画像 50 枚に混ぜ、SVM モデルを作る。その SVM モデルを使って、5050 枚の画像を分離する。この分離度合いをその質感語の recognizability(認識可能性) として評価する。 分離度が大きい場合には認識可能性が高いとし、分離度が小さい場合には認識可能性が低いとしている。平均適合率を各質感画像データセットについて計算し、これを recognizability とした。

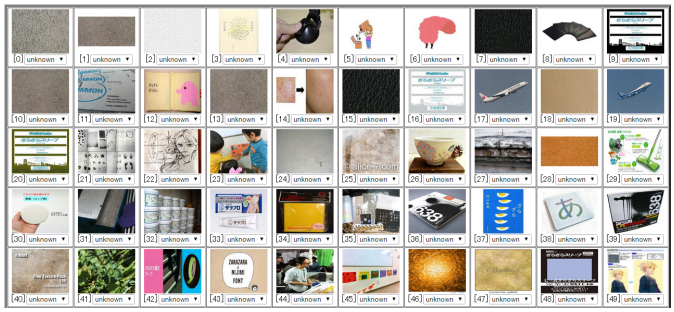

図 2　bingAPI の検索結果上位 50 枚（ざらざら）

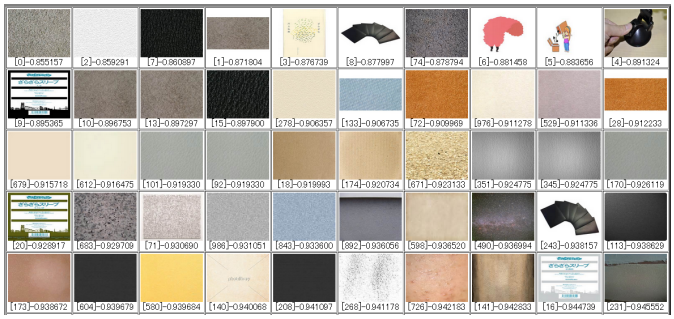

図 3　リランキング結果上位 50 枚　その 1（ざらざら）

図 4　リランキング結果上位 50 枚　その 2（ざらざら）

$$AP(n) = \frac{1}{m}\sum_{k=1}^{m} Precision_{true}(k)$$

この平均適合率は、ランキングの順位を加味した精度であり、ランキングされたデータセットを利用する本研究の目的に合っている。

3.5　画像収集システム

これまで説明した質感画像収集および認識可能性評価システムは、Web から利用可能なオンラインシステムとして実装し、簡単に、様々な語について画像 Web から画像を収集し、認識可能性を評価することが可能となっている。なお、このシステムは高速化のため多数の計算機を並列に利用し、システムへの負荷が大きいため、科研費新領域研究「質感脳情報学」のメンバーに限定して公開している。

システムは、クエリを入力すると、Bing API を使って Web 画像 1000 枚取得し、DCNN 特徴を抽出する。特徴を抽出し終わると、自動でリランキングを行い 1000 枚の Web 画像から、50 枚のクエリデータセットを構築しそれを画面に表示する。

クエリを入力してからデータセットが得られるまでには約 5 分かかる。システムで最も計算コストが必要となるのは DCNN 特徴を抽出する部分である。一台の計算機で 1000 枚の Web 画像から DCNN 特徴を抽出しようとする場合には 1 時間近くかかるが、このシステムでは、画像を集めながら DCNN 特徴を抽出し、40 台の計算機で並列に処理することで、必要な時間を約 5 分に抑えた。

また、システムは自動データセット構築のみでなく、手動でのデータセット構築にも対応している。Bing API の検索結果から、手動でポジティブ画像、ネガティブ画像を選択し、SVM を学習させることができる。リランキング結果は上位によい画像が集まるはずなので、そのリランキング結果から画像を選択することで、複数回の学習もできる。図 5 にシステム画面を示す。

図 5　system で収集した画像の例

4.　名詞＋オノマトペに対応する画像の収集と分析

第 3. 節ではオノマトペをクエリとして画像を収集したが、本節では名詞＋オノマトペのペアで画像を収集し、その認識可能性についての評価を行う。本研究では 1 つの名詞について、複数のオノマトペを用意し、これらをクエリのペアとしてそれぞれ 1000 枚 Web 画像を収集した。第 3. 節と同様の手法でリランキングし、データセットを構築する。

