社団法人 情報処理学会 研究報告
IPSJ SIG Technical Report

2006－CVIM－155 (16)
2006／9／8

# コーパスベース映像解析

佐藤真一

国立情報学研究所
〒101-8430 東京都千代田区一ツ橋 2-1-2

**概要:** 一般のユーザーが大量の映像を扱いうる環境が整備されてきており、映像の内容に基づく検索やブラウジングへの期待が高まっている。本当の意味での映像の内容検索を実現するためには、キーワードやメタデータによらない本当の映像内容検索の実現が必要であり、そのためには映像の意味内容の解析が必要だが、現状の技術水準は実応用での要求水準にはるかに及ばない。一方、音声認識、自然言語処理、文字認識、顔検出・認識などの技術は、いまや実用化の水準に達している。これらの技術の成功の要因のひとつは、コーパスベースの解析技術の利用があげられる。このアプローチを映像解析にも適用できないのであろうか。本稿では、そのための試みとして、TRECVID を紹介する。それを通して、コーパスベースのアプローチで実現可能になってきた映像解析の技術水準、コーパスベース映像解析の課題、および今後の方向性について明らかにしたい。

# Corpus-Based Video Analysis

Shin'ichi Satoh

National Institute of Informatics
2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101-8430

**Abstract:** Recent technology innovation enables ordinary people to access huge scale video archives, and thus the requirement of content-based video retrieval and browsing is raised. In order to realize "true" content-based video access, it is crucial to analyze semantic content of videos, however, currently available technologies are far below the level of the practical requirement. On the other hand, continuous speech recognition, natural language processing, character recognition, and face detection/recognition technologies were put to practical use. One of the key technical issues which enables this achievement is the use of corpus-based analysis. It is ideal if we can apply corpus-based method to video analysis. This article introduces TRECVID, one of the first approaches towards corpus-based video analysis. We would like to reveal the latest technology level of corpus-based video analysis, issues in corpus-based video analysis, and future direction of the approach.

## 1 はじめに

映像を扱うサービスの要求、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、一般の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を適切に扱うためには、映像の内容に基づく検索やブラウジングなどの技術が必要不可欠である。そのための方法のひとつとして、映像に対してメタデータを付与したり、キーワードを付与したりすることにより、テキストベースの検索手法を利用することが考えられる。しかしながらこの手法では、メタデータやキーワードを付与するための労力の問題、映像内容を記述するテキ

ストを作成する際の恣意性・主観性の問題などがあり、必ずしも映像の内容に基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、計算機による映像の意味内容の解析が必要不可欠である。しかしながら、一般に映像の意味内容の解析はきわめて困難である。その理由の一端は、映像の意味内容の多義性、あいまい性、および多様性にあると考えられる。

その一方で、音声認識や自然言語解析も、同様の困難さを有していながら、映像解析に比べて、はるかに高度な解析が可能となっている。その主たる要因は、コーペスベースの手法の成功にあると考えられる。すなわち、実データの持つ多様性を十分に反映した、正解データ (Ground Truth) つきの大規模コーパスを整備した上、統計的手法、機械学習などを用いて、コーパスに含まれる事例を適切に一般化することにより、実データのもつ多様性、あいまい性、多義性などに柔軟に対応できる技術が実現できたことが主因といえる。例えば、音声認識における大語彙連続音声コーパスの整備および HMM などの技術の開発、自然言語処理におけるテキストコーパスの利用、Probabilistic Grammar, TFIDF, LSI などの技術の開発が相当する。画像においても、手書き文字認識のために ETL 手書き文字データベース、顔検出では CMU-MIT 顔データ、HOIP 顔データなどが整備され、手書き文字認識や顔検出・認識は実用レベルの技術が実現されている。映像解析においても、コーパスベースのアプローチにより、大きな飛躍が望めると考えられる。望ましいコーパスとしては、実データの多様性を十分に反映した大規模なデータ、解析を実現するのに十分な詳細さを有する Ground Truth の整備が重要である。コーパスベース映像解析では、どのような映像を集めるか、どのような Ground Truth を用意するべきか、どのような学習アルゴリズムが利用可能か、何がシンボル (文字、音素、単語、など) か、どのような特徴量を用いればよいのか、などについて検討する必要がある。本稿では、コーパスベース映像解析に向けての世界初の本格的な試みとして、TRECVID について紹介し、そのアプローチを明らかにする。

## 2 コーパスベース解析手法

本節では、これまでに行われてきているコーパスベース解析手法の例として、音声情報に対する連続音声認識、テキスト情報に対する自然言語処理および情報検索、画像情報に対する文字認識と顔検出・認識をあげ、これまでの流れを概観する。

