図2 Bag-of-Features 異の求め方。
画像から SIFT によって特徴抽出し、あらかじめ求められた visual words に関するヒストグラムを作成する。ヒストグラムが画像の特徴量となる。

図3 学習した特徴量の重み
左端の青色の部分が色の重みを示す

図4 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）、重みが小さかったビラフ（左下）、たこ焼き（右下）の画像
図5 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

図6 領域に単語ラベルを付けた結果

図7 左上から順にcow, sheep, dog, person, bicycle, bus, car, motorbike.
水色の枠が正解データ、紫の枠がスライディングウィンドウによる認識結果
セマンティックギャップを越えて—画像・映像の内容理解に向けて—

Crossing the Semantic Gap
—Towards the Understanding of Image and Video Contents—

井手 一郎
Ichiro Ide
名古屋大学大学院情報科学研究科／国立情報学研究所
Graduate School of Information Science, Nagoya University / National Institute of Informatics.
idei@is.nagoya-u.ac.jp / idei@mii.ac.jp

柳井 啓司
Keiji Yanai
電気通信大学情報工学科
Department of Computer Science, The University of Electro-Communications.
yanai@cs.ucc.ac.jp

Keywords: semantic gap, image / video retrieval, tagging, meta-data.

1. はじめに

高速データ通信や大容量記憶装置の普及に伴い、我々の身の回りの情報は爆発的に増加している。その中でも、携帯型デジタルカメラ・映像撮影機器の普及、テレビジョン放送の多チャンネル化、画像・映像共有Webサイトの普及などに伴い、実世界に存在する画像・映像データの量が爆発的に増加している。このような背景を受け、画像・映像情報の意味内容に即して効率的・的確に検索する技術に対する要求が高まっている。

しかし、情報の表現（文字列）が文書の意味内容と直接的に結びつくテキスト情報と比較して、画像・映像情報は表現（画素値）が画像・映像の意味内容に直接的に結びつかない、あるいはさまざまな解釈が得られるために、意味内容に即して検索することが困難である。この問題は「セマンティックギャップ (semantic gap)」と呼ばれ、長年、画像・映像の低次特徴量に基づく内容理解や検索 (CBIR: Content-Based Image Retrieval) を考える際に、研究者の眼前に横たわる深い谷として知られてきた。

1990年代後半に画像・映像検索に関する研究が本格化した頃は、比較的は小規模なデータセットを対象として、低次特徴量に基づく簡単な分類・検索や、簡単なモデルに基づく内容理解が主流だったため、この問題はあまり深刻な問題として認識されていなかった。しかし、1990年代後半以降、WWWや高速データ通信の爆発的な普及に伴い、そのような方法ではインターネット上に大量に存在する多種多様な画像・映像データの検索や内容理解に対応しきれなくななることが認識され始め、深刻な問題として顕在化するようになった。なお、この研究の動向については、[Datta 08]を参照。

本稿では、セマンティックギャップを越えるとする取組みについて、1）いわば橋を架けるとする統計的アプローチと、2）いわば逆回してしまおうとするメタデータ利用によるアプローチの二つに分けて紹介する。

前者は、計算機による画像・映像の意味理解のことであり、2000年以降の急速な画像認識技術の発展によって、その実現可能性が現実に至ってきている。計算機の自動処理であるので100%の精度を期待することは難しいが、大量の画像・映像データが存在する今では有望なアプローチである。後者は、メタデータが付与された時点で意味内容の解釈がはば固定してしまうという欠点はあるが、検索対象や用途を限定した検索モデルを仮定できれば、前者よりも的確な検索を実現できるという利点がある。以後、前者について井手が3章で、後者について柳井が2章で、後者について井手が3章で分担して執筆する。

2. 統計的アプローチ

統計的アプローチとは、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重要な問題の一つである。長年、研究が続けられていながら、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年前後に起こ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、その実現が現実に至ってきている。

