マルチメディア解析・検索研究のための 大規模コーパス

(正会員) 佐藤 真一[†]

1. ま え が き

画像・映像の計算機による意味解析は、古くから画像関連研究者の夢であった。あたかも人間が見て解釈するように、画像・映像を計算機で処理して、その意味内容情報を自由自在に引き出すことができれば、さまざまな応用が可能になると考えられている。例えば、大量に撮りためたデジカメの画像、ハードディスクレコーダの番組、インターネット上の無限に存在する画像・映像から、あたかも自分で一つ一つ見て選り分けるように、今自分が欲しいと思う画像・映像を取出すことができる。これは紛れもなく、マルチメディア検索の究極の目標である。

画像・映像の意味解析は、われわれ人間には何の苦もなく実行可能である。画像・映像を見れば、そこに何が映っているか、どのような情報を伝えようとしているのか、通常、たちどころに認識できる。それは当然であり、そもそも画像・映像は人間に対し一瞥で情報を与えられるように作られているメディアであり、むしろ人間が情報を読取れない画像・映像には意味がないといえる。しかしながら、これは計算機には非常に難しく、与えられた任意の画像・映像を処理し、映っている事物や伝えようとしている情報を、人間のように柔軟に抽出できるような計算機プログラムはいまだに存在しない。これはメディア情報解析におけるセマンティックギャップのためと言われている¹⁾。画像処理や人工知能の黎明期には、人間に何の苦もなくできる画像・映像意味解析は簡単な問題であると考えられていたと言われているが²⁾、今ではもっとも実現困難な問題の一つとして認識されている³⁾。

この画像・映像の意味解析という困難な課題に対して、今、突破口が開かれようとしている。例えば、ほとんどのコンパクトデジカメは、任意の情景中で人間の顔を検出す

る機能を有しており、一部の機種、あるいはiPhotoやPicasaは、さらに人物の認識まで可能である。Google Imageの類似画像検索サービスをご存知だろうか。画像検索の結果の下に「類似の画像を探す」というリンクがあれば、その画像と類似している、すなわち、同じ物体や情景が映っている画像を検索することができる。ALIPR^{*1}というサービスでは、未知の画像に対して意味内容を表すタグを自動推定することができる。Viewdle^{*2}というサービスでは、映像中の人物名のタグを自動付与することができる。もちろんこれらはまだ画像・映像の意味解析という究極の目標からは遠いが、着実に第一歩を踏み出しているように見える。こうした技術を可能としている一つの鍵が、大規模コーパスである。本稿では、まず、大規模コーパスを用いた画像・映像(マルチメディア)解析ならびに検索研究がどのように行われるのかについて紹介し、大規模コーパスの位置づけを明確にする。また、最近、研究の最前線で用いられているマルチメディア解析・検索のためのコーパスの多くが、一般に公開されている。こうしたコーパスについて紹介する。

2. 大規模コーパスを用いたマルチメディア 解析・検索研究

本稿で対象として考える画像・映像等のマルチメディア情報としては、一般の人がデジカメで撮影した画像、家庭用ビデオカメラで撮影した映像、Web上で普通に目にする画像・映像、テレビ番組などを対象とする。検索要求としては、撮影されている事物を表す普通名詞(あるいは一般的な動詞など)、情景種別、あるいは有名人の人物名など、一般の人による要求頻度が比較的高いと思われる検索要求を想定する。すなわち、ロボットの自律ナビゲーションのための画像・映像解析、医療画像の疾患に基づく検出や分類などは対象としない。そうすると、典型的な画像・映像、ならびにその望ましい解析結果は図1のようなものとなる。

[†] 国立情報学研究所

"Recent and Technologies in Multimedia Retrieval (8): Large-Scale Corpus for Multimedia Content Analysis and Retrieval" by Shin'ichi Satoh (National Institute of Informatics, Tokyo)

*1 ALIPR, <http://www.alipr.com/>

*2 Viewdle, <http://www.viewdle.com/>

図1 典型的な画像・映像と意味解析結果の例

図2 事物に対するグラウンドトゥールズの記述の仕方

図1 (a) のような解析結果があれば、「人と犬が一緒に走っている写真」という検索が実現可能であり、また図1 (b) では「ジョージ・W・ブッシュが写っているすべての写真」という検索が可能となる。

