デジタルアーカイブにおける Deep Learning を用いたメタデータ生成

Metadata Generation Using Deep-Learning in Digital Archives

濱上知樹

HAMAGAMI Tomoki

［論文要旨］

本研究では，デジタルアーカイブ画像のメタデータを生成し類似画像検索などに役立てることを目的にしている。

一般物体認識でよく用いられている画像のヒストグラム表現手法，Bag-of-Features［4］では SIFT［2］［3］に代表される画素の濃淡分布をもとに算出された特徴点および局所特徴量が用いられるが，その一方で一般物体認識の分野で Deep Learning を用いた技術［6］が注目を集めている。

Deep Learning 手法では，画像全体を入力し，画像中に存在する主となる物体を認識させることが一般的となっており，画像中の様々な局所的な情報が欠落してしまっていた。

そこで本研究では画像をセグメントに分割し，各セグメントから Deep Learning を用いた特徴抽出を行い，クラスタリングによって分類された各セグメントのクラスタ情報を局所特徴とした Bag-of-Features を行い，ヒストグラム表現とすることで画像に存在する意味情報を反映したメタデータ生成を提案する。また，ヒストグラム間の比較にはクラスタ間の類似関係を反映した距離計算を行うことでクラスタ数が細かすぎる際に，似ている画像が類似画像として判定できない問題を解決した。

実験では，デジタルアーカイブとして小袖屏風画像［9］を用いてヒストグラム間の比較を行うことで Deep Learning［7］を用いて Bag-of-Features の応用を行うことの有効性，さらにクラスタ間の距離関係を反映した距離計算を行うことの有効性を示した。

【キーワード】デジタルアーカイブ，Deep Learning，一般物体認識，BoF，類似画像検索

❶……序論

1.1　デジタルアーカイブの増加

近年，美術品や歴史資料のデジタルアーカイブ化が進められている。デジタルアーカイブとは，現存する資料をデータ化することでコンピュータ上での半永久的な閲覧を可能にしたものであり，国立文化財機構が運営するe国宝［13］や，Googleが行っているGoogle Art Project［14］など国，企業問わずその取り組みが進められている。例に挙げたGoogle Art Projectでは東京国立博物館収蔵の国宝「観楓図屏風」を70億画素の超高解像度で撮影し，拡大することで細かい筆のタッチまで詳細に見ることができるようになっている（図1.1）。それにより実物をその場で見るよりも詳細な情報を視覚的に得ることが可能となり，デジタルアーカイブの魅力のひとつとなっている。

進められているデジタルアーカイブ化であるが，総務省によるデジタルアーカイブに関する提言によるとデジタルアーカイブ化する際にはデジタルアーカイブに蓄積されたデジタルコンテンツの性質を表す情報なしには利用も保存も難しいことが指摘されている［1］。デジタルコンテンツの性質を表すデータは，一般的に「メタデータ」と呼ばれている。

1.2　画像へのメタデータの付与

画像にメタデータを付与することでその情報を基にした類似画像検索などの応用が期待できる。検索の際に扱うメタデータによって，類似画像検索には大きく分けてText-Based Image Retrieval（TBIR）とContent-Based Image Retrieval（CBIR）が存在する。それぞれについて説明を行う。

1.2.1　Text-Based Image Retrieval（TBIR）

Text-Based Image Retrieval（TBIR）とはあらかじめ画像に対応付けられたテキストに基づいたキーワード検索である。総数 m の画像のデータセットを $I = \{i_1,\cdots,i_m\}$，それぞれの画像に関連付けられたキーワードのセットを $K = \{\mathrm{k}_1,\cdots,\mathrm{k}_m\}$ とする。単語によるクエリを W_q としたとき，I のうち W_q に関連した $\mathrm{k}_j (1 \leq j \leq m)$ をもつ i_j が検索結果として出力される。

TBIRを行うには画像にキーワードが与えられている必要があり，与えられていない場合には人

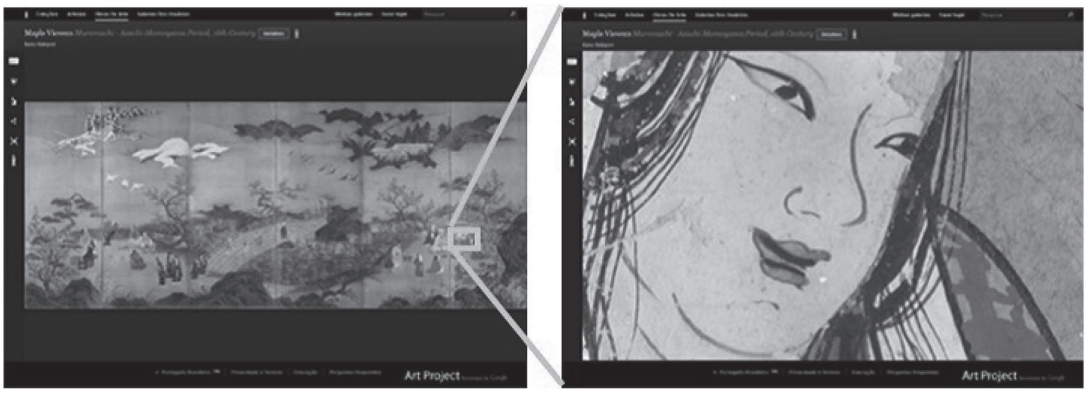


図1.1　Example of super resolution image of Google Art Project

手でキーワードを与えなければならず，膨大な手間と時間を要する。また，画像に関連付けされたキーワードがその画像を表現するのに適した言葉であるとは限らないといった問題も存在する。そこで，このような問題を解決するために画像自体がもつ特徴から類似性を比較する CBIR が研究されている。

1.2.2 Content-Based Image Retrieval（CBIR）

Content-Based Image Retrieval（CBIR）とは画像から抽出した色，模様などの画像特徴に基づいた検索である。クエリを入力してそれに関連した画像を検索する TBIR とは異なり，CBIR では類似画像を検索したい画像をシステムに与えることで検索を行う。

総数 m の画像のデータセットを $I=\{i_1,\cdots,i_m\}$，それぞれの画像から抽出した特徴のセットを $F=\{f_1,\cdots,f_m\}$ とする。クエリ画像を q，q から抽出された特徴を f_q とすると I のうち f_q に近い特徴 $f_j(1\leq j\leq m)$ をもつ i_j が検索結果として出力される。

したがって，処理の流れは以下のようになる。

1. 特徴の抽出

   画像の特徴を I の $i_1,\cdots,i_m$ それぞれに対して抽出する。

2. クエリ画像の入力

   このステップではクエリ画像 q を入力する。

3. 特徴の比較

   クエリ画像の特徴 f_q とデータセット内のすべての画像の特徴 $F=\{f_1,\cdots,f_m\}$ を比較する。

4. 検索結果の表示

   比較を行った結果，距離が小さく算出された画像 ij を検索結果として出力する。

以上の説明をふまえて TBIR と CBIR による類似画像検索の流れを図 1.2 に示す。

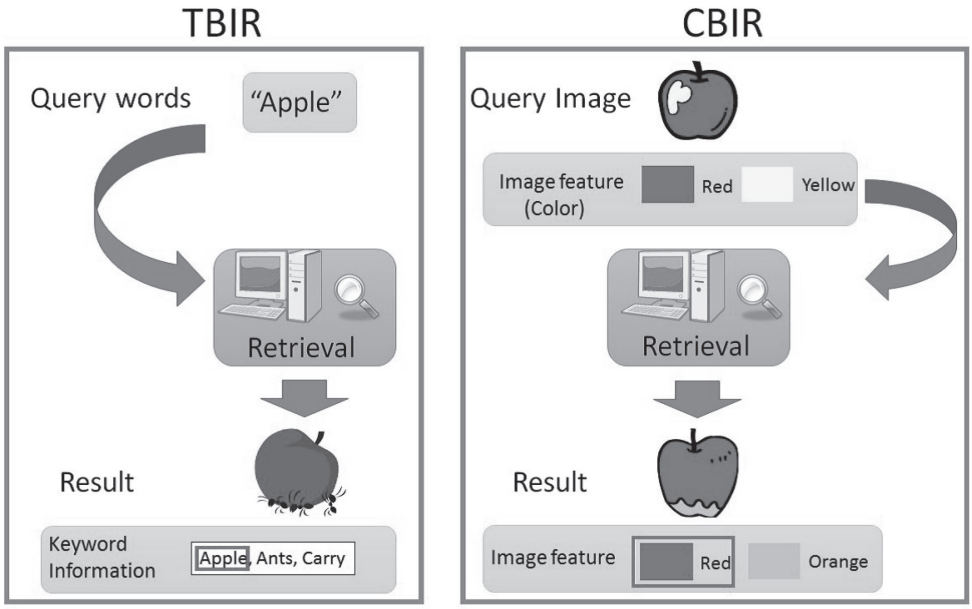


図 1.2　Flow of TBIR and CBIR

本研究ではデジタルアーカイブを対象としているため，TBIRを用いた場合，専門家でない人が検索を行う際に，適切なキーワードを入力できない可能性がある。また，CBIRは専門知識をもたない一般の人のみならず，専門家にとっても機械が提示する結果を基に新たな知見を得るきっかけになるといったメリットも存在する。したがって本研究ではCBIRについて扱っていくこととする。

CBIRを使用するには画像同士の類似性を計算するためにそれぞれの画像から特徴を抽出する必要がある。

画像から抽出した情報を基にメタデータの付与を行うには，抽出した特徴から画像中に何が写っているかを把握する一般物体認識を行う必要がある。次章以降で従来の一般物体認識に関する研究について説明する。

❷……従来の一般物体認識のアルゴリズム

一般物体認識のアルゴリズムは大きく分けて特徴抽出とその特徴を用いた画像表現というステップからなる。この章では従来用いられている特徴抽出と画像表現のアルゴリズムについて述べる。

2.1　特徴抽出

2.1.1　Scale-Invariant Feature Transform（SIFT）

Scale Invariant Feature Transform（SIFT）とはLoweによって考案された画像局所特徴量である［3］。SIFTではコーナー点やあるピクセルを中心とした周辺領域に多くの濃淡情報をもつ点を特徴点として扱う。SIFTのアルゴリズムは特徴点の検出と特徴量の記述の2段階に分けられる。

特徴点の検出では複数のDifference-of-Gaussian（DoG）画像を用いて極値の探索を行うことで特徴点の位置と特徴点を記述する範囲を表すスケールを検出する。特徴点の候補はスケールの異なるガウス関数 $G(x,y,\sigma)$ と入力画像 $I(u,v)$ を畳み込んだ平滑化画像 $L(u,v,\sigma)$ の差分（Difference-of-Gaussian, DoG画像）により求められる。したがって σ をガウシアンフィルタのスケール，x と y を注目ピクセルからの距離とすると，L と G はそれぞれ次式で表される。

$$L(u,v,\sigma) = G(x,y,\sigma) * I(u,v) \tag{2.1}$$

$$G(x,y,\sigma) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{x^2+y^2}{2\sigma^2}\right) \tag{2.2}$$

