情報処理学会論文誌　Vol. 52　No. 12　3496–3503 (Dec. 2011)

画像・長文からの潜在空間獲得による画像間類似度の改善

牛久祥孝[†1]　原田達也[†1,†2]　國吉康夫[†1]

大規模画像の効率的な利用を目指し，入力画像と扱う事物が類似した画像を検索する類似画像検索が多く研究されている．画像間類似度が重要となるが，従来の手法の多くでは画像の見た目の類似度に基づいて画像を検索するため，異なる事物を扱う画像でも類似する画像となる．本論文は，画像に文章が付随する場合に，その文章の傾向を利用して画像間類似度を改善する手法を提案する．提案手法は，文章が数百単語の長文でもよく，文章が画像のごく一部に付随する状況でも画像間類似度を改善でき，スケーラビリティにも優れている．実験では一部の画像に文章が付随したデータセットを複数用い，それぞれで提案手法が既存手法に対してより高精度に類似画像を検索できることを示す．

Improving Image Similarity Measures through Latent Space Learning between Images and Long Texts

Yoshitaka Ushiku,[†1] Tatsuya Harada[†1,†2] and Yasuo Kuniyoshi[†1]

To manage increasing multimedia data, methods for similar image retrieval are widely studied. The similarity measure of images is essential for the search. In this paper, we propose a method to improve the similarity by considering texts around images. Proposed method can improve image similarity measures based on the latent semantics obtained from the pairs of images and texts. It is notable that those texts need not be some clear tags and that long texts are applicable. Moreover, our method can improve the similarities effectively even if little portion of images has texts. Moreover, proposed method is scalabe because its computational complexity is linear on the data amount. In the experiments, we compare our method with previous methods using some datasets in which a portion of the images are annotated by texts. We show that our method can retrieve semantically similar images more precisely than existing methods.

1. はじめに

情報技術の進歩にともない，画像や音楽，動画などのマルチメディアを個人が容易に扱えるようになった．結果，そのようなデータが個人のデバイスやWeb上に大量に存在するようになり，現在も飛躍的に増大し続けている．

大量のデータを効率的に扱うためには，ユーザが所望のデータに容易にアクセスできる検索技術が必要不可欠である．画像を対象として，内容に基づいた検索を目指す研究がContent-based Image Retrieval（CBIR）として広く行われている[1)]．検索時にクエリとなるデータの形態は大きく分けて2つ存在する．1つは現在の検索エンジンでも一般的な，複数の独立した単語である．もう1つは，所望のデータと同様の事物を扱っている画像である．

例となる画像から所望の画像を検索する類似画像検索では，入力画像が扱う事物と類似した事物を扱う画像を検索する精度が重要である．ここで問題となるのが，画像のアピアランスの類似度と，それらが扱う事物の類似度の間に存在するギャップである．これはCBIRの枠組み[1)]でもセマンティック・ギャップとしてよく知られている．画像特徴量に基づいて類似画像を検索した場合，写っている事物が異なる画像も多く出力されてしまう．また逆に，扱う事物が同じであるにもかかわらず類似画像として検索されない場合もありうる．

セマンティック・ギャップを解消するためによく行われるのが，画像に付与されたメタデータと画像との関係の学習である．たとえば画像アノテーション・リトリーバルの研究では，画像の内容を表す数個のラベルと画像の関係を学習する．新規画像にも適切なラベルを推定し，ラベルによる画像検索を実現する[1)]．ただし，この枠組みでは学習画像ごとに数個のラベルが必要となる．そしてこのラベル付けは手作業に頼ることになるため，大きな労力をともなう．

ラベル付き画像をWebから自動で集める手法も多く提案されている[2)]．しかし，ラベルなどのメタデータが存在するのはFlickrのような一部の画像投稿サイトに限られ，一般的なWebページには画像が単体で存在するか，その画像にまつわる文章が存在する場合がほとんどである．そして，どの単語が画像のラベルたりうるかを一般の文章の中で自動的に判

†1 東京大学
the University of Tokyo
†2 JST さきがけ
JST PRESTO

断することは困難である．

このように「画像」と「文章つき画像」しかない，いわば教師なしの状況で類似画像検索の精度を向上させるために，文章間類似度を利用する．たとえばニュースサイトや観光旅行の情報ページなどでは，掲載されている画像の周辺に関連する文章が存在しうる．同種の事物を扱う画像に関する文章は，当然類似した表現を持つと期待できる．画像のアピアランスの類似度にこの文章間の類似度を反映させることで画像間類似度を改善できると考える．

本研究では，より精度良く同様の事物を扱う画像を検索するために，文章と画像から事物に基づく潜在的な分布を学習する手法を提案する．提案手法では，「画像」と「文章つき画像」からなるデータを対象に，「文章つき画像」から画像間類似度を学習する．文章が付随しない「画像」も含めた全画像に学習結果を適用し，「画像」と「文章つき画像」全体での類似画像検索精度を改善する．

2. 関 連 研 究

本研究の目的は，高精度な類似画像検索のための類似度学習手法の提案である．類似度学習の従来研究は，その他の分野の機械学習手法と同様，教師ありと教師なしの 2 つに分類できる．教師ありの枠組みでは，画像などのデータのカテゴリ情報をもとに，特徴量間の距離を所望の距離へと変換する．Large Marigin Nearest Neighbour[3] では，各データについて同カテゴリのデータは近く，異なるカテゴリのデータは遠くなるような線形変換を学習している．Polynominal Semantic Indexing[4] では，やはりカテゴリ情報を用いてクエリと同カテゴリのデータだけが近くなるような距離関数を多項式近似で求めている．ただし，本研究が対象とする Web の画像データでは，1 章で述べたとおり「画像」か「文章つき画像」しか存在せず，これらの教師付き学習は使えない．