マルチメディア解析・検索のためのコーパスとは、端的には、画像・映像と、その望ましい解析結果とを対にした、大量のデータセットのことを言う。望ましい解析結果、あるいは正解データのことをグラウンドトゥールズという。マルチメディア解析・検索の研究では、こうしたコーパスを、システムの開発・チューニング・学習のための訓練データ、あるいはシステムの性能評価のための評価データとして利用する。システム構築において、先験的な知識（例えば、写真の下半分が緑ならば芝生とみなす、目・鼻・口とみなしうる領域が一定の空間配置で検出できたら顔とみなす、など）を用いず、あくまで訓練データを用いて、機械学習などの手法でシステムを構築することにより、対象の多様性などに柔軟に対応した頑健な解析システムが実現可能であることが知られている⁴⁾⁵⁾。コーパスを用いたマルチメディア解析・検索でも、そうした効果を狙っている。また、コーパスを評価に用いれば、システムの性能を、実際のデータに即して効果的に評価することができる。さらに、コーパスを複数の研究者で共有すれば、それぞれのシステムの性能を公正に比較することが可能となり、当該技術の健全な推進に資する。このように、マルチメディア解析・検索研究におけるコーパスの利用は大変望ましいことであると考えられるが、効果的に利用できるコーパスには以下のような特性が求められる。

- (1) 大規模・多様：統計量（認識率など）の算出や、複雑な識別モデルの学習を安定して行うため、できるだけ大量のデータを有することが望ましい。また、実データに即した多様性を持つデータが必要となる。
- (2) 適切な分布：実データと同じようなばらつきを持つデータが必要である。例えば、顔検出の学習用に顔を含む画像だけを用意した場合、顔が写っていない画像を「知らない」顔検出器となり、どのような画像でもどこかに無理に顔を検出してしまう危険がある。
- (3) 適切・正確なグラウンドトゥールズ：タスクに応じた

適切なグラウンドトゥールズが必要となる。また、できるだけ高い精度のグラウンドトゥールズが望ましい。

このようなコーパスが準備できれば、適切なマルチメディア解析手法や機械学習アルゴリズムなどと組み合わせることにより、処理対象の複雑さや多様性に柔軟に対応した、頑健なマルチメディア解析・検索システムが構築できる。

上記のうち、規模・多様性ならびに分布は、コーパスに利用するデータにより決まる。グラウンドトゥールズは、撮影されている事物や情景種別に対する記述、あるいは人物名となると述べた。そのグラウンドトゥールズの記述法であるが、情景種別に対する記述は、画像あるいは映像セグメントに対する情景種別を表す記述であり、人物名は通常顔領域に対する人物名そのもので記述される。一方事物については、その事物を表す記述を付与する対象に応じて、図2に示すように分類できる。すなわち、事物の記述を画像・映像セグメントに対して行う場合（図2 (a)）、事物を囲む矩形領域に対して行う場合（図2 (b)）、事物の物体領域（画素単位）に対して行う場合（図2 (c)）、がある。学習データとしては、より詳細な物体領域に対する記述が与えられることが望ましい。一方、通常グラウンドトゥールズは人手にて付与しなければならない。その場合の作業は、物体領域に対する記述の作成がもっとも困難であり、画像・映像セグメントレベルの記述がもっとも容易である。

事物や情景などに付与する記述の種別をどのように決めるかという点も重要な問題となる。この種別を概念と呼ぶ場合がある。また、この概念の集合をオントロジーと呼ぶ場合がある。ここではよりわかりやすく語彙セットと呼ぶ。語彙セットをあらかじめ数十から数百程度の種類と決めておいて、グラウンドトゥールズはこれらのうちから選択して付与するという方針がありうる。一方、語彙セットは決まらずに、グラウンドトゥールズを付与する人に任せて自由に記述してもらうという方針もありうる。語彙セットをあらかじめ決定しておく場合には、解析・検索の問題の難しさを適度に制御することができる反面、実際の利用時には語彙が制限されてしまうことになり、自由度が下がってしまうという問題点がある。語彙セットを決めない場合、検索などで利用する場合の自由度が高くなる反面、マルチメデ

ィア解析問題としての難易度が飛躍的に高まってしまうという問題がある。

引き続き節で、さまざまな実際のコーパスを紹介するが、それぞれ想定しているタスク、データセットの収集法、語彙セットの決定法、グラントゥルースの付与法などが異なっている。公開コーパスを用いて研究を行うのは大変便利だが、上記の点を十分に考慮し、自分の研究の目的にもよく照らし合わせて検討する必要がある。

