© 2017 一般社団法人映像情報メディア学会

論　文

弁別的キーポイントの選択に基づく類似物体群からの特定物体認識

Discriminative Keypoint Selection for Object Retrieval from Similar Instances

渡邉之人†，正会員　入江　豪††，黒住隆行†，杵渕哲也†

Yukito Watanabe†, Go Irie††, Takayuki Kurozumi† and Tetsuya Kinebuchi†

あらまし　互いに類似した見た目でありながらも，別々の物体であるような物体群からの特定物体認識の問題を考える．類似する物体群から抽出されるキーポイント（SIFT 等）は，その多くが類似するため相互に見分けが付きにくく，結果として認識精度が劣化するという問題がおこる．本稿ではまず，類似物体間の差異となるような微小な『弁別的キーポイント』を発見するごく単純な方法を提案し，これにより類似物体群からであってもより正確に正しい物体を特定できることを示す．さらに，弁別的キーポイントの識別性能を活かしながらも高速な認識を行うためのインデクシング法として『2 レベル転置インデクス法』を提案する．この方法は，1 レベル目で全キーポイントを用いた認識を行うことで類似物体群だけに粗く絞り込み，2 レベル目で弁別的キーポイントのみを用いて正しい物体の特定を狙う．3 種類のデータセットを用いた実験を行い，本手法の有効性を明らかにする．

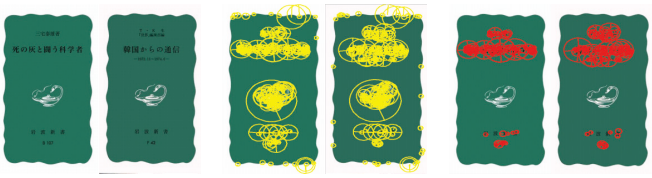
キーワード：特定物体認識，キーポイント選択，Bag-of-visual words

1. ま　え　が　き

スマートフォンやタブレット端末を始めとする高機能な小型撮像デバイスの普及を背景に，これらを活用した画像認識アプリケーションやサービスが数多く登場し，また注目を集めている．例えば Google 社は，物体にカメラをかざすことでそこに現れる商品やロゴを認識し，関連する Web ページへとアクセスを促す Google Goggles と呼ばれるアプリケーションを提供している．また，Amazon 社は，Firefly と呼ばれる画像認識アプリケーションを用いて，商品にカメラを向けることでそれを認識し，直接 amazon.com の商品販売ページへとナビゲーションするサービスを提供している．これらのアプリケーションの背後には特定物体認識と呼ばれる技術が用いられている．認識したい対象を含む画像データベースが所与の下，撮影されたクエリ画像と同一の物体が写る画像を検索する問題を解く技術であり，長らく画像処理分野における中心的なトピックの一つとなっている[7)15)]．

本稿では，画像データベース中の画像に写る異なる物体が互いに類似しているような場合における特定物体認識の問題を議論する．実際，このような状況は実応用上数多く見られる．例えば，書籍画像データベースを検索するような場合が挙げられよう．**図 1** (a) に示すような同一シリーズの書籍は，一見してよく似たデザインやレイアウトが採用されていることが多く，異なる本を見分ける上で手掛かりとなるのは，タイトル等のごく一部の領域のみである．もちろん書籍に限った話ではなく，他にも CD のカバージャケットや飲料，菓子類のパッケージ，あるいは自動車や，至る所に存在する案内板など，実世界において似て非なる物体を含むような物体群は数多く存在している．

このような似て非なる物体を含むような画像データベースに対して，通常の特定物体認識手法を適用した場合，認



(a)　(b)　(c)

図 1　類似した物体画像の例とキーポイント選択結果 *．(a) よく似た表紙を持つ異なる書籍画像．(b) DoG (Difference of Gaussian) によって検出されたキーポイント．(c) 提案法により選択された弁別的キーポイント．

2016 年 5 月 9 日受付，2016 年 10 月 26 日再受付，2016 年 12 月 2 日採録
†NTT メディアインテリジェンス研究所
（〒 239–0847 神奈川県横須賀市光の丘 1-1，TEL 046-859-2974）
††NTT コミュニケーション科学基礎研究所
（〒 243–0198 神奈川県厚木市森の里若宮 3-1）

* 画像：『死の灰と闘う科学者』，三宅 泰雄，岩波書店，『韓国からの通信——1972．11～1974．6』，T・K 生，岩波書店

(a) 全キーポイント・同物体 (33)　(b) 全キーポイント・異物体 (33)　(c) 弁別的キーポイント・同物体 (17)　(d) 弁別的キーポイント・異物体 (5)

図 2　SIFT+RANSAC による類似書籍画像のキーポイントマッチング結果 *．それぞれ左側がクエリ画像，右側がデータベース画像．括弧内の数字はマッチしたキーポイント数を表す．全キーポイントを用いた場合の (a) 同一物体のマッチング結果と (b) 異なる物体のマッチング結果．マッチしたキーポイント数にほぼ差がないことがわかる．提案法による弁別的キーポイントを用いた場合の (c) 同一物体のマッチング結果と (d) 異なる物体のマッチング結果．マッチしたキーポイント数に差が生じ，正しい認識が可能になる．