前節で述べたとおり、コーパスベースのアプローチでは、実応用の場面を反映した大規模なコーパスの構築、有効な特徴量の抽出、コーパスの事例を適切に一般化できる機械学習アルゴリズムなどの技術の開発が重要である。まずは音声認識技術を例にとる。大規模連続音声コーパスによる音声認識の近年の概要については、[1] に詳しい。音声認識では、1980 年以前は、比較的少量で、単音節の音声コーパスに基づき、数字などの単語認識が実現されていた。その裏では、Mel Frequency Cepstrum Coefficient (MFCC) 特徴量の開発、Baum-Welch アルゴリズム [2] や Viterbi アルゴリズム [3] の開発による HMM の実用化、ならびにその音声認識への適用の検討により、技術的な素地は 1990 年までに十分に整っていた。ここで、数百時間から千時間に及ぶ大語彙連続音声コーパスが整備され、上記の特徴量や機械学習アルゴリズムを適用することにより、不特定話者、語彙無制限の連続音声認識が実用化された。2000 年には、普通に売られているほとんどの PC に連続音声認識システムが搭載されており、電子メイル程度なら何の問題もなく音声で入力することが可能になっている。

テキスト処理でも、大規模コーパスに基づく解析で大きな成功を収めている。統計的自然言語処理については、[4] に詳しい。機械可読のテキストコーパスの整備は古くから行われており、著作権切れの古典のテキスト化を進めている Project Gutenberg は 1970 年代から活動を開始しており、The Wall Street Journal も 1980 年代から、日本語でも毎日新聞は 1990 年程度から大規模コーパスとして利用可能になっている。これらに基づく日本語情報検索の競争型ワークショップとして、1999 年から NTCIR が開始されている。こうした大規模コーパスを活用できる自然言語処理技術としては、確率的文脈自由文法 (Probabilistic CFG: PCFG)[5]、HMM を用いた part-of-speech (POS) tagger [6] などがある。テキスト情報検索としては、特徴量あるいは表現形式として、ベクタースペースモデル、あるいは bag of

words モデルが提案され [7]、コーパスに基づく重み付け手法として TFIDF[8]、コーパス中の単語分布による潜在的概念を用いた検索手法である Latent Semantic Indexing (LSI) [9] などがある。

画像処理においても、120 万サンプルを含む ETL 手書き文字データベースは 1990 年までに整備され、手書き文字認識技術の進展に大きく寄与しているし、顔認識のための数千サンプルを含む CMU-MIT 顔データベース [10, 11] は 1990 年代に整備され、現在のリアルタイム顔検出や顔認識研究の礎となっている。

映像コーパスについて考えると、次節で述べるとおり、TRECVID がやっと 2000 年に入ってから開始されたところであり、今現在、音声認識で言うところの単語認識からはじめ、関連する要素技術を徐々に整備していくべきところであると考えられる。

## 3 TRECVID の概要

TRECVID[1]は、2001 年より、TREC の中の Video Track として開始された。TREC は、テキストコーパスを用いた、情報検索のための競争型のワークショップであり、TREC Video Track は、その映像版として開始された。TREC2001 および TREC2002 Video Track の後、TREC 本体からは独立して、2003 年からは TRECVID として独立したワークショップとなった。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目指し、米国標準技術局 (National Institute of Standards and Technology:NIST) と Disruptive Technology Office (DTO) の主催で行われている。DTO とは、もとは ARDA として知られていた米国の研究資金配分機関であり、DARPA の一部のプロジェクトを担当している。

TRECVID では、毎年、大学の研究室や企業の研究所内の研究グループ単位での参加を募り、各参加者に同じ映像データを提供し、同じタスクについて個別に実現手法を検討させ、タスクの出力を収集してワークショップにて比較・検討を行う。ここで、特定のタスクに対してはどのような手法が有利か、データセットの選定やタスクの設定は適切であったか、評価方法は適切であったか、などについて評価を行うことにより、映像検索および映像解析技術の推進をねらっている。同じデータで同じタスクに対して異なるアプローチ間の比較検討を行うことにより、公正な比較評価ができ、より効果的に技術の進展が見込めるというのが基本的な考え方である。

NIST ならびに DTO の主催ではあるが、参加者は米国内のグループに限らず、ヨーロッパやアジアのグループが多く参加している。TRECVID2005 では、アジア/オーストラリアから 11 チーム、ヨーロッパから 17 チーム、南北アメリカから 13 チーム、ならびに米国・欧州混成の 1 チームが参加した。TRECVID の全体の企画調整は、Alan Smeaton (Dublin City University) と Wessel Kraaij (TNO Information and Communication Technology、オランダ) が、庶務については NIST の Paul Over と Tzveta Ianeva がつとめている。