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きている。
2.1 一般物体認識技術の発展

一般物体認識は、画像認識の研究が人工知能研究の一
部として行われていた 1960 年代より研究が行われてい
た。しかしながら、1980 年代までは認識のための知識
をルールとして入力によって記述していたため、認識対
象を増やすことが困難であること（人工知能研究
における「知識獲得のボトルネック」の問題）に突き当
たり、実際に画像以外の一般の画像を対象とした認識を
実現することはできなかった。その後も研究は続けられ
ていたが、画像の意味理解という観点からはあまり研究
が行われず、画像中の物体とデータベースに登録された
物体モデルの照合を行う「特定物体認識」が主に研究さ
れた。一般物体認識、特定物体認識とは異なり、認識
対象である意味カテゴリーに対応するカテゴリーの範囲
が広く、同カテゴリーに属する対象のアピール（物
体の見た目）の変化を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
トの構築が困難であり、2000 年前後までは認識技術
としての研究はほとんど行われていなかった。

一方、画像認識とは異なる研究分野として、画像データ
ベース検索という研究分野があり、見た目が類似して
いる画像を画像データベース中から検索すること。内容に基づ
く画像検索（Content-Based Image Retrieval: CBIR）
の研究が 1990 年代から行われるようになった。画像の
色やテクスチャの分布をヒストグラム化し、多点ピク
トルで画像を表現、類似画像検索することに
よって、類似画像検索が実現した。この手法は現在でも
画像データベース検索の標準的な手法として用いられて
いるが、画像色やテクスチャなどの低レベル特徴のみを用い
ているために、見た目が類似している画像や、「暖かい」、
「暗い」などの主観的な形容詞に対応する画像は検索す
ることが可能である、意味的に類似している画像を検
索することは困難であった。そのため、CBIR の研究例
は多いが、実用的に使われている例はあまり多くはない。

以上述べたように 2000 年前後までは画像の意味理解
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
ブレーキシュートが起こった。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部分を多数抽出し、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 [Schmid 97]。認識に用いる特徴的な部分の
抽出には、ステレオ三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の局所特徴抽出手法が
利用された。代表的な手法としては、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) [Sivic 07, Lowe 04] がある。SIFT
は、回転、スケール（拡大縮小）の変化に不変で、視覚の
変化やアフィン変換（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。

画像検索にもこの手法は応用されており、Sivic らはビデオ映像から視点の異なるシーケンスを検索可能なシ
ステム“Video Google”を提案した [Sivic 03]。SIFT 特
徴 [Lowe 99] をベクトル量化し 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
dていると考えた。そして、テキスト検索の手法を応用し
高速な検索を実現した。

こうした考え方を一般物体認識に応用したのが、最
近の一般物体認識の研究のアームをつくった Bag-of-
Features (BoF) [Csurka 04] である。1 枚の画像から数
百～1000 個のオーダーの多数の局所特徴量を抽出し、そ
の分布をヒストグラムで表現する新しい手法である。ヒ
ストグラムは、色に関わらず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つかないため、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、Bag-of-Features では、ベクトル
量化の考え方を用いて、学習用画像データの各画像か
ら抽出した多数の局所特徴量から代表的な局所パターン
to visual word として選び出し（図 1）、それに関するヒ
ストグラムを作成することによって画像を表現する（図
2、図解）。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。

統計的言語処理においては、文章をベクトル表現する
方法として、語順を無視して文章を単語の出現頻度ベクト
ルで表す Bag-of-Words が利用されているが、それと
図1 代表画像ベクトル（visual word）の求め方。認識対象の学習データセットのすべての画像からSIFT法により局所特徴ベクトルを多数抽出し、あらかじめ求められたk個のvisual wordsをk-means法によるクラスタリングで求めめる。

図2 Bag-of-Features表現の求め方。画像からSIFT法によって特徴抽出し、あらかじめ求められたvisual wordsに関するヒストグラムを作成する。ヒストグラムが画像の特徴量となる。

同様にBag-of-Featuresでは位置を無視して画像を局所特徴（features）の集合として表現している。実際の処理においては、代表的な局所特徴ベクトルを選び出し、すべての局所特徴ベクトルを最も近い代表ベクトルで近似してヒストグラムを構築するため、代表的な局所特徴ベクトルをvisual wordsと考えるとBag-of-Featuresは、Bag-of-Wordsの画像版と考えることができる。Bag-of-Visual-Words（BoVW）と呼ばれることがある。