4.1　分類性の評価

名詞＋オノマトペの認識においては、3. 4 節における recognizability とは異なる評価基準を追加した。3. 4 節においては、正例（データセット）画像を 50 枚、負例（ランダム）画像を 5000 枚として、ランダム画像 5000 枚を正例画像 50 枚に混ぜ、SVM モデルを作り、これを分離した。

このセクションでは、不正解画像としてランダム画像ではなく、名詞画像を用いる。名詞＋修飾語だけでなく、名詞のみでも画像を収集し、名詞データセットを構築する。一般的に使われる名詞をクエリとして得られるランキング画像の精度は非常に高精度なので、上位から 500 枚をデータセットとした。

正例（データセット）画像を 50 枚、負例（名詞）画像を 500 枚として、名詞画像 500 枚を正例画像 50 枚に混ぜ、SVM モデルを作り、これを分離した。この分離度合いを discriminability として評価する。

分離度が高いほど名詞画像と視覚性が異なり、分離度が低い場合には名詞画像との視覚性が類似していると考えられる。セクション 3. 4 と同様に平均適合率から求めた。

4.2 同一名詞クラス内でのオノマトペのマルチクラス分類

名詞＋複数のオノマトペで画像を収集すると、一つの名詞につき、複数のオノマトペデータセットができる。それぞれのクエリで 1000 枚 Web 画像を収集し、50 枚のデータセットを構築できるので、1 つの名詞クラスの内にクエリの数だけオノマトペクラスのデータセットができることになる。

これらのデータセットの画像は同じ名詞とペアで収集した画像である。オノマトペ単体で構築されたデータセットを分類することに比べて、名詞クラス内でのオノマトペの分類は、よりそのオノマトペと視覚性の関係について正しい評価が可能であると考えられる。そこで、本研究では名詞クラス内でオノマトペのマルチクラス分類を行った。

5. 画像特徴量

本研究では、リランキング、データセットの認識率を評価する際に画像認識を用いた。画像特徴量の表現は Improved Fisher Vector と Deep Convolutional Neural Network Features (DCNN features) を用い、識別器には線形 SVM を用いる。

5.1 Improved Fisher Vector (IFV) [2]

Fisher Vector は混合ガウス分布を利用したソフト量子化により、特徴量をエンコードする手法である。Improved Fisher Vector はこの Fisher Vector を L2 正規化したものであり、より精度が高い。

本研究では、SIFT 特徴量を 1000 個ランダムサンプリングしてこれをエンコードする。SIFT 特徴量は 128 次元のベクトルである。そのままエンコードしてしまうと余分な情報が混じり効果的でないので、まず PCA を使って 64 次元のベクトルに圧縮する。圧縮したベクトル群をクラスタ数 256 の混合ガウス分布を利用してソフト量子化し、エンコードし、FV とする。これを、L2 正規化して IFV とした。

次元数は、2×クラスタ数×特徴量 なので 2×64×256 = 32768 次元ベクトルとなる。

5.2 Deep Convolutional Neural Network (DCNN) [3]

近年、大規模物体認識のためのデータセットで学習した Deep Convolutional Neural Network (DCNN) の活性化信号を bag-of-features や Fisher Vector などの代わりに画像特徴ベクトルとして用いることが広く行われており、物体認識やシーン認識、属性認識など様々な認識に対して有効であることが示されている。

DCNN 自体は ImageNet Large-Scale Visual Recognition Challenge (ILSVRC) 2010 で一躍注目を浴びることとなった [6] が、当初は DCNN の学習に用いた画像カテゴリを DCNN で分類まで行う使い方が一般的であった。2013 年に Donahue らによって、ILSVRC の 1000 種類の学習データで学習した DCNN の活性化信号が学習データに含まれいないカテゴリ画像の特徴量をして有効であることが示され、Caffe, DeCaf, Overfeat などのオープンソースソフトウエアが ILSCRV データセットの学習済 DCNN パラメータとともに公開されたことにより、bag-of-features や Fisher Vector に代わるカテゴリー分類のための画像特徴表現として広く用いられるようになってきている。

本研究では、オープンソースである Overfeat [7] を利用して DCNN 特徴を抽出している。Overfeat は、ImageNet Challenge の 1000 カテゴリで pre-training した DCNN 特徴を用いており、8 層のニューラルネットからなる。5 層までが畳み込み層 (convolution layer) で、残りの 3 層が全結合層 (fully-connected layer) になっている。

