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

2005－ＳＬＰ－59（７）
2005／12／21

[招待講演] 映像情報検索の最前線

帆足 啓一郎†

†株式会社 KDDI 研究所 〒356-8502 ふじみ野市大原 2-1-15
E-mail: †hoashi@kddilabs.jp

あらまし　本稿では，近年研究が盛んに行われている映像情報検索ならびにその評価技術の最新動向について紹介する．具体的には，筆者らが参加している TREC Video Retrieval Workshop (TRECVID)の実験データ，共通タスクならびにその評価方法などについて説明し，今後の動向についての意見も述べる．

キーワード　映像情報検索，TRECVID，テストコレクション．

[Invited Paper] Frontline of Video Information Retrieval

Keiichiro HOASHI†

†KDDI R&D Laboratories, Inc.　2-1-15 Ohara, Fujimino-shi, Saitama, 356-8502 Japan
E-mail: †hoashi@kddilabs.jp

Abstract　This paper introduces the latest trends in the field of video information retrieval. Namely, activities in the TREC Video Retrieval Workshop (TRECVID), which the authors have participated in for the past few years, are explained. The explanation includes the experiment data, common tasks, and evaluation methods applied in TRECVID. The author will also present opinions regarding future directions of video information retrieval.

Keyword　Video information retrieval, TRECVID, test collection.

1. はじめに

地上波デジタル放送の開始や，HDD や DVD などの大容量メディアを搭載したレコーダーの普及にともない，一般ユーザでも大量の高画質映像データを蓄積できる環境が急速に広まっている．また，ブロードバンドネットワーク環境の拡大にともない，オンデマンド型の映像配信サービスなどが盛んになるなど，ユーザが閲覧できる映像データはここ数年急速に増加している．

しかし，ユーザが気軽に大量の映像データを蓄積することができるようになる一方，蓄積された大量のデータの中から閲覧したい映像を探しだすことが難しくなるのは言うまでもない．近年では EPG（電子番組表）に含まれるキーワードを基に自動録画する機能や番組単位での検索を可能にする機能を持つ機器も利用できるが，通常 EPG は番組単位でしか付与されていないため，たとえば録画したニュース番組の中から，興味のある話題だけを見たいなどといったニーズには対応が困難である．

また，映像データの細かい中身に対するメタデータの形式として，MPEG-7 や TV-Anytime などが規定されているが，テレビ番組などで発生する細かい事象の全てに対してメタデータを付与する作業は高コストであり，同様にしてすでに流通している映像データにメタデータを付与することも膨大な作業コストが必要となる．そこで，大量の映像データの中身を解析して検索を行う content-based な映像情報検索技術のニーズが急速に高まっている．

本稿では，近年研究が盛んに行われている content-based 映像情報検索技術，および同技術の評価方法の最新動向について解説する．具体的には，映像情報検索技術の現状を解説し，映像情報検索技術評価の最新動向として，映像情報検索のための大規模テストコレクションの代表的な取り組みである TREC Video Retrieval Workshop について、筆者らの参加[1]から得られた知見などを含めて紹介する．そして，最後に今後の映像情報検索技術の発展に向けた課題などを述べる．

2. 映像情報検索技術の現状

映像情報検索技術という観点からは，前述した content-based 映像検索やメタ情報の効率的な付与のため，映像情報のインデクシング技術の研究が盛んに行われている．これまで，映像情報のインデクシング技術としてカット点検出や無音検出などの映像データの物理構造を定義するインデクシング（構造的インデクシング）が数多く検討されており，カット点表示などは商用の映像アーカイブシステムでも一般的な機能となっているが，構造的インデクシングではより高度な検索要求，例えば特定イベントや話題などの検索を実

現することができないため，これらを可能にするために，映像データの意味構造を定義するインデクシング（意味的インデクシング）が求められている．これまで人手が介在する必要があった意味的インデクシング技術においては，映像データの増加に伴う映像制作コストの増大を抑制するため，自動化や半自動化により実現できることが望まれている．代表的な意味的インデクシングとしては，スポーツ映像からのハイライト生成[2]，イベント検出[3]，ニュース映像からのアンカーショット検出[4]，話題分割[5]，および映画のジャンル分類[6]などが挙げられる．