教師なし学習に基づく従来手法は，画像の低レベルな特徴記述を用いて画像間の類似度を計量するものがほとんどである．文献 5) では画像の特徴量に対して主成分分析（PCA）を用いて次元圧縮を行い，類似画像を検索している．PCA は特徴量次元と同じ次元数の一般化固有値問題を解くだけでよいので，学習が高速に行えるという利点を持つ．しかし一般に特徴量空間上では，それが表す事物について非線形な分布が存在すると考えられ，検索の精度には限界がある．このような非線形なデータ構造を学習するための手法として，カーネル法は頻繁に用いられる手法の 1 つである．実際に，文献 6) ではカーネル正準相関分析を用いたスペクトラルクラスタリングを経て，文章つき画像データセットの特徴量分布を改善する手法を提案している．また，この非線形な分布を低次元の多様体ととらえ，測地線距離に基づくグラフ構造を求めた後に次元圧縮を行う手法もある．特徴量空間上でグラフ構造を生成したうえで MDS を適用する ISOMAP[7] や，近傍のデータを利用して次元圧縮を行う Stochastic Neighbour embedding（SNE）[8] や Local Linear Embedding（LLE）[9] を適用した ISOSNE[10] または ISOLLE[11] を利用する手法[12] などがこれにあたる．しかしこのような非線形な距離関数を学習する手法は，一般的にデータ数に対する計算量が問題となる．

以上から，本研究が対象とする類似画像検索の要求機能として以下の 2 点があげられる．

精度　入力画像と類似した事物を扱う画像がより多く，より近い画像として出力される．

高い拡張性　数百万などといった大規模な画像を検索可能にするためには．データ数に対するスケーラビリティは重要な要素である．

本研究では，検索精度を向上させる手立てとして画像に付随する文章に注目する．1 章で述べたとおり，特に Web では画像の一部に文章が付随すると考えられる．そこで，この文章と画像から抽出した低レベルな特徴量から，潜在的な意味に基づく空間を学習する．この空間上では類似した事物に関する画像は互いに近くに位置するようになり，内容が類似した画像をより効率的に検索できる．画像と単語の関係を学習する画像アノテーション・リトリーバルの分野では，画像の特徴量と付与されたメタデータのセマンティックギャップを解消する手法が多く研究され，そのアプローチは大きくわけて 2 つに分類される．1 つは画像とラベルの関係を直接表現する手法であり，もう 1 つは潜在変数を仮定し，画像と潜在変数，ラベルと潜在変数の関係を学習する手法である．画像とラベルの間の直接的な確率分布は本来大変複雑なものであると考えられ，それ自体を推定するためには混合正規分布のあてはめなどの高コストな計算を必要とする．一方潜在変数を用いるモデルでは，潜在変数と画像，潜在変数とラベルの関係を独立と仮定することで，比較的少ない計算量での学習・認識とリーズナブルな認識性能を実現している[13]．文献 14) では，画像アノテーションを目的として，入力画像と各訓練データ間の距離をそれぞれの内容の近さで定義する Canonical Contextual Distance（CCD）を提案している．この研究では確率的正準相関分析と呼ばれる多変量解析手法を用いて，画像と単語を等価に扱いながら潜在変数を確率的に求めている．この距離の学習はデータ数に一定であり，アノテーション自体も state-of-the-art な性能を誇る．本研究の類似画像検索でも，画像の見た目（テクスチャや形状など）が重要なときと文章の見た目（各単語の頻度など）が重要なときの両方が存在すると考えられ，このように両特徴量を等価に扱える枠組みが望ましい．そこで本研究は，CCD を画像と文章からなるデータに対応する距離計量へと拡張し，スケーラブルかつ高精度な類似画像検索手法と

して提案する．

3. 文章間類似度による画像間類似度の改善

画像/文章特徴量を $\boldsymbol{x}$，$\boldsymbol{y}$ とし，それぞれ独立同分布から観測されたものと仮定する．ここで，潜在的な内容を表す変数 $\boldsymbol{z}$ を仮定し，この $\boldsymbol{z}$ から画像や文章の表現として $\boldsymbol{x}$ ないし $\boldsymbol{y}$ が互いに独立な正規分布に基づいて出現しているととらえる．

$$\boldsymbol{z} \sim \mathcal{N}(0, I_c), \boldsymbol{x}|\boldsymbol{z} \sim \mathcal{N}(W_x\boldsymbol{z} + \boldsymbol{\mu}_x, \Psi_x), \boldsymbol{y}|\boldsymbol{z} \sim \mathcal{N}(W_y\boldsymbol{z} + \boldsymbol{\mu}_y, \Psi_y). \tag{1}$$