入力画像は 231x231 のサイズである。layer-8 の出力はニューラルネットで学習したカテゴリ分類の結果なので、特徴量としては扱わない。本研究では、layer-5、6、7 の出力結果を L2 正規化し特徴量として扱う。それぞれ、36864 次元、3072 次元、4096 次元のベクトルである。特に Layer-5 の特徴量は高次元でスパースなベクトルになっている。

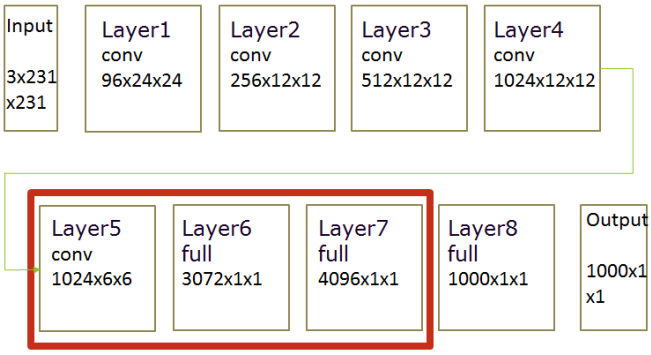

図 6 Overfeat のネットワーク

5.3 Support Vector Machine (SVM)

識別には線形 Support Vector Machine を用いる。SVM は強力な識別器であり、画像認識においても一般に使われている。機械学習によりモデルを作り、そのモデルを利用して識別を行う。SVM にはカーネルトリックを用いる手法があるが、IFV、DCNN は高次元なので線形 SVM で十分である。

SVM の出力値は以下の式から得られる。ただし、出力値を $y(x)$、入力ベクトルを x_i、学習により得られた SVM のモデルの重みベクトルを w_i、バイアスを b とする。

$$y(x) = \sum_{i=1}^{N} w_i \cdot x_i + b$$

この出力値が大きいほど、よりポジティブ画像である可能性が高いと判別されたことになる。本研究ではこの SVM の出力値を利用して、画像のリランキングを行う。

6. 実　　験

オノマトペデータセットの認識の実験と、名詞＋オノマトペデータセットの認識の実験を行った。

6.1 オノマトペの認識可能性

今回は 20 種類のオノマトペを用いて、オノマトペ画像データセットの構築、オノマトペデータセットの recognizability の評価をした。Fisher Vector、DCNN Layer-5、Layer-6、Layer-7 の 4 つの特徴量を使って実験をした。

6.1.1 リランキングの評価

ひとつのクエリー（オノマトペ）につき、BingAPI を利用して検索結果の上位 1000 枚の Web 画像を収集する。次に、Bing API の検索結果の上位 10 枚の画像を使ってリランキングを行

う。このリランキング結果の上位 20 枚の画像を使って再びリランキングを行い、ランキングの結果の上位 50 枚の画像をデータセットとした。図 7 はそのデータセットの例である。図 8 は

図 7　データセットの例

目視でデータセットを評価した結果である。Layer-6 のリランキング結果が最も良い結果となった。

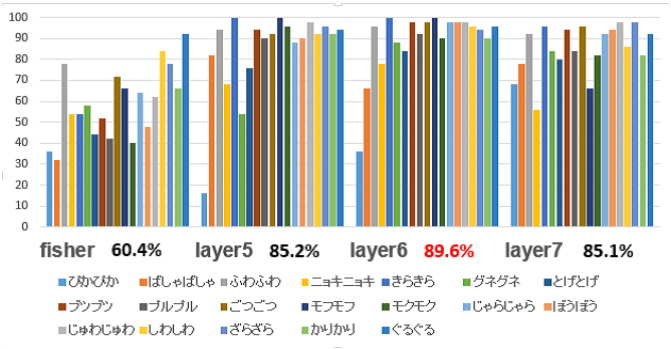

図 8　目視によるデータセットの評価

6.1.2　recognizability の評価

20 種類のオノマトペデータセットについて、IFV と DCNN の各レイヤーで評価を行った。最も結果の良かった Layer-6 でリランキングをして得られたデータセットについて、5-fold cross-varidation で実験した。図 8 に 10 種類の質感画像データセットについての IFV と DCNN の各レイヤーによる認識率を示した。今回の実験では IFV と DCNN の各レイヤーにおいて実験を行ったが、IFV と比べてどの Layer においても DCNN 特徴の精度が勝っていた。また、リランキングにおいては Layer-6 の精度がよかったが、recognizability の評価においては、Layer-7 の結果がもっとも精度がよくなった。

