特集　情報学の第4ステージ

放送映像コーパスを用いた映像解析

佐藤真一
総合研究大学院大学教授 情報学専攻／情報・システム研究機構 国立情報学研究所教授

図1　TRECVID映像コーパスで学習した解析技術による映像内容の推定結果

だれもが映像データを快適に扱うことができる環境を構築するための技術開発が進められている。映像検索・解析技術の高度化をめざした国際共同研究プロジェクトTRECVIDに期待がかかる。

映像を手軽に検索して楽しむために

テレビ放送のデジタル化、PodCastやYouTubeなどのネットワークを介した映像サービス、Google Video、Yahoo! Videoなどの映像検索サービスにより、一般ユーザーが大量の映像を気軽に利用できる環境が実現されてきている。こうした大量の映像から好きな情報に自由自在にアクセスするためには、映像の内容にもとづく検索やブラウジング（閲覧の仕方）などの技術が必要不可欠である。

そのための方法として、映像に対してメタデータやキーワードを付与することにより、テキストを基本とした検索手法を利用することが考えられる。YouTubeやGoogle Videoを含むほとんどのサービスでこの手法が用いられているが、メタデータやキーワードを付与するためにかかる労力、映像内容を記述するテキストを作成する際の恣意性・主観性などの課題があり、必ずしも映像の内容にもとづく検索が適切に実現できたとはいえない。これを解決するためには、コンピューターによる映像の意味内容の解析が必要不可欠である。しかし一般に、映像の意味内容の解析はきわめて困難である。

かぎを握るのは「コーパス」

この問題に対処するには、「コーパス」を用いた解析手法が有効と考えられている。コーパスとは、対象のデータを大量に集めたアーカイブと、これらを解析した場合の望ましい正解データ（Ground Truth）により構成される。

一般に、コーパスを学習データとすると、データから正解データへの対応付けが自動的に行える機械学習手法により、解析システムを作成できる。こうした手法の成功例としては、新聞記事を複数の人によって読み上げた大量の音声データにもとづく音声認識や、各単語の品詞や係受け関係などが付与された大量の文書にもとづく自然言語処理があげられ、いずれも実用レベルの性能を達成している。画像処理では、手書き文字の認識や顔の検出・認識において、コーパスにもとづいた手法により実用的な処理性能を実現している。

映像解析においても、コーパスを用いたアプローチにより、大きな飛躍が望めると考えられる。こうした考えにもとづき開始されたのがTRECVID（Text Retrieval Conference Video Retrieval Evaluation:トレックビッド）である。TRECVIDは、映像検索ならびにそのための映像解析技術の高度化をめざし、米国の標準技術局（NIST：National Instituteof Standards and Technology）とDTO（Disruptive Technology Office）が主催する国際共同研究プロジェクトである。世界中から映像解析・検索の研究に携わるグループが参加している。

このプロジェクトの参加者には、正解データつきの実際の放送映像（毎年100時間以上の規模）が提供され、これにもとづいて映像解析技術の研究を行うことができる。われわれの研究グループも2004年からTRECVIDに参加している。すでに合計1000時間近い映像コーパスを得ており、われわれの映像解析研究を支える土台となっている（図1）。

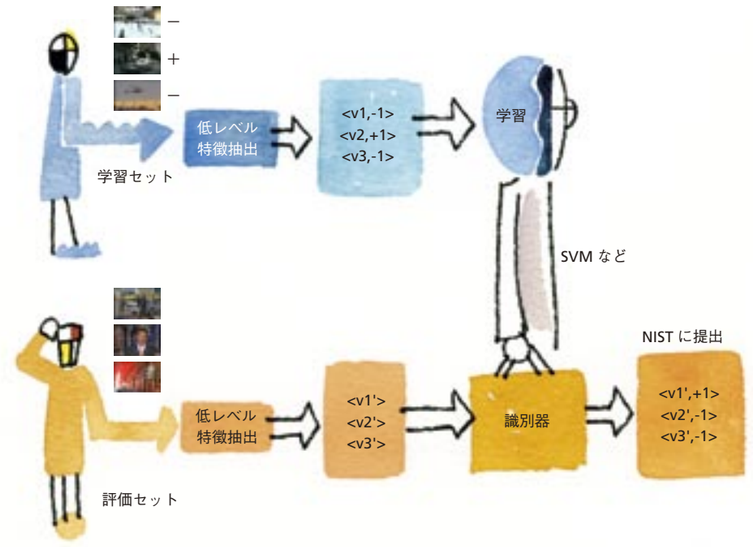

図2　典型的な高次特徴抽出システム
「車」と言う概念を学習する場合を例に取る。学習データとして、車を含む画像（＋が付与されている）と車を含まない画像（−）をできるだけ大量に集める。ついで各画像から低レベル特徴を抽出し、特徴ベクトル（v1, v2 等と標記）に変換する。これを機械学習アルゴリズムに入力すると、どのようなベクトルが車という概念に対応するかが学習される。すると、未知の画像（評価セット）から同様に低レベル特徴を抽出し、学習アルゴリズムに入力することにより、車を含む／含まないというラベルを適宜推定することができる。