3. 公開されているコーパスの紹介

3.1 TRECVID

TRECVID⁶⁾は、米国標準技術局 (National Institute of Standards and Technology: NIST) の主催で開催されている、放送映像をはじめとする映像コンテンツを対象とし、その内容解析ならびに検索の高度化を目的とした、競争型国際プロジェクトである。2001年にTREC (テキスト検索の競争型国際プロジェクト) の一部、Video Trackとして開始され、2003年からTRECVIDとして独立して開催されるようになり、今日まで毎年開催されている。本稿で紹介するほかのほとんどのコーパスと異なり、対象映像コンテンツはWebなどから自由にダウンロード可能とはなっておらず、プロジェクト参加者のみに配布されることになっている。しかし、世界中の研究グループが自由に参加可能となっており、事実上は誰でも取得可能な開かれたコーパスとなっている。

TRECVIDでは、毎年100時間程度という、相応規模の映像を配布しており、これまでで総計1000時間程度にもおよび、映像解析研究においては、TRECVIDの映像コンテンツを用いた評価がなされていないと国際会議などでは論文が採択されにくくなっているほど、標準的なコーパスとなっている。2001年と2002年は、米国政府機関関係の映像や、Open Video ProjectやInternet Archive, BBC提供などのフリーの映像素材が利用されたが、2003年からは実際の放送映像が利用された。2003, 2004年は米国の英語のニュース映像、2005, 2006年は英語・中国語・アラビア語のニュース映像、2007年～2009年はオランダ視聴覚研究所 (The Netherlands Institute for Sound and Vision) 提供のニュース、科学ニュース、ドキュメンタリー、教育番組などを含む、さまざまな種類の番組素材が利用された。2010年からはまた大きく様変わりし、Internet Archiveからクリエイティブコモンズライセンスにより自由に利用可能なインターネット上の映像を利用することになった。この他に、2005-2008年に配布されたBBC素材映像 (放送映像向きだが、編集前の素材映像)、2008年から開始した監視映像解析タスクのための映像 (ロンドンガトウィック空港の実際の監視映像) も利用可能である。これらはすべて権利処理がなされて配布され、参加者はTRECVIDのタスクのためのみならず、研究目的に自由に利用することができ、学術論文

や研究紹介のWebページなどにフレーム画像などコンテンツの一部を提示することも許されている。

コンテンツの配布を受けるには、プロジェクトへの参加が必要だが、参加者には原則としてタスクへの参加、処理結果の提出、レポートの提出が求められているという点もTRECVIDの特徴といえる。映像コンテンツの取得のための参加は認められていない。プロジェクトは年単位で運営されており、毎年、2月ごろ世界中から参加者を募り、映像コンテンツを配布し、参加者間の議論に基づきタスクの詳細を決定し、8～9月ごろ解析や検索などの結果を収集し、精度評価を行い、11月ごろにプロジェクト参加者のみ出席可能なワークショップを開催し、各参加者からのレポートを取りまとめて公開している。ウェブサイト^{*3)}に、タスクの詳細、参加者の開発した手法の詳細が記載されたワークショップの予稿集などがすべて公開されている。タスクは毎年少しずつ加除や変更が行われているが、開始当初からずっと行われているのは高次特徴抽出タスク (すなわちショットの概念クラスへの自動分類) と検索タスクである。2010年はこれまでにないほど多くのタスクが追加された。全部で6タスクあり、意味インデクシング (Semantic Indexing: SIN, 従来の高次特徴抽出タスク)、既知アイテム検索 (Known-item Search: KIS, 一度見たことのある映像をテキストで検索するタスク、従来の検索タスク)、コピー検出タスク (Content-based Copy Detection: CD)、監視映像イベント検出タスク (Surveillance Event Detection: ED)、インスタンス検索タスク (Instance Search: INS, 人物や特定物体の検索)、そしてインターネット映像に対するイベント検出タスク (Event Detection in Internet Multimedia: MED) である。