ただし，$\min\{d_x, d_y\} \geq c$，$W_x \in \mathbb{R}^{d_x \times c}$，$W_y \in \mathbb{R}^{d_y \times c}$ である．すると類似画像検索は，入力画像 $\boldsymbol{x}_q$ を条件とした $\boldsymbol{z}$ の確率分布から，訓練データ $\boldsymbol{x}_t$ と $\boldsymbol{y}_t$ のペアを条件とした $\boldsymbol{z}$ の確率分布への確率分布間距離に基づく近傍検索となる．これらの確率分布もまた正規分布となり，$p(\boldsymbol{z}|\boldsymbol{x} = \boldsymbol{x}_q) = \mathcal{N}(\boldsymbol{\mu}_q, \Phi_q)$，$p(\boldsymbol{z}|\boldsymbol{x} = \boldsymbol{x}_t, \boldsymbol{y} = \boldsymbol{y}_t) = \mathcal{N}(\boldsymbol{\mu}_t, \Phi_t)$ と表す．正規分布などの指数型分布族では確率間擬距離として Kullback-Leibler Divergence（KLD）を用いる．$p(\boldsymbol{z}|\boldsymbol{x} = \boldsymbol{x}_q)$ から $p(\boldsymbol{z}|\boldsymbol{x} = \boldsymbol{x}_t, \boldsymbol{y} = \boldsymbol{y}_t)$ への KLD は次のとおりである．

$$\begin{aligned}&\mathrm{KLD}(p(\boldsymbol{z}|\boldsymbol{x} = \boldsymbol{x}_q)||p(\boldsymbol{z}|\boldsymbol{x} = \boldsymbol{x}_t, \boldsymbol{y} = \boldsymbol{y}_t))\\ &= \frac{1}{2}\left\{\mathrm{tr}(\Phi_q\Phi_t^{-1}) - c + \log\det(\Phi_t\Phi_q^{-1}) + \left\|\Phi_t^{-\frac{1}{2}}(\boldsymbol{\mu}_q - \boldsymbol{\mu}_t)\right\|^2\right\}.\end{aligned} \tag{2}$$

Φ_t，Φ_q と c は定数なので，CCD はこれらの定数項を除いて $\mathrm{CCD}(\boldsymbol{x}_t, \boldsymbol{y}_t|\boldsymbol{x}_q) = \|\Phi_t^{-\frac{1}{2}}(\boldsymbol{\mu}_q - \boldsymbol{\mu}_t)\|^2$ と定義する．これは明らかに，訓練データと入力画像の潜在変数をそれぞれ，

$$\boldsymbol{r}_t = \Phi_t^{-\frac{1}{2}}\boldsymbol{\mu}_t, \boldsymbol{r}_q = \Phi_t^{-\frac{1}{2}}\boldsymbol{\mu}_q, \tag{3}$$

としたときの L2 距離による近傍検索と等しい．

以上から，$p(\boldsymbol{z}_t|\boldsymbol{x}_t, \boldsymbol{y}_t)$ の平均 $\boldsymbol{\mu}_t$ と分散 Φ_t，$p(\boldsymbol{z}_q|\boldsymbol{x}_q)$ の平均 $\boldsymbol{\mu}_q$ を推定する問題へと帰着された．まず，画像と文章がペアになっている訓練データ全 n ペアから正規分布のパラメータ W_x，W_y，$\boldsymbol{\mu}_x$，$\boldsymbol{\mu}_y$，Ψ_x，Ψ_y の最尤推定を行う．

$$\begin{aligned}&W_x, W_y, \boldsymbol{\mu}_x, \boldsymbol{\mu}_y, \Psi_x, \Psi_y =\\ &\mathrm{argmin}\left(\frac{(d_x + d_y)n}{2}\log 2\pi + \frac{n}{2}\log|\Sigma| + \frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}\Sigma^{-1}(\boldsymbol{\chi}_i - \boldsymbol{\mu})(\boldsymbol{\chi}_i - \boldsymbol{\mu})^T\right),\end{aligned} \tag{4}$$

$$\boldsymbol{\chi} = \begin{pmatrix}\boldsymbol{x}\\ \boldsymbol{y}\end{pmatrix}, \boldsymbol{\mu} = \begin{pmatrix}\boldsymbol{\mu}_x\\ \boldsymbol{\mu}_y\end{pmatrix}, \Sigma = \begin{pmatrix}W_xW_x^T + \Psi_x & W_xW_y^T\\ W_yW_x^T & W_yW_y^T + \Psi_y\end{pmatrix}. \tag{5}$$

この最尤推定は，確率的正準相関分析（pCCA: Probabilistic Canonical Correlation Analysis）[15] で行う最尤推定と同一のものである．この推定結果から，

$$\boldsymbol{\mu}_q = M_x^TA^T(\boldsymbol{x}_q - \bar{\boldsymbol{x}}), \Phi_q = I - M_xM_x^T, \tag{6}$$

$$\boldsymbol{\mu}_t = \begin{pmatrix}M_x\\ M_y\end{pmatrix}^T\begin{pmatrix}(I-\Lambda^2)^{-1} & -(I-\Lambda^2)^{-1}\Lambda\\ -(I-\Lambda^2)^{-1}\Lambda & (I-\Lambda^2)^{-1}\end{pmatrix}\begin{pmatrix}A^T(\boldsymbol{x}_t - \bar{\boldsymbol{x}})\\ B^T(\boldsymbol{y}_t - \bar{\boldsymbol{y}})\end{pmatrix}, \tag{7}$$

$$\Phi_t = I - \begin{pmatrix}M_x\\ M_y\end{pmatrix}^T\begin{pmatrix}(I-\Lambda^2)^{-1} & -(I-\Lambda^2)^{-1}\Lambda\\ -(I-\Lambda^2)^{-1}\Lambda & (I-\Lambda^2)^{-1}\end{pmatrix}\begin{pmatrix}M_x\\ M_y\end{pmatrix}, \tag{8}$$