DoGの結果の画像を $D(u,v,\sigma)$ とすると，DoG画像は次式で表される。

$$D(u,v,\sigma) = (G(x,y,k\sigma) - G(x,y,\sigma)) * I(u,v) \tag{2.3}$$

$$= L(u,v,k\sigma) - L(u,v,\sigma) \tag{2.4}$$

この処理を σ_0 から k 倍ずつ大きくした異なるスケール間で行い，図2.1のように複数のDoG画像を求める。

σ が一定の割合で増加し続けるとガウシアンフィルタのウィンドウサイズが増大し，処理でき

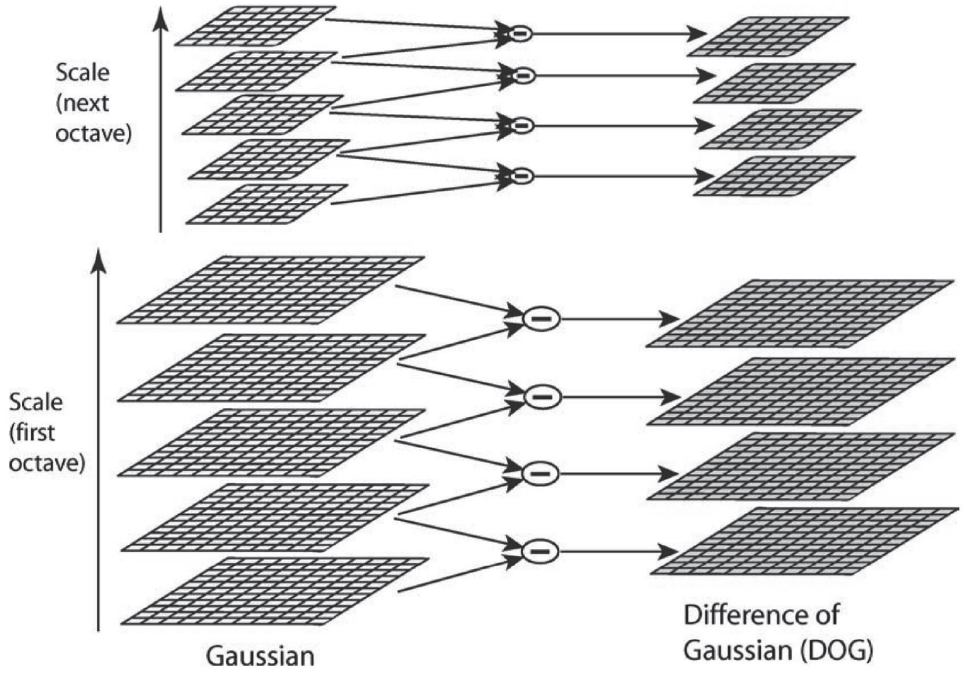


図 2.1　Flow of making DoG image.（Figure is referred from［3］）

ない端領域の拡大と計算コストの増加という問題が発生する。そこで SIFT では複数生成された平滑化画像の中から $2\sigma_0$ で平滑化された画像 $L_1(2\sigma_0)$ を 1/2 のサイズにダウンサンプリングすることでガウシアンフィルタのウィンドウサイズによる計算量の増加を防止している。1/2 のサイズにダウンサンプリングされた画像 $L_2(\sigma_0)$ と $2\sigma_0$ で平滑化を行った画像 $L_1(2\sigma_0)$ には以下のような関係が成り立つ。

$$L_1(2\sigma_0) \approx L_2(\sigma_0) \tag{2.5}$$

DoG の値が最大となる σ では特徴点はキーポイントを中心とした領域の濃淡情報をより多く含む領域に対応していると考えることが出来る。したがって，DoG 画像から極値を検出し候補となる特徴点のスケールを決定する。この処理をスケールの異なる DoG 画像の全ピクセルに対して行う。この処理の後に特徴候補から不要な候補を削除することで特徴点が決定する。

その後回転に不変な特徴を得るために特徴点のオリエンテーションを求め，そのオリエンテーションに基づいて特徴点の特徴量を記述することで SIFT のアルゴリズムは終了となる。

2.2　Bag-of-Features

一般的な SIFT を用いた画像分類手法として Bag-of-Features（BoF）［4］がある。

BoF は Bag-of-Words（BoW）というテキスト検索のモデルを導入している。BoW では，文書中の単語の出現頻度によって文書を表現しており，BoF ではそれに対応して局所特徴の出現頻度によって画像を表現している。BoF の処理は以下の流れで行われる（図 2.2）。

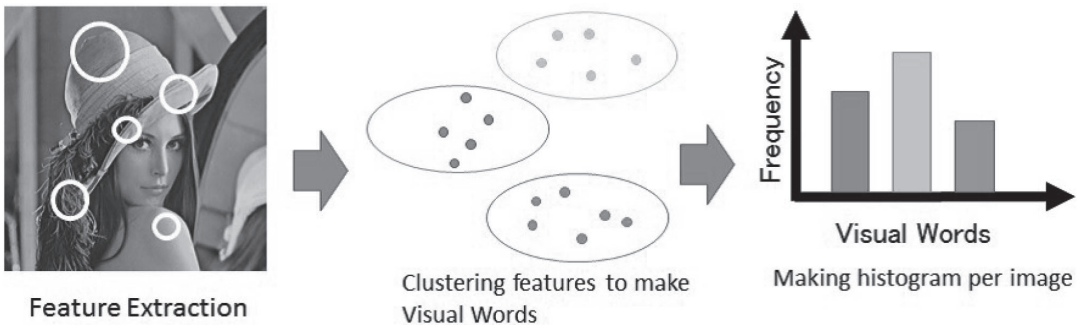


図 2.2　Flow of Bag-of-Features

1. 特徴点抽出
　特徴点を抽出し，SIFT などの局所特徴ベクトルを取得する。
2. 局所特徴量のクラスタリング
　全画像の全局所特徴ベクトルをクラスタリングし，*n* 個の visual word（クラスタの代表点）を選出する。局所特徴ベクトルを最も近い visual word に置き換えることで量子化を行う。クラスタリングには一般的に k-means 法が用いられる。
3. ヒストグラムの作成
　visual words に基づいて各画像について局所特徴のヒストグラムを作成する。

BoF に用いる特長には先ほど述べた SIFT などの画像中で濃淡変化をもつ点を特徴点とする局所特徴が用いられるが，濃淡変化の少ない画像など，どの画像に対してもその画像を表現するのに最適な特徴量とは限らない。そこで，本研究では近年物体認識において非常に高い精度を得ている Deep Learning を用いている。

❸……Deep Learning による画像特徴抽出

Deep Learning はニューラルネットを複数層重ねたネットワークを用いた機械学習法の一つである。入力画像を表現するパラメータを自動で学習するため，従来人手で設計していた画像特徴を自動で抽出できることが大きな強みである。本章では近年の画像分類のコンテストで Deep Learning 手法の中でも特に高い精度を上げている Convolutional Neural Network（CNN）について説明を行う。

3.1　Convolutional Neural Network（CNN）

CNN とは，畳み込み層とプーリング層を交互に重ねた構造をもつ多層ニューラルネットワークで図 3.1 のような構造をもつ。2012 年の ImageNet が毎年開催している 120 万枚の画像から 1,000 クラスのカテゴリ識別の精度を競う The ImageNet Large Scale Visual Recognition Challenge（ILSVRC）[5] において上位 5 つの候補に正解が含まれていない Top-5 エラー率で AlexNet [6] と呼ばれるネットワークが従来手法から大幅な改善となる 16.4% という低いエラー率を示し CNN

が画像認識に広く使われることとなった。

畳み込み層（図 3.2）は局所領域において複数のフィルタを畳み込む構造をもち，n–1 層の特徴マップ k からの出力を h_k^{n-1}，フィルタを w_k^n，活性化関数を ϕ，バイアスを b_k とすると，畳み込み層 n 層目の特徴マップ j への出力は式 3.1 で表される。

$$h_j^n = \phi(\sum_{k=1}^{K} h_k^{n-1} * w_k^n) + b_k \tag{3.1}$$

活性化関数には Rectified Linear function（ReL）（式 3.2）がよく用いられている。これは，ReL が単純で計算量が小さく，学習が高速で層が多層になっても勾配が減衰せずに伝わるためである。

$$\phi(x) = \max(0, x) \tag{3.2}$$

プーリング層（図 3.3）は局所領域における情報を一部捨てることにより，位置変化の影響を吸収するもので，n–1 層の局所領域 N 内の $(\overline{x}, \overline{y})$ において，輝度値の最大のもののみを n 層に伝える最大プーリング（式 3.3）がよく用いられる。

$$h_j^n(x, y) = max_{\overline{x} \in N(x), \overline{y} \in N(y)} h_j^{n-1}(\overline{x}, \overline{y}) \tag{3.3}$$

