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.

[招待講演] 画像・映像の認識と意味的検索

柳井　啓司†

† 電気通信大学 大学院情報理工学研究科 総合情報学専攻
〒 182–8585 東京都調布市調布ヶ丘 1–5–1
E-mail: yanai@cs.uec.ac.jp

あらまし　近年，画像・映像データの爆発的な増大によって，大量の画像・映像に対する検索技術の必要性が高まってきている．Web 上ではタグと呼ばれるユーザによって付与されたテキストを手がかりにテキストベースの検索をすることが一般的であるが，研究としては，画像・映像に対してその内容の自動認識を行ない検索を行う研究が広く行われている．本講演では，現在行われている画像および映像に対する認識技術および検索技術の研究について解説を行う．
キーワード　画像認識, 映像認識，TRECVID，マルチメディア映像検索

[Invited Talk] Image/Video Recognition and Retrieval

Keiji YANAI†

† Department of Informatics, Graduate School of Informatics and Engineering,
The University of Electro-Communications
Chofugaoka 1–5–1, Chofu-shi, Tokyo, 182–8585 Japan
E-mail: yanai@cs.uec.ac.jp

Abstract　As the amount of image and video data increases exponentially, the needs to search a large amount of visual data are getting larger. Currently it is commom that images and videos on the Web have text tags attached by hand for tag-based search. On the other hand, image/video recognition and retrieval are being studied very actively. In this talk, we explain the current state on image/video recogntion and retrieval research.
Key words　object recognition, video recognition, TRECVID, image and video retrieval

1. はじめに

デジタルカメラ・ビデオの普及，記録メディアやハードディスクの大容量化によって，一般の個人が大量にデジタル写真を撮影し，それを蓄積することが容易に可能となり,．さらに Web 上には，個人が撮影した写真やビデオなどを不特定多数の人と共有する Flickr, Picasa, Youtube などの social media と呼ばれる画像・動画共有 Web サイトが普及し，誰もが自分の撮影した画像・映像を公開し，他人の撮影した画像・映像を閲覧することが可能となった．このように我々の周りにはデジタル画像・映像データが溢れている一方で，大量の画像・映像データに対して検索を行なうには，人手によって付けられたタグと呼ばれるテキスト情報や，自動で付けられた撮影日時の情報を用いているのが現状であり，画像・映像が表すシーンや対象物の意味内容を計算機が理解し，それに基づいて意味内容による検索を行なうことはほとんど実用化されていない．これは，一般の画像・映像が表すシーンや含まれている物体，動作，イベントなどの多様性が極めて高く，人間の顔などの一部の対象を除いて，実用的な認識精度を達成することが困難なためである．

しかしながら，一般的な画像・映像に対する認識の研究は，近年になって急速に進歩しており，認識技術を用いた意味的な検索の実用化まであと僅かの段階まで近づいてきている．これは，一般的な物体に対する認識技術である「一般物体認識」技術の発展に負うところが大きい.「一般物体認識」とは，制約のない一般的な実世界シーンの画像に対して，計算機がその中に含まれる物体もしくはシーンを一般的な名称で認識することで，画像認識の研究において最も困難な課題の一つとされている．なぜなら，制約のない画像における「一般的な名称」が表す同一カテゴリーの範囲が広く，同一カテゴリーに属する対象の見た目の変化が極めて大きいために，(1) 対象の特徴抽出，(2) 認識モデルの構築，(3) 学習データセットの構築，が困難なためである．そのため，2000 年以前はほとんど研究が行われていなかったが，近年の，(1) 新しい画像表現の提案，(2) 機械学習手法の進歩，(3)Web の普及によるデータセット構築の容易化，(4) 計算機の高速化・大容量化，によって研究が大きく発展し，一般物体認識の実用化が現実のものとなってきている．一般物体認識技術を用いることで，画像・映像に自動的にテキストを付与することが可能となり，例えば,「飛行機」や「自転車」が写っている画像や映像中の部分を，事前に人手によるキーワード付けをすることなく，検索することが可能となる．