## 4 提供されるデータ

TRECVID では、対象とする映像としては、放送用の映像素材を考えている。したがって、参加者間で、相応の規模の放送映像素材を共有しなければならない。各研究グループが例年 2 月ごろに TRECVID への参加表明をすると、4 月ごろに NIST あるいは Linguistic Data Consortium (LDC) よりハードディスクが送られてくる。これに共有すべき映像が納められている (もっとも、映像を手元にコピーした後には、ハードディスクは返送しなければならない)。大部分の映像は、MPEG-1 フォーマットのファイルとして格納されている。提供される映像は、手法の (主として機械学習手法のための) 訓練データとして使ってもよい映像 (トレーニングセット) と、最終的にタスクの結果を出力するために用いる評価用データとしての映像 (テストセット) とに分かれる。もちろん、テストセットは手法の訓練に用いてはならず、開発者は事前にテストセットを見ることは許されない (が、事前に評価用にデータは送られてくるので、この辺は紳士協定)。

提供される映像の内容を表 1 に示す。2001 年と 2002 年に利用された映像は、NIST より提供された米国政府機関関係の映像、Open Video Project[2]および Internet Archive[3]で提供されているフリーの映像素材、および BBC から提供された映像素材であった。これらの映像を用いて、ショット分割、特徴

[1] http://www-nlpir.nist.gov/projects/trecvid/

[2] http://www.open-video.org/

[3] http://www.archive.org/details/movies/

表 1: TRECVID で提供される映像

| 年 | ワークショップ名 | 分量 | 内容 |
|---|---|---|---|
| 2001 | TREC2001 Video Track | 11 時間 | NIST Video, Open Video Project, BBC Stockshot |
| 2002 | TREC2002 Video Track | 69 時間 | The Internet Archive, Open Video Project |
| 2003 | TRECVID2003 | 120+13 時間 | ニュース (ABC+CNN)+会議など (C-SPAN) |
| 2004 | TRECVID2004 | 70 時間 | ニュース (ABC World News Tonight, CNN Headline) |
| 2005 | TRECVID2005 | 170+50 時間 | ニュース (英語、中国語、アラビア語)+BBC Rushes |
| 2006 | TRECVID2006 | 159+50 時間 | ニュース (英語、中国語、アラビア語)+BBC Rushes |

抽出、検索などのタスクが課せられた。しかしながら、これらの映像は内容や品質の点でもばらつきが大きく、実際に利用要求の高い放送映像などの完成度の高い映像とも大きく異なるものであり、これらの点は問題として認識された。これを受けて、2003年からは、LDC との連携により、実際に放送されたニュース映像が提供されるようになった。2003 年と2004 年は、ABC と CNN のニュース映像が提供され、2005 年と 2006 年は、英語のニュース映像 (CNN, NBC, MSNBC) に加え、中国語 (CCTV, NTDTV, Phoenix) とアラビア語 (LBC, ALH) のニュース映像も提供されている。ニュース映像は、もちろん放送に供するだけの品質を持っており、明確な話題の分割点、キャスターショット (アンカーショットと呼ばれる)、天気図などによる時間的な明解なパターンを持ち、その構造については比較的解析しやすいという特性を持つ。その一方、ニュース内の素材映像には森羅万象の映像が利用され、解析が困難な上、通常の検索要求は音声情報 (発話内容、すなわちテキスト情報) で対応できてしまい、映像検索ならではの解析や検索の特性が出にくいという問題があった。そこで、2005 年からは、BBC Rush と呼ばれる、英国 BBC から提供された、放送映像を作るために撮影された素材映像群が対象映像として提供されるようになった。これらには、もちろん、ナレーションもなく、編集もされていないため時間パターンも持たず、検索などに供するためには、本当に映像内容を解析するしかない対象である。

このようにして各年に配布される映像間には、基本的には重なりはなく、ニュース映像については、これまでに延べ 500 時間あまりが提供されたことになる。以下本稿では、TRECVID の主たる映像素材として、ニュース映像を中心に述べる。