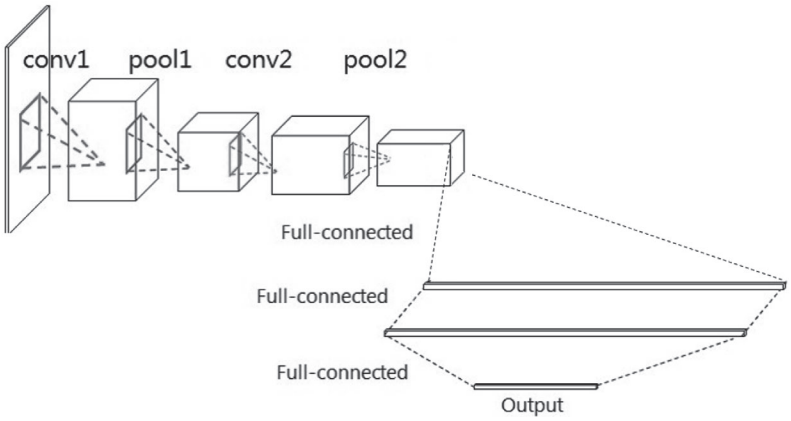


図 3.1　Example of structure of Convolutional Neural Network

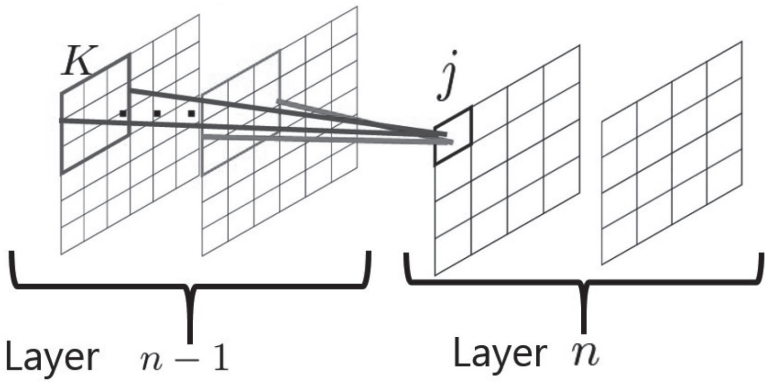


図 3.2　Convolution

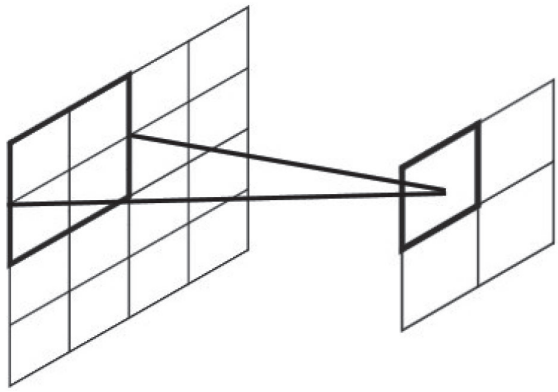


図 3.3　Pooling

CNN は教師あり学習の手法の一つである誤差逆伝播法により，出力が正しくなるようにパラメータを調整していくことで入力画像を表現するのに適切なフィルタを学習していく。多層のニューラルネットで誤差逆伝播法を可能にしているのは畳み込み層による局所領域のみの接続や，プーリング層による情報量の削減などの処理である。

3.2　Deep Learning で学習させる際の課題

Deep Learning を用いて画像中の意味的情報を学習するには大量の画像データが必要となる。しかし，対象となるデジタルアーカイブは貴重な芸術品や歴史資料をデータ化したものであり，大量に存在するデータとは言えない。したがって本研究ではこの課題を解決するために他のデータセットで学習済みの結果をもとに対象画像を分類するアプローチを取る。

他のデータセットを利用した代表的な手法として，Deep Convolutional Activation Feature（DeCAF）［8］がある。DeCAF はあらかじめ別のデータセットで学習しておいたネットワークを特徴抽出器として用いた手法で，出力層に近い 4,096 次元のベクトル（図 3.4）を特徴量としている。DeCAF は別のデータセットで学習したネットワークから抽出した特徴であるが，所望のドメインでの分類に対しても有効であることが分かっており［8］，データ数が少ない本研究に有効な特徴量であると言える。

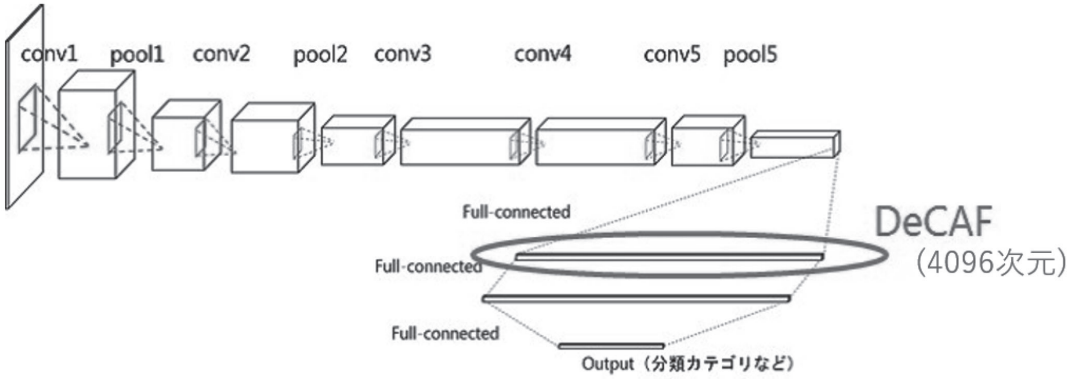


図 3.4　DeCAF

❹……提案手法

本研究では画像検索に有用なメタデータを生成する。提案手法は大きく，Deep Learning によって得られた特徴を用いたヒストグラム表現，ヒストグラムの比較の二つに分かれる。

4.1　Deep Learning によって得られた特徴を用いたヒストグラム表現

通常 Deep Learning は画像全体をネットワークに入力してその画像を表現するのに適した特徴を得るが，画像全体から得られる情報だけではその画像の大まかな特徴しか把握できず，高精細なデジタルアーカイブ画像に含まれている数多くの細かい情報が失われてしまう。そこで本研究では画像を複数の解像度でセグメントに分割し，分割したセグメント一つ一つから DeCAF［11］を抽出する。その後抽出した DeCAF をクラスタリングし，各分割画像がどのクラスに属しているかでヒストグラムを作成する（図 4.1）。これにより，画像中の細かい情報も逃さず取得することが可能となる。さらに，2 章で述べた従来の BoF では SIFT 特徴という点情報からヒストグラムを作成していたが，本手法では分割画像自身から特徴を得てヒストグラムを作成しているために従来手法と

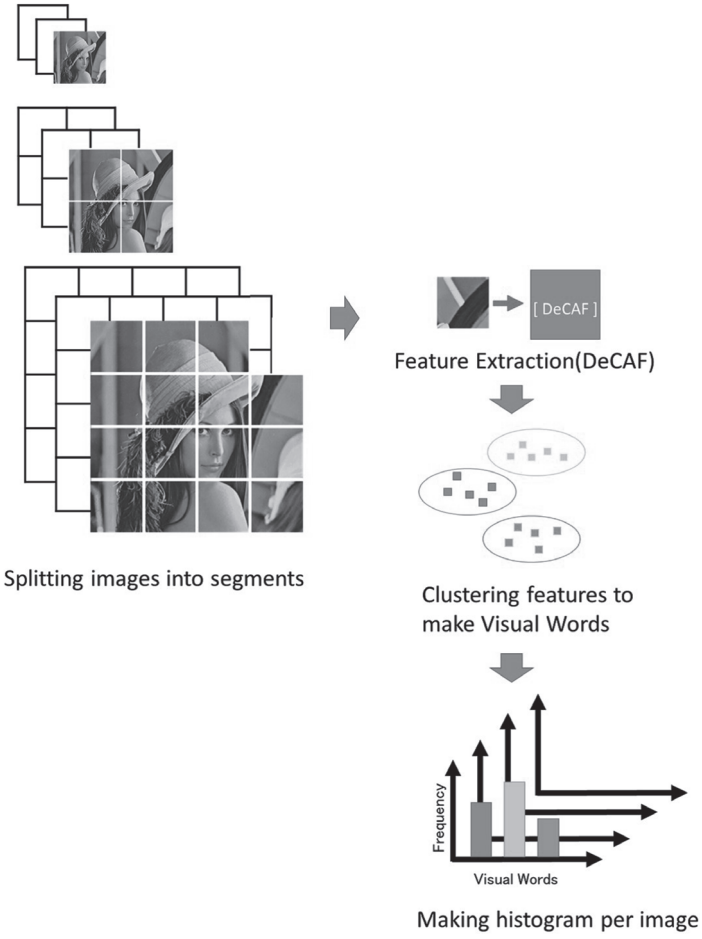


図 4.1　Proposed method

比較して画像中の局所特徴を表現するのに適したヒストグラムを得ることが可能となる。提案手法の流れをまとめると以下のようになる。

1. 分割画像の作成
　複数の解像度の画像から分割画像（以降，セグメントと呼ぶ）を作成する。
2. 画像特徴（DeCAF）の抽出
　各セグメントからDeCAFを抽出する。
3. 特徴量のクラスタリング
　全画像の全セグメントから抽出されたDeCAFをn個のクラスタにクラスタリングする。
4. ヒストグラムの作成
　visual wordsに基づいて各画像についてセグメント画像がどのクラスタに属しているかを表したヒストグラムを作成する。

3.の特徴量のクラスタリングに用いる手法の代表的なものとして，k-means法がある。k-means法はあらかじめクラスタ数kを決定しておき，処理の初めに特徴空間中にランダムにクラスタの母点を配置する必要がある。それゆえ特徴空間におけるデータの分布を考慮したクラスタリングができず，クラスタ数kを増やした際に画像のクラス分類に有用でない画像のクラスタ数も無意味に増加してしまう。クラス分類では，識別に有用でない画像のクラスタはひとかたまりに，識別に有用な画像のクラスタほどより細かくクラスが分けることができることが好ましい。したがって，本研究ではクラスタリングに階層的クラスタリングを用いた。階層的クラスタリングでは，それぞれのデータを1つのクラスタとして扱い，クラスタ間の類似度を計算して最も類似したクラスタを最終的には1つのクラスタになるまで併合していくクラスタリング手法である。階層的クラスタリングの結果はデンドログラムで表現することが出来る（図4.2）。グラフの高さはクラスタ間の距離を表しており，高さが低い状態で併合されているクラスタ同士は類似した特徴をもつクラスタであるといえる。デンドログラムによって画像間の画像特徴の類似関係を把握することが可能となる。

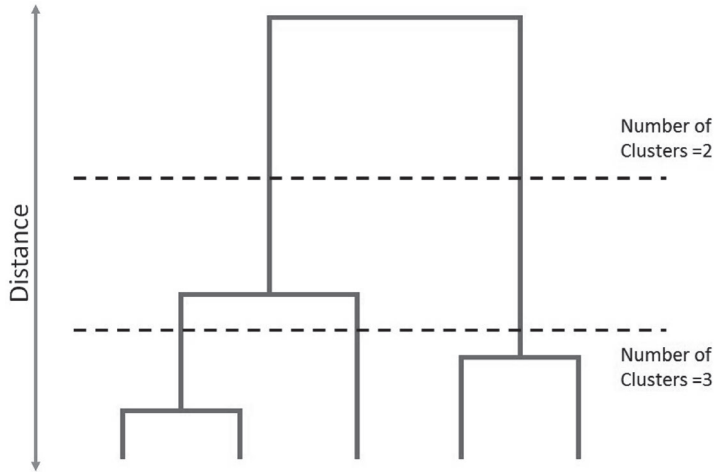


図4.2　Example of dendrogram

4.2　ヒストグラムの比較

ヒストグラムの比較には，histogram intersection が多く用いられる。histogram intersection は，2 つのヒストグラムを H_1 と H_2，ヒストグラム H_n の i 番目のビンを $H_n[i]$，ヒストグラムのビンの数を n として，式 4.1 によって算出する。

$$intersection = \sum_{i=1}^{n} \min(H_1[i], H_2[i]) \tag{4.1}$$

しかし，histogram intersection のような対応するビン同士だけを見る距離尺度では，2 つの画像 A, B を比較する際に，画像 A のビン i の高さ A_j と画像 B のビン j の高さ B_j の類似関係は考慮されない。そのため，A_i が大きくて A_j が小さく，また，B_i が小さくて B_j が大きい場合，たとえビン i とビン j がお互いに類似したセグメント画像を含むクラスタであったとしてもまったく類似していない画像として相関が計算されてしまう。この場面は特にクラスタリングの際にクラスタ数を細かくした際に発生する。クラスタ数を細かくすることでひとつひとつの画像を異なる画像として区別できるようにはなるが，類似した画像を同じクラスタとしてグループに分けることには適さない。これは，類似した画像であっても別のクラスタに分類されるためである。一方でクラスタ数を減らすと今度はヒストグラムによる画像の表現力が落ち，画像の比較に影響が生じる。この設定するクラスタ数による影響を減らすため，ビンとビンの類似性を考慮した距離計算を取り入れる。