一般物体認識の研究が発展する前の 2000 年以前は，画像・映像に対する「認識」と「検索」の研究は，別々に行われてきた．大量の画像・映像データに対する「検索」は，見た目が類似している画像や映像中のシーンを検索することが主な目的で，色やテクスチャパターンなどの特徴量分布に基づいて，画像同士の類似度を定義し，類似している画像やシーンの検索を行った．こうした類似検索による画像・映像検索を content-based image/video retrieval (CBIR/CBVR) と呼ばれていた．なお，CBIR/CBVR では，検索のクエリーが画像で，出力も画像となる．例えば,「自転車」のようなキーワード検索は，データベース画像に人手によって事前にキーワードが付与されていない限り不可能であった．このような単純な画像特徴量の類似に基づく画像検索は，特徴量が画像の意味内容に直接結びつかないため，見た目が類似している画像は検索できても，意味的に類似している画像を検索することは難しかった．例えば，ライオン画像に対して，類似画像として黄色い車の画像が出力されるようなことが一般的に起こっていた．こうした問題は,「セマンティックギャップ」と呼ばれており，セマンティックギャップを越えるには，画像・映像の「認識」を行なうことが不可欠である．CBIR の研究例は多いが，CBIR で用いられている技術ではセマンティックギャップを越えることは困難で，そのため実用的に使われている例はあまり多くはない．現在においては，一般物体認識技術の発展により，画像・映像検索におけるセマンティックギャップは徐々に狭められつつあると言える．

このように意味内容に応じた画像・映像「検索」実現のためには，その前段階として意味内容の「認識」を行なうことが一般的である．例えば，映像検索の競争型国際ワークショップである TRECVID では，200 時間近い映像データに対して予め決められた “hand” や “singing” などの概念を認識する semantic indexing task (2009 年までの high level feature extraction task) と，“sun setting into the clouds” や “someone playing electric guitar in their living room” などのシーン記述文に対応するシーンを検索する known-item search task (2009 年までの search task) が課題として設定されており，semantic indexing は主に一般物体認識技術を用いてシーンを認識し，複数の概念の認識結果を複合させて search task ではシーン検索を行なう．

本稿では，まず第 2 節で画像・映像検索のための基礎的な技術である物体認識技術を説明し，さらにその映像認識への応用についても述べる．第 3 節では，映像検索の事例として TRECVID における取り組みを説明し，最後に第 4 節で全体をまとめる．

2. 物体認識技術の発展

物体認識技術の目指すところは，まさに計算機による画像・映像の意味理解のことであり，人工知能研究が始まった当初からの重要な問題の一つである．長年，研究が続けられていたが，つい最近までは制約のない一般的な画像・映像の意味理解は実用化が困難であると考えられていた．ところが 2000 年前後に起ったブレークスルーにより，画像・映像の意味理解のための技術は近年，急速に進歩を遂げており，その実現が現実味を帯びてきている．

画像の意味理解の研究は，画像認識の研究分野で，**一般物体認識（generic object recognition）** と呼ばれ，近年活発に研究が行われるようになって来ている [1]．具体的には，写真画像中の「山」「海」などの一般的なシーンの意味カテゴリーや，「椅子」「ライオン」など画像に含まれる物体の意味カテゴリーを認識する研究が行われている．一般物体認識技術を用いることで，画像に対する自動キーワード付けや，画像の意味内容による分類や検索などが可能となることが期待できる．また，一般物体認識の技術は，画像のみならず映像に対しても手法が拡張され,「歩く」「立ち上がる」などの一般的な動作の認識に加えて，物体の認識と組み合わせた「携帯電話で通話する」「自動車に乗り込む」などの複雑なシーンを認識する研究も最近では行われている．

このような画像中の対象のカテゴリーを認識する**一般物体認識**に加えて，画像中の同一の対象を認識する**特定物体認識**という認識もある．**特定物体認識**は,「東京タワー」などの特定のランドマークや「iPhone」などの特定の工業製品のようなまったく同じ形状の物体に対する認識技術で，大量の画像データベースに対して高速な検索を行うことが研究課題となっている．図 1 に 2 つの認識ついての処理の流れについて記す．特定物体認識では，例えば，多くの時計の写真をデータベースに登録しておいて，同一の時計が入力画像に存在するかを調べる．認識対象の画像中の局所パターンとほぼ一致する局所パターンをもつ画像をデータペース中から検索することによって認識を行うため，物体の位置検出も同時に行うことができる．一方，一般物体認識の例では，入力画像が「ライオン」か「トラ」かを判定するが，その際に入力画像とまったく同じライオンの写真が学習画像になくても，学習による汎化によって，それがライオンであると認識する必要があり，特定物体認識より困難な問題であると言える．特定物体認識と異なり，局所パターンの直接の対応でなく，その分布を用いて認識を行うため，位置まで特定する場合はさらに一段難しい問題となる．