識に失敗することが多い．類似する物体間では，抽出されたキーポイントのほとんどが類似してしまうため，異なる物体間であるにも関わらずマッチしてしまうようなキーポイントが多数存在し，これらが誤認識を招く要因となるためである．さらに，これら似て非なる物体はそのレイアウトも類似していることが多く，RANSAC のような方法に基づく幾何検証[12] を行っても精度が改善されにくい．先に示した書籍を例に取り，最もポピュラーな局所特徴量である SIFT[11] および RANSAC を用いてキーポイントマッチングを行った結果を図 **2**(a)，(b) に示す．マッチしたキーポイント数を比較すると，同一物体同士のマッチ数と，異なる物体間でのマッチ数がほぼ同数となってしまい，結果的に互いに判別できない結果になってしまうことがわかる．

よく似た物体を見分けることを志向した研究領域として，fine-grained visual categorization (FGVC) と呼ばれるタスクが存在する．これは，鳥，犬，花などの特定の物体カテゴリ（大カテゴリ）に属する画像の中から，“種”のようなより粒度の小さなカテゴリへ画像を分類する問題である[18]．しかし，FGVC は飽くまで分類を目的とした一般物体認識に属するタスクであり，同一の物体を認識する特定物体認識とはその問題設定やアプローチが大きく異なる．著者らが知る限り，その重要性にも関わらず，特定物体認識において，陽に似て非なる物体を見分けることを目的としたアプローチは現在まで報告されていない．

以上の観測に基づき，本稿では，似て非なる物体画像を含む画像データベースからの特定物体認識法を提案する．我々のアプローチは，類似する物体群の間で差異となるようなキーポイント（本稿では『弁別的キーポイント』と呼ぶ）を特定し，これを用いて検索を実行することである．具体的には，「似て非なる物体間において共通して出現するようなキーポイントは弁別的ではない」という仮説に基づき，まずデータベース画像のおのおのを仮想的にクエリ画像と見做した場合に検索される物体画像を類似物体画像と見做し，これらに共通するキーポイントを特定して，除去する．さらに，発見した弁別的キーポイントを用いた効率的な特定物体認識を実現すべく，転置インデクス (inverted file index: IVF)[15] を拡張した『2 レベル転置インデクス法』を提案する．3 種類のデータセットを用いた実験により，従来の方法よりも高い精度が得られることを示す．

以降，本論文の構成は次の通りとする．まず，2 章で本稿に関連する研究について概観し，3 章で提案する弁別的キーポイント選択法，4 章で 2 レベル転置インデクス法について説明する．5 章で評価実験結果を報告し，最後に 6 章で本論文の結論をまとめる．

2. 関 連 研 究

本稿に関連する特定物体認識およびキーポイント選択の各研究領域について概観する．

2.1 特定物体認識

特定物体認識における基本的なアプローチはキーポイントマッチングに基づくものである[7)11)15)]．各画像から検出した多数のキーポイントを，SIFT などの局所特徴量によって表現し，クエリ画像と各データベース画像との間でマッチしたキーポイントの数が多ければ多いほど，検索結果の上位にランキングする．しかし，このような方法は，多数のキーポイント同士の比較が必要となるため，大規模な画像データベースに対して適用することは計算量的に困難である．

そこで，より効率的な検索を実現する方法が検討されてきた．最もよく知られるアプローチは，bag-of-visual-words (BoW)[4)15)] に基づくものであろう．BoW では，各々の局所特徴量を，予め学習しておいたコードブック（学習には k-means や product quantization (PQ)[7] といったクラスタリング法が用いられることが多い）を用いて，visual word と呼ばれる符号に量子化し，各 visual word の出現頻度によるヒストグラムを求めることで，各画像をベクトルの形式で表現する．BoW は量子化による損失の影響が大きいため，これを回避するためにサイズの大きいコードブック（例えば visual word 数にして 10,000 以上）を利用する必要がある．大きなコードブックを用いる場合，見かけ上の表現は非常に高次元なものとなるが，各画像から抽出される visual word の種類には偏りが生じるため，各 visual word

からそれを保持する画像を逆引きできる転置インデクスを用いることで，効率的に検索を実行することができることが知られている[6)12)]．

2.2　キーポイント選択

キーポイント選択は，3 次元再構成やカメラキャリブレーション，特定物体認識など，多岐にわたる画像処理タスクにおいて重要な要素技術である．本稿では特定物体認識問題におけるキーポイント選択について概観する．

古くから，“より重要な”キーポイントだけを選択することで，認識時間を増大させずに認識精度を高めようとする試みがなされてきた．最も典型的な方法は逆文書頻度 (inverse document frequency: IDF) である．IDF では，多くの画像において頻繁に出現するような visual word は認識に寄与しない，すなわち重要ではないという考えの下，画像データベース中における visual word の出現頻度に従ってキーポイントの重要度を定める重みを低下させる．IDF は非常に単純であり，経験的には検索精度を高める効果があることから頻繁に用いられる手法となっている．しかし，必ずしも visual word の出現頻度が物体の見分けやすさを反映しているとは限らないため，似て非なる物体を見分ける上では最適な選択方法であるとは言い難い．

Torii ら[16)] は，建物の窓や樹木の葉など，多様なコンテキストにおいてしばしば観測される繰り返し構造の影響を考慮した方法を提案している．このような繰り返し構造はよく似たキーポイントを意図せず多数出現させてしまい，結果として画像間の類似度の評価に悪影響を及ぼす．Torii らは繰り返し構造上のキーポイントを検知して少数の visual word にまとめることにより，認識の精度と効率を向上させた．本稿で提案する方法は，繰り返し構造ではなく，見た目の類似性に基づいてキーポイントを選択するものであるため，この方法とは相補的である．Knopp ら[9)] は，画像に付与された地理的な位置情報を付記したメタデータ (geotag) を利用し，異なる地点で撮影された類似する画像間からキーポイントを選択することにより，高精度な地点認識を実現している． 我々の方法はメタデータを利用する方法ではなく，アプローチが異なる．

