

ニュース映像中のモノログシーン検出による発言集の自動作成

関岡 直城[†] 高橋 友和[†] 井手 一郎^{†,††} 村瀬 洋[†]

[†] 名古屋大学大学院情報科学研究科 〒464-8603 愛知県名古屋市千種区不老町

^{††} 国立情報学研究所 〒101-8430 東京都千代田区一ツ橋 2-1-2

E-mail: [†]{nsekioka,ttakahashi,ide,murase}@murase.m.is.nagoya-u.ac.jp

あらまし 大容量 HDD の普及による映像の大量蓄積に伴い、ユーザの所望するシーンを自動検出し、提示する技術が求められている。これを受けて本報告では、ニュース映像における演説やインタビューなどのモノログシーン検出手法の提案、及びそれらを用いたニュース発言集の自動作成を試みる。モノログシーン検出では、画像・音声・テキスト情報を用いた統合メディア処理により既存の手法を効果的に組み合わせることで、入力映像のみを情報源とした自動検出手法を提案する。発言集の作成には、クローズドキャプション中の人名による、各モノログシーンの名前の対応付けにより、登場人物の発言集の作成を試みた。実験の結果、モノログ数の最も多い上位 3 名の登場人物に対して、平均再現率 37 %、平均適合率 52 % の発言集の正答率が得られた。

キーワード ニュース映像、モノログシーン

Automatic Creation of a Speech Archive by Monologue Scene Detection in News Videos

Naoki SEKIOKA[†], Tomokazu TAKAHASHI[†], Ichiro IDE^{†,††}, and Hiroshi MURASE[†]

[†] Graduate School of Information Science, Nagoya University

Furo-cho, Chikusa-ku, Nagoya, Aichi, 464-8603, Japan

^{††} National Institute of Informatics

2-1-2 Hitotubashi, Chiyoda-ku, Tokyo, 101-8430, Japan

E-mail: [†]{nsekioka,ttakahashi,ide,murase}@murase.m.is.nagoya-u.ac.jp

Abstract According to accumulation of large amount of videos by the spread of large capacity HDDs, automatic detection and presentation of the scene that a user desires are requested. In this report, we propose a method of detecting monologue scenes in news videos, such as speeches or interviews, and of creating a news speech archive. In the monologue scene detection, existing techniques are effectively combined by media integration using image, audio, and text information. By this process, we propose a method of automatic monologue scene detection using only input videos as the source. To create a news speech archive, monologue scenes named by person names that appear in the closed caption text are used. As a result of experiments, we obtained recall of 37% and precision of 52% as a percentage of correct answers for persons with the top three large speech clusters.

Key words news video, monologue scene

1. はじめに

近年、大容量 HDD の普及により映像を大量に蓄積して利用する機会が増えつつある。これに伴い、ユーザの所望するシーンを映像中から自動で検出し、提示する技術が求められている。特にニュース映像は人間の活動の記録であり、その内容の重要性や資料的価値の点からも、このような技術の需要が高い。

本報告では、このようなニュース映像から演説やインタビュー

などのモノログシーンを検出し、それらを用いたニュース発言集の自動作成を試みる。作成された発言集からユーザの関心のある過去数年間の発言を閲覧したり、番組製作者側の編集作業の支援などへの様々な利用が考えられる。

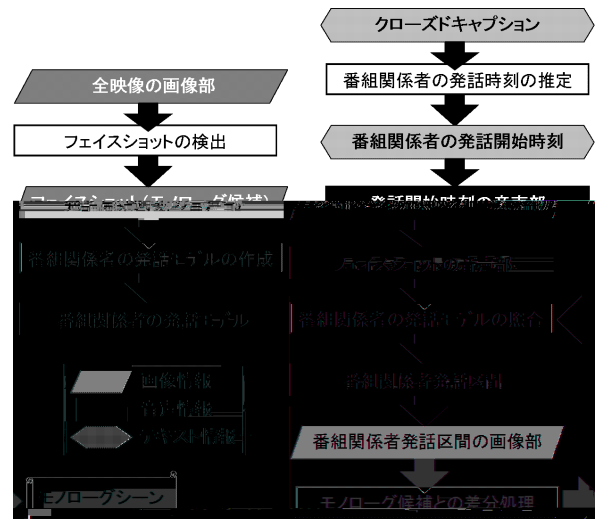
本報告ではまず、2. でモノログシーン検出の従来研究についてまとめ、3. で提案手法によるモノログシーンの検出過程の紹介とその評価実験を行う。4. ではモノログシーンの名前の対応付けによるニュース発言集の作成手法の紹介と本報告で

