D-003

コンテンツ非依存特徴量に基づく動画話題分割手法
Video Story Segmentation based on Content-Independent Features

帆足 啓一郎† 菅野 勝† 松本 一則† 菅谷 史昭†
Keiichiro Hoashi Masaru Sugano Kazunori Matsumoto Fumiaki Sugaya

1. はじめに

デジタル放送の開始や，デジタルハードディスクレコーダーの普及などにより，個人でも大量のデジタル動画データを所有することが一般的になりつつある．本研究は，大量の動画データの閲覧を容易にするための重要な要素技術である「動画話題分割」，すなわち，動画を意味的な話題単位に自動的に分割する技術に関するものである．本研究では，従来手法とは異なり，あらゆる動画コンテンツに適用可能な話題分割手法を提案し，TRECVIDデータに基づく評価実験を通じその有効性を実証する．

2. 問題

ニュース番組の話題分割に関する研究は，Merlinoらの研究 [1] など，多数報告されている．しかし，これらの既存研究のほとんどは，分析対象コンテンツ，すなわちニュース番組特有の特徴量や現象に基づき話題分割を行っている．たとえば，Merlinoらは，アンカーショットの出現，アンカーからレポーターへの話者転換など，ニュース番組特有の現象を抽出し，その結果を話題分割に有効な “cue” として利用し，話題分割を行っている．こうした手法は，ニュース番組の話題分割における既存手法の主流であるといえる．しかし，この手法では，分析対象コンテンツに特化した特徴量などを基に話題分割を行うため，他のコンテンツの話題分割への適用が難しい．

3. 提案手法

本研究では，上記従来手法の問題点を勘案し，コンテンツに依存しない動画像の低レベルな特徴量のみを基に話題分割を行う手法を提案する．具体的には，抽出した特徴量を基に，動画像の各ショットをベクトル化し，得られたショットベクトルを入力とした「話題分割点識別器」を構築することにより，話題分割点が出現するショットを識別する．本手法は，話題分割の対象となるコンテンツに対する事前分析を必要としないため，従来手法と比較して効率的な手法である．また，あらゆるコンテンツへの適用が可能であるため，従来手法と比較して汎用的な手法である．以下，本手法で抽出する特徴量ならびに話題分割手法に関する詳細説明を示す．

3.1 特徴量

本手法で動画像から抽出する特徴量の一覧を表1に示す．このうち，オーディオクラスとは，各オーディオフレームのオーディオ情報を silence, speech, music, noise の4クラスのいずれかに分類し [2]，対象ショット内での各クラスに属するフレームの出現率を算出することにより得る．オーディオクラスを含め，表1に示されている全ての特徴量は，MPEG から直接抽出することが可能であり，かつ動画のコンテンツ種別に非依存であることから，あらゆるコンテンツ種別の動画からの抽出が可能な汎用的特徴量であるといえる．

3.2 SVMによる話題分割

前節で示された特徴量を基に構築されたショットのベクトルを利用し，話題分割点が出現するショットを抽出するための識別器をサポートベクターマシン (SVM)[3] で構築する．まず，学習データから話題分割点が出現するショットならびに出現しないショットを抽出し，話題分割点出現ショットを正例，それ以外のショットを負例として学習を行う．この学習によって構築された識別器に，分析対象動画のショットベクトルを入力することにより，話題分割点が出現するショットを抽出し，該当ショットの開始点を話題分割点とする．

また，上記手法の他，連続する N 件のショットを1つのベクトルとして表し，話題分割点が出現する連続ショットを識別する手法も提案する．連続ショットを表すベクトルは，該当の N 件のショットベクトルを接続することにより生成する．すなわち，各ショットベクトルが k 次元だとすると，連続ショットベクトルの次元数は $k \times N$ となる．

4. 評価実験

4.1 実験データ

提案手法の有効性を確認するため，TRECVID 2003[4] の話題分割 (story segmentation) 実験用データに基づく評価実験を行う．この実験データは，約4ケ月間に放送されたニュース番組 (ABC World News Tonight, CNN Headlines, 約120時間) より構成されている．実験データ内の全ての動画に対し，個々の話題の開始時刻の情報が付与されており，本実験ではこの正解データを基に学習および評価を行う．本実験では，前半2ケ月間に放送された動画を学習データとし，後半2ケ月間の動画に対する話題分割精度を評価する．また，本実験では実験データに対するショット検出は行わず，TRECVIDから提供されている common shot boundary 情報を利用する．

4.2 結果

本実験は，TRECVIDが定めた評価基準にしたがい，正解話題分割点の前後5秒以内に検出された話題分割点を正解とみなし，適合率 (Precision) と再現率 (Recall) を算出する．

表2に，ABC と CNN のそれぞれに対する提案手法の適合率，再現率ならびに F-measure(F1) を示す．なお，表中の “1-shot”, “N=2”, “N=3” は，それぞれ単独ショットならびに N=2, 3 とした連続ショットを SVM への入力とした場合の結果を表す．

表2より，全般的に ABC に対する話題分割の方が高いことが明らかである．CNN では，当日のニュースを，レ

† （株）KDDI 研究所，KDDI R&D Laboratories, Inc.