画像をBag-of-Featuresによって表現することによって、Bag-of-Words表現されたテキスト文書と同様に扱うことが可能となり、画像分類が文書分類と同じ問題として取り扱うことが可能となった。そのため、Bag-of-Features表現された画像の分類は、テキスト分類と同様にサポートベクターマシン（SVM）が一般的に用いられ、カーネル関数としてはヒストグラム特徴に有効であるといわれているカイ2乗RBFカーネルが用いられることが多い。また文書分類のための確率的トピック抽出の手法として提案されたprobabilistic Latent Semantic Analysis（pLSA）[Hofmann 01]、Latent Dirichlet Allocation（LDA）[Blei 03]などが画像分類に応用されている。

Bag-of-Features（BoF）はシンプルな手法であるため、その拡張がさまざまな面において試みられている。コードブック作成法の工夫、局所特徴の位置情報の利用、動画像への拡張、スライディングウィンドウによる位置特徴、領域分割との融合などである。Dollarらは静止画像の局所特徴を時間軸方向に拡張した三次元の時空間局所特徴ベクトル量化化する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た[2005]。

2.2 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、「すみれ」、「パンダ」のような特徴的な色をもつし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、エッジ特徴などさまざま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、形状などの多様な特徴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[Varma 07]。car、face、airplaneなど一般的な意味カテゴリーに対応する画像を101種類9144枚、256種類30607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テル着工科大学のCaltech-101、Caltech-256[2]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によって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。

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[上東 09]。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3、口絵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ーを示す。図4、口絵に色特徴の重みが大きかったオムライス、エビチリおよび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さまざま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、映像からは映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、ショットごとに認識が行われる。

映像検索に関する国際ワークショップのTRECVID[8][佐藤 08]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

*2 http://www.vision.caltech.edu/Image_Datasets/Caltech256/
*3 http://www.nlpirl.nist.gov/projects/trecvid/
 lerの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。

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年度版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-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、主催者より提供される映像の分割単位であるショットを対象に物体認識、シーンを含む候補のショット数を最大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質から抽出したBoF,色ヒストグラム、テクスチャ特徴量に加えて、映像特徴の動き情報、音、音声をテキスト化した音声認識特徴のBag-of-Wordsベクトルなどさまざまな特徴量を抽出し、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の線形和、multiple kernel learningなどさまざまな手法が試みられている。図5に、口絵に2008年度の1位のチームであるアムステルダム大学による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。Singing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確である。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。

図5 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

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、「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、ランキング付けで解答する。

2.3 対象位置の検出

これまで説明したCaltech-101/256やTRECVIDは、画像やショット中に対象が含まれるか1/0で分類し認識を行った。それに対して、対象が画像中や映像中どこに含まれているかを検出する位置検出を伴う認識に関する研究も行われている。方法としては、画像を領域ごとに分割して領域ごとに認識する手法と、画像の一部にウィンドウを設定し、それを拡大縮小してスライドさせながら、各ウィンドウに対して画像全体を分類するのと同様の方法で分類を行い、画像全体から該当物体の検出を行うスライディングウィンドウと呼ばれる方法の二つが存在する。

領域に基づく方法で最も有名な方法がBarnardらによるword-image-translation model [Barnard 03]がある。彼らは、あらかじめ画像全体に対して数億のキーワードが付けられている画像データを用いて、領域分割された画像の領域への自動アノテーションを行った。画像と単語の対応のみで、領域と単語の対応付けがされていない学習データを用いて、領域分割された各画像領域と単語の対応付けを統計的に推定する手法を統計的機械翻訳手法を画像に適用することによって実現した（図6,口絵）。

一方、スライディングウィンドウによる位置検出は、一般物体認識のベンチマークワークショップのPASCAL Visual Object Classes Challengeの4種類のタスクのうちの一つにdetection課題として含まれている。与えられた学習画像を用いて学習し、与えられたテスト画像のどこに物体が含まれているか検出する。20種類の物体（person, bird, cat, cow, dog, horse, sheep, aeroplane, bicycle, boat, bus, car, motorbike, train, bottle, chair, dining table, potted plant, sofa, tv/monitor）について、個別に検出を行うが、そ
の精度（平均適合率）は20種類の平均で22％、最高でも4割程度に留まっている（図7、図8）。

最近では、画像全体の分類に関しては新たな手法によってある程度の分類精度が得られるようになっているため、今後はより困難な課題である、画像からの認識対象物体の切り出しに研究の中心が移りつつある。