図 2 に参考までに現在研究されている一般物体認識の主なタスクを 5 種類示す．**画像全体のカテゴリー分類**が最も標準的なタスクで，複数のカテゴリーラベルを画像に付与する**画像アノテーション**，領域分割された画像の各領域に対してカテゴリーラベルを付与する**画像ラベリング**，長方形の矩形で画像中の物体の存在位置を検出する**カテゴリー物体検出**, 物体の領域を正確に切り出す**カテゴリー領域抽出**などのタスクが研究課題として扱われている．この中の主に，画像全体のカテゴリー分類と，複数ラベルを付与する画像アノテーションが画像・映像検索に応用されている．

以上のように一般物体認識と特定物体認識は，その目的や処理や認識モデルは異なっているが，最も基礎となる認識対象の特徴表現手法はほぼ同一である．どちらも，画像から多数の対象に特徴的な局所的なパターンを局所特徴量として抽出して認識に利用する．

本節では，画像・映像の意味的検索の実現のために必要となる画像の意味理解に関する研究の現状について,（1）新しい画像特徴表現手法である bag-of-features,（2）特徴統合による画像・映像認識，について解説する．

2.1 物体認識の画像表現: Bag-of-features

2000 年前後までは，物体認識，特に画像の意味理解といえる一般物体認識問題は困難な問題として考えられていて，画像・映像は自動認識を期待しないで積極的にメタデータを付与するという動きが MPEG-7 や Semantic Web の提案とともに出てくるようになった．それと前後して，90 年代の後半から 2000 年代の前半にかけて，一般物体認識に関するブレークスルーが

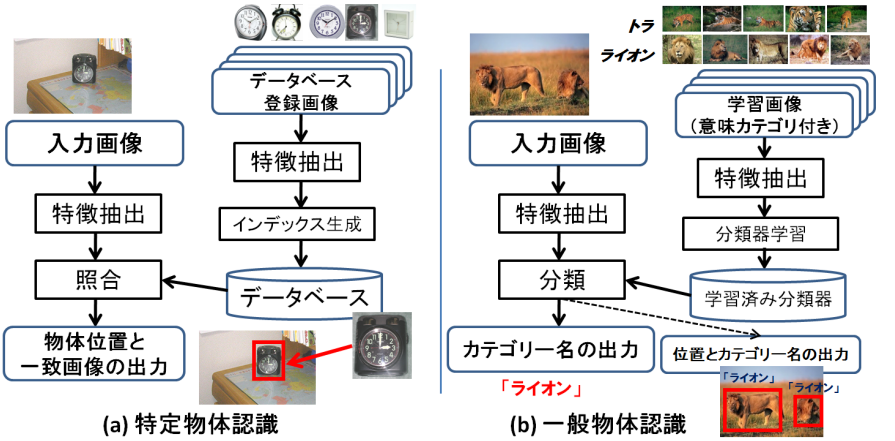

図 1　特定物体認識と一般物体認識の違い．同一物体をデータベース中から検索するのが特定物体認識，意味カテゴリーを当てるのが一般物体認識．

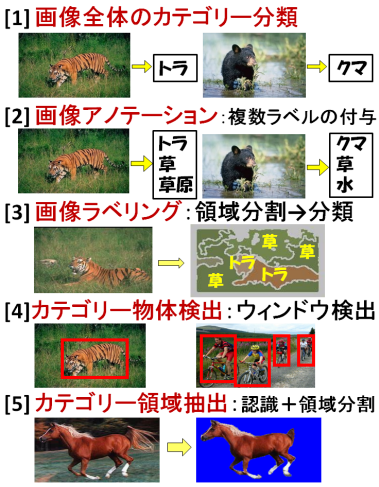

図 2　一般物体認識 (カテゴリー認識) の主要な 5 種類のタスク．

起った．それに関する重要な研究は，(1) 局所特徴の組合せによる画像の表現，(2) 局所特徴の表現法，そして，(3) 局所特徴のヒストグラム表現である bag-of-features である．

まずは 1990 年代後半に，認識対象全体を用いるのではなく，認識対象の特徴的な局所パターンを多数抽出し，その組合せによって，画像検索および特定物体認識を行う方法が提案された [2]．認識に用いる特徴的な部分の抽出には，元々はステレオ 3 次元復元やパノラマ画像生成に必要な複数画像の対応点検出のために研究されてきた局所特徴抽出手法が利用された．代表的な方法としては，特徴点検出と特徴ベクトルの抽出法をセットにした SIFT（Scale Invariant Feature Transform）法 [3] がある．