2. 従来研究

これらの従来研究では共通して、Web ページ等の入力映像以外の外部情報や事前知識を情報源として利用している。しかし映像によっては、これらの情報を放送局などから取得できない可能性や、事前に人手で学習データを収集する際にかかるコストなどの問題が挙げられる。これをうけて本研究では、画像・音声・テキストを用いた統合メディア処理により、既存の手法を効果的に組み合わせることで、これらの問題に対処する。特に、テキスト情報を手がかりに番組関係者の発話モデルを作成することで、入力映像のみを情報源とするモノローグシーンの自動検出手法を提案する。

3.1 処理の概要

を得て、音声情報によりそれらの人物の発話モデルを作成する．これにより、番組関係者の発話区間を除去することでモノログシーンを自動検出する．図 2 にモノログシーンの検出過程を示す．



3.2 モノログシーン候補の検出

3.2.1 フェイスショットの検出手順

3.2.1 フェイスショットの検出手順

はじめに前処理として、RGB ヒストグラムによるカット検出を行い、各ショットから顔領域を検出する．ここでは、現時点で最も性能の良いオブジェクト検出手法とされている Haar-like 特徴による高速オブジェクト検出手法 [4] [5] を適用した．Intel 社がオープンソースで公開しているコンピュータビジョン関連のライブラリ (Open Source Computer Vision Library, OpenCV) [6] にこのアルゴリズムが実装されているため、本研究ではこれを使用した．演説やインタビューなどを撮影する場合、カメラマンは被写体となる話者をフレーム内に収め、さらに話者の表情を的確に捉えようとする．そのため、モノロースcene中にみられる顔の多くは、比較的大きな正面顔かつ画面の中央付近に現れる傾向がある．そこで、フェイスショットを検出する際の条件として、以下の3つを設定した．なお、正面顔の向きは、OpenCV に付属する正面顔検出器によって検出できる範囲内とする．

- 顔の向きは正面。
- 顔の大きさは画像全体の 8 % 以上。
- 顔の重心は図 3 の灰色部分内。

3.2.2 フェイスショットの検出実験

— 2 —

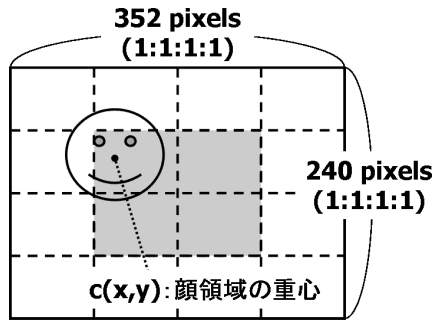


図 3 顔領域検出の位置条件

子などのオクルージョンによる影響が主な原因であった。しかし、検出されたフェイスショットには、モノログではないシーンが数多くみられた。その内訳として、アナウンサやレポートの顔のクローズアップショット、映像中の人物と発話者が一致していないショットが挙げられる。後者は、番組関係者が映像中の人物に関する情報や、その発言内容に間接的に言及する場合であり、ニュース映像に頻繁にみられるシーンである。

表 1 ニュース映像の仕様

ニュース映像	「NHK ニュース 7」
フレームレート	30 (frame/sec)
画素数	352 × 240 (pixel)
総時間	30 (min) × 31 本 (1ヶ月分)
期間	2004 年 1 月 1 日 ~ 1 月 31 日