Hartmann ら[5)] は，マッチするキーポイントとしないキーポイントの事例から識別器を学習し，マッチしやすいキーポイントのみを判別して選択する方法を提案している．我々の方法は識別器を用いない．

提案法に最も関連するアプローチとしては，Turcot らによる方法[17)] がある．この方法では，同一物体を含む複数の画像間で，幾何的に整合したキーポイントのみを選択して保持することにより，物体固有のキーポイントに基づく検索を実現しようとしているものである．我々の方法は，幾何的な整合性ではなく，似て非なる物体画像に跨って出現するようなキーポイントを選択除去する点で異なっている．端的な差異に言及すると，Turcot らの方法は同一物体の画像を常に複数枚要求するが，我々の方法は一物体あたり 1 枚であっても動作する．

以上概観する限り，重要なキーポイントを選択することによって，特定物体認識の精度を向上しようとする試みは多く報告されているが，似て非なる物体を見分けることを目的としたキーポイント選択法に関する取り組みは未だ報告されていない．本稿では，似た物体群からの特定物体認識を目的とした新たなキーポイント選択法，および，選択したキーポイントに基づく効率的な認識を実現するためのインデクシング法を提案する．

3.　弁別的キーポイントの選択とマッチング

画像データベースに n 枚の画像 $\{I_1, \ldots, I_n\}$ が登録されているとし，I_i から抽出された m_i 個のキーポイントの局所特徴量ベクトル (SIFT など) を $\{\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,m_i}\}$ と表す．簡単のため，画像データベース中には物体ごとに画像は 1 枚のみしか存在しないと仮定するが，本手法は物体ごとに複数枚存在する場合にも容易に適用可能である．

3.1　最近傍距離比に基づくマッチング

キーポイントマッチングは，一般に最近傍距離比 (nearest neighbor distance ratio: NNDR)[11)] と呼ばれるマッチング基準に基づいて実施される．2 枚の画像 I_i と I_j 間でのマッチングを考える．画像 I_i のキーポイント $\mathbf{x}_{i,a}$ に対して，これに最も近い画像 I_j のキーポイントが $\mathbf{x}_{j,b}$，二番目に近いキーポイントが $\mathbf{x}_{j,c}$ であったとすると，その距離比が閾値 τ 以上の場合であるときのみマッチすると判定する．厳密には，

$$d(\mathbf{x}_{i,a}, \mathbf{x}_{j,b}) > \tau d(\mathbf{x}_{i,a}, \mathbf{x}_{j,c}) \tag{1}$$

となる場合に，$\mathbf{x}_{i,a}$ と $\mathbf{x}_{j,b}$ はマッチしたと判定する．ここで，d は距離関数であり，通常は ℓ_2 距離が用いられる．τ は距離比に対する閾値であり，本稿では 11) に倣い $\tau = 0.8$ と固定することとする．提案法では NNDR 基準に基づき，任意の異なる 2 つのキーポイントの同一性を判断する．

3.2　弁別的キーポイント選択

提案する弁別的キーポイント選択法の目的は，データベースに登録された画像のそれぞれについて，その画像に写る物体を正しく見分ける上で重要なキーポイントを選択することである．

まず，背後にあるアイディアを説明する．あるデータベース画像 I_i に着目し，I_i の弁別的キーポイントを選択することを考える．理想的には，I_i のみに出現し，その他のデータベース画像 $\{I_j\}_{j \neq i}$ には表れないようなキーポイントのみを選択すれば，画像 I_i に写る物体固有のキーポイントを得ることができるであろう．しかしながら，現実的にはこのようなキーポイントを得ることは難しい．キーポイントはごく微小な領域の輝度変化（テクスチャ等）を記述しているにすぎず，他のあらゆる物体に対して同じような微小領域が含まれないことを仮定するのは困難であるからである．図 1 に示した書籍の場合を例に直観的に説明しよう．

表 1　BSC-200 におけるマッチング方向の比較．データセットの詳細は 4 章に記載する．

方法	精度 (%)
キーポイント選択なし	78.9
弁別的キーポイントのみ（クエリ方向）	77.5
弁別的キーポイントのみ（データベース方向）	83.3

図 1 の 2 つの書籍を比較した場合，タイトルの文字領域が相互に弁別的であることは明らかである．一方，全ての書籍を見渡した場合に，もしこれらのタイトルに含まれる文字と同一の文字が出現した場合，その文字上のキーポイントはもはやその本固有のものではなくなってしまう．つまり，ある本が他の全ての書籍と比して固有のキーポイントを持つには，他のどの書籍にも含まれないような文字をタイトルに持っている必要があり，およそ非現実的な状況を仮定しなければならなくなってしまう．

以上の知見に基づき，提案する弁別的キーポイント選択法では，あるデータベース画像の弁別的キーポイントを選択する際，対象とする範囲を類似するデータベース画像のみに絞り，この範囲で固有であると見做すことができるようなキーポイントのみを選択する．より具体的には，各データベース画像のそれぞれを仮想的にクエリ画像と見做して一度検索を実行し，検索された上位のデータベース画像に現れるキーポイントとマッチしているものを除外することにより，当該クエリ画像と見做したデータベース画像の弁別的キーポイントを得る．アルゴリズムは下記の通りである．