これらの既存文献では，実証実験に使用した映像データにおいてはある程度の結果が得られることが報告されているが，少数の映像データにおける限定された評価手法の範囲内での結果しか示されていないものが殆どである．これは，例えばあらゆるニュース映像に対して十分な精度を保証するための映像データおよび正解（ground truth）の入手が困難であること，さらに意味的インデクシング技術の評価は一般的に主観的側面に立つことが多いため，評価尺度も確立されていないことに起因する．また，スポーツ映像やTV映像を実験に用いている場合は，著作権の問題からそれらを共通的に利用することができないため，方式の優位性を示すための比較実験を行うことも困難である．従って，大規模かつ共通的な映像データを実験に利用できるための環境を構築することが望まれている．

3. 映像情報検索技術評価の最新動向

前述の背景により，テキスト情報検索分野におけるTRECやNTCIRのような，共通の評価プラットホームが映像情報検索分野にも必要となっている．そこで，近年TREC Video Retrieval Workshop (TRECVID) [1]で提供されるテストコレクションが，映像情報検索技術の性能を評価するためのデータとして，デファクトスタンダードになりつつある．本節では，映像情報検索技術評価に関する最新動向の代表例として，同ワークショップでの取り組みについて紹介する．

3.1. TREC Video Retrieval Workshop

TRECVIDは，当初TRECの中の1つのサブタスクとして，TREC2001とTREC2002に含まれていたが，2003年からは，独立したワークショップとして開催されており，今年のTRECVID 2005[7]で4回目の開催となる．映像情報検索研究への関心の増加にともない，TRECVIDの規模も年々拡大しており，TRECVID 2005には48組織が参加を表明している（主催者情報による）．

TRECと同様，TRECVIDも大量の実験用データとともに，共通のタスクおよび各タスクに対する評価基準を設定している．参加者は，TRECVIDが配布する実験用データに基づいてタスクに取り組み，与えられた期限までに各タスクの実験結果を主催者であるNIST（National Institute of Standards and Technologies）に送付する．NISTでは，タスクごとに設定された評価基準に従って全ての実験結果を評価し，その結果を参加者に返送するとともに，TRECVIDワークショップで報告を行う．

TRECVIDは，映像情報検索技術評価のためのテストコレクションとしては，世界で初めてのものである．データ量も豊富な上，データに含まれる映像データの（研究目的での）利用権も許諾されていることなどから，映像情報検索研究を進めるためには理想的な実験データであるといえる．また，TRECVIDのタスクは，構造的インデクシングであるカット点検出にはじまり，意味的インデクシングとしての高次特徴抽出に至るまで，非常に幅が広く，かつタスクの内容や評価基準もTRECVID参加者の間での議論を経て決定されている．以上の点から，映像情報検索技術の分野においては，TRECVIDのテストコレクションがデファクトスタンダードとしての地位を確立しているといえる．

以下，TRECVIDのテストコレクションに含まれる映像データ，および共通タスクとその評価基準について，それぞれ紹介する．

3.2. TRECVID 実験データ

TRECVIDでは，各タスクの実験を行うための映像データとして，ニュース番組を中心とした大量のMPEG-1ファイルが用意されている．TRECVID 2003では，1998年1月～6月に米国で放送された「ABC World News Tonight」および「CNN Headline News」の2つのニュース番組（241本，約120時間分）と，米国議会での議論の模様などが収録されているCSPAN（25本，約13時間分）の映像から構成されている．また，TRECVID 2004では，TRECVID 2003の実験データに加え，新たに1998年10月～12月に放送された「ABC World News Tonight」と「CNN Headline News」（128本，約64時間分）が追加されている．

TRECVID 2005では，新たな実験データが用意された．具体的には，2004年11月に放送された，英語・中国語・アラビア語のニュース番組の映像（合計約170時間分）から構成されている．

TRECVIDでは，上記の映像データを放送日に従って2つのグループに大別している．この2グループの内，時系列的に古い方を「学習データ(development data)」，

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

新しい方を「評価データ(test data)」と定義している．後述する TRECVID のタスクの多くでは，学習データを基に実験システムの開発ならびにチューニングを行い，評価データに対する実験結果を提出する手順を採っている．