2.4 学習データの作成

次に、学習データについて必要なデータセットの作成の問題について触れ、実用を目指した認識を行う場合、大規模な学習データセットが不可欠である。現在は、Caltech-101/256もPASCAL ChallengeもTRECVIDもすべて人手によって学習データおよび評価データが作成されている。Caltech-101/256は画像枚数がそれぞれ9,000枚、30,000枚程度なのでCaltech7CSのグループが独自に構築したが、画像データがさらに多くなると単なるグループが構築することは困難である。TRECVIDでは会議者が共同で、学習用映像から切り出された約40,000枚の画像に対して20種類の物体とシーンのアノテーションを行い、学習データを作成している。また、TRECVIDの2005年80時間分のニュース映像データについて、手作って449種類の学習データを作成し公開している、IBM、CMU、Columbia大を中心としたLSCOM（Large-Scale Concept Ontology for Multimedia）*5 [Naphade 06]というプロジェクトもある。

以上は、画像全体について特定物体が含まれるかどうかラベル付けされたデータセットであるが、大まかに領域分割された画像のそれぞれの領域にラベル付けされたデータセットをWeb上のボランティアによって構築するLabelMeプロジェクト*6[Russell 05]というのもある。

こうしたデータは、画像全体にアノテーションされたものにより構築に手間が掛かり、画像中の各物体の位置の検出まで含めた認識のための研究データとして利用価値が高い。

ほかには、次節で詳しく説明する画像へのアノテーションの作業をオンラインゲーム化したESP game[von Ahn 04]や、オンライン上で有料で作業をしてもらうAmazon Mechanical Turkを用いた大規模一般画像画像データベースの作成[Deng 09]などの、ネットワーク上に不特定多数の人々作業を行ってもらう試みもある。

2.5 Webマイニングによる認識のための知識収集

一方、一般物体認識や映像認識のためのWebから自動知識獲得の研究も試みられている。特に近年はYahoo画像検索APIやFlickrAPIなどの画像を容易にWebから収集するためのWeb APIサービスが提供されるようになっており、こうした研究を行うための環境が整えている。

柳井[Yanai 03,柳井 04]はWebから画像を自動収集し、それを一般物体認識のための学習データとして利用することを提案した。近年、Webからの知識獲得（Webマイニング）の研究が盛んに行われているが、この研究はその画像版ということで「Web画像マイニング」と呼ばれている。

Web上的知識は人手によって構築されたデータセットとは異なり、常に誤った知識（ノイズ）が含まれている。例えば、ライオン画像をWebから収集しても、収集した画像の適合度は決して完璧な7～8割程度にしかならない。そこで、こうしたノイズを含むWeb上のデータを利用しては、ノイズの除去が必要である。柳井ら[Yanai 05,柳井 07]はEMアルゴリズムを応用した繰り返し手法によって、モデル学習時にノイズの影響を少なくする方法を提案している。

一方、ノイズが含まれしていても大量にデータがあれば、十分に学習データとして利用できるという研究結果も示されている。Torralbaらは、8,000万枚もの大量のWeb画像を収集し32×32の画像に縮小して、単純なk-最近傍分類で画像分類を行った。その結果、Webから収集したノイズが含まれた画像データを学習データとしてもその量が十分に多ければ、単純な手法であってもBag-of-Featuresなどの新しい手法に匹敵する一般的物体認識が実現できることを示した[Torralba 08]。
3. メタデータ利用によるアプローチ

メタデータを直接利用してセマンティックギャップを越えようとするアプローチは、低次画像特徴とセマンティクスの間にある（かもしれませんが）関連に欠けることを除く。人が与えたタグなど、画像・映像に付随するテキスト情報、あるいは外部に存在する関連テキスト情報を利用するものである。ひとたびテキストでセマンティクスが表現されると、それ以後は従来のテキストによる概念の範囲で問題を考えることができる。このように、セマンティックギャップの問題をいわば迂回することで、比較的高次のセマンティクスを扱えるが、タグを与えた瞬間、あるいはテキスト情報の利用規則が確定した瞬間に、特定の解釈がほぼ固定化してしまうため、狭義のセマンティックギャップの問題に対しては比較的低い弱いである。しかし、対象映像と応用場面を事前に関連に設定すれば、十分有効である。

このアプローチの採用には、何らかの方法で人が効率的にタグを与える仕掛けを用意するか、外部の関連コンテツからの必要に応じる方法、以下両者に分け、いくつかの試みを紹介する。