Step 0. $i = 1$ に初期化する．

Step 1. データベース内の画像 I_i をクエリ画像と見做し，画像データベース $\{I_j\}_{j \neq i}$ に対して NNDR 基準に基づく検索を実行し，ランキング上位 s 枚の画像を類似画像として得る．

Step 2. I_i の全てのキーポイント $\{\mathbf{x}_{i,1}, \ldots, \mathbf{x}_{i,m_i}\}$ について，s 枚の類似画像のキーポイントとマッチした回数でランキングし，回数の少ない方から $\alpha\%$を画像 I_i の弁別的キーポイントとして選択する．

Step 3. $i = n$ ならば終了し，そうでなければ $i \leftarrow i + 1$ として Step 1. に戻る．

上記のアルゴリズムによって弁別的キーポイントを選択した結果が図 1 (c) である．ここに示されているように，相互に固有のキーポイントが選択できていることが分かる．

なお，物体ごとに 2 枚以上の画像が存在するようなデータベースに適用する場合には，Step 1 において得られる類似画像の中から，クエリ画像と同一の物体を含むものを除去すればよい．

3.3　弁別的キーポイントによるマッチング

上記弁別的キーポイント選択に基づき，データベース画像はいずれも選択された弁別的キーポイントによって表現されることになる．一方で，クエリ画像の方は依然として弁別的であるかないかに関わらず全てのキーポイントを含んでいるため，誤対応となるようなマッチが多数発生してしまい，誤認識の原因となる．

この問題を回避するため，提案法では NNDR マッチングの方向に制約を与える．式 (1) の定義が示す通り，NNDR は有向なマッチング基準であり，I_i と I_j どちらをクエリ画像と取るかにより，異なるマッチング結果を返す．仮に I_i をクエリ画像とした場合（これをクエリ方向と呼ぶ），NNDR の距離 $d(\mathbf{x}_{i,a}, \mathbf{x}_{j,b})$ および $d(\mathbf{x}_{i,a}, \mathbf{x}_{j,c})$ は，いずれも必ずしも弁別的ではないクエリ画像のキーポイント $\mathbf{x}_{i,a}$ を基準に計算されるため，弁別的キーポイント以外の点とのマッチが多数得られてしまう．そこで，本稿では必ず I_i をデータベース画像に取ることで（これをデータベース方向と呼ぶ），弁別的キーポイントを基準としたマッチングが行われるようにする．

図 2(c)，(d) は，この方式で実施した弁別的キーポイントによるマッチング結果である．異なる物体間の場合に比べ，同一の物体間の方がマッチ数が多くなっており，望ましい結果が得られていることが確認できる．**表 1** に示す通り，実際の検索精度も向上していることが確認できる．なお，NNDR に基づくマッチングの計算時間は向きに依らず同一である．

4.　2 レベル転置インデクス

2.1 に述べた通り，キーポイントマッチングに基づく特定物体認識は非効率的であり，大規模な画像データベースに適用するのは現実的ではない．そこで我々は，弁別的キーポイントの利点を活かしつつ，さらに効率的な認識を実現するためのインデクシング法を提案する．

この方法は一般的に利用される visual word 量子化と IVF を用いたフレームワーク[7)15)] に基づいている．所与のクエリ画像に対して，トップ m 枚 ($m \ll n$) のデータベース画像を IVF によって高速に絞込み，これらの候補画像に対してのみキーポイントマッチングによる精細なランキングを適用することで，データベース画像の枚数に対して線形に処理時間が増大していくことを防ぐ．

一方で，提案法による弁別的キーポイントを用いる場合，単純な IVF の適用は好ましくない．通常，特定物体認識においてはデータベース画像から抽出されたキーポイントに基づいてコードブックを学習することが多い．提案法においては，データベース画像は弁別的キーポイントにより表現されているから，これらに基づいてコードブックを学習することになる．しかし，このように学習されたコードブックは，クエリ画像に含まれるような弁別的ではないキーポイントの情報を取り込むことができないため，これらのキーポイントは量子化誤差の大きい visual word に割り当てられることになり，結果的に不正確なマッチングが起こる原因となってしまう．また，弁別的キーポイントのみを用いて IVF を構成し検索を実行した場合，検索精度が低下する可能性もある．図 1 の例に挙げているような，同一のシ

リーズに属するような書籍を見分ける上では，弁別的キーポイントのみを用いた検索が有効であることは既に示した通りである．一方で，あらゆる書籍を見渡した場合，同一のシリーズに属するような書籍はむしろごく一部であり，他の多くの書籍は見た目の全貌が大きく異なっている．そういった書籍同士を識別するような場合には，弁別的ではないキーポイントも有益な手掛かりとなりうるため，弁別的キーポイントのみでは過度に欠落した情報によって検索を実行することになってしまう．

これらの問題を解決すべく，2 レベル転置インデクス (TIVF: two-level inverted file index) を提案する．TIVF は，その名の通り 2 層の IVF によって構成される．1 レベル目の IVF (IVF^1) では，クエリ画像に対して全貌が類似するようなデータベース画像を上位にランキングすることを目的としたものであり，コードブックは全てのキーポイント（弁別的ではないキーポイントも含む）を用いて学習される．端的に言えば，15) などで用いられている一般的な IVF と等価なものである．2 レベル目の IVF (IVF^2) は，IVF^1 で検索された全貌が類似するような似て非なるデータベース画像の集合の中から，正確にクエリ画像と同一の物体を含むデータベース画像を検索することを目的として構成されるものであり，コードブックは弁別的キーポイントのみを用いて学習される．TIVF を用いることで，弁別的ではないキーポイントが持つ情報の欠落を防ぎつつ，さらに似て非なる物体も見分けることが可能になる．