また，TRECVID の映像データには，以下の情報が提供されている．

- **共通カット点情報**（Common shot boundary）
  全ての映像ファイルに対し，カット点検出処理を行った結果が提供されている．このカット点検出結果によって分割された個々の「ショット」が，後述する Feature extraction および Search の各タスクの評価の単位として利用されている．また，各ショットには，キーフレーム（各ショットを表す代表的なフレームの静止画像）も併せて付与されている．
- **音声認識結果**
  個々の映像ファイルから抽出されたオーディオ情報に対し，音声認識処理を実行した結果が付与されている．TRECVID 2005 の実験データには，英語以外の言語（中国語，アラビア語）の映像も含まれているため，これらの映像の音声認識結果に対しては，英語への自動翻訳結果も提供されている．

さらに，TRECVID 2003 と TRECVID 2005 では，参加者の有志によって，実験データに含まれる映像ファイルに対しアノテーション情報を付与する作業（common feature annotation）が行われている．この作業で付与されるアノテーション情報は合計 133 種類に及んでいる．実際の作業では，参加者有志で TRECVID 2003 学習データ（約 60 時間）をファイル単位で分担し，IBM Research が開発したアノテーションツールを利用して行われた．その結果，アノテーションが付与された映像データとしては，世界最大規模のデータが構築された．また，TRECVID 2005 においても，同様の取り組みが行われたが，TRECVID 2003 でのアノテーション付与作業の負荷が大きかったことなどから，TRECVID 2005 では，LSCOM (Large Scale Concept Ontology for Multimedia Understanding) ワークショップで議論されている，ニュース番組解析のために有効な特徴量（"concept"）39 件をアノテーションとして付与している．

3.3. TRECVID タスク

前述の通り，TRECVID では毎年複数のタスクを設定しており，タスクごとに評価基準を設けている．以下，筆者らが参加した TRECVID 2004 ならびに 2005 において設定されたタスクの概要について，それぞれ簡単に紹介する．

3.3.1. Shot boundary detection（カット点検出）

Shot boundary detection（以下，SBD）タスクの目的は，実験対象映像ファイルに含まれるカット点，すなわち，各ショット間の切替点を自動的に検出することである．TRECVID の SBD タスクでは，カット点を「瞬時カット」と「特殊カット」の 2 種類に大別している．瞬時カットとは，ショットが瞬間的に切り替わるカット点のことであり，特殊カットは，複数フレームにまたがってショットが切り替わるカット点である．具体的には，ショット切り替え前のショットの映像がフェードアウトしながら次のショットの映像がフェードインするディゾルブ型のカット点と，ワイプなどの映像効果をともなうカット点が特殊カットに含まれる．

SBD タスクの評価基準としては，全正解ショット切替点の内，検出することができたショット切替点の割合（再現率，Recall）と，検出されたショット切替点中の正解ショット切替点の割合（適合率，Precision）が採用されている．また，特殊カットに対しては，上記の指標に加え，Frame-recall と Frame-precision という指標も採用されている．Frame-recall および Frame-precision は，カット点の有無だけでなく，区間としてどの程度正確に検出できているかを示す指標である．

また本タスクでは，カット点検出の精度を示す上記の各指標に加え，カット点検出に必要な処理時間（complexity）についての参加者からの報告が必要となっている．具体的には，カット点検出に要した総処理時間，MPEG デコード処理時間，および実験で利用した計算機の CPU の情報が報告事項として求められている．

3.3.2. Story Segmentation（話題分割）

Story segmentation タスクの目的は，映像データを意味的な単位，すなわち「話題」に分割することである．TRECVID 2004 では，実験データとしてニュース番組を利用している．ニュース番組は，一般的には複数の話題によって構成されているが，本タスクでは，評価データに含まれる個々のニュース番組を構成する話題の境界（話題分割点）を自動的に検出することが目的である．従来，テキスト情報検索の分野でも話題分割の研究は進められてはいるが，本タスクでは，テキスト情報に基づく従来の話題分割技術に対し，映像から得られる特徴の利用方法について評価することが目的の 1 つであるとされている．