SIFT 法は (1) 特徴点とその点の最適スケールの検出，(2) 特徴点の周辺パターンの輝度勾配ヒストグラムによる 128 次元ベクトルによる記述，の 2 つの処理を含んだアルゴリズムである．画像中のエッジやコーナーなどの特徴的な部分が特徴点として自動的に検出され，さらにその周辺パターンに基づいてパターンのスケールと主方向が決定され，回転，スケール変化 (拡大縮小)，明るさ変化に不変な形でその周辺パターンが特徴ベクトルとして記述される．SIFT 特徴量は，回転，スケール変化，明るさ変化だけでなく，一定の範囲内のアフィン変換 (視点の移動) にも頑健であることが実験によって示されている．つまり，図 3 に示すように，1 枚の画像で特徴点が抽出されベクトルで記述されると，もう一枚の回転，縮小，明るさ変化を加えた画像でも，同じ場所から特徴点が抽出され，その点のベクトルの値もほぼ等しくなる．そうすることにより，SIFT 法で抽出した特徴ベクトルの探索のみで，異なる画像間の対応点が検出できることになる．また，SIFT 法は濃淡画像の輝度勾配を特徴量としていて，色情報を一切使っていないため，色が異なっていても濃淡の変化が似ているなら類似パターンと見なされることも特徴である．

抽出する特徴点の数はパラメータによって制御可能であるが，通常は多くの対応点の候補が多数あった方がより処理が頑健になるので数百から数千個の特徴点を抽出する．そのため，多数の対応点が得られ，多少の誤対応や，部分的な隠れによる対応点の減少が起っても，ある程度の範囲内なら，物体の対応をとることが可能となる．以上が特定物体認識の基本原理である．

SIFT 法のアルゴリズム自体は実装は容易であるとは言えないが，提案者の D. Lowe 自らによるものを初め，いくつかのソフトウェアが Web 上に公開されており，手軽に利用可能となっている．なお，SIFT に関する日本語の解説としては，中部大の藤吉先生による解説 [4] が詳しい．SIFT 以外にも同様の局所特徴量は数種類提案されており，特に SURF [5] は，オープンソース画像認識ライブラリである OpenCV のバージョン 1.1 以降にライブラリ関数として取り込まれているため，手軽に利用可能である．また，[6] の HP で公開されているソフトウェアのように，SIFT 特徴に加えて，後述する bag-of-features 特徴量まで 1 つのコマンドで抽出可能な公開ソフトウェアも存在する．

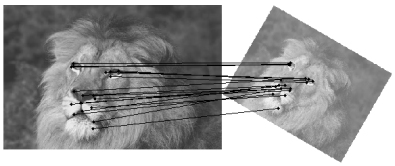

図 3　SIFT 特徴量を使った局所パターンのマッチングの例．

2.2　Visual Words と Bag-of-features

SIFT 法に代表される局所特徴量による認識は，高精度で頑健な特定物体認識を可能としたが，1 つの画像から数百から数千のもの多数の局所特徴量を抽出すると，多数の画像に対して特徴点を高速に照合することが困難になる．そこで，1 枚の画像から多数抽出される局所特徴ベクトルをベクトル量子化し，代表ベクトルである code word に置き換えて，対応点の検索を行う手法が提案された [7](図 4)．代表ベクトルは visual word とも呼ばれ，特定物体認識を行う場合はその数はデータベースのサイズに応じて，1 万から 100 万程度の値が選ばれる．この visual words の考え方を用いると，画像から抽出された局所特徴ベクトルは単語 (visual word) に変換されるので，1 つの画像は数百から数千の単語の集合によって表現されることになる．つまり，画像は，文章や Web ページなどと同じで，単語の集合として表現されることになる．実際に visual words を提案した論文 [7] では，映像検索にこの手法を応用した．

J. Sivic らはビデオ映像から視点の異なる同一シーンを検索可能なシステム “Video Google” を提案した [7]．SIFT 特徴 [8] をベクトル量子化し visual word を作成し，ビデオ中の各フレーム画像は多数の visual word を含んでいると考えた．そして，テキスト検索の手法を応用し高速な特定物体認識を実現し

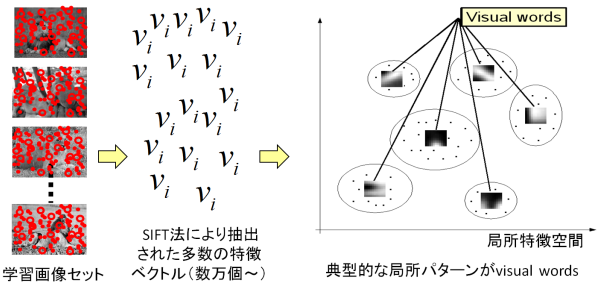

図 4　代表局所パターン (visual word) の求め方．認識対象の学習データセットから局所特徴特徴ベクトルを抽出し，クラスタリングで visual words を求める．