TIVF による検索処理は，以下の 2 ステップからなる．

Step 1. IVF^1 を用いてトップ l 枚の候補画像を検索する．

Step 2. IVF^2 を用いて，l 枚の候補画像をランキングし，トップ m 枚 ($m \leqq l$) の候補画像を返す．これは，l 枚の候補画像のみを対象として，IVF^2 によって検索を実行することで実現できる．

最終的に，トップ m 枚の候補画像を対象にキーポイントマッチングを実行する．また，この際 RANSAC などによる幾何検証を実行しても構わない．

5. 実　験

提案法による特定物体認識精度を評価する．

5.1 データセット

本稿の実験では，下記の 3 種類のデータセットを用いる．なお，いずれのデータセットにおいても全ての画像は長辺 640 ピクセルにリサイズして用いる．

Books with Similar Covers (BSC). 似て非なる物体の画像を含む画像データベースから，正しく特定物体認識できることを確認すべく準備したデータセットであり，1,200 冊分の書籍画像によって構成される (各書籍 1 枚) データベース画像と，90 冊分のクエリ画像で構成される．データベース画像 1,200 冊のうち 200 冊は類似した表紙を持つ同一シリーズの書籍であり，クエリ画像は全てこの 200 冊のうちのいずれかを撮影したものである．本データセットの評価に用いる IVF および TIVF のコードブックサイズ（visual word 数）は 65,536 とする．

Stanford Mobile Visual Search (SMVS). 1,200 枚のデータベース画像と 3,300 枚のクエリ画像で構成された公開データセットである[2]．CD やランドマークなど，8 カテゴリに属する複数種類の物体を含む．本データセットの評価に用いる IVF および TIVF のコードブックサイズは 100,000 とする．

Mobile. 400 枚のデータベース画像と 2,500 枚のクエリ画像から構成される公開データセットである[19]．本や雑誌などを携帯端末から撮影した画像によって構成されている．本データセットの評価に用いる IVF および TIVF のコードブックサイズは 100,000 とする．

5.2 プロトコル

本稿では，一貫してキーポイント検出法として DoG，局所特徴量として SIFT を用いる．また，NNDR の距離関数 d は常に ℓ_2 距離を用いることとする．認識精度の評価尺度としてはトップ 1 適合率を用いることする．すなわち，クエリ画像と同一の物体を含むデータベース画像を検索結果 1 位にランクすることができたクエリ画像のみ認識が成功したと判断し，全クエリ画像数に対する認識が成功したクエリ画像数の割合を報告する．提案法の弁別的キーポイント選択については，類似画像数 s および選択率 α をそれぞれ 30，50 とする（これらのパラメータの影響については次節で評価する）．また，IVF および TIVF により選択される候補画像数については，一貫して l を 50 (IVF^1)，m を 30 (IVF^2) とする．候補画像に対しては，RANSAC によるリランキングを適用し，検索結果を改善する[3]．RANSAC は，2 枚の画像間でマッチしたキーポイントの内，同一物体上のマッチのみを抽出する統計手法である．同一物体上のキーポイント群であれば，その幾何的な配置は撮影視点のみに依存して線形に変化するという合理的な仮定の下，線形に変化しているキーポイント群のみを統計的に抽出する．この結果を用い，同一物体上にあると判断されたキーポイントのマッチ数によって，検索結果をリランキングする．詳細は 3) に詳しいのでこちらを参照されたい．

5.3 BSC における結果

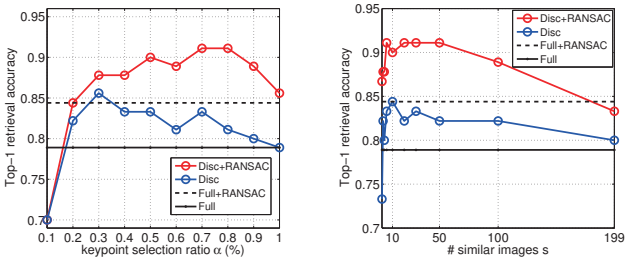
データベース画像として類似した表紙を持つ書籍のみを含む 200 冊 (BSC-200) を用いた場合，それ以外の書籍も含む 1,200 冊全て (BSC-1200) を用いた場合の 2 通りの場合について評価を行う．次の 6 種類の手法を比較する：全キーポイントを用いたキーポイントマッチングによるもの (Full)，弁別的キーポイントのみを用いたキーポイントマッチングによるもの (Disc)，Full に IVF を適用して効率化したもの (Full+IVF)，Disc に IVF を適用して効率化たもの (Disc+IVF)，Disc に TIVF の 1 レベル目の IVF (IVF^1) のみを適用したもの (Disc+IVF^1)，Disc に TIVF

表 2　BSC における結果．トップ 1 適合率 (%) を記載する．

BSC-200		
Method	w/o RANSAC	w/ RANSAC
Full (w/o IVF)	78.9	84.4
Disc (w/o IVF)	83.3	90.0
Full+IVF	77.8	82.2
Disc+IVF	78.9	82.2
Disc+IVF^1	82.2	86.7
Disc+TIVF	**87.8**	**91.1**
BSC-1200		
Method	w/o RANSAC	w/ RANSAC
Full+IVF	74.4	77.8
Disc+IVF	76.7	78.9
Disc+IVF^1	76.7	82.2
Disc+TIVF	**82.2**	**87.8**