4.2.1　Earth Mover's Distance（EMD）

Earth Mover's Distance（EMD）［10］は，2 つの分布間の距離を測るのに用いられる距離である。分布は特徴量と供給量（または需要量）の集合（シグネチャ）からなる（図 4.3）。EMD は線形計画問題のひとつである輸送問題の解に基づいて計算される。

m 個の供給地をもつ供給地集合 A と，n 個の需要地をもつ需要地集合 B はそれぞれ式 4.2，式 4.3 で表される。

$$A = \{(\boldsymbol{a}_1, w_{\boldsymbol{a}_1}), \cdots, (\boldsymbol{a}_m, w_{\boldsymbol{a}_m})\} \tag{4.2}$$

$$B = \{(\boldsymbol{b}_1, w_{\boldsymbol{b}_1}), \cdots, (\boldsymbol{b}_m, w_{\boldsymbol{b}_m})\} \tag{4.3}$$

A については a_i は i 番目の供給地を表す特徴ベクトルで w_{ai} は i 番目の供給地が有する供給量である。B については b_j は j 番目の需要地を表す特徴ベクトルで w_{bj} は j 番目の需要地が有する需要量である。供給地と各需要地間の単位あたりの輸送コスト D = [d_{ij}] は a_i と b_j の差を表し，一般的には a_i と b_j のユークリッド距離（式 4.4）が用いられる。

$$d_{ij} = \| a_i - b_j \| \tag{4.4}$$

供給地から需要地への輸送量をフロー F = [f_{ij}] とする。総コストを最小にする適切なフロー F が輸送問題の解である。単位あたりの輸送コスト d_{ij}，輸送量 F = [f_{ij}] より，総コスト $WORK(A,B,F)$ は

以下のように表される。

$$WORK(A,B,\boldsymbol{F}) = \sum_{i=1}^{m}\sum_{j=1}^{n} d_{ij} f_{ij} \tag{4.5}$$

ただし，総輸送コストを計算する際，以下の制約条件を満たさなければならない。

1. 供給地から需要地の1方向のみにしか輸送されない。

$f_{ij} \geq 0, (1 \leq i \leq m, 1 \leq j \leq n)$

2. 供給地 i から供給できる要領は供給量 w_{ai} を超過しない。

$$\sum_{j=1}^{n} f_{ij} \leq w_{\boldsymbol{a}_i}, (1 \leq i \leq m)$$

3. 需要地 j が受け取れる要領は需要量 w_{bj} 以下である。

$$\sum_{j=1}^{m} f_{ij} \leq w_{\boldsymbol{b}_j}, (1 \leq i \leq n)$$

4. 供給地から移動できる最大総輸送量

$$\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij} = \min\left(\sum_{i=1}^{m} w_{\boldsymbol{a}_i} \sum_{j=1}^{n} w_{\boldsymbol{b}_j}\right)$$

$EMD(A,B)$ は輸送問題の解である最適な F により得られる $\min(WORK(A,B,F))$ を総フローで割ることで求められる。

$$EMD(A,B) = \frac{\min(WORK(A,B,\boldsymbol{F}))}{\sum_{i=1}^{m}\sum_{j=1}^{n} f_{ij}} \tag{4.6}$$

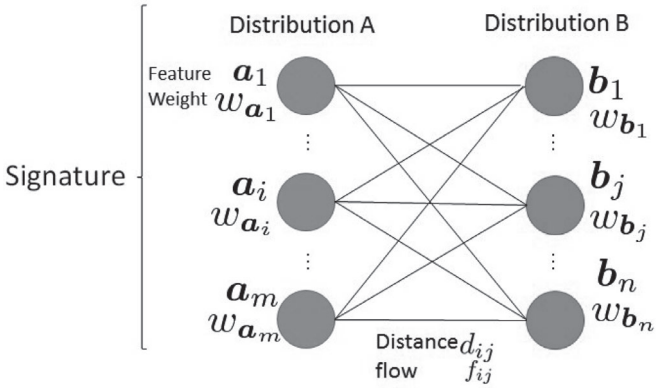


図 4.3　Situation of calculation EMD

4.2.2　Earth Mover's Distance の適用

本研究では，作成したデンドログラムを基にヒストグラムを作成している。デンドログラムでは隣接したクラスタ同士は類似度が高く，離れているクラスタほど類似度は低い。EMD を計算する際の距離にはクラスタ間の距離（図 4.4），そして重みには各クラスタのビンの高さを用いた。

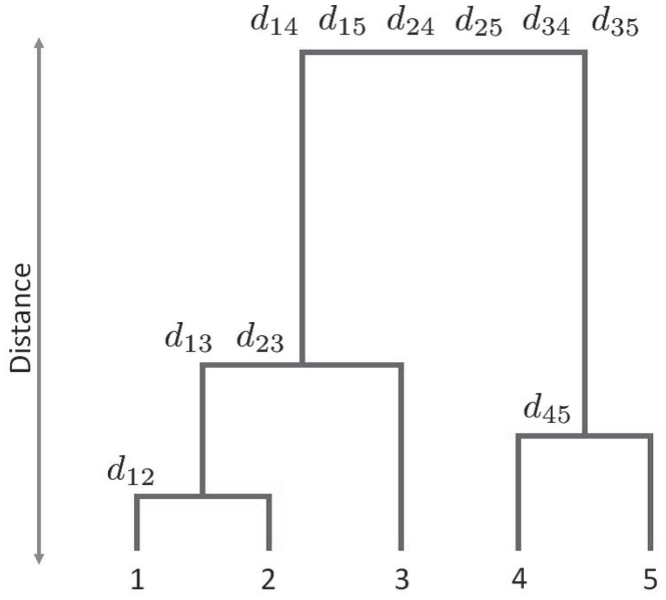


図 4.4　Distances for calculating EMD

❺……デジタルアーカイブを用いた実験

提案手法の目的は画像中から意味を反映した画像特徴を取り出し，対象の画像群の中から意味的に類似した画像を取得することにある。本章では実際にデジタルアーカイブデータを使用した実験を行い，提案手法の有効性を確認した。

5.1　使用したデジタルアーカイブ

本研究ではデジタルアーカイブとして，国立歴史民俗博物館から提供された小袖屏風の高解像度画像を使用している。

小袖屏風（図 5.1）とは状態の芳しくない小袖裂（こそでぎれ）の保存と鑑賞を両立させるために，昭和のはじめ頃，染織美術の蒐集家である野村正治郎によって考案された屏風である。小袖屏風には貴重な裂類が多数貼り付けられており，それらは資料不足に悩む近世の染織研究にきわめて重要な位置を占めている。

この実験では小袖屏風間の意味的関係性を把握するための意味情報抽出を行う。これにより，データ間の新たな関係性の発見の支援や抽出した特徴を用いた歴史研究支援といった効果が見込まれる。

小袖屏風には構図や技法など，様々な構成要素があるがその中でも最も重要なのが小袖に描かれているモチーフ（文様）である（図 5.2，図 5.3）。このモチーフには大きなものから非常に注目してみなければ分からないものまで様々であり，そのような点で複数の解像度で特徴を抽出する本手法が有効であるといえる。