と，所望の平均および分散を解析的に求められる[⋆1]．$\bar{\boldsymbol{x}} \in \mathbb{R}^{d_x}$，$\bar{\boldsymbol{y}} \in \mathbb{R}^{d_y}$ はそれぞれ特徴量 $\boldsymbol{x}$，$\boldsymbol{y}$ の平均である．また，M_x と M_y は $M_xM_y^T = \Lambda$ かつスペクトルノルムが 1 未満となるような任意の行列である．ここでは，文献 14) と同様に $0 < \beta < 1$ なる β で $M_x = \Lambda^\beta$，$M_y = \Lambda^{1-\beta}$ とする．pCCA は正準相関分析（CCA: Canonical Correlation Analysis）の確率的な解釈である．CCA は画像特徴量ベクトル $\boldsymbol{x}$ と文章特徴量ベクトル $\boldsymbol{y}$ に対し，正準変量 $\boldsymbol{s} = A^T(\boldsymbol{x} - \bar{\boldsymbol{x}})$ と $\boldsymbol{t} = B^T(\boldsymbol{y} - \bar{\boldsymbol{y}})$ の相関を最大にする行列 A と B を求めるもので，以下の一般固有値問題を解けばよい．

$$R_{XY}R_Y^{-1}R_{YX}A = R_XA\Lambda^2, R_{YX}R_X^{-1}R_{XY}B = R_YB\Lambda^2, \tag{9}$$

ここで，Λ^2 は固有値を対角要素とする対角行列で，$R_X = \sum_{i=1}^n(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T$，$R_Y = \sum_{i=1}^n(\boldsymbol{y}_i - \bar{\boldsymbol{y}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^T$，$R_{XY} = \sum_{i=1}^n(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{y}_i - \bar{\boldsymbol{y}})^T = R_{YX}^T$ である．

このように CCA は特徴量次元と同じ次元数の一般化固有値問題を解けばよく，計算量がデータ数に依存しない点でスケーラビリティに優れている．また，CCD は CCA によって得られる 2 つの正準変量 $\boldsymbol{s}$，$\boldsymbol{t}$ それぞれでの距離をうまく融合させた距離計量ともとらえられる．実際に式 (7) は，2 つの正準変量の各次元が，画像と文章の相関の多寡に基づいて重みづけられながら 1 つのベクトルに融合される形になっている．このような確率的アプローチをとらず，CCA をカーネル化した KCCA を用いている研究[6] もあるが，解の安定性や計算量の大きさが問題である．また，KCCA を用いた際の確率的な枠組みは明らかではなく，CCD のような正準相関の低い次元の自動縮小は行えない．一方，確率的に潜在変数をとらえる枠組みとして，確率的潜在意味分析（pLSA）[16] や潜在的ディリクレ配分法（LDA）[17] などもあるが，これらは画像と文章のような 2 つの特徴量間の相関を扱えない．pCCA は

⋆1 ここで，$\boldsymbol{\mu}_t$ と Φ_t の定義の一部で，文献 15) と符号が異なっている点に注意されたい．我々の計算の限りでは，元の文献にある式が誤りである．

特徴量間の相関を陽に扱える点で，類似画像検索性能において優れた性質を持つと考えられる．また計算量にしても，反復計算による最適化が不要で，データ数に対し一定の計算量で大域的最適解が求まる点で pCCA の方が優れている．

全体の流れは以下のようになる．まず，画像と文章がペアになっている全訓練データで CCA を行い，変換行列 A と B，正準相関 Λ を計算する．これらを用いて潜在的な意味に基づく各データの座標 $\boldsymbol{r}$ を獲得する．画像検索時は，L2 距離に基づいた近傍検索によって類似画像を獲得できる．

4. 特徴量抽出手法と実験設定

この章では，本研究が用いる特徴抽出手法，データセット，評価手法について説明する．なお，特徴量はそれぞれ必要に応じて PCA による次元圧縮を行っている．

4.1 特徴量抽出手法の選定

画像特徴量は一般的な手法を複数採用する．特徴量ベクトルを 1 つずつまたは複数つないで用い，任意の組合せで提案手法が検索精度を向上させることを示す．画像特徴量としては，画像の形状情報とテクスチャ情報の利用が求められる．そこで，おもに形状情報を対象とした特徴量として 5 章では SIFT [18]，6 章では SURF [19]，テクスチャ情報を対象とした特徴量として HLAC [20]，Gist [21]，LBP [22] を採用する．SURF はよく用いられる Bag of Features（BoF）[23] 表現で，SIFT は近傍の複数のコードブックで局所特徴量を表現する Locality-constrained Linear Coding（LLC）[24] によって画像全体の特徴量を得る．

文章特徴量としては，各文にどのような単語が出現したかという情報を利用する．具体的には，文全体に出現した単語が 1 つの成分に対応するベクトル空間モデルを採用する．文に出現しない単語に対応する成分は 0 になる．出現する単語に対する重み付けとしては，tfidf 重み付け[25] を採用する．各文においてその単語がいかほどの割合で出現するかの頻度（term frequency）のみでは “is” や “a” のような意味を持たない単語への重み付けが行われてしまう．そこで，文書頻度の逆数（inverted document frequency）を乗ずることで，多くの文に出現する単語の重みを下げられる．