表 3　コードブックサイズの認識精度への影響．Disc+TIVF+RANSAC の IVF^1，IVF^2 のコードブックサイズをそれぞれ独立に変化させた場合のトップ 1 適合率 (%) を記載する．

コードブックサイズ	2^{11}	2^{12}	2^{13}	2^{14}	2^{15}	2^{16}
IVF^1 (IVF^2 は 2^{16} 固定)	55.6	58.9	68.9	73.3	73.3	87.8
IVF^2 (IVF^1 は 2^{16} 固定)	53.3	57.8	67.8	70.0	76.7	87.8



(a) α　　(b) s

図 3　(a) キーポイント選択率 α，(b) 類似画像数 s の変動に対する精度への影響．

(IVF^1+IVF^2) を適用したもの (Disc+TIVF)．いずれの方法においても，RANSAC による幾何検証の有無それぞれについて評価した．BSC-1200 については，データベース画像枚数が多いため，IVF（または TIVF）を適用した 4 手法についてのみ評価を実施する．

BSC-200 における結果．まず，BSC-200 における結果を**表 2** の上段に示す．RANSAC の有無にかかわらず我々の Disc は Full を上回っており，提案法が類似する物体群からの特定物体認識に適していることが確認できる．次に，IVF と組み合わせた場合も，Disc+IVF は Full+IVF よりも高い精度を示している．しかし，その差は比較的小さく，また Disc よりも精度は大きく劣化してしまっていることが確認できる．これに対して，提案法である TIVF と組み合わせることで (Disc+TIVF)，精度が大きく改善されることが確認できる．1 レベル目の IVF^1 のみでランキングした場合 (Disc+IVF^1) と比べ精度が改善されていることからも，提案法の狙い通り，2 レベル目の IVF^2 により似て非なる物体を精細に見分けられることができていることが裏付けられている．Disc+TIVF の Full+IVF に対する最終的な精度差は，RANSAC 無しの場合で 10.0%，RANSAC 有りの場合で 8.9%に達した．以上のことより，弁別的キーポイント選択および TIVF により，類似した物体画像を含む画像データベースから，より正確に同一物体の画像を検索できることが確認できたと言える．なお，Disc に比べて Disc+TIVF の方がよい精度が得られているが，これは TIVF の効果によるものと考えられる．すなわち，4 章で議論したように，弁別的キーポイントのみを用いて認識を実行する場合，弁別的でないキーポイントや全貌が似ていない物体画像の影響により，不正確なマッチが起こる可能性があるが，TIVF では IVF^1 の導入により弁別的ではないキーポイントも含めたコードブックによる検索がなされるため，この影響が回避されたものと考えられる．

BSC-1200 における結果．続いて，BSC-1200 の場合の結果を表 2 の下段に示す．こちらについても BSC-200 と同様の傾向が確認できる．Disc+IVF が Full+IVF よりも高い精度が得られており，弁別的キーポイントの有効性が確認できる．加えて，Disc+IVF^1 の精度も上回っていることから，TIVF を用いることによってさらに精度が改善されていることが分かる．Disc+TIVF の Full+IVF に対する最終的な精度差は依然として大きく，RANSAC 無しの場合で 7.8%，RANSAC 有りの場合では 10.0%に達している．これらのことから，提案法は，見た目の全貌が異なるような物体画像が含まれているような場合であっても，より高精度な検索が可能であることがわかる．

コードブックサイズに対する影響．一般に，認識精度はコードブックサイズに依存する．提案法については，IVF^2 で似て非なる物体間の微小な差異を捉える必要があり，十分な量子化粒度が得られない場合，精度低下が想定される．この精度低下の程度を評価すべく，IVF^2 のコードブックサイズを 2^{11} から 2^{16} まで 6 段階に変化させた場合の認識精度を調査した．比較のため，IVF^1 のコードブックサイズも同様に評価した．結果を**表 3** に示す．この結果，IVF^1 と IVF^2 のコードブックサイズに対する性能変化の傾向に大きな差異がないということが見て取れる．

パラメータに対する影響．パラメータ依存性を評価すべく，Disc におけるキーポイント選択の類似画像数 s および選択率 α を変動させた場合の結果を**図 3** に示す．結果から，いずれのパラメータに対しても多少センシティブであることが分かる．これらの決定は探索的に行う必要があるが，グリッドサーチ等により自動化可能である．一方，非常に広い範囲で，従来法である Full，Full+RANSAC の精度を上回っていることから，提案法の安定した有効性を確認できる．

認識処理時間．BSC-200 における Full+IVF と Disc+TIVF の認識処理時間を**表 4** に示す．TIVF では，IVF による検索が 2 回発生することになるが，これによる処理時間の増加は処理時間全体の 10%程度であり，ボトルネックにはな

表 4　認識時間 (秒) の比較.