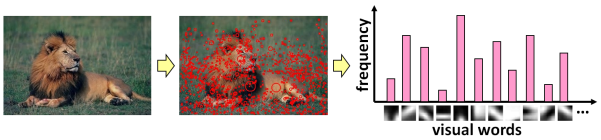

図 5　Bag-of-features 表現の求め方．すべての局所特徴量を visual words に対応させ，ヒストグラムを作成する．

た．この研究では，visual word の考え方を導入することで，クエリー画像と同一物体を含む画像や映像の検索が，テキスト検索と同様に転置インデックスを用いて高速に実行することができることを示した．

Visual words の最初の論文は特定物体認識を目的としていたため，それだけでは一般物体認識への適用は不可能であった．局所特徴量および visual words を一般物体認識に応用すること可能としたのは，bag-of-features 表現 (BoF) [9] である．

文章をベクトル表現する方法として，語順を無視して単語の出現頻度ベクトルで文章を表現する bag-of-words(BoW) 表現が言語処理や情報検索の分野で用いられているが，それとまったく同様に，各特徴点の画像中での位置，つまり visual words の位置を無視して，visual words を bag-of-words 化したのが，bag-of-features 表現である (図 5)．そのため，bag-of-visual-words（BoVW）と呼ばれることもある．なお，BoW と同様に TF-IDF で要素を重み付けする方法や，各要素を 1 か 0 の 2 値にしてヒストグラムを構成する方法も試みられているが，一般には，単純に visual words の出現回数をカウントしてヒストグラムを作成し，BoF ベクトルとする場合が多い．なお，BoF ベクトルは，各画像の visual words の合計頻度に差がある場合は，L1 正規化することが一般的である．

Bag-of-features は，結局，画像から抽出された局所特徴量の分布を visual words のヒストグラムで表現しているということである．ヒストグラムは，色に関しては従来より画像表現の一つとして利用されてきたが，色ヒストグラムは似た色の画像の検索には有効であったものの，色は物体のカテゴリーとは必ずしも直接結び付かないために，カテゴリー認識を目的とした一般物体認識においてはあまり有効ではなかった．それに対して，局所パターンは物体のカテゴリーと関係が深く，その分布のヒストグラムである bag-of-features は多くの一般物体認識の研究においてその有効性が示されている．さらに，注目すべきは bag-of-features の元となる局所特徴量は濃淡変化のみに注目していて bag-of-features には色に関する情報はまったく含まれていないにも関わらず，従来の色などの特徴量よりも高い精度でカテゴリー分類が可能となっており，物体のカテゴリー認識には色情報は重要ではないということが実験結果から示された形になっている．

Bag-of-features 表現はヒストグラム表現であるため，各局所パターンの位置の情報が完全に捨てられてしまっているが，逆にその潔さが表現の簡潔さにつながり，現在，一般物体認識において標準的な画像表現手法として広く使われるに至っている．

なお，一般物体認識においては同じカテゴリーに属する物体の細かな差異が吸収されることが望ましいので，visual words のサイズは，特定物体認識ほどは大きくせずに数百から数千程度である．一方，特定物体認識では，まったく同じ局所パターンだけが 1 つの visual word に割り当てられることが望ましいので，数万から百万程度のサイズが一般的である．特定物体認識においては visual words は広く用いられているが，局所パターンの分布が類似していることよりも，一致する局所パターンが一定数存在するということの方が重要であるので，分布を表現した bag-of-features ベクトルを機械学習手法で学習するようなことは特定物体認識においては行われない．

Bag-of-features の画期的な点は，bag-of-features 表現に変換された画像は文章とまったく等価に扱うことができる点である．そのため，bag-of-features が提案された直後は，競って言語処理の分野で提案された手法が画像認識に応用されるということが起った．特にカテゴリー分類においては，bag-of-words 表現が数千から数万次元もの高次元になるテキスト分類で定評のあったサポートベクターマシン（SVM）が同様に数百から数千次元になる bag-of-features ベクトルに対しても幅広く用いられている．

なお，SIFT 法などの局所特徴量抽出手法は，特徴点の検出の処理も含んでいるが，第一段階の処理の特徴点検出を用いずに，決められたピクセル毎の格子点 (グリッド) やランダムに選ばれた点を機械的に特徴点とする方法も一般物体認識においては広く用いられている．特徴点検出手法では，空や道路の路面のような均一な領域からは特徴点が得られないが，物体カテゴリーの認識においては，テクスチャのない均一な局所特徴も重要な情報であるため，画像の内容にかかわらず機械的に特徴点の位置およびスケールを選択する方法も有効であるとされている [10], [11].