TRECVIDでは、まずは利用可能な映像データセットを定義している。例えば、ある年のある一月分のCNNのニュース映像、というように用意している。一方で、タスクは可能な限り一般性が高く、データセットへの依存性は少ないように設計している。しかしながら、そのため、例えば高次特徴抽出タスクにおける学習データで、極端に事例数の少ない概念カテゴリーが出現したりする問題が起こりうる。学習用のメタデータは、通常、参加研究グループで手分けをして手作業で付与しているが、上記のとおり用意するデータセットに恣意性がほとんどないことと、タスクの一般性が比較的高いことから、概念ラベル付与などのメタデータ付与が明確に行えないあいまいなショットなども多いため、メタデータには若干のノイズを含む。高次特徴抽出タスクのメタデータは、これまでショット単位のラベルとして提供されていたが、2010年のインスタンス検索タスクではじめて、物体単位 (矩形領域ならびに物体領域) の検出ならびに精度評価が行われる見込みである。

*3) TRECVID, <http://www-nlpir.nist.gov/projects/trecvid/>

3.2 CALTECH 101/256

CALTECH 101/256は、カリフォルニア工科大の研究者らにより構築された、画像意味分類研究のためのコーパスである。一般画像認識研究では定番のコーパスであり、最も古くからあるコーパスだが、いまだに広く使われている。メタデータは画像単位に付けられており、物体単位ではない。

CALTECH 101⁷⁾は、2003年に構築・配布が開始されたコーパスであり、概念ラベルとして101カテゴリーを有するため、CALTECH 101と呼ばれる。もともとは四つの概念カテゴリー分類のために整備した画像データセットCALTECH 4を拡張し、あらたに97カテゴリーの画像を追加して構築された。ウェブサイト^{*4)}にて、データセットも含み各種の情報が公開されている。CALTECH 101（ならびにCALTECH 256）は、公開されているデータセットを使って各研究者が自由に研究し、論文発表をするようになっており、特段のタスク設計やワークショップなどはない。公開されているデータセットは、画像データとラベルとからなっており、カテゴリーは相互に排他的となっている（すなわち、各画像には一つのカテゴリーにのみ所属する）。画像データは、カテゴリーにつき40～800画像あり、合計9,144画像からなる。データセット構築は、最初に概念カテゴリーを決定し、次いでそれらに属する画像を決定するという手順により行われている。具体的には、まずは辞書から視覚的に表現しうる具体的な事物を表現する単語を101選択した。ついで、これらの単語に対応しうる画像をGoogle画像検索を用いてインターネット上から選び出し、さらに、人手により実際に単語に対応している画像を選別することにより構築している。したがって、画像セット自体は（インターネット上の画像などの）実際の画像の分布に従っているわけではなく、選択された101のカテゴリーに従った偏った分布を持っている。一方、人手で画像を選別しており、メタデータのノイズは存在しない。タスク設定やワークショップなどは行われていないが、一般画像認識に関する多くの論文で利用されており、そうした中で標準的な評価法が確立して行った。例えば、文献8) などでは、各カテゴリーからランダムに選んだ30画像を学習画像として利用し、それ以外の画像で識別性能を評価するという方法が取られおり、多くの論文で採用されている。提供される画像はカラー画像だが、ほとんどの研究ではこれらを濃淡画像に変換して利用している。それでも、各カテゴリー30画像で学習したケースで80%近い認識率が達成されている。加えて、色情報を用いればさらに高い認識率が達成できることが知られている。しかしながら、提供データで認識率を達成している概念識別器でも、任意の画像の意味分類は十分な精度では実現できていない。いくつか理由が考えられるが、一つには、CALTECH 101の画像は、対

象物以外の背景部分が比較的単純で、かつ対象物のみが中央に大きく撮影されており、大きさや向きもほぼ同一に正規化されてしまっていて、問題を過度に簡易化しすぎているといわれている。

CALTECH 256⁹⁾は、CALTECH 101の問題点を受けて再構築され、2007年より配布が開始されている。概念カテゴリーが256種類となり、より難しくなっていて、背景もより複雑となり、正規化もなされていない、より多くの画像が提供される。全部で30,607画像となり、ウェブサイト^{*5)}にて配布されている。概念セットの選び方、画像の収集法ともにCALTECH 101とほぼ同様である。CALTECH 101と同様、メタデータ付きの画像セットが配布されているのみであるが、昨今の研究論文ではCALTECH 256が盛んに利用されている。濃淡画像を用いた認識では、カテゴリーごとに30画像を用いた学習で40%内外、60画像で学習した場合で50%弱であり、やはりCALTECH 101よりも難しい対象であるといえる。いずれのデータセットも、一般画像認識研究における定番のコーパスである。