手法	特徴抽出	IVF[1] visual word 量子化	IVF[1] 検索	IVF[2] visual word 量子化	IVF[2] 検索	キーポイントマッチング	合計
Full+IVF+RANSAC	0.737	0.412	0.006	-	-	4.999	6.154
Disc+TIVF+RANSAC	0.737	0.412	0.006	0.412	0.006	3.246	4.820

表 5　SMVS における結果．トップ 1 適合率 (%) を記載する．

手法	DVDs	Videos	Books	CDs	Prints	Cards	Art	Landmarks	平均
Liu[10]	73.8	84.3	89.0	68.5	29.5	54.0	47.8	10.2	57.1
Voctree[10]	70.0	82.5	86.2	66.3	29.0	53.7	47.0	11.2	55.7
Affine[14]	96.3	95.8	93.3	94.3	60.8	80.3	83.8	7.6	76.2
Epipolar[14]	**99.5**	**99.8**	97.8	**98.3**	76.0	97.3	89.3	55.0	89.1
Lowe's GHT[14]	**99.5**	**99.8**	**98.8**	98.0	74.8	97.0	90.1	38.6	87.1
Shimamura[14]	**99.5**	**99.8**	**98.8**	98.0	82.3	97.0	93.7	**63.4**	91.6
Full+IVF+RANSAC	98.5	95.8	96.5	92.5	98.0	97.5	92.3	26.4	87.2
Disc+IVF[1]+RANSAC	99.3	99.5	98.3	91.0	98.3	97.5	93.1	48.9	90.7
Disc+TIVF+RANSAC	**99.5**	**99.8**	**98.8**	94.3	**99.0**	**98.8**	**96.2**	54.3	**92.6**

表 6　Mobile における結果．トップ 1 適合率，およびトップ 10 適合率を記載する．

手法	トップ 1 適合率 (%)	トップ 10 適合率 (%)
Wang[19]	83.8	89.6
Zheng[20]	96.0	98.8
Full+IVF+RANSAC	92.1	99.6
Disc+IVF[1]+RANSAC	93.9	99.7
Disc+TIVF+RANSAC	**96.6**	**99.9**

らない．一方，弁別的キーポイント選択によるキーポイント数の減少により，キーポイントマッチングは Full+IVF よりも少ない時間で処理を終えることができる．結果的に，認識処理時間全体では Full+IVF の場合と同等またはそれ以下の時間で処理を終えることが可能である．

5.4　SVMS および Mobile における結果

次に，SVMS，Mobile データセットにおいて本手法を様々な手法と比較する．

SMVS における結果． まず SMVS では，本手法と[10] および[14] で報告されている精度を比較する．我々が知る限り，本データセットで最も高い精度を達成しているのは Shimamura ら[14] による方法である．結果を**表 5** に示す．多くのカテゴリにおいて，本手法 (Disc+TIVF+RANSAC) は他の手法を上回った．Landmark においては Shimamura らの手法が最も精度が良いが，これは建物などの 3 次元物体を考慮した幾何検証を行っているためであると想定できる．また，本手法は Full+IVF+RANSAC の精度を上回っており，類似する物体が支配的でないデータベースの場合においても有効であるといえる．

Mobile における結果． 次に，Mobile データセットにおける結果を**表 6** に示す．本手法を，Wang ら[19] および Zheng ら[20] の手法と比較する．本手法のトップ 1 ，トップ 10 精度はそれぞれ 96.6%，99.9%であり，いずれも他の手法を超える精度を達成できている．

提案法の有効性． SMVS，Mobile データセットにおいても，提案法が最も良い精度を示す理由を分析する．比較している従来法は，いずれも提案法のように似て非なる物体を誤認識する原因となるような特徴を積極的にフィルタするようなものではない．一方，SMVS，Mobile データセットでは，従来法では誤認識を引き起こすほどによく類似した物体を含む画像が，BSC ほどでは無いにせよ少なからず存在する．結果，これらを正しく認識できるようになったことを理由に精度が改善されていると考えられる．このことは，**図 4** の認識結果例からも見て取れる．クエリ画像 (1 列目) に対して，従来法 (2 列目) の認識結果は，レイアウトや一部文字列が同一ではあるが物体としては異なるような雑誌やポスターを結果として返している．提案法 (4 列目) で

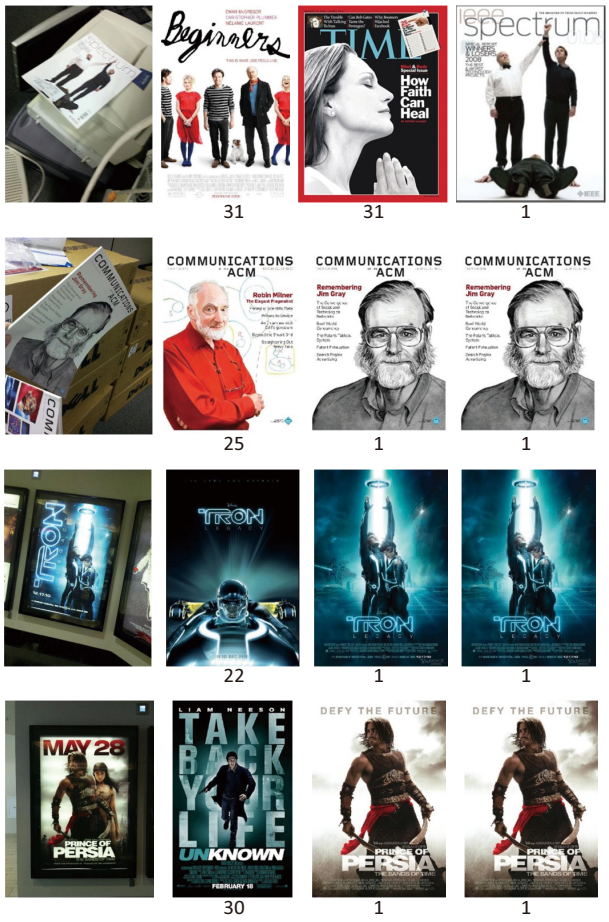


図　4　Mobile における認識結果例．1 列目がクエリ画像，2〜4 列目がそれぞれ Full+IVF+RANSAC，Disc+IVF[1]+RANSAC，Disc+TIVF+RANSAC の認識結果．各認識結果下の数字は，正解のランキングを表す．