Bag-of-features（BoF）はシンプルな手法であるために，その拡張が様々な面において試みられている．F. Jurie ら [12] による k-means によるクラスタリングの代りの，オンラインクラスタリングと mean-shift [13] に基づいたコードブック作成法，F. Perronnin ら [14] による GMM および EM アルゴリズムによる確率的クラスタリングによるコードブック作成法，J. Weijer ら [15] による色情報の追加，時空間局所特徴を用いた動画像への拡張などである．

2.3　時空間特徴量

近年，動画の解析のために時空間特徴が注目を集めている．時空間特徴とは，動画から抽出される特徴の一つで動き情報と視覚情報を同時に表現することが可能な特徴である．

Dollar らは静止画像の局所特徴を時間軸方向に拡張した 3 次元の時空間局所特徴をベクトル量子化する bag-of-video-words を提案し，歩く，走るなどの人間の動作の分類に成功している [16]．Dollar らの研究によって動作認識も物体認識と同じ BoF で手軽に扱えることが示されたため，特に，動作認識への BoF の応用は近年急速に広まっている．

主要な時空間特徴抽出手法として，cuboid と呼ばれる立方体を抽出し，それを特徴化する手法がある (図 6)．Dollar らは空

間軸にガウシアンフィルター，時間軸にガボールフィルタを適応することでこの cuboid を抽出する手法を提案した [16].

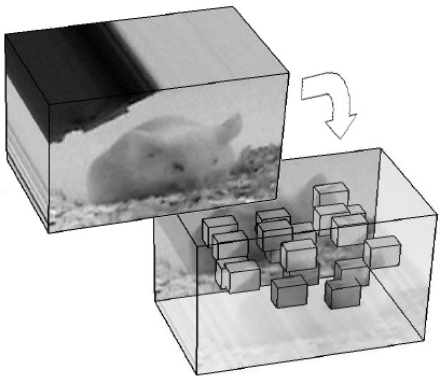

図 6　時空間特徴量である Cuboid.

Web 動画に対する動作認識を行った研究も近年登場している．Cinbis らは Web 上から動作を自動学習する手法を提案し，Youtube データを動作認識に利用している [17]．この研究では，まず query word をもとにして画像を収集し，その画像から特徴を抽出することで動作モデルを構築し，実際の映像において認識を行っている．しかしこの手法の学習データは Web 上から収集された静止画像であり，動作の記述も全て動画像ではなく静止画像ベースで行われている．

他に Web 動画に対する認識の研究として，Liu らの研究が挙げられる [18]．この手法は特徴量として [16] で提案された時空間特徴を，視覚特徴として SIFT 記述子を利用し，Adaboost に基づき統合する手法を提案している．またこの研究では，Page Rank に基づく重要な特徴の選択を行っている．

また，野口らは，動作を認識するために新たな時空間特徴を提案し，動き特徴，静止画像特徴と Multiple Kernel Learning によって統合することによって，Web 動画の分類性能に関して [18] を上回る結果を出している [19].

2.4　特徴統合による画像・映像認識

Bag-of-features（BoF）による画像特徴量は画像の意味理解に有効であるが,「ひまわり」「パンダ」のような特徴的な色を持っている認識対象も存在するため，局所特徴量に基づく BoF のみでなく，色やテクスチャ，エッジ特徴など様々な特徴量を対象に応じて選択的に利用することで認識精度がさらに向上できることが近年示されている．

Varma らは，BoF に加えて，テクスチャ，色，形状などの多様な画像特徴を画像から抽出し，認識対象に応じた適切な重みを Multiple Kernel Learning の手法を用いて推定し，一般物体認識を行うことを提案した [20]．Car，Face，Airplane など一般的な意味カテゴリに対応する画像を 101 種類 9,144 枚，256 種類 30,607 枚含んだ一般物体認識のアルゴリズムの性能比較のための標準的なベンチマークデータセットであるカルフォルニア工科大学の Caltech-101，Caltech-256 (注1) を用いた画像分類実験で，単一の特徴量ではそれぞれ約 65%，約 35%であった分類精度が，Multiple Kernel Learning による特徴統合によって約 90%，約 60% にそれぞれ向上した．2004 年には Caltech-101 の分類精度が 20%未満であったことから，ここ数年で急速に技術が進歩していることがうかがえる．

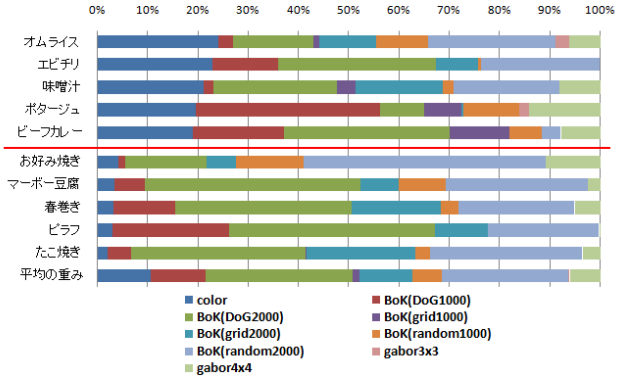

図 7　学習した特徴量の重み．左端の青色が部分が色の重みを示す.