ニュース映像には、映像以外にも、映像に関連するさまざまな情報が提供される。後述する高次特徴抽出および検索タスクでは、ショット単位で、高次特徴の存在や、検索課題への適否の評価が求められる。このとき、ショット分割を各参加者ごとに個別に実行したのでは、ショット分割結果が参加者ごとにまちまちになり、タスクの結果の評価に支障をきたす。そのため、提供された映像に対して、共通に利用すべきショット分割点の情報として、共通ショット境界リファレンス情報が提供される。各ショットには ID が振られ、これにより高次特徴抽出や検索タスクの結果を記述することになる。ショット情報に付随して、各ショットの代表フレームの情報が JPEG 画像つきで提供される。少なからぬ参加者は、高次特徴抽出や検索タスクにおいて、各ショットの視覚的特徴量として、映像から特徴量を算出するのではなく、この代表フレーム画像から、静止画の特徴量を算出して使っている。また、テキストデータとして、映像中の音声情報に対し音声認識を適用したトランスクリプト情報が与えられる。トランスクリプトはすべて英語で与えられる。英語の映像については英語音声認識を適用しているが、中国語とアラビア語については、まずそれぞれの言語の音声認識を適用し、しかる後に各言語から英語への機械翻訳を適用し、自動的に得ている。日本でも提供が開始されているが、米国では特にほとんどの番組に文字字幕情報 (クローズドキャプション) が付与されているが、TRECVID ではこれを利用せず、あくまで音声も含む映像のみが与えられた情況を想定している。音声認識には当然認識誤りが混入し、中国語やアラビア語の映像についてはさらに翻訳の誤りも混入することになり、問題を難しくしているが、TRECVID ではこれに対応することも求められている。図 1 に与えられるトランスクリプト情報の例を示す。英語についてはまず問題ないが、中国語とアラビア語については、ほとんど言語として成立していない。しかし、キーワードの羅列としては使えそうに見える。

54 3043 Federal and city manager Susan Presley Roman is watching the MSNBC on the MS Ireland and Japan as soon as MS and a new line cinema has announced it had as soon as it is now when you hold measures when your decision townhouse on the future and have as much an MSNBC and
3104 5279 the and has a chair and terrorism R. F. I. S. identified as a campaign during its final weekend and only seven and then live in Spain -- a campaign stop in outlook and was nonsense that indicate a blast of the bush administration for allowing bean wanting to Stanford for more in December two thousand one

(a) English

<text_track id="0x4c57584c" name="Language Weaver Translation">
<text record_id="1">
<timespan in_msec="10290" in_smpte="00:00:10:08" out_msec="38440"
out_smpte="00:00:38:12"/>
The station, such as good overlapping Lebanese this week and on Tuesday heading George W. Bush and John Kelley to gain confidence of the American people four years to come appear to be critical in wars, particularly in the Middle East, where many American results will determine the fate of the regulations and other organizations, including Syria and Lebanon and Syria target effects of Accountability Act </text>

(b) Arabic

<text_track id="0x4c57584c" name="Language Weaver Translation">
<text record_id="1">
<timespan in_msec="7230" in_smpte="00:00:07:07" out_msec="23940"
out_smpte="00:00:23:27"/>
<prop_list/>Members welcome good news programs watching to see detailed reports downtown Tel Aviv, Israel an open market 1 October explosion killed at least 4 people killed more than 30 people were injured this the Palestinian National Authority Chairman Yasser Arafat to France after the first place in the attack against Israeli attacks </text>

(c) Chinese

図 1: トランスクリプトの例

トランスクリプト情報の関連情報としては、音声認識結果ならではの副産物として、発話人物同定結果、発話中の人物名の検出結果も与えられている。

この他、後述の高次特徴抽出タスクでは、トレーニングセットの各ショットについて正解 (Ground Truth) が提供されるので、Ground Truth つきの大規模映像コレクションが提供されることになる。参加者は、トレーニングセットのショットから特徴量を抽出し、Ground Truth にしたがって分類器を学習し、テストセットの分類結果を NIST に提出すればよい。その結果は、世界中の研究グループの提出した結果と比較検討され、自分たちの成果の世界の中での位置づけや、他の手法と比較した特性などが明らかになる。このように、TRECVID ではすぐに研究に着手できる状況のデータが提供され、分量も映像データとしては大規模なものであり、かつ実験結果は他の研究グループの結果と比較検討される。映像解析・検索研究の素材ならびに環境としては、申し分ないものと思われる。

# 5 主なタスクの概要

TRECVID で参加者に課せられるタスクは、年により、また対象の映像の種類により、少しずつ異なっている。ニュース映像に関していうと、これまでにショット検出、話題境界検出、低次特徴量抽出 (カメラワーク検出)、高次特徴量抽出、検索といったタスクが課せられた。ここでは、それらのタスクのうち、主なものについて概要を述べる。

## 5.1 ショット検出 (Shot boundary detection)

ショット検出タスクでは、与えられた映像中のショット境界を検出する。また、これらが瞬時ショット境界 (cut) であったか、暫時ショット境界 (gradual) であったかの判定も求められている。性能は主として precision と recall で評価される。これに加え、検出に要する処理時間も評価項目となっている。