は，混同しやすい部分の特徴を避ける効果により，正しい結果を返すことができていることが分かる．以上のことから，提案法の有効性を確認することができたと言える．

6. むすび

類似する物体群からの特定物体認識の問題を考え，特定の物体にのみ存在する弁別的キーポイントを選択する方法，および，よりこれを用いて効率的な認識を行うための TIVF を提案した． 3 種類のデータセットを用いた評価により，他の手法を上回る精度を達成した．

〔文　献〕

1）S. Cao and N. Snavely: ”Minimal Scene Descriptions from Structure from Motion Models”, In CVPR (2014)
2）V. Chandrasekhar, D. M. Chen, S. S. Tsai, N. Cheung, H. Chen, G. Takacs, Y. A. Reznik, R. Vedantham, R. Grzeszczuk, J. Bach, and B. Girod: ”The Stanford Mobile Visual Search Data Set”, In ACM Multimedia Systems (2011)
3）O. Chum, J. Matas, and Š. Obdržálek: ”Enhancing RANSAC by generalized model optimization”, In ACCV (2004)
4）G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray: ”Visual categorization with bags of keypoints”, Workshop on Statistical Learning in Computer Vision, ECCV, pp. 1–22 (2004)
5）W. Hartmann, M. Havlena, and K. Schindler: ”Predicting Matchability”, In CVPR (2014)
6）H. Jégou, M. Douze, and C. Schmid: ”Improving bag-of-features for large scale image search”, IJCV, **87**, 3 (2010)
7）H. Jégou, M. Douze, and C: ”Schmid, Product Quantization for Nearest Neighbor Search”, IEEE Trans. PAMI., **33**, 1, pp. 117–128 (2011)
8）A. Khosla, N. Jayadevaprakash, B. Yao, and F. F. Li: ”Novel dataset for fine-grained image categorization”, In First Workshop on Fine-Grained Visual Categorization, CVPR (2011)
9）J. Knopp, J. Sivic, and T. Pajdla: ”Avoiding Confusing Features in Place Recognition”, In ECCV (2010)
10）X. Liu, B. Lang, Y. Xu, and B. Cheng: ”Feature Grouping and Local Soft Match for Mobile Visual Search”, Pattern Recognition Letters, **33**, 3, pp. 239–246 (2012)
11）D.G. Lowe: ”Distinctive Image Features from Scale-Invariant Keypoints”, IJCV, **60**, 2, pp. 91–110 (2004)
12）J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman: ”Object retrieval with large vocabularies and fast spatial matching”, In CVPR (2007)
13）R. Raguram, J. Tighe, and J.-M. Frahm: ”Improved Geometric Verification for Large Scale Landmark Image Collections”, In BMVC (2012)
14）J. Shimamura, T. Yoshida, and Y. Taniguchi: ”Geometric Verification Method to Handle 3D Viewpoint Changes”, In MIRU (2014)
15）J. Sivic and A. Zisserman: ”Video Google: A Text Retrieval Approach to Object Matching in Videos”, In ICCV (2003)
16）A. Torii, J. Sivic, T. Pajdla, and M. Okutomi: ”Visual Place Recognition with Repetitive Structures”, In CVPR (2013)
17）P. Turcot and D. Lowe: ”Better Matching with Fewer Features: The Selection of Useful Features in Large Database Recognition Problems”, In ICCV (2009)
18）P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona: ”Caltech-ucsd birds 200”, California Institute of Technology (2010)
19）X. Wang, M. Yang, T. Cour, S. Zhu, K. Yu, and T. X. Han: ”Contextual Weighting for Vocabulary Tree based Image Retrieval”, In ICCV (2011)
20）L. Zheng, S. Wang, Z. Liu, and Q. Tian: ”Packing and Padding: Coupled Multi-Index for Accurate Image Retrieval”, In CVPR (2014)

渡邉　之人（わたなべ　ゆきと） 2013 年 阪大大学院博士前期課程了．同年日本電信電話（株）入社．現在，NTT メディアインテリジェンス研究所研究員．パターン認識，画像処理の研究開発に従事．

入江　豪（いりえ　ごう） 2004 年 慶大・理工卒，2006 年 同大院修士課程了．同年日本電信電話（株）入社．現在，NTT コミュニケーション科学基礎研究所研究主任．2012〜2013 米コロンビア大学客員研究員．マルチメディア検索と解析の研究に従事．博士（情報理工学）．正会員．

黒住　隆行（くろずみ　たかゆき） 1997 年 都立大・理・物理卒．1999 年 北陸先端科学技術大学院大学情報科学研究科博士前期課程了．同年日本電信電話（株）に入社．2007 年 北陸先端科学技術大学院大学情報科学研究科博士後期課程了．博士（情報科学）．現在，NTT メディアインテリジェンス研究所主任研究員．メディア探索，パターン認識，画像処理の研究に従事．

杵渕　哲也（きねぶち　てつや） 1997 年 東北大学大学院・博士課程前期了．同年日本電信電話（株）入社．現在，NTT メディアインテリジェンス研究所主幹研究員．パターン認識，画像処理の研究開発に従事．