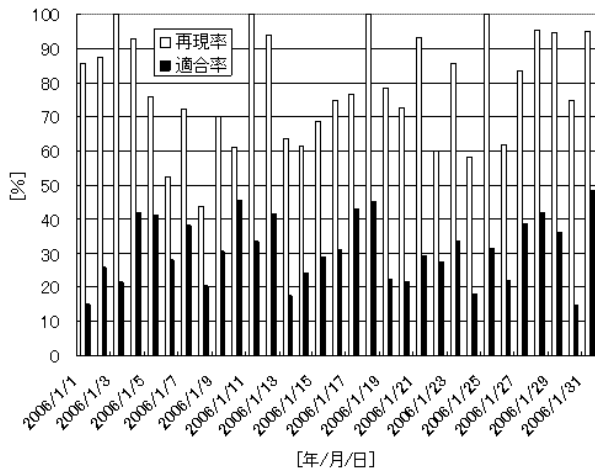


図 4 モノログシーン候補 (フェイスショット) の検出結果

3.3 モノログシーン候補の絞り込み

フェイスショットに含まれるモノログではない区間に共通する傾向として、音声画像中の人物ではなく、番組関係者である点が挙げられる。図 5 のように、フェイスショットには、番組関係者の発話区間や無音区間が含まれている。そこで提案手法では、テキスト情報と音声情報の併用により、各番組関係者の発話モデルをそれぞれ動的に作成し、モノログシーン候補 (フェイスショット) の音声部とモデル照合する。これにより、音声が番組関係者である発話区間を除去し、正確にモノログシーンを検出する。

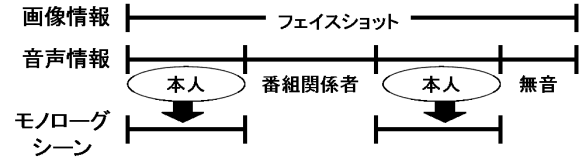


図 5 フェイスショットとモノログシーンの関係

3.3.1 音声学習サンプルの自動収集

クローズドキャプション (Closed Caption, CC) とも呼ばれる文字放送字幕テキストを利用して、各番組関係者の発話モデル作成のために学習サンプルの収集を行う。CC には、音声の書き下しテキスト (発話文) とその音声が発せられた時刻が記述されているため、各番組関係者の発話文を推定することができれば、その発話時刻を手がかりに入力映像から直接、音声学習サンプルを収集することができる。なお、実際に発せられた音声と CC に記述された発話文及び発話時刻には若干のズレが生じているが、本研究では、事前にそれらの同期処理が施された CC を使用している。

アナウンサはニュース映像の放送開始後、最初に発話する人物と考えられるため、CC の冒頭に記述された発話文をアナウンサのものとして仮定する。また、レポートの発話文の推定には、その直前のアナウンサの発話文の内容に注目する。評価データとは別の 20 日分のニュース映像において、レポートが発話する直前の文に含まれる語の出現頻度を調べたところ、「記者」、「取材」、「中継」の 3 つの語が多く含まれることを確認した。ここでは、この 3 つの語に対するキーワード検索と文法的特徴を組み合わせた以下の 2 つの条件を満たす発話文を、レポートが発話する直前のアナウンサのものとして推定する。

- 条件 1. 呼びかけ：文の末尾が「固有名詞 + “さん”」。
- 条件 2. 文が過去形でない and “記者”、“取材”、“中継”のいずれかを含む。

条件 1 はアナウンサによるレポートへの呼びかけを想定している。また、条件 2 はアナウンサとレポートとのやりとりがリアルタイムで行われることを想定している。この 2 つの条件のいずれかを満たす場合に限り、学習サンプルとしての収集の対象とみなす。なお、品詞分解と時制判定のために、形態素解析システム「JUMAN ver.5.1」[8]と構文解析システム「KNP ver.2.0」[9]を使用した。

3.3.2 番組関係者の発話区間の除去

入力映像から収集された各番組関係者の音声学習データに対し、FFT パワーによる有音区間の検出後、発話モデルをそれぞれ作成し、モノログシーン候補 (フェイスショット) の音声部とベクトル量子化歪によるモデル照合を行う。発話モデルの作成とモデル照合の処理過程を図 6 に、分析条件を表 2 に示す。