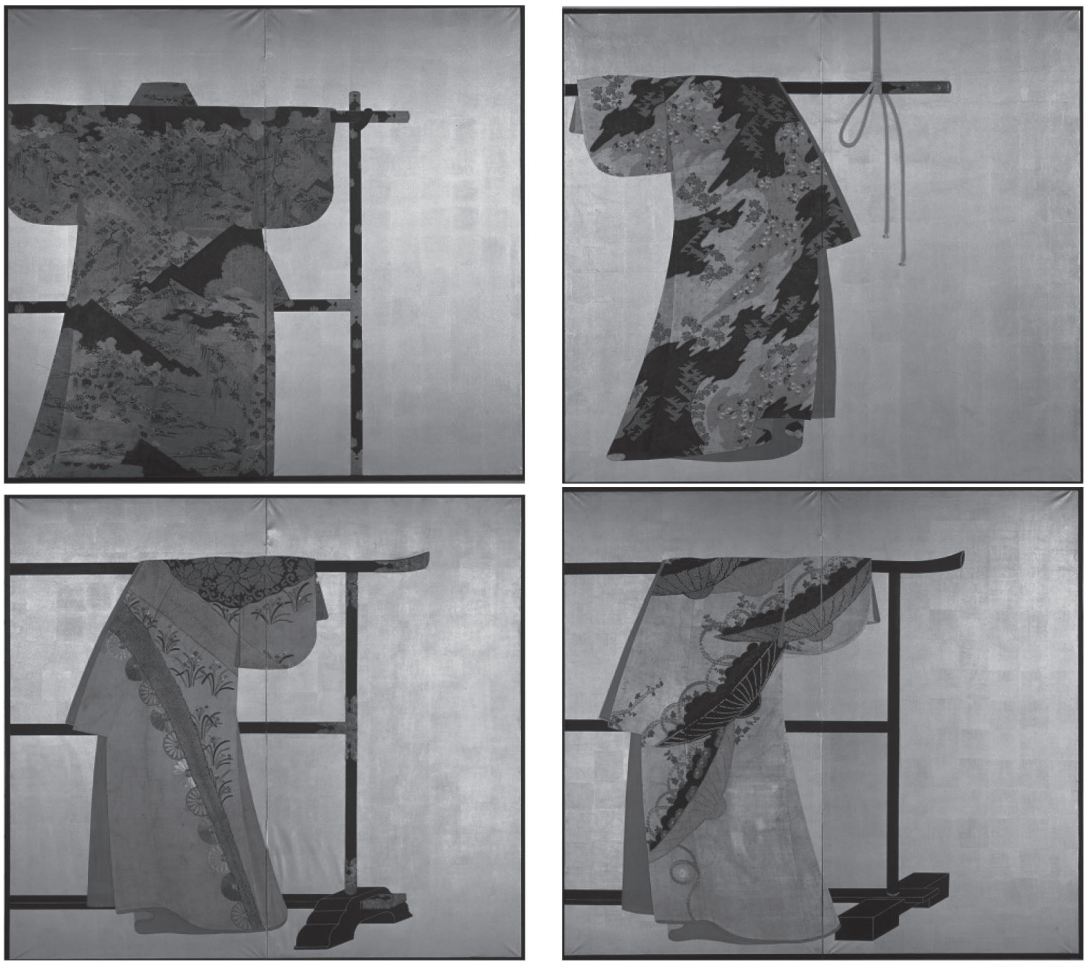


図 5.1　Example of kosode byobu

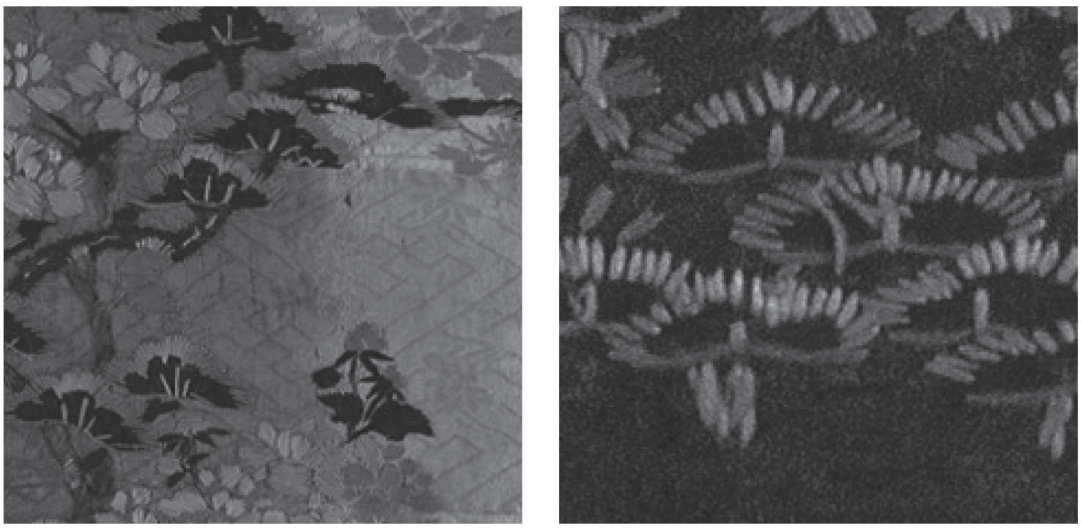


図 5.2　Example of motif（matsu）

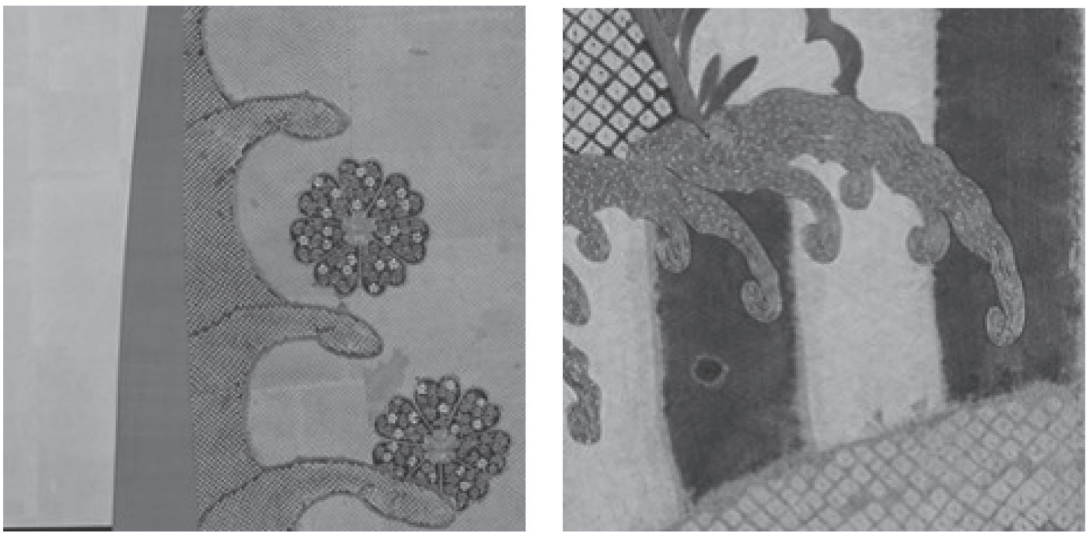


図 5.3　Example of motif（nami）

5.2　ヒストグラムの作成

5.2.1　用いたセグメント

用いたセグメントの詳細は表 5.1 の通りである。小袖屏風画像を分割した画像の中から，正方形になっていない端の部分を除いたセグメントを 227 × 227 にリサイズしている。

ただし 6 分割と 20 分割の画像に対しては正方形になっていない画像でも分割数が少ないためにその画像が占める割合が大きいのでその画像も 227 × 227 と強制的に正方形にしている。

表 5.1　Segment specifications

the number of segmengs/kosode byobu	the number of used segmengs/kosode byobu	percentage of each segment in a kosode byobu image	sum of the number of segments
6	4	22.16%	414
20	15	5.54%	1,558
80	63	1.39%	6,561
320	285	0.35%	29,659

5.2.2　Deep Learning による特徴抽出

DeCAF を抽出する前に学習させておく大規模データセットには一般物体認識のコンテスト［5］に用いられる ImageNet を利用した。ImageNet は 120 万枚の訓練画像, 1,000 カテゴリからなるデータセットであり，自然物，人工物の多種多様な画像を含んでいるため，植物から船まで幅広く描かれている小袖屏風画像にも有用な特徴抽出が期待できるデータセットである。ネットワークはライブラリ Caffe［12］にて提供されている ImageNet のデータセットをすでに学習したネットワークを使用した。ネットワークの構成を表 5.2 に示す。

表5.2　Structure of network

layer	map size	function
Input	227 × 227 × 3	
Conv 1	55 × 55 × 96	ReL
Pool 1	27 × 27 × 96	
Conv 2	27 × 27 × 256	ReL
Pool 2	13 × 13 × 256	
Conv 3	13 × 13 × 384	ReL
Conv 4	13 × 13 × 384	ReL
Conv 5	13 × 13 × 256	ReL
Pool 5	6 × 6 × 256	
Full Connected 6	1 × 1 × 4,096 (DeCAF)	ReL
Full Connected 7	1 × 1 × 4,096	ReL
Full Connected 8	1 × 1 × 1,000	softmax

5.2.3　クラスタリング

小袖屏風画像のセグメントを用いて階層的クラスタリングを行った結果を図5.4に示す。図5.4から分かるように小袖屏風画像のセグメントは大きく2つのクラスタに分かれており，それぞれのクラスタ間で類似度に大きな差があることが分かる。左のクラスタについてクラスタに属するセグメントを出力すると，小袖屏風の屏風部分の金箔の部分などのモチーフに関係のないセグメントにより構成されたクラスタとなっていた（図5.5)。したがって，左のクラスタのセグメントは扱わず，以降では右のクラスタについて扱っていくこととする。

5.2.3.1　クラスタ数の決定

クラスタ数はクラスタ数を決定する手法であるelbow methodを参考に，距離に対してのクラスタ数をプロットし（図5.6)，増加の仕方が変わる点を距離の閾値の目安とした。このグラフから，距離が2,500前後で大きくクラスタ数の増減が変化することが分かる。また，クラスタ数の増減が大きく変わる箇所について拡大したグラフ（図5.7）を見ると，そのエルボー付近に相当する距離2,500におけるクラスタ数は180であることが分かる。この結果より，クラスタ数は180に決定した。

5.2.4　クラスタリング結果の定量評価

画像特徴を基にクラスタリングした結果が人の直観に沿っているかを評価するために，ラベル付データを使用し各クラスタに含まれるモチーフの割合を調べた。データは国立歴史民俗博物館で小袖屏風の研究をしている専門家に小袖屏風のどの領域にどのモチーフが存在しているかラベルをつけてもらい，その情報とセグメントを対応付けた。専門家からのラベル付データは一部ラベルの付与が省略されており小袖屏風中の全てのモチーフを網羅したものが得られなかったため，分母はクラスタに属するセグメント数ではなく各クラスタに属するラベル付けされたセグメント数として割合を算出した。