4.2 データセット

提案手法が，類似した事物を扱う画像を精度良く検索できることを示すために，あらかじめいくつかのカテゴリに基づいて画像と文章が収集された複数のデータセットで実験を行う．いずれのデータセットでも，基本的には 9 割の画像を「文章つきの画像」と見なして類似度を改善し，残る 1 割を「画像」のみのデータと見なして検索のクエリに用いる．

5 章では，20 カテゴリの画像とキャプションからなるデータセットで実験を行う．学習データがカテゴリあたり 45 ペアと少ない状況でも，提案手法が類似画像検索の精度を向上させることを示す．6 章では，平均 800 語の長文と画像からなるデータセットで実験を行う．5 章で用いるデータセットに含まれる文章は，画像の内容を直接述べる簡潔なものである．対して一般的な Web ページでは，画像に付随する文章はより長く，画像の内容と関係の薄い記述も含むものが多いと考えられる．そこで長文を含むデータセットを独自に構築し，そのようなデータセットでも提案手法が検索精度を改善できることを示す．

4.3 評価方法

用いた評価指標は Mean Average Precision（MAP）と呼ばれるもので，入力画像に対して得られたランク付き画像検索結果から計算される Average Precision（AP）を平均したものである．AP は N 点のランク付き被検索画像から，$\mathrm{AP} = (\sum_{r=1}^{N} \mathrm{Precision}(r) \times \mathrm{Relevance}(r))/\mathrm{N_r}$ と計算される．ここで $\mathrm{N_r}$ は関連する画像の総数，$\mathrm{Precision}(r)$ は r 番目までの検索結果において関連画像が含まれている割合（適合率）である．また，$\mathrm{Relevance}(r)$ は r 番目の検索結果が入力画像と同一のカテゴリに属するものであれば 1，そうでなければ 0 をとる．MAP 値は検索結果として得られるランクから Recall-Precision 曲線を描いたときにその面積を表す指標であり，0 から 1 の値をとる．ここで，データセットのカテゴリ情報は MAP 値算出のためだけに用いられる点に注意されたい．提案手法はあくまで「画像」と「文章つき画像」を対象としたもので，カテゴリ情報を含まない状況で学習が可能である．

5. 多様な事物を扱うデータセットでの実験

この章では，さまざまなカテゴリに基づいて収集された画像とそのキャプションからなるデータセットを用いて，提案する類似画像検索手法の評価実験を行う．

5.1 PASCAL Sentence データセット

この章では，実験用データセットとして PASCAL Sentence [26] を用いる．PASCAL Sentence は画像認識ワークショップの PASCAL で配布されたデータセットの一部，20 カテゴリの画像を 50 枚ずつ抽出し，各画像に対してクラウドソーシングの Amazon mechanical turk を利用して 5 文前後の文を付与している．画像とキャプションの例を**図 1** に示す．各文は平均で 12 単語弱の長さであり，画像に写っている内容を直接記述している．本実験では 900 ペア（カテゴリそれぞれで 45）の画像・文章データを訓練用データ，残る 100 ペアの画像・文章データを評価用データとする．

図 1　PASCAL Sentence データセットの画像例
Fig. 1　An example of an image and captions of PASCAL Sentence dataset.

表 1　類似画像検索精度の比較
Table 1　Comparison of MAP scores about similar image retrieval.

画像特徴量	HLAC	gist	LBP	SIFT + LLC
PCA + L2 距離	0.1109	0.0804	0.0829	0.0932
CCD	0.1356	0.1234	0.1224	0.1257

5.2　類似度改善効果の検証

各画像特徴について，次の 2 つの手法による検索精度を調べた．

PCA + L2 距離　訓練データの画像特徴量のみに PCA を行い，テストデータも含めた全データを主成分空間上で近傍検索した際の検索精度．

CCD　訓練データの画像・文章ペアから潜在空間を獲得し，テストデータの画像も含めた全データを潜在空間上で近傍検索した際の検索精度．

画像特徴量のみで評価指標が最も高くなるパラメータと，CCD に用いたときに評価指標が最も高くなるパラメータを探索した．CCD の成績はパラメータ β，画像・文章量特徴量の主成分分析における打ち切り次元，pCCA の打ち切り次元の 4 パラメータに左右される．具体的には，β については 0.1 から 0.2 刻みで 0.9 まで，pCCA の打ち切り次元は画像・文章特徴量で次元数が少ない方に対し 2 割刻みで 2 割〜10 割，画像・文章それぞれの特徴量の打ち切り次元については 20 次元刻みで最適なパラメータを探索した．

訓練データを画像特徴量のみで検索した場合，文章特徴と統合した CCD で検索した場合それぞれの精度を**表 1** に示す．画像特徴量それぞれを単独で用いるよりも，文章の類似度を統合する CCD による検索のほうが精度良く訓練データを検索できている．

6.　長文を含むデータセットでの実験

5 章で扱ったデータセットでは，画像に付随する文章はそれぞれの画像の内容を直接記述するものであった．一般の Web ページでは，画像の内容を直接記述しないような文章が，より長文の形で存在する場合が多いと考えられる．そこでこの章では，そのようなデータセットを用いて提案手法の有効性を検証する．また，非線形な距離学習手法との比較や，分布改善の視覚的な検証も行う．

図 2　New York Times データセットの画像例
Fig. 2　Image examples of New York Times dataset.