3.3 PASCAL VOC

PASCAL VOCとは、EU国際共同研究プロジェクトPASCAL (Pattern Analysis, Statistical Modelling and Computational Learning) の一部の、PASCAL Visual Object Classes Challengeという課題に携わるグループにより、整備・運用されている、画像中の物体認識のためのコーパスで、ウェブサイト^{*6)}から、データセット、メタデータ、各種ツールなどがダウンロードできる。2005年から毎年開催されており、データならびに関連ツールは一般に公開される。例年ワークショップが開催されており、参加研究グループの成果発表ならびに表彰が行われる。通常ICCVないしはECCVの併設ワークショップとして行われる。ワークショップには一般の研究グループも参加可能である。

例年配布されるデータセットは、Flickrなどの10,000規模の画像（2009年は14,737画像）で構成される。PASCAL VOCにおいて特筆すべきは、画像に含まれる事物による画像分類のみならず、画像中の物体の位置検出ならびに領域抽出までタスクとしており、メタデータには画像に対するラベルに加え、物体の外接矩形領域ならびに画素単位のラベルも含まれている。これらは人手で付与されており、高精度である。タスクとしては、分類・検出 (Classification/Detection)、セグメンテーション (Segmentation)、人物レイアウト (Person Layout) がある。分類・検出では、画像に対するラベル推定ならびに物体を含む矩形領域の検出を行う。セグメンテーションでは、物体単位のセグメンテーションならびにセグメント単位の物体ラベル推定を行う。

*4 CALTECH 101, http://www.vision.caltech.edu/Image_Datasets/Caltech101/

*5 CALTECH 256, http://www.vision.caltech.edu/Image_Datasets/Caltech256/

*6 PASCAL VOC, <http://pascallin.ecs.soton.ac.uk/challenges/VOC/>

人物レイアウトでは、人物を構成する頭部、手、ならびに足領域の検出を行う。分類・検出、ならびにセグメンテーションで用いる概念カテゴリーの数は、CALTECHに比べると圧倒的に少なく、数個から数十程度である(2010年は20カテゴリー)。これらとは別に、2010年には、後述のImageNetデータを用いたLarge Scale Visual Recognition Challengeというタスクも行われる予定である。このタスクでは、10,000,000画像を用いた、10,000概念カテゴリーへの分類を行う。

3.4 Labeled Faces in the Wild (LFW)

顔の検出や認識など、画像・映像中の顔画像処理は極めて重要な研究課題であり、FERET, Yale Face Database, CMU PIEなど、多くの研究用顔データベースが整備され利用されてきた。しかしながらその多くが、照明条件や表情、顔の向きなどを制限したものであった。それに対し、一般の写真や映像などに現れる顔を対象にした、さまざまな制限がほとんどない顔画像のコーパスが提供され始めている。Labeled Faces in the Wild (LFW)¹⁰⁾は、マサチューセッツ大学アマースト校の研究者らにより整備された、制限のない条件下での顔画像認識研究のためのコーパスで、ウェブサイト^{*7}からダウンロード可能である。Webから収集した13,000画像を有し、Viola-Jones顔検出器で検出した顔を対象とし、人手により、フルネームの名前があいまい性なくメタデータとして付与されている。顔検出器により検出されたほぼ正面向きの顔のみが含まれているが、一定の範囲で顔の向きもさまざまであり、照明条件、人種、表情、画像のもともとの利用目的などはきわめて多様である。合計5,000人以上の名前が付与されており、1,680人については2回以上出現しており、最多はGeorge W. Bushであって530回出現している。

なお、これに先立ち、別の研究者らによるFaces in the Wild¹¹⁾というデータセットも公開されている^{*8}。これはYahoo Newsの画像から顔検出器により検出した30,281顔画像により構成されており、LFWと同様の特性の顔画像からなるが、メタデータとしての名前は文献¹¹⁾の手法で自動推定されたものが付与されており、精度は80%程度である。

3.5 そのほかのマルチメディアコーパス

LabelMe¹²⁾は、MITの研究者らにより整備されている、一般画像認識のためのコーパスであり、ウェブサイト^{*9}からダウンロード可能である。このコーパスは大変ユニークであり、数万画像が公開されているが、メタデータ付与も一般公開されており、誰もが自由にメタデータを付与し、付与結果がまとめて公開されるという方式をとっている。上記のWebページでメタデータ付与用のインタフェースが公開されており、物体を囲む多角形を書込み、その物体名