5.2.4.1　モチーフ領域に関係のある矩形領域の選択

専門家によるラベリングデータはモチーフを取り囲む多角形の頂点の座標情報となっている。モ

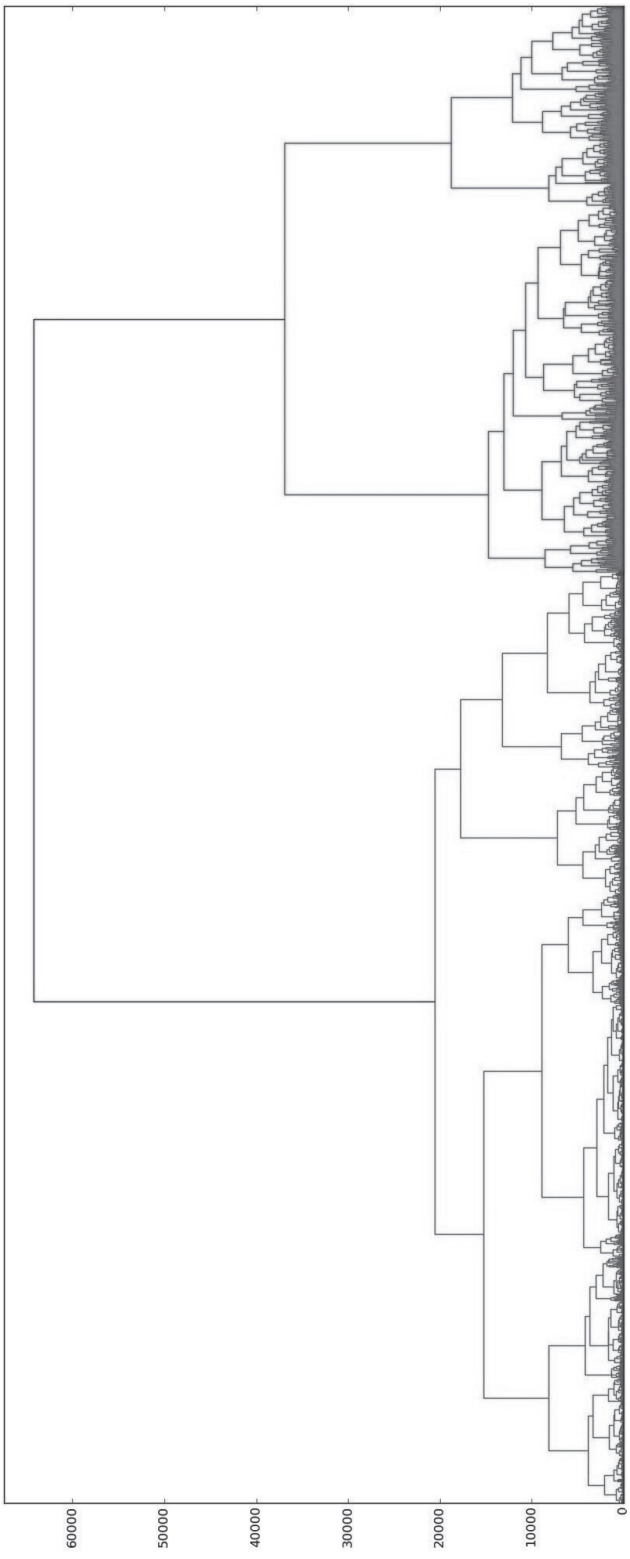


図 5.4　Dendrogram of kosode images

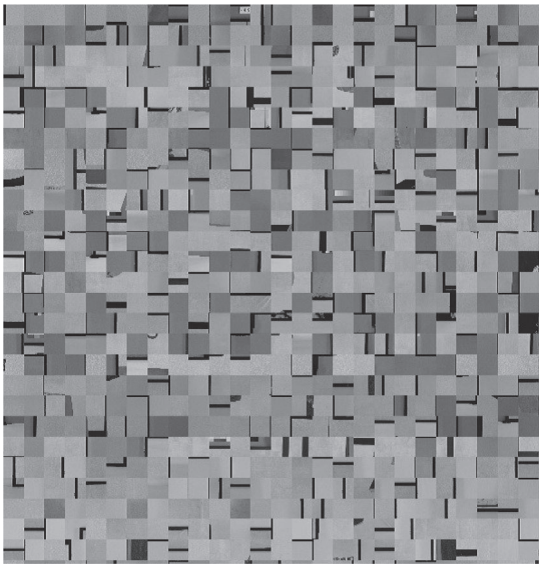


図 5.5　Segments of left cluster

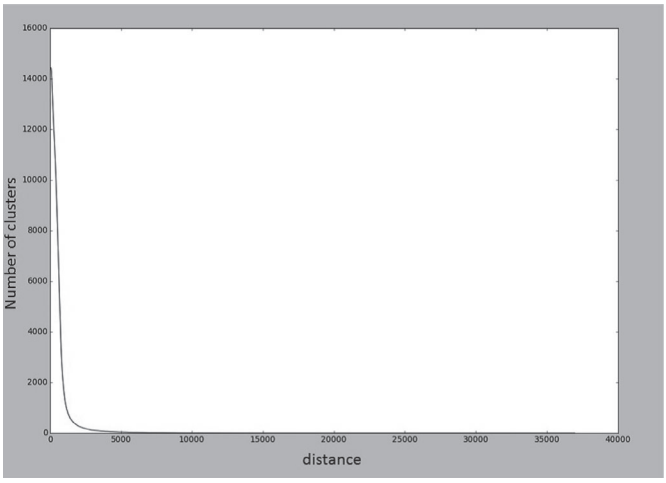


図 5.6　Graph shows the number of clusters for distance

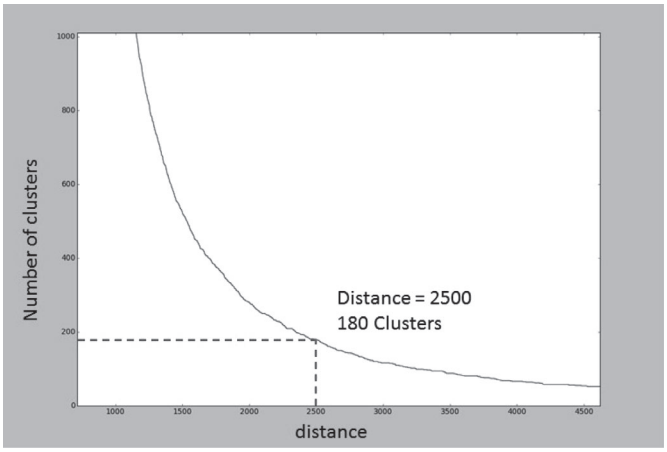


図 5.7　Zoomed graph（Fig.5.6）

チーフに外接する矩形領域（図 5.8）に重なったセグメント画像をモチーフ領域に関係のある画像として選択（図 5.9）し，画素単位での重複度を算出する。重複度が 50% 以上，つまりモチーフ画像中の画素がセグメント画像の画素の半分以上重複（図 5.10）していれば，モチーフ画像のラベルを割り振ることとした。

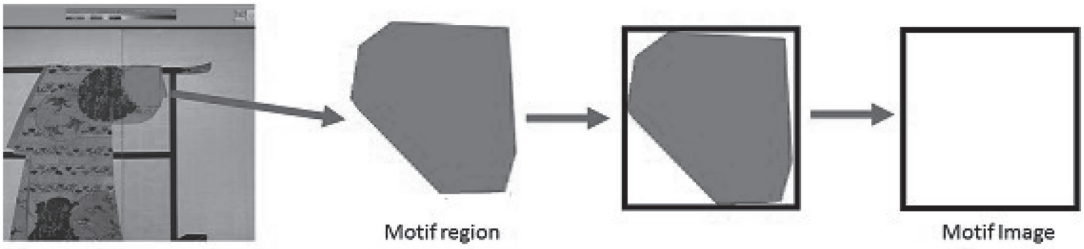


図 5.8　Motif image

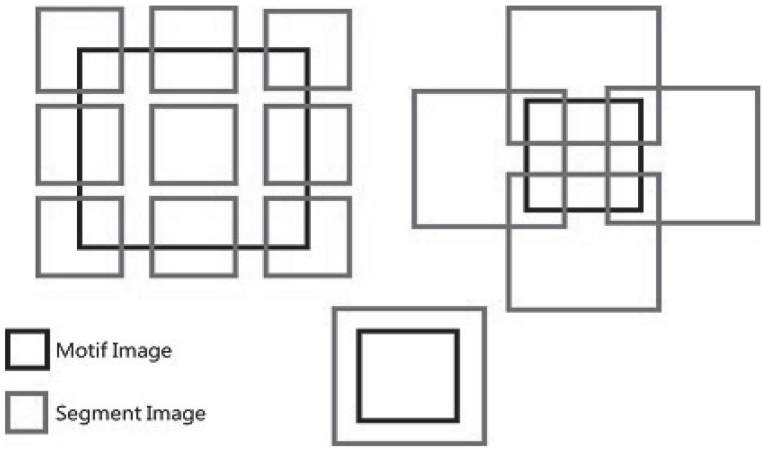


図 5.9　Selecting image related with motif region

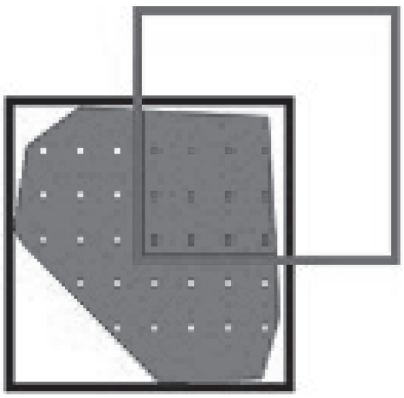


図 5.10　labeling to a motif image

5.2.4.2　クラスタリング結果の定量評価の結果

クラスタリング結果の定量評価の結果を表 5.3 に示す。モチーフは各クラスタで割合が多い順に上位 3 件まで出力している。

表5.3　Motif per cluster

Cluster No.	Motif name					
1	no motif					
2	no motif					
3	no motif					
4	no motif					
5	no motif					
6	no motif					
7	no motif					
8	no motif					
9	tsudumi	1.75%				
10	no motif					
11	ohmihakkei	1.75%				
12	no motif					
13	no motif					
14	no motif					
15	kaede	2.33%	tsuta	1.16%	ro	1.16%
16	kaede	3.49%	sakura	2.91%	kiku	1.74%
17	tanzaku	88.89%	sakura	36.11%		
18	yatsuhashi	3.45%	ume	2.87%	yamabuki	2.30%
19	ohmihakkei	39.23%	byobu	0.77%	toukaidougojuusantsugi	0.77%
20	kaede	53.57%	sidareyanagi	8.93%	ashi	1.79%
21	sidareyanagi	16.67%	sohshi	11.11%	matsu	7.41%
22	sakura	2.09%	takegaki	1.05%	takaka	0.52%
23	takegaki	19.40%	moji	19.40%	manmaku	4.48%
24	ume	3.75%	moji	3.75%	yukiwa	2.50%
25	takegaki	100.00%	moji	3.70%		
26	yatsuhashi	9.60%	takegaki	5.65%	kaki	4.52%
27	senmen	1.48%	hashi	1.48%	byobu	1.48%
28	kui	2.60%	kaede	1.30%	takaka	1.30%
29	manmaku	0.52%	botan	0.52%		
30	moji	3.23%	nami	1.61%	kaede	1.61%
31	no motif					
32	syougikoma	4.17%				
33	no motif					
34	no motif					
35	no motif					
36	manmaku	1.68%	ume	1.12%	jakago	0.56%
37	juunishiban	26.67%	shidareyanagi	1.67%		
38	ume	4.79%	kiri	4.79%	hagi	3.59%
39	yatsuhashi	3.64%	moji	1.82%		
40	seigaiha	30.56%	taki	27.78%	kiku	8.33%
41	moji	12.12%	ran	3.03%		
42	ran	6.84%	waha	2.56%	ume	1.71%
43	yatsuhashi	4.62%	iwa	1.73%	nami	1.73%
44	no motif					
45	no motif					
46	ami	28.57%	nami	11.43%		
47	no motif					
48	no motif					
49	no motif					
50	no motif					
51	no motif					
52	no motif					
53	no motif					
54	no motif					
55	no motif					
56	no motif					
57	no motif					
58	no motif					
59	no motif					
60	no motif					