主たる手法としては、基本的には各フレームから得られる特徴量の大きな変化により検出する手法があげられるが、急激な物体の動きやカメラワーク、フラッシュなどに対応するためにさまざまな特徴量が提案されている。色ヒストグラム、時空間画像特徴量、ガボール特徴量などが試されている。また、処理時間を少なくするため、与えられた MPEG ファイルを完全にデコードせず、DCT 係数などを直接参照する方法も提案されている。フレームごとにショット境界か否かを判定するために SVM などの機械学習を利用するグループも多く、高い性能を上げている。この場合、訓練時には、当然ながら非ショット境界のフレームのサンプルがショット境界フレームのサンプルよりも圧倒的に多いことになり、正例と負例の数が極端に異なった学習データから学習しなければならない。ショット検出タスクの精度は概して高く、瞬時ショット境界では F 値で 0.9 以上、暫時ショット境界でも 0.8 程度を達成している。

## 5.2 高次特徴抽出 (High-level feature extraction)

本当の意味での映像検索を実現するためには、映像中の意味的な特徴の有無を自動的に検出する必要がある。このように、映像から「意味」を抽出することを目的としているのが高次特徴抽出タスクである。TRECVID2005 で求められた高次特徴は、(1) People walking/running, (2) Explosion or fire, (3) Map, (4) US flag, (5) Building exterior, (6) Waterscape/waterfront, (7) Mountain, (8) Prisoner,


図 2: 高次特徴抽出タスクのトレーニングセットの例

(9) Sports, (10) Car の 10 種類であった。また、TRECVID2006 では、39 種類の高次特徴の抽出結果の提出が求められている (図 5 に示す)。本タスクでは、テストセット中の各ショットにつき、上記の特徴が存在するか否かを識別することが求められている。学習用には、トレーニングセットの各ショットに対し、上記の特徴がある/ないという Ground Truth が与えられる。トレーニングセットの例を図 2 に示す。各特徴ごとにラベル付けがされたショットにより学習し、未知のショットをクラスわけするという問題であり、パターン認識・機械学習問題ととらえることができる。課題としては、各ショットからどのような特徴量を抽出するか、どのような学習アルゴリズムを用いるか、などとなる。特徴量としては、視覚的特徴量、音響特徴量、テキスト特徴量などが用いられており、特に視覚特徴量については、多くのグループがショットの代表フレームに対する静止画の特徴量を利用している。

高次特徴の選定において、2005 年より、前出の DTO 主催の Large Scale Concept Ontology for Multimedia (LSCOM)[4][12] というワークショップと協力してタスクを定義している。LSCOM では、1000 程度の概念 (高次特徴に相当する) を定義し、TRECVID2005 のトレーニングセットを元に Ground Truth を作成し、マルチメディアデータ解析用のオントロジーの構築を目指している。実際、TRECVID2005 および 2006 の高次特徴抽出タスクは、LSCOM の部分集合となっている。TRECVID2006 では、LSCOM で現在までに作成の終わっている 400 あまりの概念についての Ground Truth も利用してよいことになっている。

高次特徴抽出タスクは、通常 Mean Average Precision (MAP) で評価される。MAP とは、各 recall で平均した precision であり、簡単に言うと、precision-recall 曲線と座標軸とで囲まれる部分の面積である。高次特徴によるが、最高の精度で 0.1 から 0.5 程度を達成している。

## 5.3 検索 (Search)

検索タスクは、TRECVID の最終目標ともいえるタスクで、高次特徴抽出の結果をフルに使い、与えられた問い合わせに合致するショットを同定するというタスクである。問い合わせに合致するショットを報告するという点では、高次特徴抽出タスクと似ているが、ひとつ大きく異なる点は、高次特徴抽出タスクでは抽出すべき高次特徴があらかじめ知らされているのに対し、問い合わせは結果提出の直前まで知らされない点である。当然、問い合わせ内容をもとにシステムを作ることは許されない。

各問い合わせは、自然文による説明と、Web 上の画像およびトレーニングセット中の映像による視覚的事例により構成される。2006 年の問い合わせは 24 種類であり、2006 年 8 月 10 日に公開された。自然文と画像の事例の一部による検索タスクの例を図 3 に示す。

