■佐藤真一　コンテンツ科学研究系教授

【タイトル】
動画がもつ意味を理解できる視覚をつくる

【本文】
私たちと同じように世界を見ることのできる「人工の目」をつくりたいと思っています。これは、テレビドラマに興じながら、世間話もしていられるような視覚システムです。また、人間の赤ちゃんのように、目から入ってくる情報の意味を理解して学習する。さらには、職人さんの動作を映したドキュメンタリーからスキルを学び取る、そんな視覚システムの実現を目指しています。

顔の映像から名前を判別する
そのためにはまず、学習用に大量の動画データを収集する必要があります。NII では 2001 年 3 月からニュース番組やドキュメンタリーなどのテレビ映像を大量に収集・蓄積しています。これを学習用データ（コーパス）として使うには、個々の動画が何を意味しているかをあらかじめ定義しておかなければなりません。私たちの研究グループで行った、映像につけられた字幕を手がかりに、登場する人の顔から名前を判別する「Name-IT」というプロジェクトでは、7 割がたは正しく対応付けができました。
でも、これでは映像の意味を理解しているとは言えません。字幕など使わずに、映像だけからその意味付けを理解することを考えています。また、ニュース番組だけでは動画が示す情報の範囲が限られてきます。対象をもっと広げて、トレンドドラマにあるようなごく日常的な生活を映した動画までを目標にしています。

写っているモノやコトで動画を検索可能に
動画検索の技術は、テキスト検索とは違って基本的なレベルからやるべきことがたくさん残っています。
動画の映像を意味付けするには、その前に、映像からどんな意味を読み取るか、どういう検索処理をしたいかを考える必要があります。A という動画と B という動画が同じかどうかを調べたり、CM 部分だけを取り除いたりする技術はすでにあります。ところが、何が映っているのか、何をしているのか、という意味を探るにはどんな検索をしたらいいのか、まだよくわかっていません。もっと言えば、そもそも、画像についてどんな検索ができたら価値があるのか、という点では研究者の中でも結論が出ていません。その議論を並行させながら、各研究者ごとに、こんな検索ができたらいいはずだと考えて、研究を進めているのが実情です。
この議論を少し収束させながら、動画検索技術の進歩を図る 1 つの試みとして、米 NIST（National Institute of Standards and Technology）が「TRECVID」（トレックビッド）という動画検索技術の研究開発プロジェクト（ワークショップ）を進めています。
TRECVID では、映像の意味（分類）のセットを決めて、実際の映像にその分類をつけた動画データをコーパスとして提供しています。例えば、空港の駐機場のショットとか、空を飛行機が飛んでいるショットがあって、そのショットに写りこんでいるものの名前（1 ショットに 1 つとは限りません）や、シーンの状況などをタグ付けしてあります。
各研究者はそれぞれが開発した視覚システムに TRECVID のコーパスを学習させた後、テスト用データの意味解析を行わせ、その結果を提出します。正答率が示されるだけでなく、参加しているほかの研究グループと比較できるようにもなっています。つまり第三者が仮に基準を決めて、それに沿った意味付けをした動画データを基に、各研究者が正答率の高さを競いあうことで、動画検索の技術を一緒に磨いていこうという企画なのです。私たち NII グループのほかに、米 IBM や米 Microsoft、日本から KDDI や電通大なども参加しています。
こういった場で研鑽しながら、私としてはまず数年以内に動画に映っているもの（主語）を指定して検索（認識）できる技術を確立したいと考えています。主語にあたるものがきちんと認識

できるようになれば、おそらくその延長線上で、その主語がいったい何をしているのか、という述語的な情報を検索する技術を生み出せるだろうと思っています。

（取材・構成　齋藤　淳）