3.4 モノログシーンの検出実験

3.3 節の絞り込み処理により、モノログシーンの検出実験を行った。評価データは表 1 に示したニュース映像と同じで、3.2.2 節の実験結果として求められたフェイスショットを絞り込みの対象とした。実験の結果を図 8 に示す。全体では、平均再現

必要となる．ニュース映像のトピック分割に関する研究として，Ide ら [12] は，長時間規模のニュース映像を対象とした意味内容に基づく知的構造化手法を提案し，テキスト特徴を用いたトピック分割・追跡手法を実現している．ここでは，これらの2つの手法により解析され，人物名詞とトピック境界がタグ付けされたCCから，名前候補の推定を行う．

なお，ひとつのモノログシーンには複数の名前候補が対応付けされると考えられるため，スコアにより各候補を順序付ける．スコアは，主語になる人物名詞 (+5)，語の出現頻度 (+出現回数 × 2) と設定した．

4.1.2 名前の対応付け実験

これまでの処理により，モノログシーンへの名前の対応付け実験を行った．本手法では，原理的にCCに名前の記述がないモノログシーンには名前を対応付けすることが不可能なため，本実験では，CCに名前の記述があるモノログシーンのみを対象とする．評価データは表1に示したニュース映像と同じで，名前の対応付けの正誤判定は，スコアの上位3位までに正しい名前が対応付けられれば，成功とみなした．実験の結果，平均正答率約62%と，十分な精度には達していないものの，テキスト情報のみを用いた結果としては，比較的良好と考える．実験結果から，本手法ではモノログシーンの名前候補として，頻繁に登場する人物名，つまり，話題の中心人物のスコアが全体的に高くなる傾向がみられた．

図8 モノログシーンの名前の対応付け結果

4.2 モノログシーンのクラスタリング

登場人物のニュース発言集を作成するために，モノログシーンに付与された名前候補とそのスコアを手がかりとしたクラスタリングを行う．具体的には，各モノログシーンの名前候補のスコアを特徴ベクトルとして，最近傍 (Nearest Neighbor, NN) 法によりクラスタリングする．各クラスターの代表ベクトルの更新は，新規にクラスターに属するモノログシーンのスコアとそれ以前の代表ベクトルのスコアとの比較により，その上位3位の名前候補を新規クラスターの代表ベクトルとする．また，各モノログシーンに付与された名前候補の論理積 (AND) をとり，共通する名前候補のスコアを乗算した総和をモノログ間類似度として算出する．類似度の算出例を図9に示す．

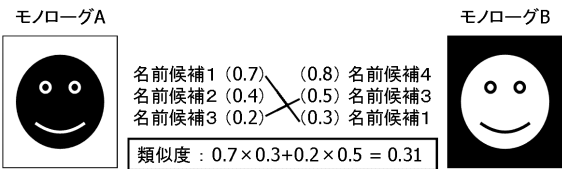


図9 類似度の算出例

表3 ニュース発言集の作成結果

発言集名		小泉 総理大臣	ブッシュ 大統領	福田 官房長官
全モノログ数		49	9	11
収集されたモノログ数		38	13	11
正分類モノログ数		10	5	4
誤分類	非モノログ数	18	8	7
	名前の対応付けの 分類誤りモノログ数	10	0	0
再現率 [%]		20	56	36
適合率 [%]		36	62	57