3.1 画像・映像に対するタグの付与

画像・映像のタグは、データベース構築時に入手で集中する方法もあるが、大量のデータ入手で付与するのは現実的ではない。MPEG-7やセマンティックWebといったメタデータ記述方法が提案されているが、入手による付与が見込まれて応用が少ないうち、普及していことがある。そこでここでは、大量の画像・映像に付与するタグ付け方法として、画像・映像を含むコンテツから自動抽出する方法と、近年盛んにフォルクソノミーによる方法に分け紹介する。

§ 1 画像・映像を含むコンテツからの自動抽出

1990年代後半以降、インターネット上の画像・映像を検索する技術が求められるようになり、FrankelらはWebページ中の画像に対する自動タグ付け手法による画像検索システム"Webseer"[Frankel 96]を作成した。このシステムは、HTML文書中の画像ファイルに関する記述と関連が強い語をタグとして抽出するもの。具体的には、関連が強いと考えられる順に、画像データのディレクトリファイル名、同一ブロック内のテキスト、画像タグ中のALTテキスト、HTML文書のTITLEテキスト、画像ページされたリンクのアッケスト、その他のテキストをおのおのの重み付けして抽出する。その後、米国Google社が2001年に画像検索サービス"Google Image"の、2005年に映像検索サービス"Google Video"を公開。今日では他の社の同種サービスを含め、日常的に利用されている。これからのサービスにおけるタグ付け方法は公開されていないが、おおむね上記のシステムと同様と考えられる。

画像・映像を含むコンテツからの自動抽出は、大量のデータに対するタグ付けを自動化しやすい反面、精度に問題がある。

§ 2 フォルクソノミーによる付与

2000年代に入り、インターネット上のデータ共有や、ブログなどがソーシャルネットワーキングサービスの普及に伴い、大勢の利用者がさまざまなデータを共有し、タグやコメントを付与するしくみ（フォルクソノミー）が広まる。

カナダLudicorp社（後米Yahoo!社が買収）は2004年に画像共有サービス"Flickr"を公開。このサービスでは、テキストタグ以外に、ジョタグと呼ばれる位置情報をはじめ、撮影時の種々のパラメータも付与できる。また、von Ahnらは、"ESPゲーム"と呼ばれるネットワーク対戦型ゲームを開発した[von Ahn 04]。このゲームは、ネットワーク上の対戦相手と同一の画像を見ながら画像内容を描写し合い、先に相手と同じ描写をしたほうが勝つルールになっている。一方、利用者がゲーム楽しんでいる間に、システムはさまざまな視点から参加者が描写したタグを蓄積している。一般に人手による大量データへのタグ付けは現実的でないと考えられるが、簡単なゲームを通じて無理なく実現するしくみを考案した点で、このアプローチは革新的である。さらに、DengらはWordNetの構築に基づき大量の画像を分類し、タグを付与する際に、米国アマゾン社が提供するフォルクソノミーを仲介するサービス"Amazon Mechanical Turk"を利用することで、人手による大量の修正作業を実現した[Deng 09]。

一方、映像については、米国YouTube社（後にGoogle社が買収）が2005年に映像共有サービス"YouTube"を公開した。同様にYamamotoらは、
ブログサービスと連動して任意のユーザが映像中の任意の時刻をフレーム内領域へコメントを付与する機能をもつ映像共有・アノテーションシステム“Synvivie”*14（図8）を開発した [Yamamoto 08]。その後、類似した機能を提供する商用サービスとして、我が国のニコン社が2007年に、映像中の任意の時刻をコメントを付与する機能をもつ映像共有サービス“ニコンコ動画”*15を公開し、爆発的な人気を誇っていることから、適切なスクショ考えがあれば、手入による大量データへのタグ付けも現実的であることが実証された。

フォルクソノミーによる方法は、おおむね正確なタグが得られるが、少なからず雑音も含まれることや、画像内容とは直接関係ない情報も含むことがある。

3-2 関連コンテンツの利用

前節で紹介した方法はいずれも汎用的な利用を想定したうえで画像・映像中の具体的な事象を記述することを目的としている。一方、事象の時間的因果関係など、より高次のセマンティクスを含む映像の内容を理解したいという要望がある。そのためには、映像と関連するコンテンツのテキスト情報を利用したタグ付けや、セマンティクスに対応する構造を抽出する方法が有効である。ただし、関連するコンテンツの入手やセマンティクスの記述方法を考えると、一般に対象映像の種類を限定する必要がある。