Cluster No.	Motif name					
61	no motif					
62	keshi	1.79%				
63	no motif					
64	jakago	8.49%	nami	8.49%	sotetsu	2.83%
65	jakago	7.14%				
66	no motif					
67	kui	2.03%	ume	0.51%	kaede	0.51%
68	ohmihakkei	6.67%	yukiwa	2.13%	kaede	1.87%
69	kiku	5.08%	waha	3.39%	yukiwa	3.39%
70	kiku	1.23%	senmen	1.23%	ami	0.62%
71	no motif					
72	moji	0.80%	yatsuhashi	0.80%	matsu	0.80%
73	no motif					
74	no motif					
75	juunishiban	0.88%				
76	no motif					
77	no motif					
78	ro	2.78%	yamabuki	2.78%	hashi	1.39%
79	manmaku	2.53%	aboshi	2.02%	senmen	1.52%
80	hansen	1.98%	tsuta	0.99%	jakago	0.99%
81	senmen	0.93%	matsu	0.93%	takegaki	0.93%
82	no motif					
83	no motif					
84	yukiwa	0.55%	kaki	0.55%		
85	no motif					
86	no motif					
87	no motif					
88	no motif					
89	no motif					
90	yukiwa	0.81%				
91	tanzaku	0.29%	kiku	0.29%	ohmihakkei	0.29%
92	tanzaku	1.39%				
93	no motif					
94	hashi	5.13%	moji	2.56%	iwa	0.85%
95	kaede	0.73%	yukiwa	0.73%	yatsuhashi	0.37%
96	ran	3.28%	yatsuhashi	3.28%	senmen	3.28%
97	kiku	13.86%	ume	7.83%	iwa	6.02%
98	yatsuhashi	30.91%	yukiwa	20.00%	waha	9.09%
99	nami	11.28%	ro	6.39%	katawaguruma	6.02%
100	seigaiha	100.00%				
101	kiku	33.33%				
102	yamabuki	21.85%	botan	17.65%	ume	9.24%
103	karamatsu	16.06%	senmen	13.14%	yukiwa	11.68%
104	kiku	33.33%	hashi	20.00%	kikkou	4.44%
105	seigaiha	100.00%				
106	taki	2.63%				
107	sotetsu	58.54%	kaede	2.44%	nami	2.44%
108	taki	30.30%	kui	30.30%	seigaiha	6.06%
109	takaka	24.19%	katawaguruma	10.48%	kakitsubata	8.87%
110	nami	10.29%	matsu	8.82%	jakago	5.88%
111	taki	8.14%	kui	6.98%	botan	5.81%
112	yatsuhashi	9.89%	yamabuki	6.59%	ume	5.49%
113	yukiwa	14.94%	moji	14.94%	karamatsu	9.20%
114	taki	21.05%	katawaguruma	13.16%	ume	13.16%
115	takaka	10.32%	shiorido	8.73%	kiku	7.14%
116	ume	2.70%				
117	tsudumi	51.67%	yatsuhashi	3.33%	taki	1.67%
118	no motif					
119	taki	40.74%	tsudumi	29.63%	bohoku	18.52%
120	no motif					
121	no motif					

Cluster No.	Motif name					
122	hagi	14.29%				
123	kiri	50.00%				
124	yukiwa	21.05%	jakago	19.30%	moji	19.30%
125	jakago	37.50%	moji	33.33%	kiri	12.50%
126	uchiwa	8.00%	kiku	1.33%	haneuchiwa	1.33%
127	juunishiban	47.83%	kiri	21.74%		
128	juunishiban	47.46%	hanaguruma	3.39%	aboshi	1.69%
129	no motif					
130	no motif					
131	tsudumi	100.00%				
132	no motif					
133	ume	44.44%	waha	7.41%	kaki	7.41%
134	huji	37.84%	hagi	10.81%	mushikago	10.81%
135	yamabuki	50.00%	moji	16.67%	manmaku	7.41%
136	sakura	80.49%	ran	2.44%		
137	karamatsu	73.91%	takegaki	56.52%	iwa	8.70%
138	hagi	97.22%	kaki	66.67%		
139	kaki	100.00%	hagi	43.75%		
140	tsudumi	73.47%	ume	14.29%	ran	6.12%
141	ume	85.71%	kaki	39.29%	moji	7.14%
142	kaki	15.91%	hagi	13.64%	matsu	11.36%
143	yamabuki	26.53%	hansen	12.24%	noshi	10.20%
144	yamabuki	87.50%	sakura	12.50%		
145	matsu	13.89%	yamabuki	13.89%	huji	8.33%
146	ryusui	2.17%				
147	ume	97.10%	sakura	1.45%		
148	ume	12.90%	botan	3.23%	kiku	1.61%
149	sakura	3.33%				
150	yukiwa	2.47%	tsuru	1.23%		
151	no motif					
152	iwa	3.41%	kikukarakusa	2.27%	kiku	1.14%
153	ume	12.86%	huji	2.86%	matsu	2.86%
154	ran	0.81%				
155	ume	58.06%				
156	yukiwa	9.84%	ume	8.20%	ran	4.92%
157	yamabuki	66.67%				
158	yukiwa	10.34%	kiku	6.90%	senmen	6.90%
159	senmen	30.26%	ran	7.89%	kiku	3.95%
160	kaede	88.89%	tsuta	13.89%		
161	ryuusui	4.94%	kaede	2.47%	kiku	2.47%
162	kaede	52.83%	sakura	20.75%	tanzaku	5.66%
163	kaede	53.85%	tsuta	12.09%	suisen	7.69%
164	ashi	4.29%	sakura	4.29%	hansen	4.29%
165	byobu	46.05%	toukaidougojuusantsugi	15.79%	toukaidougojuusantsugi	9.21%
166	kiku	66.67%	magaki	23.08%	sakura	7.69%
167	kiku	35.71%	ichou	5.36%		
168	shidarezakura	91.23%	manmaku	68.42%	kiku	5.26%
169	sakura	100.00%	tanzaku	35.71%		
170	karamatsu	42.11%	kiku	10.53%	iwa	5.26%
171	jakago	50.00%				
172	kiku	73.68%	magaki	31.58%		
173	ume	28.30%	kikukarakusa	9.43%	karamatsu	7.55%
174	kiku	20.27%	chidori	6.76%	ume	5.41%
175	kiku	7.45%	kaede	2.13%	yukinoshita	2.13%
176	kiku	11.35%	matsu	6.38%	sakura	5.67%
177	kiku	8.51%	sakura	6.38%	kaede	4.26%
178	matsu	42.42%	ume	3.03%		
179	kiku	4.46%	kaede	2.68%	toukaidougojuusantsugi	2.68%
180	yuri	11.11%	kiku	9.40%	tanzaku	4.27%

5.2.4.3　得られたクラスタの例

得られたクラスタの例を示す。図 5.11 は枝垂桜のモチーフを多く含むクラスタとなっている。図 5.11 は単一の小袖屏風から得られたセグメントであるが，そのクラスタから最小距離の関係にあるクラスタ（図 5.12）も桜を含むクラスタとなっており，モチーフを反映した特徴を抽出出来ているためにクラスタ間の類似関係を反映したデンドログラムの構築が成功している例といえる。

図 5.11　cluster 168

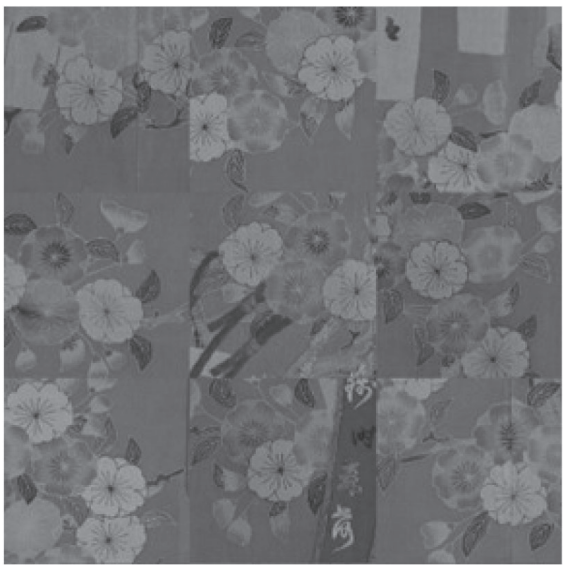


図 5.12　cluster 169（the nearest cluster to cluster 168）

続いてモチーフを反映したクラスタとならなかった例を示す。図 5.13 は，見た目は似ているが意味としての判別が正しく行えていなかったケースである。セグメントを見ると，文字や波のモチーフの一部，さらに小袖をつるす道具である衣桁の一部のセグメントが同じクラスタとして現れていた。このようなケースは見た目が似ているが故に機械的に意味を判別することが難しく，今後の課題といえる。

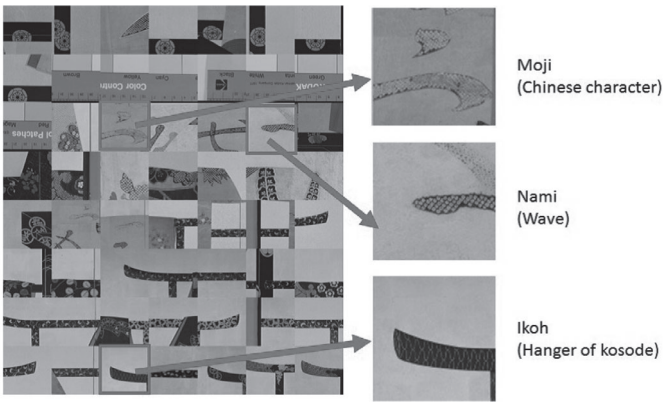


図 5.13　cluster 30

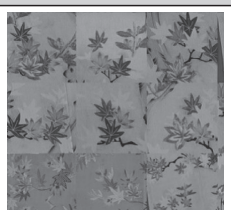
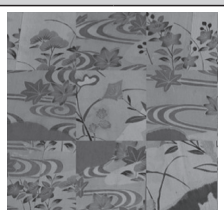
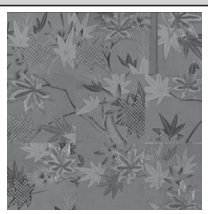
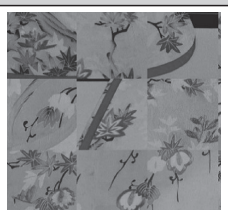
5.3　従来手法との比較

提案手法と従来手法とで比較を行った。従来手法では SIFT を用いて BoF を行い，ヒストグラムの比較には histogram intersection を用いた。各小袖屏風画像のサイズは 1,951 × 1,551 とし，各画像から 1,000 個の SIFT 特徴点を抽出して k-means によるクラスタリングにより 100 個のクラスタに分割した。結果の例を表 5.4 に示す。従来手法ではクエリ画像と共通のモチーフをもつ小袖屏風画像を出力できていないのに対し，提案手法では「楓」のモチーフを共通してもつ小袖屏風画像を出力できていることが分かる。提案手法で検索結果上位の小袖屏風画像が属しているクラスタの中から一部を表 5.5 に示す。表 5.5 を見ると，「楓」を多く含むクラスタが形成されているために検索結果において類似した小袖屏風画像を上位に出力できたものと思われる。