ただし、どのデータセットにおいても Layer-7 の結果がもっともよかったわけではなかった。特に、「ぶつぶつ」や「ざらざら」などのテクスチャに近い画像で構築されているデータセットについては Layer-5 の結果が Layer-7 の結果より精度がよかった。Layer-7 はより 1000 クラス分類の結果に近く、一般の物体認識にチューニングされていると考えられる。それでこのような結果になったのではないかと考えられる。テクスチャクラスなどの異なるクラスでファインチューニングによる追加学

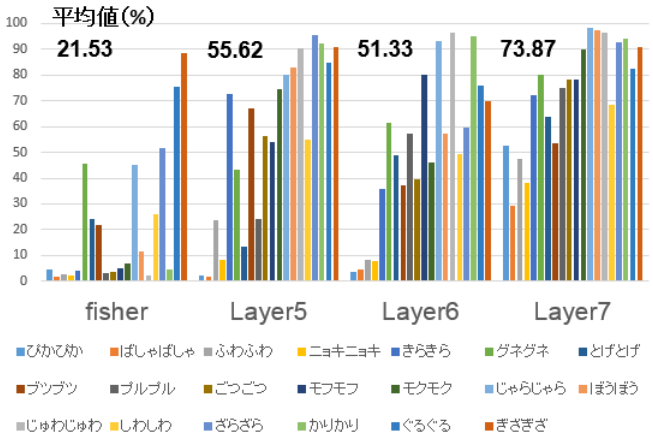

図 9　recognizability の評価

習をすることでまた違った結果が期待できる。

6.2　名詞＋オノマトペの分類

名詞＋オノマトペで Web 画像を収集し、データセットを構築した。今回の実験においては、名詞としてケーキ、花を用いて、これに対して複数のオノマトペで画像を収集し、認識を行った。また、今回は名詞とオノマトペのみの場合に限定するとクラスが少なくなってしまうので、名詞＋形容詞などでも画像を収集した。図 13 に今回用いたクエリのペアと集めた画像の例を示した。

実験は Layer-6 でリランキングを行いデータセットを構築した。それぞれのデータセットについて、上位 25 枚と、上位 50 枚について recognizability と discriminability を求めた。また、名詞＋オノマトペクラスのマルチクラス分類を行った。評価においても Layer-6 を用いた。評価は、5-fold cross-varidation で求めた。

	recognizability		discriminability		confusion matrix							
	25	50	25	50								
ゴロゴロ ケーキ	87.5	96.6	49.5	63.8	31	3	10	2	3	1	0	62.0
パサパサ ケーキ	84.4	92.0	76.0	84.0	3	26	3	7	5	6	0	52.0
サクサク ケーキ	86.0	96.9	75.4	79.3	8	4	24	8	2	3	1	48.0
ふわふわ ケーキ	95.0	99.2	89.7	92.9	1	2	2	42	3	0	0	84.0
なめらか ケーキ	72.5	96.3	33.2	28.5	3	2	3	3	37	2	0	74.0
濃厚 ケーキ	90.9	96.8	77.1	78.3	1	0	2	1	0	46	0	92.0
淡い ケーキ	55.6	89.9	30.4	54.5	0	0	0	0	3	0	47	94.0
					66.0	70.3	54.5	66.7	69.8	79.3	97.9	72.3

図 10　ケーキ＋オノマトペの認識結果

どのデータセットにおいても高い recognizability を示している。特に、上位 50 枚における recognizability を 6.1.2 と比較すると、名詞＋オノマトペにおけるデータセットは、オノマトペ単体で収集した画像のデータセットより精度が高くなっていることがわかる。

今回のケーキ名詞におけるマルチクラス分類の結果は 72.3%、花名詞におけるマルチクラス分類の結果は 84.6%となった。オノマトペのみのクラスに限定すると、精度は 61.5%、71.5%になる。名詞としてケーキ、花を選んだ場合では、オノマトペの認識は形容詞の認識と比べて難しいようだった。しかし、決して悪くない精度で認識できている。オノマトペと視覚性には関