検索タスクでは、interactive, manual, および automatic という実験条件が設定されている。automatic では、システムへの入力は上記の自然文・画像と映像の事例・ショットの例しか許されず、これらをもとにテストセット中から適切なショットを選ばなければならない。manual では、まず人間が問い合わせを見ることができ、これらをもとにシステムを操作し (適切なキーワードを入力する、画像特徴量のうち色よりもレイアウトに重きを置く、など)、結果を得ることができる。interactive では、さらに、システムの出力を見て、人間がシステムを再調整して結果を精緻化することができる。manual および interactive では、問い合わせが与えられてからシス

[4]http://www.ee.columbia.edu/dvmm/lscom/

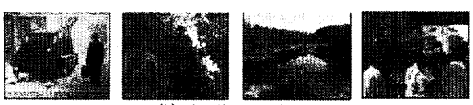
174 Find shots with a view of one or more tall buildings (more than 4 stories) and the top story visible

177 Find shots of a daytime demonstration or protest with at least part of one building visible

179 Find shots of Saddam Hussein with at least one other person's face at least partially visible

図 3: TRECVID2005 の検索タスクの例

(a) Interviews



(b) miscellaneous footages

図 4: BBC Rush で用いる映像の例

テムを操作して最終結果を得るまでの時間が最大で15分と制限されている。このように、異なる実験条件で結果を出力できる点も高次特徴抽出タスクと異なる部分である。

結果は precision-recall 曲線および MAP で評価される。問い合わせ、ならびに実験条件によるが、TRECVID2005 の interactive の場合、MAP で 0.1 から 1.0 近くまで達成できている。

### 5.4 BBC Rush

BBC Rush タスクでは、実は明確なタスクは定義されておらず、与えられた編集前映像を使って何ができるかを探ることが求められている。処理対象となる映像の例を図 4 に示す。このようなさまざまな素材映像が提供される。与えられた映像を使えば、基本的には何をやってもよいが、最小の要求用件が二つあげられている。

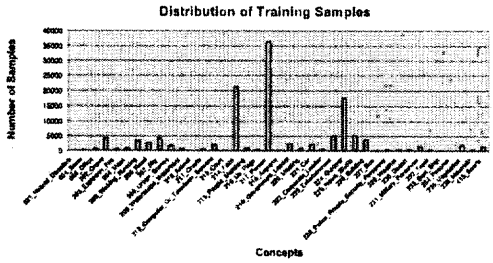



図 5: トレーニングセット中の高次特徴の正例の分布

- ショットの取り直しや同じシーンの取り流し素材映像などによる映像の冗長性を除くこと
- 最低 6 種類の特徴量を用い、素材を効率よくブラウズすること

特徴量も何を使ってもよいことになっているが、参考として、また DTO 主催の Video Analysis and Content Extraction Program Phase III (VACE III) で策定した Broad Agency Announcement (BAA) という、映像解析研究の方向性についての提言の中で触れられている特徴量 (例えばインタビューか否か、カメラ固定か否か、など) があげられている。

## 6 コーパスベース映像解析に関する関連研究

ここまで述べてきたように、TRECVID では、大量の実際の放送映像、LSCOM による数百種類の概念に基づくアノテーション、音声認識によるトランスクリプトなどが与えられ、単に TRECVID のタスクを遂行するのみならず、コーパスベース映像解析に向けての研究として、さまざまなアイデアが試せる場が提供される。本節では、TRECVID のコンテンツを使ったアプローチを中心に、コーパスベース映像解析に関連する研究をいくつか紹介する。

コーパスベースのアプローチでは、機械学習やデータマイニングに関する研究も重要である。ひとつには、前出のように、ショット検出でも高次特徴抽出でも、正例と負例の分量が極端に偏った学習サンプルからの学習が求められる。図 5 には TRECVID2006 での高次特徴と、それぞれにつき与えられた正例のショット数を表している。トレーニングセットに含

まれる全ショット数がおよそ62,000ショットなので、これに比べるとFaceやOutdoorなど一部の高次特徴を除き、1,000以下(一部は100以下)の大変少数の正例のみしか与えられていないことが分かる。したがって、少数の学習データからの学習、および負例に比べて極端に小数の正例のみ利用可能な場合の学習手法が求められる。[13]などに関連するアプローチが述べられている。

高次特徴抽出は、各高次特徴の種類ごとにショットの識別器を作成すればよく、そのように考えると単純なパターン認識・機械学習の問題といえる。しかし、LSCOMなどにより数百におよぶ種類のラベルが与えらていることを考えると、トレーニングセットの中で、異なるラベル間のサンプルの分布の相関を観察することができ、これによりラベル間の依存関係が推定できる。[14]では、Restricted Bolzman Machines (RBM) およびConditional Random Field (CRF)により高次特徴間の依存関係を推定し、高次特徴抽出の精度を上げている。[15]でもクラスタリングを用いて高次特徴間の類似度を求めている。[16]では、高次特徴の時間パターンを用いてショット分類の精度を上げている。映像はもちろんマルチモーダルな情報だが、[17]では、高次特徴抽出のための機械学習において、画像特徴やテキスト特徴などの異なる種類の特徴を適切に融合するアーキテクチャについて検討している。検索については、[18]では高次特徴抽出で得られるような情報を映像検索に活用する方法を検討しており、[19]ではTRECVIDコンテンツを用いた質問応答検索を提案している。[20]では、検索に有効な高次特徴の分類法について検討している。