6.1　New York Times データセット

長文を含む画像で構成されるデータセットは存在しないため，本研究ではニュースサイトを利用して長文つき画像データセットを独自に用意する．New York Times API を利用して，同サイトから election，baseball，basketball，terrorism，weddings の 5 カテゴリに属するニュース文と画像のペア 16,152 対を収集した．画像の一部を**図 2** に示す．画像は解像度も見た目も分散が大きく，文章は平均で 800 単語程度と長文である．

6.2　類似度改善効果と文章の割合

データセットに含まれる文章の割合を変えながら，類似画像検索の精度を比較した．前章の PCA + L2 距離と CCD に加え，通常の CCA + L2 距離とも比較した．CCA では画像の正準空間と文章の正準空間 2 つが獲得され，それぞれの空間でデータの座標は $A^T(\boldsymbol{x}_t - \bar{\boldsymbol{x}})$，$B^T(\boldsymbol{y}_t - \bar{\boldsymbol{y}})$ と表現される（式 (7)）．CCA + L2 距離による検索では，テストデータの画像も含めた全データを画像の正準空間上で近傍検索した．

まず，この学習用画像データの特徴量に対して寄与率が 99% となるように主成分分析を行った．文章特徴量に対しても主成分分析を行い，打ち切り次元は画像特徴量と同次元となるようにした．その後，提案手法を用いて全画像データを同一の潜在空間上に分布させた．CCA は上位の正準空間で打ち切るので，今回も 10 段階で最良の結果が出る次元を調べた．提案手法の β の値については 0.1 から 0.9 まで 0.1 刻みでチューニングを行った．

表 2　文章量に対する類似画像検索精度
Table 2　MAP scores of similar image retrieval for each text amount.

文章の割合	PCA + L2 距離	CCA + L2 距離	CCD
10%	0.0732	0.0914	0.0998
90%	0.0737	0.0895	0.1806

文章が付随しない画像をクエリとして類似画像を検索し，MAP 値を計算したものを**表 2** に示す．画像特徴量の類似度のみで検索するより CCA の正準空間上での近傍検索がより高精度であった[⋆1]．さらに，通常の CCA による画像側の正準空間よりも良い精度で画像を検索できている．通常の CCA でも文章特徴量の分布から画像特徴量の分布を改善しているが，

- 画像の正準空間と文章の正準空間を単一の潜在空間に融合できる，
- 潜在空間の各次元を正準相関に基づいて適切に重み付けしている（式 (7)），

という点で，CCD は CCA と L2 距離の組合せより優れている．特に CCA は上位の次元で打ち切らないと検索精度が低下するが，CCD では低位の次元の重みが自動で下がる．

また，文章が付随する画像の割合がより低い場合を想定し，全データ中の 1 割のみに文章が付随しているとして同様に類似度検索を行った結果もあわせて示す．このように文章がごく一部にしかない状況でも，他のアプローチより精度良く類似画像が検索できている．

6.3　任意の画像特徴量の組合せの検証

画像特徴量の任意の組合せについて評価を行い，一般的な画像特徴量に対する提案手法の有効性を検証する．この実験でも，**PCA + L2 距離**：画像特徴量空間，**CCA + L2 距離**：画像特徴量空間から CCA で得た正準空間，**CCD**：提案手法による潜在空間それぞれで，全 16,152 データをそれぞれクエリとした場合の近傍検索を行う．それぞれでの類似画像検索の精度を**図 3** に示す．この図が示すように，すべての特徴量の組合せについて，提案手法は画像特徴量のみで類似画像検索を行うよりも大きく精度を改善した．

6.4　その他の手法との比較

提案手法と他の手法との比較を行う．全 16,152 データ中の 9 割の画像に文章が付随しているという設定で，各データを入力として他のデータとの類似画像検索精度を比較する．

ここまで行われてきた比較は，提案手法も含めてすべて線形な手法であった．この節では，非線形な手法として画像特徴量空間内の多様体を仮定するアプローチとの比較を行う．

⋆1 表 2 の PCA では文章をまったく用いていないが，画像のみを用いる場合でも，PCA を行う際に使用する画像の量が全体の 10%/90%と変化するため，検索精度も同一にはならない．

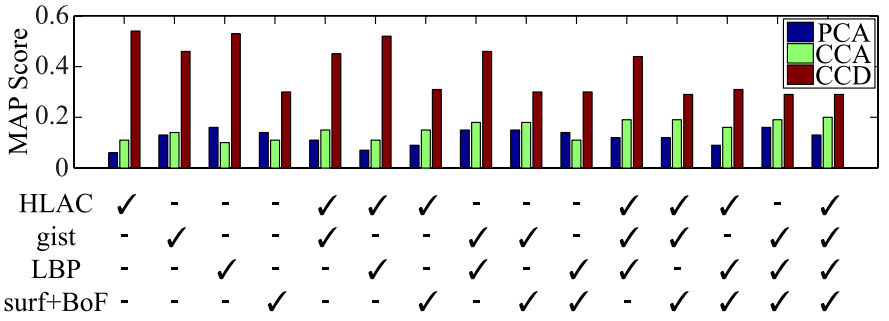

図 3　特徴量の全組合せでの各手法の比較．画像特徴量の主成分空間（PCA + L2 距離），画像側の正準空間（CCA + L2 距離），提案手法による潜在空間（CCD）での MAP
Fig. 3　Comparison using arbitrary combinations of image features. Scores are MAP on image feature space with PCA, on image canonical space with CCA, and on latent space with proposed CCD, respectively.