本タスクでは，以下の 3 つの実験条件を設定してお

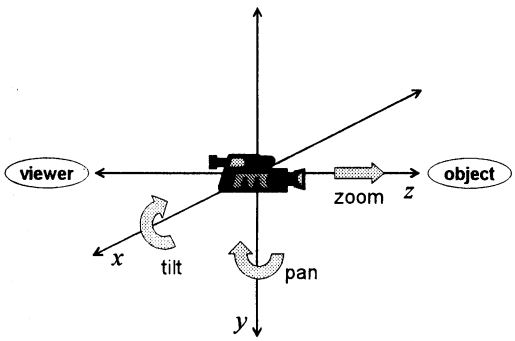

図 1　カメラモーションの説明

り，参加者は，上記 3 条件のそれぞれに該当する実験結果を送付することが義務付けられている．

- Audio + Video
  映像から抽出される audio-video 特徴量のみに基づく話題分割．
- Audio + Video + ASR
  上記 audio-video 特徴量に加え，音声認識結果から得られるテキスト情報を利用した話題分割．
- ASR Only
  音声認識結果から得られるテキスト情報のみを利用した話題分割．

本タスクでは，参照話題分割点の前後 5 秒以内，合計 10 秒の区間内に検出された話題分割点を正解とみなし，再現率と適合率をそれぞれ算出し，評価基準として採用している．

3.3.3. Low level feature extraction（低次特徴抽出）

TRECVID 2005 において新たに開始されたタスクである．本タスクの目的は，ショット内のカメラモーションの検出である．具体的には，「パン」「チルト」「ズーム」の，それぞれのカメラモーションが出現するショット（共通カット点情報に基づく）を検出することが目的である．エラー! 参照元が見つかりません。に，カメラモーションの説明を示す．

本タスクでの評価基準は，カメラモーション検出結果に対する再現率と適合率である．ただし，カメラの手ぶれにともなうカメラモーションなど，判断が曖昧なショットについては，あらかじめ評価の対象外としている．

3.3.4. Feature extraction（高次特徴抽出）

Feature extraction タスクの目的は，指定された事象(feature)が出現する箇所を，分析対象映像データから検出することである．具体的には，前述の common shot boundary によって設定されたショットの中から，課題として与えられた feature が出現するショットを検出するタスクである．TRECVID 2005 では，「People walking/running(2 人以上の人間が歩いている／走っている)」など，前述のアノテーション付与作業にて設定された 39 件のアノテーションのうち，10 件が本タスクの課題として設定されている（
参照）．

表 1 TRECVID 2005 feature 一覧

ID	Feature
38	People walking/running
39	Explosion or fire
40	Map
41	US flag
42	Building exterior
43	Waterscape/waterfront
44	Mountain
45	Prisoner
46	Sports
47	Car

本タスクの評価基準は，送付された検索結果（スコアによって順位付けられた最大 1000 件のショット)の再現率，適合率，および平均適合率 (Average precision) である．これらの指標の算出に必要な正解データは，TREC でも採用されている「プーリング方式」によって構築されている．具体的には，参加者が送付した実験結果に含まれるショットのみを評価対象として，TRECVID の評価者が目で検出対象 feature の出現有無を確認した結果が，正解データとして採用されている．

3.3.5. Search（検索）

Search タスクの目的は，与えられたクエリを満たすショットを効率的に検索するシステムの開発である．タスクの目的自体は，feature extraction タスクと類似しているが，本タスクでは，検索実験に人間が介在していることが前提となっている点が，FE タスクと大きく異なっている．すなわち，検索精度とともに，検索システムのインタフェースも評価の対象となっているのが，本タスクの特徴といえる．また，検索課題にあたる Topic も，Feature extraction のそれよりも複雑であるほか，Topic 自体がテキスト文に加え，正解ショットのサンプル（動画と静止画）によって提示されている点も本タスクの特徴である．図 3に，TRECVID 2004 における検索課題の例を示す．この例では，検索要求を表すテキスト文の他に，正解の例を示す静止画 1 件と，動画（ショット）2 件が与えられている．

本タスクでは，Manual，Interactive，Full automatic の 3 つの実験条件が設定されている（図 2 参照）．Manual 実験では，与えられた検索課題（Topic）を基に，システム利用者が検索システムに入力するクエリ