TRECVIDのように大量の映像が与えられると、ショット間の単純な関連性を検出し、そのパターンをアーカイブ全体で評価することにより、映像マイニング的な技術が実現可能となる。有効なショット間の単純な関連性の検出法として注目を集めているのは、テロップなどのみが違うまったく同じ映像素材同士、あるいは同じシーンを同じ時点に異なった場所から取ったショット同士などを同定する、Near Duplicate Detection[21–24]と呼ばれる技術である。Near Duplicate Detectionで検出されるような関連性は、任意に集めた数時間程度の映像では数件程度しか検出できず、そこからマイニングのような形で構造などの高次情報を抽出するには、少なくともTRECVIDで提供されるような百時間規模の映像が必要である。このような大規模な映像アーカイブから抽出した単純な関連性に基づくマイニング技術に関する検討としては、[25–27]などがあげられ、ニュース映像の構造解析、映像ブラウザへの応用、ニュースアーカイブの話題構造の解析などが行われている。

近年、映像解析に限らず、機械学習は画像解析に広く使われており、画像を撮影したシーンの距離の推定[28]、セグメントの法線方向の大まかな推定[29]、画像データベース中の画像へのセグメントレベルのアノテーション[30,31]などが実現されている。同様のアプローチが映像アーカイブにも適用されつつあり、教師なしでショットの意味づけを行う研究[32,33]が行われている。また、映像中の顔と名前の対応付けについても、いまだにさまざまなアプローチが提案されている[34–37]。

## 7 おわりに

本稿では、コーパスベース映像解析に向けての世界初のアプローチのひとつとして、TRECVIDについて紹介した。TRECVIDは、映像解析および検索の高度化を目指した研究グループ参加型の研究プログラムであり、品質、分量ともに実応用での要求に近い映像アーカイブを参加者間で共有し、大規模なGround Truthも共有した上、共通のタスクについて競争的なワークショップを開催し、戦略的な研究の高度化を目指している。TRECVIDにより、映像解析研究の技術水準が大幅に向上し、映像解析のためのコーパスベースのアプローチの有効性が実際に示された。TRECVIDのタスクに限らず、大規模な映像アーカイブが利用可能になって初めて実現できたさまざまな研究の方向性も試みられ始めている。しかしながら、コーパスベース映像解析の試みはまだ緒についたばかりであり、TRECVIDにもまだ不足している点や改善しなければならない点も多く存在する。コーパスベース映像解析の今後の動向に注目していきたい。

## 参考文献

[1] S. Furui: "Recent progress in corpus-based spontaneous speech recognition", IE-

ICE Trans. on Information and Systems, **E88-D**, 3, pp. 366–375 (2005).

[2] L. E. Baum, T. Petrie, G. Soules and N. Weiss: "A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains", Annals of Mathematical Statistics, **41**, 1, pp. 164–171 (1970).

[3] A. J. Viterbi: "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm", IEEE Trans. on Information Theory, **IT-13**, pp. 260–269 (1967).

[4] C. D. Manning and H. Schütze: "Foundations of statistical natural language processing", MIT Press (1999).

[5] T. L. Booth and R. A. Thomson: "Applying probability measures to abstract languages", IEEE Trans. on Computers, **C-22**, pp. 442–450 (1973).

[6] F. Jelinek and R. Mercer: "Probability distribution estimation from sparse data", IBM Technical Disclosure Bulletin, 28 (1968).

[7] G. Salton and M. E. Lesk: "Computer evaluation of indexing and text processing", Journal of the ACM, **15**, 1, pp. 8–36 (1968).

[8] K. S. Jones: "A statistical interpretation of term specificity and its application in retrieval", Journal of Documentation, **28**, pp. 11–21 (1972).

[9] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer and R. Harshman: "Indexing by latent semantic analysis", Journal of the American Society for Information Science, **41**, pp. 391–407 (1990).

[10] H. A. Rowley, S. Baluja and T. Kanade: "Neural network-based face detection", IEEE Trans. on PAMI, **20**, 1, pp. 23–38 (1998).