図 11　ケーキ＋オノマトペの画像例

	recognizability		discriminability		confusion matrix							
	25	50	25	50								
ボンボン 花	49.0	94.9	30.3	56.7	35	3	2	2	4	0	4	70.0
ふわふわ 花	63.0	80.9	66.5	61.9	7	36	2	3	2	0	0	72.0
フレッシュ 花	74.0	91.7	54.9	60.2	2	1	38	7	1	0	1	76.0
メイン 花	83.2	97.3	73.3	63.5	0	2	3	44	1	0	0	88.0
ブルー 花	48.1	93.9	29.0	45.1	1	0	0	0	49	0	0	98.0
黄色い 花	100.0	100.0	24.9	37.9	0	0	0	0	0	50	0	100.0
赤い 花	84.1	93.7	52.2	47.3	3	0	0	0	3	0	44	88.0
					72.9	85.7	84.4	78.6	81.7	100.0	89.8	84.6

図 12　花＋オノマトペの認識結果

図 13　花＋オノマトペの画像例

連があるのではないかと考えられる。また、マルチクラス分類ではオノマトペのうちで誤認識の多いものがあり、ケーキでの分類においては、「サクサク」「ゴロゴロ」のうち 14〜16%が「ふわふわ」に分類されている。花での分類においては、「ふわふわ」の 14%が「ぽんぽん」に分類されている。

オノマトペのマルチクラス分類の精度は形容詞の分類に比べて精度が低くなったが、discriminability は高い値になっている。一方で、形容詞の discriminability はそれほど高くない。discriminability は、recognizability と名詞画像との類似度を加味した値になっている。recognizability が高いのに、discriminability が高いということは、そのデータセットと、名詞画像データセットの類似度が低いということになる。オノマトペとともに用いられるケーキは一般のケーキと比べて、特徴的な見た目をもっているといえるかもしれない。ただし、単にケーキにおいてはオノマトペより、形容詞と一緒に使われることが多く、それらと類似する画像が名詞画像データセットに多いだけということも考えられる。

7.　まとめ と 今後の課題

今回の実験では、オノマトペ単体で収集した画像の認識と、名詞＋オノマトペで収集した画像の認識を行った。オノマトペ単体で収集した画像の認識からリランキングには DCNN 特徴が有効であることがわかった。また、DCNN 特徴の各レイヤーによって精度は異なり、画像の傾向によって違いがあるようだった。一般の物体に近いデータセットの認識においては Layer-7 の精度がよかったが、一般の物体認識から遠ざかっているテクスチャに近い画像のデータセットなどでは Layer-5 のほうが精度が高くなる可能性もある。今回は Layer-5 から Layer-7 までを調べたが、Layer-5 より浅いレイヤーの特徴量を用いた場合の結果を調べてみる必要もあると考えられる。

名詞＋オノマトペで収集した画像のデータセットは、オノマトペ単体で収集した画像のデータセットより精度が高かった。マルチクラス分類の精度も悪くなかった。オノマトペと視覚性には関連があることが分かった。

今回はケーキと花についてのオノマトペ画像のみを集め、マルチクラス分類をしてみたが、名詞＋オノマトペでの画像収集の研究はまだ発展できる可能性がある。名詞＋オノマトペで画像を収集することは単にそれだけで、名詞画像のデータセットの拡張につながる。今回はケーキについてのオノマトペと形容詞合わせて七つのペアで画像を収集した。名詞のみで収集した画像を合わせれば、合計 8000 枚のケーキ画像が集まったことになる。また、異なる名詞と共通するオノマトペで集めた画像の関係を調べてみることも興味深く、今後の課題である。

謝辞　本研究は文部科学省科学研究費新領域研究「質感脳情報学」公募研究 25135714 の助成を受けたものです。

文　　献

[1] C. Liu, L. Sharan, E. Adelson, and R. Rosenholtz. Exploring features in a bayesian framework for material recognition. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2010.

[2] F. Perronnin, J. Sanchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *Proc. of European Conference on Computer Vision*, 2010.

[3] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell. DeCAF: A deep convolutional activation feature for generic visual recognition. 2014.

[4] M. Cimpoi, S. Maji, I Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *Proc. of IEEE Computer Vision and Pattern Recognition*, 2014.

[5] M. Ozeki and T. Okatani. Understanding convolutional neural networks in terms of category-level attributes. In *Proc. of Asian Conference on Computer Vision*, 2014.

[6] A. Krizhevsky, I. Sutskever, and G E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2012.

[7] S. Pierre, E. David, Z. Xiang, M. Michael, F. Rob, and L Yann. Overfeat: Integrated recognition, localization and detection using convolutional networks. In *Proc. of International Conference on Learning Representations*, 2014.