表 5.4　Example of search results by conventional and proposed method

	Rank 1 (query)	Rank 2	Rank 3
Conventional			
motif	left: aboshi, shidarezakura, nami, manmaku right: ashi, oshidori, kaede, kiku, nami	taki, bohoku, moji, yamabuki	kakine, kiku, sohshi, nanten
Proposed			
motif	left: aboshi, shidarezakura, nami, manmaku right: ashi, oshidori, kaede, kiku, nami	kaede, tsuta	left: kaede, kikyoh, susuki, hujibakama, ryuusui right: kikyoh, kiku, keshi, susuki, chidori, hagi, yuri

表 5.5　Example of clusters and segments

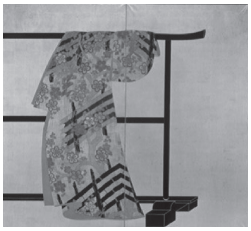
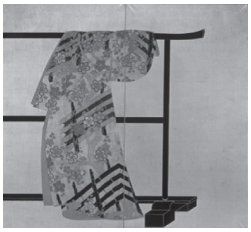
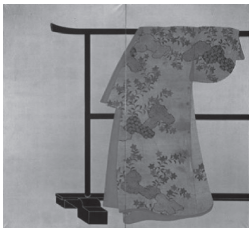
Cluster 160		Cluster 161		Cluster 162		Cluster 163	
							
Evaluation with labeled data by an expert							
motif	*percentage*	*motif*	*percentage*	*motif*	*percentage*	*motif*	*percentage*
kaede	88.89%	ryuusui	4.94%	kaede	52.83%	kaede	53.85%
tsuta	13.89%	kaede	2.47%	sakura	20.75%	tsuta	12.09%

5.4　ヒストグラム距離算出法の比較

クラスタ間の関係性を考慮した EMD によるヒストグラム比較の有効性を確認するために，階層的クラスタリングによって作成した同じヒストグラムを従来手法（histogram intersection）と提案手法（クラスタ間の関係性を考慮した EMD）とで比較を行った。結果の例を表 5.6 に示す。表から，従来手法ではクエリ画像に共通するモチーフを出力できていないのに対し，提案手法では「垣」のモチーフをもつクエリ画像に対して，最も似ているものとして「竹垣」のモチーフをもつ小袖屏風画像を出力できていた。「垣」と「竹垣」は，共に垣という意味で共通したモチーフであり，妥当な結果が得られているといえる。提案手法で得られた「垣」と「竹垣」のモチーフを含む小袖屏風画像のセグメントから得られたクラスタの例を表 5.7 に示す。これを見ると，cluster 23 には竹垣のモチーフ，cluster 26 には竹垣と垣のモチーフが多く含まれていることが分かる。

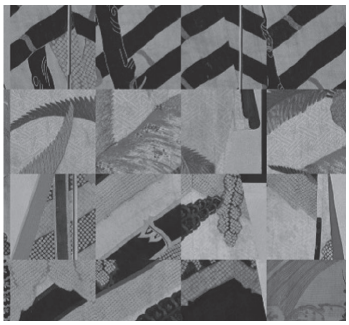
「垣」と「竹垣」のモチーフを含む小袖屏風画像について，180 あるクラスタの中から cluster 23

表 5.6　Example of search results by conventional and proposed method

	Rank 1 (query)	Rank 2	Rank 3
Conventional			
motif	ume, kaki, moji	iwa, kaede, hashi	kashiwaba, takaka, takanoha
Proposed			
motif	ume, kaki, moji	takegaki, budoh	iwa, kikukarakusa, yuri

と cluster 26 のある箇所を一部抜き出してヒストグラムを示したのが図 5.14 である。これらはクラスタの分布に共通点が少なく，image 68 は cluster 26 のセグメントを多く含んでいるのに対し，image 70 は cluster 23 を多く含んでいる。cluster 23 と cluster 26 はクラスタ番号は違うが，表 5.7 に示したように，共に垣の意味で共通したモチーフを含むクラスタであり，これらは類似したクラスタである。ビンとビンの対応関係のみを見る histogram intersection では，これらは類似性がないものとみなされるが，EMD による距離計算によって距離が小さく算出されたためにこのようなヒストグラムをもつ画像同士でも類似画像として出力されたと考えられる。

表 5.7　Example of clusters and segment

Cluster 23		Cluster 26	
			
Evaluation with labeled data by an expert			
motif	*percentage*	*motif*	*percentage*
takegaki	19.4%	yatsuhashi	9.6%
moji	19.4%	takegaki	5.65%
manmaku	4.48%	kaki	4.52%

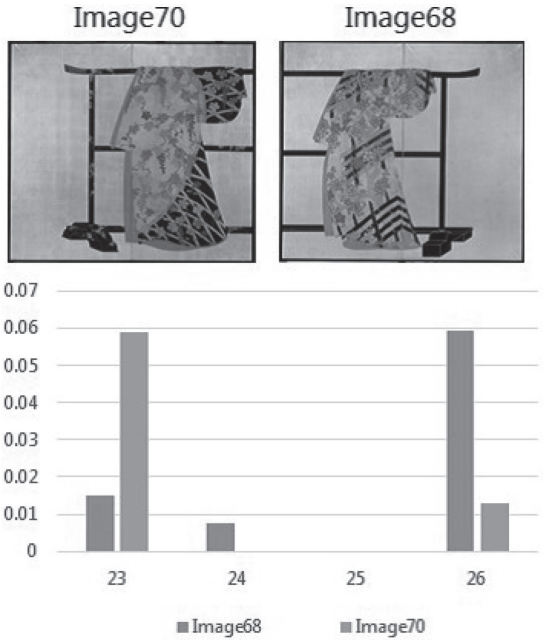


図 5.14　Example of histogram

❻……結論

本論文では，意味情報を反映した画像特徴を取り出し，類似画像検索に有効なメタデータとしてクラスタを取得し，BoF のようにヒストグラムを作成することを提案した。

既存手法では濃淡情報を基に特徴を記述しているため，画像のもつ意味情報を反映できていなかった。

一方で，提案手法では画像をセグメントに分割し，Deep Learning によって各セグメントから得られる画像特徴を基にヒストグラムを作ることで，画像のもつ意味情報の分布を表現することが出来るようになった。さらに，EMD 算出のコストにデンドログラムによって表現されるクラスタ間の距離を適用することでクラスタ間の類似性を反映したヒストグラム比較を実現した。これにより，ヒストグラムの比較の際に類似している特徴をもつ画像であっても類似度が小さく出てしまう問題を解消した。実験では，提案手法の有効性を，デジタルアーカイブを用いた検証によって示した。実験の結果から，DeCAF によって同じ意味をもつ画像に対し共通の特徴が得られ，かつクラスタ間の距離を反映した EMD によるヒストグラムの比較によって共通の対象が描かれている画像を類似画像として得られることが示された。今後の展望としては，見た目が似ているものの異なる意味の判別や，適切なセグメントの分割方法の検討などが挙げられる。

参考文献

［1］総務省：“知のデジタルアーカイブ―社会の知識インフラの拡充に向けて―提言”，2012，http://www.soumu.go.jp/main_content/000156248.pdf（参照 2015-12-15）.

［2］藤吉弘亘ほか：“コンピュータビジョン最先端ガイド2”，アドコム・メディア株式会社，pp.5–11，2010.

［3］D. Lowe: “Distinctive Image Features from Scale-Invariant Keypoints”, International Journal of Computer Vision Volume 60, Issue 2, pp.91–110, 2004.

［4］J. Sivic and A. Zisserman: “Video Google: A Text Retrieval Approach to Object Matching in Video”, Proc. The International Conference on Computer Vision, Volume 2, pp.1470–1477, 2003.

［5］ImageNet: “Imagenet Large Scale Visual Recognition Challenge（ILSVRC）”, http://image-net.org/challenges/LSVRC/（参照 2015-12-21）.

［6］A. Krizhevsky, I. Sutskever, G. E. Hinton: “ImageNet Classification with Deep Convolutional Neural Networks”, Neural Information Processing Systems, 2012.

［7］岡谷貴之：“機械学習プロフェッショナルシリーズ深層学習”，講談社，pp.10–11，2015.

［8］J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, T. Darrell: “DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition,” Proceedings of The 31st International Conference on Machine Learning, pp.647–655, 2014.

［9］国立歴史民俗博物館，国立歴史民俗博物館資料図録 2　野村コレクション　小袖屏風，2002.

［10］Y. Rubner, C. Tomasi and L. J. Guibas: “The Earth Mover’s Distance as a Metric for Image Retrieval”, International Journal of Computer Vision, Volume 40, Issue 2, pp.99–121, 2000.

［11］Y. Jia, E. Shelhamer: “Caffe” http://caffe.berkeleyvision.org/（参照 2016-2-3）.

［12］国立歴史民俗博物館：“データベースれきはく館蔵野村正治郎衣裳コレクション” http://www.rekihaku.ac.jp/doc/t-db-index.html（参照 2016-2-8）.

［13］国立文化財機構：“e 国宝―国立博物館所蔵 国宝・重要文化財” http://www.emuseum.jp/（参照 2019-7-16）.
［14］Cultural Institute:“アートプロジェクト” https://www.google.com/intl/ja/culturalinstitute/about/artproject/（参照 2019-7-16）.

（横浜国立大学大学院工学研究院，国立歴史民俗博物館共同研究員）

（2019 年 3 月 14 日受付，2019 年 8 月 5 日審査終了）

Metadata Generation Using Deep-Learning in Digital Archives

HAMAGAMI Tomoki

The goal of this study is to generate "metadata" of the images in digital archives and to use them for similar image retrieval system. Generally, bag-of-features [4], which is a histogram representation method for object recognition in images, uses feature points and local features calculated based on the grayscale distribution of pixels by using SIFT [3]. On the other hand, the deep learning approach [6] has attracted attention in the field of general object recognition as end-to-end learning, not local features oriented learning. The typical deep learning recognizes the whole structured object in the image; however, it misses significant sub-parts in the image. To overcome the issue, we divide an image into segments, extract features from each segment using deep learning, then apply bag-of-features using the clustering for the local features, and finally represents it as histogram expression reflects the metadata in the image. Furthermore, by using distance calculation representing the similarity of the cluster as comparing the histograms, the discriminant precision of a similar image could be improved, when the number of clusters is too small. The experiment result shows the effectiveness of generating metadata using BoF with deep learning, the distance evaluation method reflecting the relationship between clusters.

Key words: deep learning, general object recognition, bug of feature, similar image retrieval