を記述するようになっている。語彙には制限がなく、ラベル付けする人の自由となっている。これまでに数千語彙付与されたが、複数回出現するのは100~1,000語彙程度である。データセットは一般に公開されているが、ダウンロードするにはラベル付けに協力することが望まれている。

Microsoft Research Asia Multimedia Dataset (MSRA-MM)は、MSRAの研究者らにより整備されている、マルチメディア情報の意味解析や検索のためのコーパスである^{*10}。Microsoft Live Searchにより取得した画像・映像からなる。画像については、代表的な68クエリーと、それぞれに適合する各1,000画像程度、合計65,443画像が提供され、映像については、165クエリーにより10,277映像が提供されている。意味解析向きのグランドトゥールズとしては、画像には50概念のラベルが付与されており、映像は10にカテゴリーわけされている。元の画像・映像は権利問題のため提供されず、代わりに、特徴量のみが配布される。

Oxford Buildings Dataset¹³⁾は、一般画像認識と並んで、画像中の特定の物体の同定を行う特定物体認識も盛んに研究されている中で、建物などのランドマーク物体の同定を行う際の定番のデータセットで、ウェブサイト^{*11}からダウンロード可能である。Flickrから得た5000画像からなり、11のランドマーク(オックスフォード内の建物)によるグランドトゥールズが付与されている。各ランドマークにつき5画像は、そのランドマーク領域の外接矩形領域つきで提供され、クエリーとして利用することが想定されている。そのほかの全画像については、各ランドマークの写り具合について4段階のラベルが付与されている(Good: ランドマーク自体の写真~Junk: ランドマークが写っていない)。

Hollywood Human Actions Dataset¹⁴⁾^{*12}は、実際の映画の中の人物動作ならびにシーンクラスに関するグランドトゥールズが付与されているコーパスである。32の映画から抽出した映像に対し、人間動作でタグ付けされている。Hollywood 2¹⁵⁾^{*13}は、69映画から抽出した合計20時間以上の映像を含み、12種類の人物動作と10種類のシーンクラスによりラベル付けされている。

TinyImage¹⁶⁾^{*14}は、非常にユニークな画像コーパスである。約8,000万画像、合計200GBを有するが、32×32画素と極めて低解像度である。Wordnet中の全具象名詞を概念として使い、75,000種の内容意を用意し、画像は複数の画像検索エンジンでWebから取得したものである。8ヵ月かけてヒットした全画像をダウンロードしている。人手による精査は行われておらず、ラベルは必ずしも正確とは言えない。

*10 MSRA-MM関連ページ, <http://research.microsoft.com/apps/pubs/default.aspx?id=79942>など。

*11 Oxford Buildings Dataset, <http://www.robots.ox.ac.uk/~vgg/data/oxbuildings/>

*12 Hollywood, <http://irisa.fr/vista/Equipe/People/Laptev/download.html>

*13 Hollywood 2, <http://www.irisa.fr/vista/actions/hollywood2/>

*14 TinyImages, <http://groups.csail.mit.edu/vision/TinyImages/>

*7 LFW, <http://vis-www.cs.umass.edu/lfw/>

*8 Faces in the Wild, <http://www.tamaraberg.com/faceDataset/>

*9 LabelMe, <http://labelme.csail.mit.edu/>

ImageNet¹⁷⁾*15は、一般画像認識用の大規模画像コーパスであり、前述の通りPASCAL VOC Large Scale Visual Recognition Challengeタスクでも利用される。1,000万画像を含み、人手により付与された1万概念(複合名詞も含み、synsetと呼ばれている)のグラントゥールズを有する。画像の著作権などの権利に配慮し、画像自体は配布せず、そのURLのリストを配布している。