表 1: 動画像から抽出する特徴量一覧

オーディオ	動き	色	時系列
- 平均 RMS	- 動きベクトル水平成分	- 開始フレーム色配置	- ショット長
- 開始 n フレーム平均 RMS	- 動きベクトル垂直成分	- 中央フレーム色配置	- ショット密度
- オーディオクラス	- 動きベクトル大きさ	- 最終フレーム色配置	
(silence, speech, music, noise)	- 動き強度	(6 * Y, 3 * Cb, 3 * Cr)	

表 2: 提案手法による適合率と再現率

	ABC			CNN		
Method	Prec	Rec	F1	Prec	Rec	F1
1-shot	0.603	0.590	0.596	0.516	0.530	0.523
$N = 2$	0.615	0.597	0.606	0.540	0.526	0.533
$N = 3$	0.615	0.594	0.604	0.556	0.536	0.546

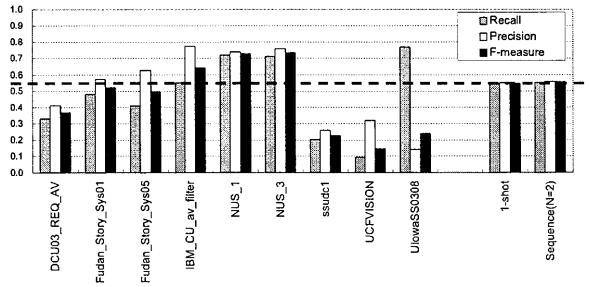

図 1: TRECVID 2003 公式実験結果と提案手法の比較

ポート画像を交えてダイジェストで紹介するなど，ABC と比較して番組の構成が複雑であり，したがって話題分割が難しく，精度が低下したものと推測される．また，単独ショットと連続ショットの両手法を比較した場合，ABC, CNN ともに連続ショットの方が若干高い精度が得られているが，有意な差ではない．

4.3 TRECVID 結果との比較

提案手法の話題分割精度を検証するため，TRECVID 2003 の話題分割タスクに提出された実験結果との比較を行う．図 1 に，提案手法 (1-shot, N=2) と，TRECVID に提出された公式実験結果のうち，“Audio + Video” 条件 (=音声認識結果などのテキスト情報を利用しない条件) の実験結果の適合率，再現率，F-measure を示す．

図 1 から明らかな通り，TRECVID 公式実験結果のうち，提案手法を上回る F-measure を達成したのは `NUS_1,3` と `IBM_CU_av_filter` の 3 件である．`NUS_1,3` では，ABC と CNN のそれぞれで事前に定義されたカテゴリ (ABC:12, CNN:17 カテゴリ) へのショット自動分類結果に基づき，話題分割点の検出を行っている [5]．また，`IBM_CU_av_filter` は，アンカーショット，CM の検出結果などを含めた多数の特徴量を基に話題分割を行っている [6]．すなわち，これらの手法では，高い精度が得られているものの，それぞれ分析対象コンテンツに特化した特徴量などを基に話題分割を行っていることが明らかである．

ここで，番組の構成が複雑な CNN に対する話題分割の再現率を比較すると，提案手法では，表 2 の通り 0.526 ～0.536 の再現率が得られているのに対し，`NUS_1,3`, `IBM_CU_av_filter` の再現率はそれぞれ 0.734, 0.733, 0.494 である．`NUS_1,3` では，前述の通り，CNN に特化したカテゴリへのショット分類結果を利用して話題分割を行っているため，高い再現率が得られている．しかし，`IBM_CU_av_filter` は，ABC の再現率 (0.728) に対し，CNN の再現率が大きく下回っている．この結果より，`IBM_CU_av_filter` は，CNN で表れるような多様な話題分割には対応できず，したがって検出されなかった話題分割点が増加したことがわかる．これに対し，提案手法では，汎用的特徴量を利用しているにも関らず，ABC, CNN で比較的安定した再現率が得られており，CNN では `IBM_CU_av_filter` も上回る再現率が得られていることから，多様なコンテンツに対するロバスト性が高いことがわかる．さらに，上記以外の全ての TRECVID 実験結果に対しては，提案手法の話題分割精度の方が優位であることが図 1 より明らかである．以上の実験結果より，提案手法はコンテンツ特有の特徴量を一切利用していないにも関らず，高い精度での話題分割が可能であり，かつ汎用的な手法であることが実証された．

5. まとめ

本研究では，あらゆる動画コンテンツから抽出できる汎用的な特徴量に基づく動画像話題分割手法を提案した．TRECVID データに対する評価実験の結果，多様なコンテンツに対し，高い精度での話題分割が得られていることが実証され，提案手法の有効性が確認された．

参考文献

[1] A. Merlino et al: Broadcast news navigation using story segmentation, Proceedings of ACM Multimedia, pp 381-391, 1997.

[2] Y. Nakajima et al: A fast audio classification from MPEG coded data, Proceedings of ICASSP '99, Vol 6, pp 3005-3008, 1999.

[3] V. Vapnik: Statistical learning theory, A Wiley-Interscience Publication, 1998.

[4] A. Smeaton et al: TRECVID 2003 - An Introduction, Proc. of TRECVID 2003, `http://www-nlpir.nist.gov/projects/tvpubs/papers/tv3intropaper.pdf`, 2003.

[5] L. Chaisorn et al: Two-level multi-modal framework for news story segmentation of large video corpus, Proc. of TRECVID 2003, `http://www-nlpir.nist.gov/projects/tvpubs/papers/nus.final.paper.pdf`, 2003.

[6] A. Amir et al: IBM Research TRECVID-2003 video retrieval system, Proc. of TRECVID 2003, `http://www-nlpir.nist.gov/projects/tvpubs/papers/ibm.smith.paper.final2.pdf`, 2003.