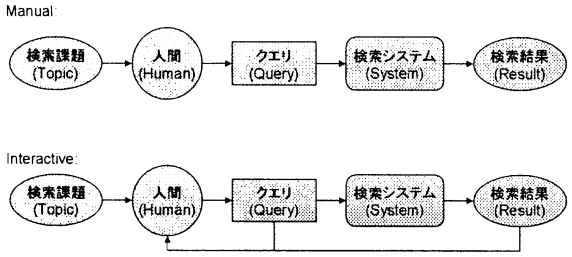

図 2　TRECVID Search タスク実験条件概要

を作成し，得られた検索結果の精度の評価を行う．一方，Interactive 実験では，Manual 実験と同様，与えられた Topic を基に，システム利用者がクエリを作成し，検索システムに入力するが，その結果得られる検索結果などを基に，システム利用者がクエリを修正し，1 度だけ再検索を行うことができる．検索の精度評価は，再検索の結果に対して行われる．Full automatic 実験では，上記の 2 条件と異なり，検索時に一切人間が介在しない条件である．

本タスクの評価基準は，Feature extraction タスクと同様，再現率・適合率・平均適合率である．また，評価のための正解データも，Feature extraction タスク同様，プーリング方式によって構築されている．

3.3.6. BBC Rushes

TRECVID 2005 での非公式タスクとして"BBC Rushes" が開催された．"BBC Rushes"とは，英国の放送局 BBC が番組制作時に撮影した映像の素材を集めたものであり，ナレーションや字幕スーパーなどといった効果がほとんど施されていない，生の素材に近い映像データである．BBC Rushes はほぼ未加工な映像データゆえ，音声認識処理を実行しても有益なテキスト情報を得ることは難しいと想像されるが，逆に，映像データから得られる視覚的・聴覚的特徴量のみに基づく映像情報検索研究への取り組みを促進させる実験データとなりうることで，TRECVID 2005 の試験的なタスクとして開催された．

本タスクでは，特に明確な課題・評価基準等は設定せず，BBC Rushes データを利用した実験やその結果などについての報告が行われた．

4. TRECVID に見る映像情報検索の動向

TRECVID で設定されているタスクのうち，映像の意味的コンテンツに無関係なカット点検出ならびに低次特徴抽出の 2 つ以外のタスクでは，音声認識結果から得られるテキスト情報を利用することが可能である．たとえば，話題分割タスクにおいては，音声認識結果から得られるテキストに基づき，TextTiling アルゴリズムなどによって話題分割点を検出する方法を採用している参加者も見られる．また，高次特徴抽出や検索といったタスクでも，検索対象映像の中からクエリに適合するショットを検出するために，クエリに含まれるキーワードを，当該映像に対する音声認識結果に含まれるテキストの中から検索する方式が，TRECVID 2004 までは主流となっており，テキスト情報検索結果に対し，映像解析技術を付加する形の実験システムが多く報告されていた．

しかし，TRECVID 2005 においては，非英語映像コンテンツの追加にともなうテキスト情報の劣化（音声認識に加え，機械翻訳処理が行われたため，テキスト文の誤りが増加）という背景もあり，映像解析をメインアルゴリズムとして，テキスト情報を付加的に利用する研究例が多数報告されている．たとえば，検索タスクにおいては，IBM Research がテキストのみ・映像のみ・テキスト＋映像の 3 条件による比較実験を行い，映像のみを利用した検索システムの精度がテキストのみの精度を上回る結果が得られたという報告を行っており，映像解析技術が急速に進歩している様子がうかがえた．

その一方，検索クエリの種別によっては，映像解析技術のみでは高精度な検索結果を得ることは困難であると考えられる．たとえば「ブッシュ大統領」など，特定の人物を検索したい場合は，汎用的な映像解析技術だけでは高精度な検索結果を得ることは難しく，音声認識結果や字幕スーパーに対する OCR 結果から得られるテキスト文の方が有効な情報と思われる．

TRECVID のそもそもの目的が映像情報検索技術の向上であるため，映像解析によって検索可能なクエリの比重が高く，したがって上記のような報告が増えてきていると考えられる．しかしながら，検索対象映像コンテンツの中からの特定人物検索など，実際に映像検索を行う際のニーズとして，テキスト情報の有用性が高いものも当然存在する．そのため，テキスト・音声（オーディオ）・映像の各モダリティに対する解析技術の効果的な統合が，実用的な映像情報検索システムの開発において有効であることはいうまでもない．TRECVID においては，音声認識結果やテキスト情報の利用は必ずしも主流ではないが，現実的なニーズを考慮した場合，今後もこれらの情報の利用方法については研究を進める必要があるだろう．