よく用いられる[27] ISOLLE，ISOSNE などの手法は，いずれも特徴量分布からグラフ構造を生成し，ノード間の距離を保ったまま低次元空間上にグラフ構造を展開するものである．つまり，これらの手法による類似画像は，グラフ構造上での近傍検索に等しい．この実験では，ISOLLE の原著論文[11] に従い，次のような手順で類似画像を検索した．特徴量次元上で全画像 16,152 枚をノードとし，ある距離 ε より近いノードどうしを接続した．この ε は全ノードが 1 つのグラフ構造を構成する最短のものである．このようにして得られたグラフ上で Dijkstra 法[28] によって全ノード対全ノードの最短経路を求めた．

また，提案手法は画像と文章から潜在的な意味に基づく分布を獲得しているが，3 章の終わりでも述べたとおり同様の目的の手法がいくつかある．そこで，pLSA を画像アノテーションに用いた研究[29] に従い，次のような手順で類似画像を検索した．まず，訓練データの画像特徴量と文章特徴量それぞれで pLSA により潜在空間を獲得した．得られた 2 つの潜在変数をつなぎ合わせた新たな潜在変数上でもう 1 度 pLSA を行い，画像と文章を統合した潜在空間を獲得した．画像・文章ペアをクエリとする近傍検索はこの統合された潜在空間で，画像のみをクエリとする近傍検索は画像特徴から得た潜在空間で行った．

各手法を用いた場合の検索精度を，**表 3** に示す．やはり，提案手法が最も優れた検索能力を有している．ISOLLE はグラフ構造によって，PCA のみでは扱えない画像特徴の多様体構造を扱えるが，付随する文章をまったく扱えない点で提案手法や pLSA に劣り，計算量も大きい．pLSA は通常の CCA による近傍検索よりも優れた検索精度を誇るが，画像と文章

表 3　提案手法と ISOLLE，pLSA との比較
Table 3　Comparison between proposed method, ISOLLE and pLSA.

method	all	election	baseball	basketball	terrorism	wedding
PCA + L2 距離	0.074	0.083	0.075	0.073	0.038	0.100
ISOLLE	0.145	0.104	0.099	0.075	0.043	0.406
pLSA	0.245	0.124	0.172	0.121	0.211	0.378
CCD	0.542	0.369	0.708	0.579	0.360	0.692

図 4　特徴量分布改善の概観図．左：元の画像特徴量の主成分空間（主要な 2 次元），右：提案手法による潜在空間（主要な 2 次元）
Fig. 4　Overview of improved feature distribution. Left: Main 2D space of the original image feature. Right: Main 2D space of the latent space.

の共起や相関を陽には扱わない．この点で，提案手法より劣る性能となったと考えられる．

6.5　視覚化による提案手法の有用性の検証

提案手法による画像分布の改善効果を視覚的に示す．画像特徴量のみで主成分分析を行って上位 2 次元の主成分空間で画像を配置したものを**図 4** 左側，提案手法によって獲得された潜在空間の上位 2 次元の正準空間で画像を配置したものを図 4 右側に示す．特に wedding に関する画像の分布が著しく改善されている．提案手法が得る潜在空間では，画像と文章の相関が最大化される．潜在空間で上位な次元ほど，画像と文章に共起する潜在的な意味を反映した分布になっているため，図 4 のように分布改善効果を視覚的に確認できる．

7. おわりに

Web 上の大量の画像を扱うには，高精度に類似画像を検索するスケーラブルな類似度学習手法が必要となる．本論文では，画像の類似度と付随する文章の類似度を統合し，類似画像検索の精度を向上させる手法を提案した．複数のデータセットを用いた実験から，提案手法が精度良く扱う内容の類似した画像を検索できることを示した．

提案手法は類似度の統合をデータ数に依存しない計算量で学習できるが，検索時は全データと類似度を計算するため計算量が線形オーダになる．今後はより高速な近似近傍検索手法について検討を行っていく．また，今回のようなニュースサイトだけでなく，種々の形態の Web サイトから画像と文章を抽出した際の性能評価も行う．

参　考　文　献

1) Datta, R., Joshi, D., Li, J. and Wang, J.Z.: Image retrieval: Ideas, Influences, and Trend of the New Age, *CSUR*, Vol.40, No.2, pp.1–60 (2008).
2) Wang, X.-J., Zhang, L., Liu, M., Li, Y. and Ma, W.-Y.: ARISTA – Image Search to Annotation on Billions of Web Photos, *Proc. CVPR*, pp.2987–2994 (2010).
3) Weinberger, K.Q., Blitzer, J. and Saul, L.K.: Distance Metric Learning for Large Margin Nearest Neighbor Classification, *Proc. NIPS*, pp.1473–1480 (2005).
4) Bai, B., Weston, J., Grangier, D., Collobert, R., Sadamasa, K., Qi, Y., Cortes, C. and Mohri, M.: Polynomial Semantic Indexing, *Proc. NIPS*, pp.64–72 (2009).
5) Moghaddam, B., Tian, Q., Lesh, N., Shen, C. and Huang, T.S.: Visualization and user-modeling for browsing personal photo libraries, *IJCV*, Vol.56, No.1, pp.109–130 (2004).
6) Blaschko, M.B. and Lampert, C.H.: Correlational Spectral Clustering, *Proc. CVPR*, pp.1–8 (2008).
7) Tenenbaum, J.B., Silva, V. and Langford, J.C.: A global geometric framework for nonlinear dimensionality reduction, *Science*, Vol.290, No.5500, pp.2319–2323 (2000).
8) Hinton, G. and Roweis, S.: Stochastic neighbor embedding, *Proc. NIPS*, pp.857–864 (2003).
9) Roweis, S. and Saul, L.: Nonlinear dimensionality reduction by locally linear embedding, *Science*, Vol.290, No.5500, pp.2323–2326 (2000).
10) Nguyen, G. and Worring, M.: Similarity based visualization of image collections, *Proc. AVIVDiLib* (2005).
11) Varini, C., Degenhard, A. and Nattkemper, T.: ISOLLE: LLE with geodesic distance, *Neurocomputing*, Vol.69, No.13-15, pp.1768–1771 (2006).
12) Nguyen, G.P. and Worring, M.: Interactive access to large image collections using similarity-based visualization, *JVLC*, Vol.19, No.2, pp.203–224 (2008).
13) Lavrenko, V., Manmatha, R. and Jeon, J.: A Model for Learning the Semantics of Pictures, *Proc. NIPS* (2003).