図 8　色特徴の重みが大きかったオムライス（左上），エビチリ（右上），重みが小さかったピラフ（左下），たこ焼き（右下）の画像.

上東らは，同じ手法を 50 種類の食事画像の分類に応用した [21]．実験では 50 種類のマルチクラス分類を行い，平均分類率 61.34%を達成した．単体の特徴では最高でも 34.64%であったので，特徴統合によって大きく性能が向上した．図 7 に学習した特徴量の重みを色特徴の重み（青色）の上位下位 5 種類ずつの食事カテゴリを示す．図 8 に色特徴の重みが大きかったオムライス，エビチリと色特徴の重みが小さかったピラフ，たこ焼きの画像を示す．オムライス，エビチリは，それぞれ黄色と赤色が特徴的な色であるが，ピラフやたこ焼きは様々な色が含まれていて特徴的な色がないため認識には色特徴は重要ではなかった．

一方，映像認識の場合は，映像からは画像に比べると動きや音声などさらに多くの特徴を抽出することが可能であるため，特徴統合はより重要な課題となる．映像認識については，次節の TRECVID で述べる．

3.　TRECVID

映像検索に関する国際ワークショップの TRECVID (注2) [22] では，大量のテレビ映像（2009 年の場合は 97,150 ショットからなる 280 時間の映像）から決められた 20 種類の物体もしくはシーンを含むショットを選び出す高次特徴抽出課題（high-level feature extraction task）およびその結果を用いた検索課題（search task）が実施されている．映像は通常，多数のシーンが含まれているので，映像を同一シーンからなるショットに分割して，ショット毎に認識が行われ，TRECVID では，主催者より公式のショット分割境界の情報が与えられ，各課題では該当シーンをショット単位で回答する．

前者の高次特徴抽出課題は一般物体認識や動作認識そのものであり，2009 年度は，一般物体認識での認識対象としては一般的な chair，telephone，bus などの物体，classroom，traffic intersection，cityscape などの静的シーンに加えて，person-playing-

(注1)：http://www.vision.caltech.edu/Image_Datasets/Caltech256/

(注2)：http://www-nlpir.nist.gov/projects/trecvid/

図 9　2008 年度の 1 位のチームによる classroom（左上），cityscape（右上），telephone（左中），singing（右中），airplane_flying（左下），bus（右下）の 4 位までの認識結果．

a-musical-instrument，person-in-the-act-of-sitting-down のような動作を伴うシーン，singing のような音声を伴うシーンなどの映像ならではのシーンが認識対象として含まれている．高次特徴抽出課題では，主催者から提供される映像の分割単位であるショットを対象に対象物体，シーンを含む候補のショットを最大 2000 まで解答する．参加チームは，各ショットの代表フレーム画像から抽出した BoF，色ヒストグラム，テクスチャ特徴量に加えて，映像独自の動き情報，音，音声をテキスト化した音声認識テキストの bag-of-words ベクトルなど様々な特徴量を抽出し，それらを統合して認識を行う．統合手法としては，単純なベクトルの結合，boosting，個々の特徴の SVM による認識結果の重み付き線型和 (AP weighted fusion)，multiple kernel learning など様々な手法が試みられている．図 9 に 2008 年度の 1 位のチームであるアムステルダム大による 6 種類 (classroom, cityscape, telephone, singing, airplane_flying, bus) の高次特徴の 4 位までの認識結果を示す．Sining は 2 枚，airplane_flying は 1 枚，bus は 3 枚それぞれ誤りが含まれているものの他はすべて正解である．特に，telephone は映像中には小さくしか写っていないにもかかわらず認識が成功している．

後者の検索課題は，高次特徴抽出の認識結果を組み合わせて複雑な映像検索を行うタスクで，例えば，classroom，person-playing-a-musical-instrument，singing の結果を組み合わせて，「教室で楽器を演奏している人と歌っている人がいるシーン」を検索し，ランキングを付けて解答する．

4. ま と め

近年，画像・映像データの爆発的な増大によって，大量の画像・映像に対する検索技術の必要性が高まってきている．そこで，本稿では，現在行われている画像および映像に対する認識技術および TRECVID を中心とした映像検索技術の研究について解説を述べた．

文　献

[1] 柳井啓司: 一般物体認識の現状と今後, 情報処理学会論文誌: コンピュータビジョン・イメージメディア, Vol. 48, No. SIG16 (CVIM19), pp. 1–24 (2007).