4. む す び

画像・映像などのマルチメディア情報の意味解析、ならびに検索の研究のため、大規模な実際のマルチメディア情報により構成されたコーパスに関し、コーパスを用いた研究の流れ、コーパスの性質や設計方法の概要、さらには、主たる公開コーパスについて紹介した。この他にも、紙面の都合で紹介しきれなかった多くのコーパスがある。MIR-FLICKR-25000¹⁶⁾、Microsoft Research Cambridge¹⁷⁾、Image CLEF¹⁸⁾、MUSCLE VCD¹⁹⁾、ImageEval²⁰⁾などはいずれも興味深いコーパス、ならびに、それに基づく研究プロジェクトやワークショップである。本稿で述べたようなアプローチに興味をもたれた方は、ぜひ、これらのコーパスを用い、より実際に「使いもの」になるレベルに近い、マルチメディア解析・検索研究に挑戦していただき、今一番熱いこの分野の突破口を開いていただきたい。(2010年5月10日受付)

〔文 献〕

- 1) 馬場口：“マルチメディア検索の技術動向”，映情学誌，pp.58-63 (2010)
- 2) 金出：“素人のように考え、玄人として実行する：問題解決のメタ技術”，PHP研究所 (2003)
- 3) 鷺見，内田，佐藤，佐藤，日浦，福井，馬場口：“パターン認識・メ

ディア理解の10大チャレンジテーマ”，信学誌，92，8，pp.665-675 (2009)

- 4) 佐藤真一：“コーパスベース映像解析”，電子情報通信学会技術研究報告，106，229，pp.111-120，PRMU2006-75 (2006)
- 5) 佐藤，齋藤：“石頭なコンピュータの眼を鍛える：コーパスで人間の視覚にどこまで迫れるか”，丸善 (2009)
- 6) A.F. Smeaton, P. Over and W. Kraaij: "Evaluation Campaigns and Trecvid", MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, New York, NY, USA, ACM Press, pp.321-330 (2006)
- 7) L. Fei-Fei, R. Fergus and P. Perona: "One-shot Learning of Object Categories", IEEE Transactions on Pattern Analysis and Machine Intelligence, 28, pp.594-611 (2006)
- 8) K. Grauman and T. Darrell: "Pyramid Match Kernels: Discriminative Classification with Sets of Image Features", Proc. of ICCV (2005)
- 9) G. Griffin, A. Holub and P. Perona: "Caltech-256 Object Category Dataset", Technical Report 7694, California Institute of Technology (2007)
- 10) G.B. Huang, M. Ramesh, T. Berg and E. Learned-Miller: "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments", Technical Report 07-49, University of Massachusetts, Amherst (2007)
- 11) T.L. Berg, A.C. Berg, J. Edwards and D.A. Forsyth: "Who's in the Picture", Proc. of Neural Information Processing Systems (NIPS) (2004)
- 12) B.C. Russell, A. Torralba, K.P. Murphy and W.T. Freeman: "Labelme: A Database and Web-based Tool for Image Annotation", Int. J. Comput. Vision, 77, 1-3, pp.157-173 (2008)
- 13) J. Philbin, O. Chum, M. Isard, J. Sivic and A. Zisserman: "Object Retrieval with Large Vocabularies and Fast Spatial Matching", Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2007)
- 14) I. Laptev, M. Marszalek, C. Schmid and B. Rozenfeld: "Learning Realistic Human Actions from Movies", Proc. of CVPR (2008)
- 15) M. Marszalek, I. Laptev and C. Schmid: "Actions in Context", Proc. of CVPR (2009)
- 16) A. Torralba, R. Fergus and W.T. Freeman: "80 Million Tiny Images: A Large Dataset for Non-parametric Object and Scene Recognition", IEEE Trans. On PAMI, 30, 11, pp.1958-1970 (2008)
- 17) J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei: "ImageNet: A Large-Scale Hierarchical Image Database", CVPR09 (2009)

*15 ImageNet, <http://www.image-net.org/>

*16 MIRFLICKR-25000, <http://press.liacs.nl/mirflickr/>

*17 Microsoft Research Cambridge, <http://research.microsoft.com/en-us/projects/ObjectClassRecognition/>

*18 Image CLEF, <http://www.imageclef.org/>

*19 MUSCLE VCD, <http://www-rocq.inria.fr/imedia/civr-bench/data.html>

*20 ImageEval, http://www.imageval.org/e_presentation.html

ま・と・う しん・いち
佐藤 真一 1987年，東京大学工学部電子工学科卒業。1989年，同大学工学情報工学科修士課程了。1992年，同大学専攻博士課程修了。現在，国立情報学研究所教授。映像解析およびマルチメディアデータベースなどの研究に従事。工学博士。正会員。