5. おわりに

本稿では，映像情報検索技術の最新動向として，主に TRECVID での取り組みについて紹介した．TRECVID におけるテストコレクションのような大規

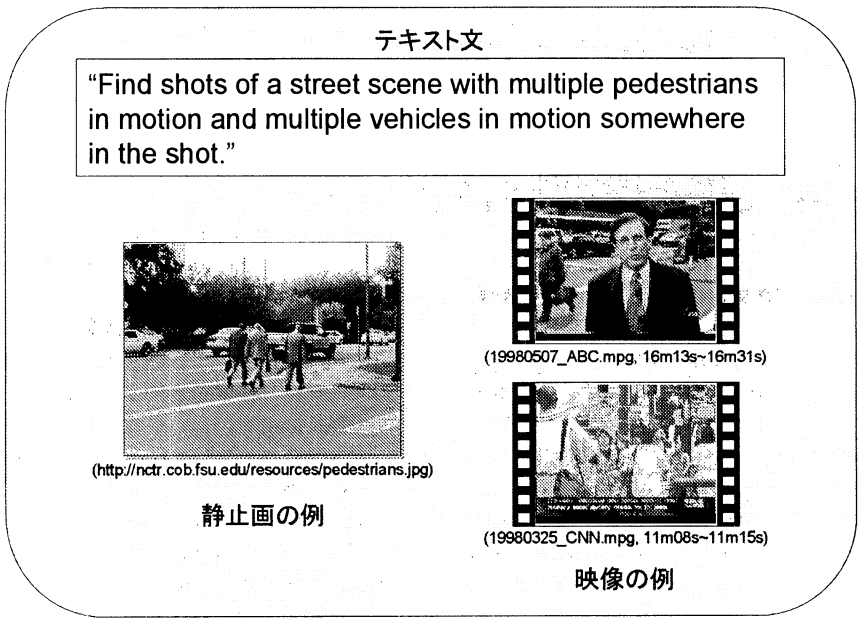

図 3 TRECVID 2004 Search タスクにおける課題例(Topic 125)

模な実験データは，映像情報検索だけでなく，さまざまな研究を進めるために有益なデータであると考えられる．したがって，国内の研究機関にも TRECVID への参加を勧める．また，TRECVID のようなテストコレクション構築の試みが国内でも本格的に開始されることを祈念する．

謝辞

日頃ご指導いただく KDDI 研究所・浅見徹所長，および中島康之執行役員に感謝します．また，TRECVID への参加においてご尽力いただいている KDDI 研究所松本一則氏，内藤正樹氏，菅野勝氏に感謝申し上げます．

文 献

[1] K. Hoashi et al: "Shot Boundary Determination on MPEG Compressed Domain and Story Segmentation Experiments for TRECVID 2004", Proceedings of TRECVID 2004, http://www-nlpir.nist.gov/projects/tvpubs/tvpapers04/kddi.pdf

[2] A. Hanjalic, "Generic Approach to Highlight Extraction from a Sports Video," IEEE ICIP 2003, Vol. 1, pp. 1-4, September 2003.など

[3] 新田，馬場口：放送型スポーツ映像の意味内容獲得のためのストーリー分割法，電子情報通信学会論文誌(D-II), Vol. J86-D-II, No. 8, pp. 1222-1233, August 2003 など

[4] Q. Huang, Z. Liu, et al., ``Automated Generation of News Content Hierarchy by Integrating Audio, Video, and Text Information," IEEE ICASSP 99, Vol. 6, pp. 3025-3028, March 1999 など

[5] S. Boykin et al: ``Improving broadcast news segmentation processing'', Proceedings of IEEE Multimedia Systems, pp 744-749, 1999.

[6] M. Sugano, M. Furuya, Y. Nakajima and H. Yanagihara, "Shot Classification and Scene Segmentation Based on MPEG Compressed Movie Analysis," IEEE PCM 2004, Vol. I, pp. 271-279, November-December 2004 など

[7] P. Over et al: "TRECVID 2005 An Introduction", Proceedings of TRECVID 2005 Workshop, http://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/tv5intro.pdf, 2005.