関連コンテンツを利用した映像へのタグ付けに関する研究として、映画・ドラマ映像と台本 [柳沼 96]、講義映像とプレゼンテーション資料 [Wang 08]などさまざまな試みがあるが、本稿ではニュース映像と料理映像に関する著者らの研究を中心に紹介する。

§1 ニュース映像へのタグ付け

ニュース映像へのタグ付けに関する研究は、米国カーネギーメロン大学における先駆的な研究プロジェクト“Informedia” [Wactlar 99] をはじめ、音声認識結果から

*14 http://synvivie.net/
*15 http://www.nicovideo.jp/
聞の類似度を測ることで，映像間の意味的関連性を調べる方法もある．この方法により，ニュース映像間の類似度に時系列的前後関係を加味することで，トピックの変遷を表現するフレーム構造を作成して可視化し，そのうえで映像を閲覧できるようにした "mediaWalker"（図10）[井手 08]のような研究がある。

§ 2 料理映像へのタグ付け

関連コンテンツを利用して一般的な映像にタグ付けする場合，通常大きな順序の入替えが生じないため，おおむね線形で対応付けによりタグ付けができる．しかし料理映像を対象とし，レビューサンプルを利用してタグ付けする際にには，調理手順が入替え可能であることを考えると，工夫が必要である．三浦らは，調理手順の依存関係を解析することで，その流れを非順序グラフ結晶として記述したうえで，映像中の調理動作や調理場所の画像特徴と，レビューサンプル中の記述を対応付ける手法を提案した[三浦 03]．また，浜田らはこの結果を用いて，映像の調理支援インタフェース "HappyCooking" を作成した[浜田 06]。

図11 料理映像にレビュの調理手順をタグ付けすることで，映像を用いた調理支援を実現する "HappyCooking" インタフェース。右上欄の調理手順構造から，必要な箇所の映像を選択的に再生しながら作業できる

4. おわりに

本稿では，画像・映像の内容理解に向けて，セマンティックギャップを越えようとする取組みについて，自動処理によるアプローチを，人手により与えられたタグを利用するアプローチを紹介した．これらのアプローチは，前者が大量データに強いが精度が不十分，後者が精度は高いがデータに対しては難しいという，それぞれ短所，長所があり，今後は，この二つのアプローチを相補的に利用することによってセマンティックギャップをより浅くしていくことが可能になると予想される。従来は，前者は画像認識研究，後者はWebマインニングやテキスト・知識処理研究として長い間独立に研究が行われており融合した研究は多くはなかったが，今後は二つのアプローチを融合した新しい人工知能研究の研究が大いに期待される。

謝辞

"Synvie"の画面例をご提供下さった，名古屋大学の鈴木憲氏および名古屋工業大学の山本大介氏に感謝する。

◇ 参考文献 ◇

[井手 08] 井手玲子，井手一郎，佐藤真一，坂井修一：マルチメディア調理支援ソフトウェア「HappyCooking」，第2回デジタルコンテンツシンポジウム（2006）
[Russell 08] Russell, B. C., Torralba, R., Murphy, K. P. and Freeman, W. T.: LabelMe: A database and Web-based tool for


[上東 09] 上東太一，著足 稔，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)


2009 年 6 月 19 日 受理

著者紹介

井手 一郎（正会員）
1994年東京大学工学部電子工学科卒業。2000年同大学院工学系研究科電気工学専攻博士課程修了。博士（工学）。同年国立情報研究機構助手。2004年名古屋大学大学院情報科学研究科助手教授，国立情報研究機構客員教授兼任。2007年同准教授。現在に至る。電子情報通信学会、情報処理学会、情報処理メディア学会、IEEE Computer Society、ACM 各会員。

柳井 啓司（正会員）
1995年東京大学工学部計数工学科卒業。1997年同大学院工学系研究科情報工学専攻修士課程修了。同年電気通信大学情報工学科助手，2003～04年文部科学省在外講師として米国アリゾナ大学に滞在。2006年電気通信大学情報工学科教授。現在に至る。博士（工学）。電子情報通信学会、情報処理学会、IEEE Computer Society、ACM 各会員。