4.3 ニュース発言集の作成実験

これまでの実験に用いてきた評価データ (1ヶ月分のニュース映像) に，さらに1ヶ月分のニュース映像を加えた評価データに対して，モノログシーンの検出，及びモノログシーンの名前の対応付けを行い，ニュース発言集の作成実験を行った．モノログ数の最も多かった上位3位の登場人物に対する実験結果を表3に示す．その結果，平均再現率37%，平均適合率52%の発言集の完成度が得られた．これは，モノログシーンの検出からモノログシーンのクラスタリングまでの完全自動処理の結果であり，表中の全モノログ数にはCCに名前の記述のないモノログ数も含まれている．誤分類の原因としては，モノログシーンの誤検出の占める割合が多く，前述した録画中継におけるレポーターの影響が大きい．他の原因は，モノログシーンの名前の対応付け誤りによるものである．表3の結果をみると，「小泉総理大臣」の対応付け誤りのモノログ数が他に比べ，極端に多いことが分かる．前述の通り，頻繁に登場する話題の中心人物の名前を候補として誤って推定してしまう影響がこの結果からもみられた．実験により作成されたブッシュ大統領のニュース発言集について，図10に正分類の結果を，図11に2つの原因による誤分類の結果を示す．

5. ま と め

本研究では，ニュース映像中のモノログシーンの検出，及び登場人物のニュース発言集の作成を試みた．モノログシーンの検出では，画像・音声・テキスト情報を用いた統合メディア処理により，既存の手法を効果的に組み合わせることで，入力映像のみを情報源とするモノログシーンの自動検出手法を提案した．

モノログシーン候補の検出では，およそその候補を検出することができたが，顔の向きや帽子などのオクルージョンによる検出漏れが少数みられたため，横顔への対応などが今後必要となる．

モノログシーン候補の絞込みには、テキスト情報を手がかりに作成した番組関係者の発話モデルによるモデル照合を行うことで、番組関係者の発話区間を除去した。その結果、アナウンサ発話区間は良好に除去できたものの、録画中継におけるレポータの発話区間の除去が不十分であった。これは、アナウンサとのやり取りが生じないためであり、CC 中から文法的特徴が抽出できず、発話モデルを作成することができなかったためである。このように、突発的なレポータ発話に関しては、テキスト情報のみでの対処は困難なため、画像情報によるレポータシーンの検出などが今後の課題として挙げられる。

モノログシーンの名前の対応付けでは、CC に出現する人物名を利用し、名前の対応付けを行った。テキスト情報のみを用いた結果としては比較的良好であったため、オープンキャプションの認識結果との併用などにより、さらなる改善が期待できる。

登場人物によるニュース発言集の作成では、名前候補とスコアを手がかりとした最近傍法によるモノログシーンのクラスタリングを行った。クラスタリングの精度は、モノログシーンの名前の対応付けの精度に大きく依存するため、これとは独立した手がかり、例えば、顔や音声によるクラスタリング手法の検討が必要不可欠であると考える。

実際にこれらの手法を用いて、発言集を作成した。精度としてはまだ十分な結果は得られていないものの、発言集作成までの個々の手法についての上記の改善手法により、さらなる精度の向上が期待できる。