[11] K. K. Sung and T. Poggio: "Example-based learning for view-based human face detection", IEEE Trans. on PAMI, **20**, 1, pp. 39–51 (1998).

[12] M. Naphade, S.-F. Chang, A. Hauptmann and J. Curtis: "Lange-scale concept ontology for multimedia", IEEE Multimedia, pp. 86–91 (2006).

[13] N. V. Chawla, N. Japkowicz and A. Kotcz: "Editorial: special issue on learning from imbalanced data sets", SIGKDD Explorations, **6**, 1, pp. 1–6 (2004).

[14] R. Yan, M.-Y. Chen and A. Hauptmann: "Mining relationship between video concepts using probabilistic graphical models", Proc. of ICME (2006).

[15] M. Koskela and A. Smeaton: "Clustering-based analysis of semantic concept models for video shots", Proc. ICME (2006).

[16] J. R. Kender and M. R. Naphade: "Video news shot labeling refinement via shot rhythm models", Proc. ICME (2006).

[17] C. G. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek and A. W. Smeulders: "The challenge problem for automated detection of 101 semantic concepts in multimedia", ACM Multimedia (2006).

[18] S.-Y. Neo, J. Zhao, M.-Y. Kan and T.-S. Chua: "Video retrieval using high level features: Exploiting query matching and confidence-based weighting", Proc. of CIVR (2006).

[19] H. Yang, L. Chaisorn, Y. Zhao, S.-Y. Neo and T.-S. Chua: "Videoqa: question answering on news video", ACM Multimedia (2003).

[20] W.-H. Lin and A. Hauptmann: "Which thousand words are worth picture? experiments on video retrieval using a thousand concepts", Proc. ICME (2006).

[21] D.-Q. Zhang and S.-F. Chang: "Detecting image near-duplicate by stochastic attributed relational graph matching with learning", ACM Multimedia (2004).

[22] 山岸, 佐藤, 浜田 : "大規模映像アーカイブのための映像断片照合の高速化", 情報技術レターズ, 第一回情報科学技術フォーラム (FIT2002), pp. 157–158 (2002). LI-17.

[23] F. Yamagishi, S. Satoh, T. Hamada and M. Sakauchi: "Identical video segment detection for large-scale broadcast video archives", Proc. of International Workshop on Content-Based Multimedia Indexing (CBMI'03), pp. 135–142 (2003).

[24] M. Takimoto, S. Satoh and M. Sakauchi: "Identification and detection of the same scene based on flashlight patterns", Proc. ICME (2006).

[25] S. Satoh: "News video analysis based on identical shot detection", Proc. of International Conference on Multimedia and Expo (2002).

[26] F. Yamagishi, S. Satoh and M. Sakauchi: "A news video browser using identical video segment detection", Proc. of Pacific-Rim Conference on Multimedia (PCM2004), Vol. II, pp. 205–212 (2004).

[27] Y. Zhai and M. Shah: "Tracking news stories across different sources", ACM Multimedia (2005).

[28] A. Torreba and A. Oliva: "Depth estimation from image structure", IEEE Trans. PAMI, **24**, 9 (2002).

[29] D. Hoiem, A. A. Efros and M. Hebert: "Geometric context from a single image", Proc. ICCV (2005).

[30] J. Winn, A. Criminisi and T. Minka: "Object categorization by learned universal visual dictionary", Proc. ICCV (2005).

[31] P. Duygulu, K. Barnard, J. F. G. de Freitas and D. A. Forsyth: "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary", Proc. ECCV (2002).

[32] P. Duygulu, M. Bastan and D. A. Forsyth: "Translating images to words for recognizing objects in large image and video collections", Towards Category-Level Object Recognition (Eds. by J. Ponce, M. Hebert, C. Schmid and A. Zisserman), Springer Verlag (2006).

[33] M. Bastan and P. Duygulu: "Recognizing objects and scenes in news videos", Proc. CIVR (2006).

[34] S. Satoh, Y. Nakamura and T. Kanade: "Name-It: Naming and detecting faces in news videos", IEEE MultiMedia, **6**, 1, pp. 22–35 (1999).

[35] T. Miller, A. C. Berg, J. Edwards, M. Maire, R. White, Y. W. Teh, E. Learned-Miller and D. A. Forsyth: "Faces and names in the news", Proc. CVPR (2004).

[36] J. Sivic, M. Everingham and A. Zisserman: "Person spotting: video shot retrieval for face sets", Proc. CIVR (2005).

[37] J. Yang, R. Yan and A. G. Hauptmann: "Multiple instance learning for labeling faces in broadcasting news video", ACM Multimedia (2005).