[2] Schmid, C. and Mohr, R.: Local Grayvalue Invariants for Image Retrieval, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, No. 5, pp. 530–535 (1997).

[3] Lowe, D. G.: Distinctive Image Features from Scale-Invariant Keypoints, *International Journal of Computer Vision*, Vol. 60, No. 2, pp. 91–110 (2004).

[4] 藤吉弘亘: Gradient ベースの特徴抽出 – SIFT と HOG –, 情報処理学会研究会報告: コンピュータビジョン・イメージメディア研究会, No. CVIM-160, pp. 211–224 (2007).

[5] Bay, H., Tuytelaars, T. and Van Gool, L.: SURF: Speeded up robust features, *Proc. of European Conference on Computer Vision*, pp. 404–415 (2006).

[6] van de Sande, K. E. A., Gevers, T. and Snoek, C. G. M.: Evaluating Color Descriptors for Object and Scene Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2010).

[7] Sivic, J. and Zisserman, A.: Video Google: A Text Retrieval Approach to Object Matching in Videos, *Proc. of IEEE International Conference on Computer Vision*, pp. 1470–1477 (2003).

[8] Lowe, D. G.: Object recognition from local scale-invariant features, *Proc. of IEEE International Conference on Computer Vision*, pp. 1150–1157 (1999).

[9] Csurka, G., Bray, C., Dance, C. and Fan, L.: Visual categorization with bags of keypoints, *Proc. of ECCV Workshop on Statistical Learning in Computer Vision*, pp. 59–74 (2004).

[10] Fei-Fei, L. and Perona, P.: A Bayesian Hierarchical Model for Learning Natural Scene Categories, *Proc. of IEEE Computer Vision and Pattern Recognition*, pp. 524–531 (2005).

[11] Nowak, E., Jurie, F., Triggs, W. and Vision, M.: Sampling strategies for bag-of-features image classification, *Proc. of European Conference on Computer Vision*, pp. IV:490–503 (2006).

[12] Jurie, F. and Triggs, B.: Creating Efficient Codebooks for Visual Recognition, *Proc. of IEEE International Conference on Computer Vision*, pp. I:604–610 (2005).

[13] Comaniciu, D. and Meer, P.: Mean Shift: A Robust Approach toward Feature Space Analysis, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 25, No. 5, pp. 603–619 (2002).

[14] Perronnin, F., Dance, C., Csurka, G. and Bressan, M.: Adapted vocabularies for generic visual categorization, *Proc. of European Conference on Computer Vision*, pp. IV:464–475 (2006).

[15] Weijer, J. v. d. and Schmid, C.: Coloring local feature extraction, *Proc. of European Conference on Computer Vision*, pp. II:334–348 (2006).

[16] Dollar, P., Rabaud, V., Cottrell, G. and Belongie, S.: Behavior Recognition via Sparse Spatio-Temporal Features, *Proc. of ICCV WS on Visual Surveillance and Performance Evaluation of Tracking and Surveillance(VS-PETS)* (2005).

[17] Cinbins, R. I., Cinbins, R. and Sclaroff, S.: Learning action from the web, *Proc. of IEEE Computer Vision and Pattern Recognition* (2009).

[18] Liu, J., Luo, J. and MSha, M.: Recognizing realistic action from videos, *Proc. of IEEE Computer Vision and Pattern Recognition* (2009).

[19] 野口顕嗣, 下田保志, 柳井啓司: 動作認識のための時空間特徴量と特徴統合手法の提案, 画像の認識・理解シンポジウム (MIRU 2010) (2010).

[20] Varma, M. and Ray, D.: Learning the discriminative power-invariance trade-off, *Proc. of IEEE International Conference on Computer Vision*, pp. 1150–1157 (2007).

[21] 上東太一, 甫足創, 柳井啓司: Multiple Kernel Learning による 50 種類の食事画像の認識, 画像の認識・理解シンポジウム (MIRU 2009) (2009).

[22] 佐藤真一: 映像内容検索における TRECVID の取組み, 電子情報通信学会誌, Vol. 91, No. 1, pp. 55–59 (2008).