謝辞

本研究の一部は 21 世紀 COE プログラム、科学研究費補助金による。また、実験のデータとして使用したニュース映像を提供して頂いた国立情報学研究所に感謝する。

文 献

- [1] <http://www-nlpir.nist.gov/projects/trecvid/>
- [2] A. Hauptmann, R.V. Baron, M.-Y. Chen, M. Christel, P. Duygulu, C. Huang, R. Jin, W.-H. Lin, T. Ng, N. Moraveji, N. Papemick, C.G.M. Snoek, G. Tzanetakis, J. Yang, R. Yan, and H.D. Wactlar, “Informedia at TRECVID 2003: Analyzing and Searching Broadcast News Video”, Online Proc. TRECVID, Nov. 2003.
- [3] A. Amir, M. Berg, S. Chang, W. Hsu, G. Iyengar, C. Lin, M. Naphade, A. Natsev, C. Neti, H. Nock, J.R. Smith, B. Tseng, Y. Wu, and D. Zhang, “IBM Research TRECVID-2003 Video Retrieval System”, Online Proc. TRECVID, Nov. 2003.
- [4] R. Lienhart and J. Maydt, “An Extended Set of Haar-like Features for Rapid Object Detection”, Proc. IEEE 2002 International Conference on Image Processing, vol.1, pp.900-903, Sep. 2002.
- [5] A. Kuranov, R. Lienhart, and V. Pisarevsky, “An Empirical Analysis of Boosting Algorithms for Rapid Objects with an Extended Set of Haar-Like Features”, Intel Tech. Rep. MRL-TR-July02-01, July 2002.
- [6] Open Source Computer Vision Library, <http://www.intel.com/technology/computing/opencv/>
- [7] N. Katayama, H. Mo, I. Ide, and S. Satoh, “Mining Large-scale Broadcast Video Archives Towards Inter-video Structuring”, Proc. 5th Pacific Rim Conf. on Multimedia Part II, Lecture Notes in Computer Science, Springer-Verlag, vol.3332, pp.489-496, Dec. 2004.
- [8] 日本語形態素解析システム JUMAN ver.5.1, <http://www.kc.t.u-tokyo.ac.jp/nl-resource/juman.html>
- [9] 日本語構文解析システム KNP ver.2.0, <http://www.kc.t.u-tokyo.ac.jp/nl-resource/knp.html>
- [10] S. Satoh, Y. Nakamura, and T. Kanade, “Name-It: Naming and Detecting Faces in News Videos”, IEEE Multimedia, vol.6, no.1, pp.22-35, Jan-March 1999.
- [11] 井手一郎, 浜田玲子, 坂井修一, 田中英彦, “テレビニュース字幕の語義属性解析のための辞書作成”, 電子情報通信学会論文誌 (D-II), vol.J85-D-II, no.7, pp.1201-1210, July 2002.
- [12] I. Ide, H. Mo, N. Katayama, and S. Satoh, “Topic-based Inter-video Structuring of a Large-scale News Video Corpus”, Proc. 2003 IEEE Intl. Conf. on Multimedia and Expo, vol.3, pp.305-308, July 2003.

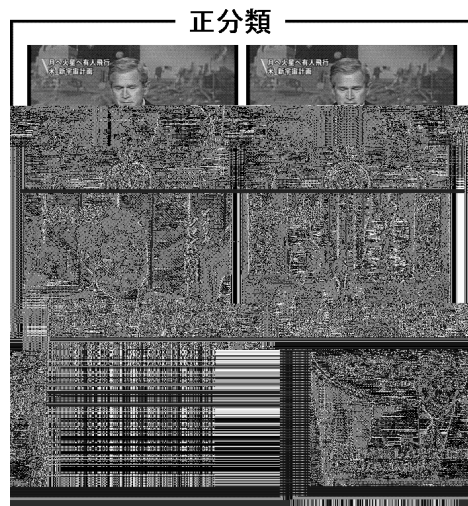


図 10 ブッシュ大統領の発言集 (正分類)

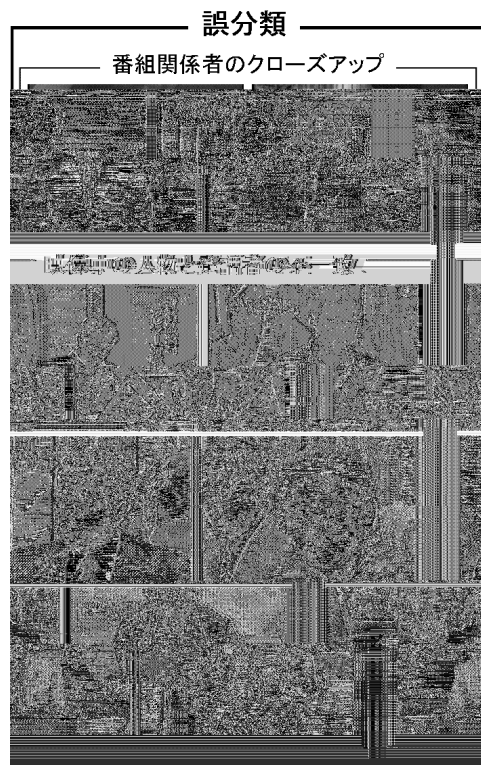


図 11 ブッシュ大統領の発言集 (誤分類)