検索・解析技術の高度化をめざす

実際の放送映像の規模や多様性を考えると、1000時間の映像コーパスは必ずしも十分とはいえない。また、TRECVIDで提供される放送映像は、英語（英・米）、中国語、アラビア語、オランダ語であり、日本の映像は含まれていない。そのため、日本の放送映像の解析に適切なコーパスなのかわからないという問題もある。

そこで、われわれは、東京地区で受信可能な地上波7チャンネルを蓄積し続け、映像コーパスとして利用しようと考えている。これまで、特定のニュース番組（NHKニュース7）は2001年から、7チャンネルの全番組についても2009年8月より蓄積してきた。2010年夏には6万時間の映像コーパスができあがる。

このコーパスには、人手による正解データ付与が行われていないため、実質上正解データがないという問題があるが、単に大量にデータがあるだけで十分解析に役立つことも考えられる。ひとつの例が、Google検索エンジンで利用されているPageRankアルゴリズムである。これは、正解データなしの大量のウェブ情報にもとづき、リンク関係の解析により各ウェブページの重要度を適切に決定する。その有効性はGoogleの成功により証明済みといえる。

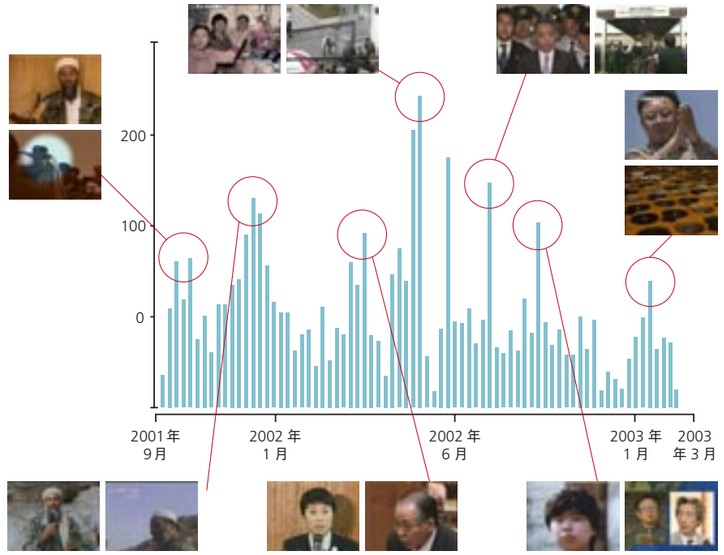

図3　ニュース映像から発見した重要ショット
1年半分のニュース映像から繰り返し見られた資料映像をすべて検出し、その数を調査したところ、いくつかの特定の期間について極端に多くの繰り返し資料映像が見られた。そうした期間では、図示したような重要なイベントに対応する映像が繰り返し利用されていた。

映像コーパスでも、映像が大量にあるだけでまったく新たな解析が可能となった例がある。年単位のニュース映像をみると、同じ資料映像を複数回利用している例が多く観察できる。そこで、1.5年分のニュース映像について繰り返し利用された資料映像約4000 ショットをすべて検出し、その観測パターンに従い、特定の時期に発生した重要なイベントに対応する映像を発見することに成功している（図3）。

この放送映像アーカイブは大量であるため、コンピューターによる解析だけでなくメンテナンスにも苦労する。しかし、これは宝の山であると確信している。今後、TRECVID のような正解データ付きの映像コーパスにより、映像中の事物の認識に加え、動作や関係性などの高次の意味情報まで扱えるようになっていくと期待できる。加えて、われわれの研究グループ独自のアプローチとして、大量の放送映像コーパスを活用した、これまでにない柔軟な映像解析手法について検討し、映像中の隠された意味、映像制作者やテレビ局などの映像提供者の意図なども解析可能になると考えている。これらを通して、映像意味解析のブレークスルーを達成したいと考えている。

佐藤真一（さとう・しんいち）
映像解析を始めたのは、15 年前に在外研究員として CMU Informedia デジタル映像図書館プロジェクトに参加したのがきっかけ。出始めた 4GB のハードディスクに CNN ニュースを 10 日分詰め込み、ああ、これでいろいろ楽しいことができると思ってから、その魅力にとりつかれている。いまや 50TB に 1 年分の番組を詰め込み、学生や同僚も巻き込んで、もっと楽しいことをしようと思っている。