14) Nakayama, H., Harada, T. and Kuniyoshi, Y.: Canonical contextual distance for large-scale image annotation and retrieval, *Proc. ACM MM Workshop on Large-Scale Multimedia Retrieval and Mining*, pp.3–10 (2009).
15) Bach, F.R. and Jordan, M.I.: A Probabilistic Interpretation of Canonical Correlation Analysis, Technical Report 688, Department of Statistics, University California, Berkeley (2005).
16) Hofmann, T.: Probabilistic Latent Semantic Analysis, *Proc. UAI*, pp.289–296 (1999).
17) Blei, D., Ng, A. and Jordan, M.: Latent dirichlet allocation, *JMLR*, Vol.3, pp.993–1022 (2003).
18) Lowe, D.G.: Distinctive Image Features from Scale-Invariant Keypoints, *IJCV*, Vol.60, No.2, pp.91–110 (2004).
19) Bay, H., Tuytelaars, T. and Van Gool, L.: SURF: Speeded up robust features, *Proc. ECCV*, pp.404–417 (2006).
20) Otsu, N. and Kurita, T.: A New Scheme for Practical Flexible and Intelligent Vision Systems, *Proc. IAPR Workshop on Computer Vision*, pp.431–435 (1988).
21) Oliva, A. and Torralba, A.: Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope, *IJCV*, Vol.42, No.3, pp.145–175 (2001).
22) Ojala, T., Pietikäinen, M. and Harwood, D.: A Comparative Study of Texture Measures with Classification Based on Feature Distributions, *Pattern Recognition*, Vol.29, No.1, pp.51–59 (1996).
23) Csurka, G., Dance, C., Fan, L., Willamowski, J. and Bray, C.: Visual Categorization with Bags of Keypoints, *Proc. ECCV Workshop on Statistical Learning in Computer Vision*, pp.22–37 (2004).
24) Wang, J., Yang, J., Yu, K., Lv, F., Huang, T. and Gong, Y.: Locality-constrained Linear Coding for Image Classification, *Proc. CVPR*, pp.3360–3367 (2010).
25) Salton, G. and Yang, C.S.: On the Specification of Term Values in Automatic Indexing, *Journal of Documentation*, Vol.29, No.4, pp.351–372 (1973).
26) Farhadi, A., Hejrati, M., Sadeghi, M.A., Young, P., Rashtchian, C., Hockenmaier, J. and Forsyth, D.: Every Picture Tells a Story: Generating Sentences from Images, *Proc. ECCV*, pp.15–28 (2010).
27) Heesch, D.: A survey of browsing models for content based image retrieval, *Multimedia Tools and Applications*, Vol.40, No.2, pp.261–284 (2008).
28) Dijkstra, E.W.: A note on two problems in connection with graphs, *Numerische Mathematik*, No.1, pp.269–271 (1959).
29) Romberg, S., Horster, E. and Lienhart, R.: Multimodal Plsa on Visual Features and Tags, *Proc. ICME*, pp.414–417 (2009).

(平成 23 年 4 月 11 日受付)
(平成 23 年 9 月 12 日採録)

牛久　祥孝

2011 年 3 月東京大学大学院情報理工学系研究科知能機械情報学専攻修士課程修了．同年 4 月より同専攻博士課程に進学し現在に至る．日本ロボット学会，電子情報通信学会，人工知能学会，IEEE，ACM 等の学生会員．

原田　達也（正会員）

2001 年 3 月東京大学大学院工学系研究科機械工学博士課程修了．2000 年 1 月から 2001 年 12 月まで日本学術振興会特別研究員．2001 年 12 月東京大学大学院情報理工学系研究科助手，2006 年 4 月同講師，2009 年 4 月同准教授となり現在に至る．博士（工学）．IEEE，日本ロボット学会，人工知能学会，電子情報通信学会等の会員．

國吉　康夫（正会員）

1991 年東京大学大学院工学系研究科情報工学専攻博士課程修了，工学博士，同年電子技術総合研究所入所，1996 年より 1 年間米国 MIT AI Lab. 客員研究員，2001 年東京大学助教授，2005 年同教授，現在に至る．身体性に基づく認知の創発と発達，人間の行為の観察・理解システム，ヒューマノイドロボット等の研究に従事．佐藤記念知能ロボット研究奨励賞，IJCAI Outstanding Paper Award，日本ロボット学会論文賞，ゴールドメダル「東京テクノ・フォーラム 21 賞」等受賞．日本ロボット学会，人工知能学会，IEEE，日本赤ちゃん学会等の会員．日本学術会議連携会員．