ニュース映像における人物領域の分離による場面推定

Scene identification by character region segmentation in news video

井手一郎†[1]　浜田玲子‡　坂井修一‡　田中英彦‡
† 国立情報学研究所[2]　‡ 東京大学大学院工学系研究科[3]
Ichiro Ide†　Reiko Hamada‡　Shuichi Sakai‡　Hidehiko Tanaka‡
†National Institute of Informatics　‡Graduate School of Engineering, The University of Tokyo

概要: ニュース映像の再利用や検索の要請が高まるなか，筆者らは映像中の音声やテキストを用いて，画像内容との対応を考慮した索引付けを行う機構を提案している．本稿では，そのなかの画像内容解析に必要な場面推定手法の提案と，実際の映像に適用した結果を紹介する．提案手法の特徴は，ニュース映像に人物像が多く出現する傾向をふまえ，推定の際に人物領域を分離して背景領域の画像特徴量のみを参照する点にある．この結果，分離しない時と比べて，7%から 8%の性能向上が見られ，提案手法の有効性が示された．

Abstract: We have been proposing an automatic news video indexing system that considers the correspondence between image and keyword. In this paper, we will introduce the scene identification method required for the image content analysis portion of the system, and also the result of its application to actual video. The method is characteristic that it takes advantage of frequent human figure existence in news video, by segmenting character region from background when identifying a scene. As a result, the method showed 7% to 8% of improvment in identification ability.

1　はじめに

映像資源の効率的な再利用や検索への要請が高まるなか，Informedia プロジェクト [12] をはじめ，映像データベースを構築するための自動索引付けに関する研究が盛んに行われている．しかし，既存の一般的な自動索引付け手法は，主に映像中の音声やテキストから抽出される重要語を取捨選択せずに利用しており，画像内容との対応を考慮したものは，人物の顔と名前の対応を考慮した手法 [11] などに散見されるに過ぎない．

筆者らはこれらの要請と問題点を考慮し，映像中のテキストから得られる情報を利用しつつ，画像内容との対応を考慮した自動索引付け機構を提案している．索引付け対象としては，将来の再利用や検索という点で利用価値が高いニュース映像を扱う．具体的には，図 1 に示すように，画像とテキストの各々から得られる情報について，いわゆる 4W (「いつ (When)」…時相,「どこで (Where)」…場面,「誰が (Who)」…人物,「何を (What)」…行為) と呼ばれる属性毎の内容

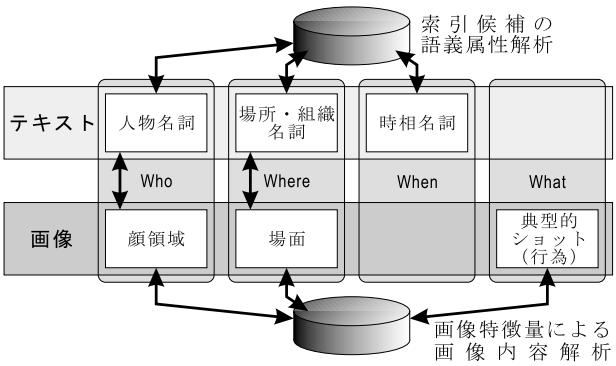

図 1　映像内容との対応を考慮した索引付け機構の概要

解析を行い，各々の対応に基づく索引付けを行うことを目指している．ニュース映像に対して，これらの属性による検索を想定することは自然であり，このような限定は妥当であると考える．本稿では，この機構のなかの画像内容解析部について取り上げ，ニュース映像中に人物像が頻出することを考慮し，人物領域の分離による内容解析，特に背景領域からの場面推定手法

[1] 電子メール：ide@nii.ac.jp
[2] 〒 101-8430 東京都千代田区一ツ橋 2-1-2
[3] 〒 113-8656 東京都文京区本郷 7-3-1

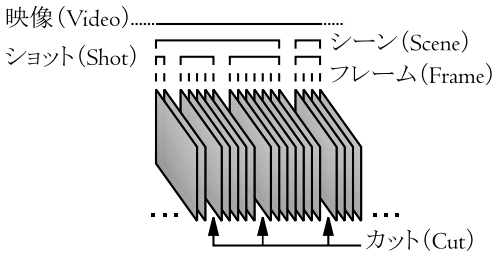

図 2　映像の構成と用語の定義

を提案する．また，[4] において報告した予備実験の規模を拡大した評価実験により，提案手法の有効性を評価する．

図 2 に映像の構成と用語の定義を示す．映像はフレームと呼ばれる静止画像から構成され，画像的に連続なフレーム群をショットと呼ぶ．また，ショット間の不連続点をカットと呼ぶ．図 1 に示した機構ではショット単位での索引付けを目標としている．また，画像的・内容的に連続するショット群をシーンと呼び，ニュース映像において，内容的なショットは話題に対応する．

2　人物領域分離による画像内容解析

2.1　手法の概要

従来より，領域分割による画像内容の解析は行われてきたが，一般性の維持を考慮して，伸縮は行われるものの，矩形領域への定型的な分割が主流だった．しかしニュース映像は，主に人間社会に生起する事象に関する情報提供手段であるため，一般に人物像が大きく写ることが多い．そのため，ニュース映像の画像内容解析を行う際には，人物像の存在を考慮することが重要と考えられる．また，そのような性質から，人物という特定の事物の存在を前提とした処理を導入しても，一般性が大きく損なわれることはない．このような人物像に注目した領域分割により，従来の画像全体や定型的な矩形領域への分割による画像内容解析の際に排除しきれなかった，前景に出現する人物像による画像特徴量の変動の影響を排除し，より正確な解析が可能になることが期待される．

人物像の存在，位置，人数などを基準とした，典型的ショット分類によるニュース映像の大雑把な内容推定手法も存在し [2, 7]，図 1 の機構でも「行為 (What)」の推定に採用している．しかし，このような画像内容解析は，あくまでも画像全体が対象であり，4W のうちでも特に,「人物 (Who)」と「場面 (Where)」の解析を行うためには，画像中の人物領域 (人物像) と

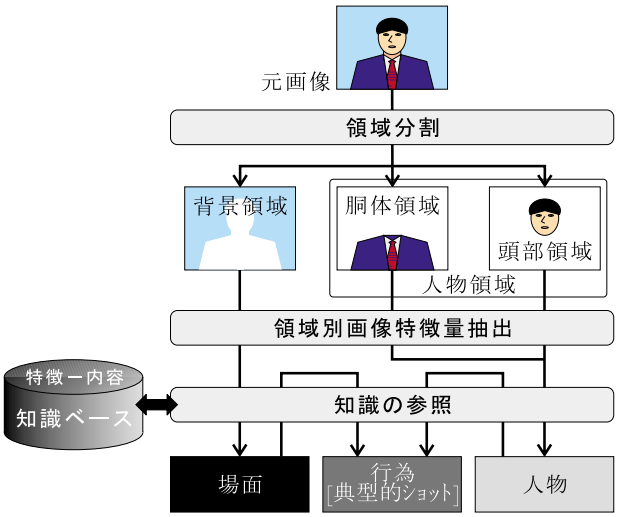

図 3　人物領域分割による画像内容解析

背景領域を分離して個別に解析する必要がある．また，画像全体を対象とする際にも，これらを個別に把握することにより，より正確な内容解析が可能になると思われる．

そこで本研究では，図 3 に示すように，顔領域の大きさや向きなどが一定条件を満たすような人物領域 (顔領域を含む頭部領域と胴体領域からなる) が画像中に存在する場合，人物領域とそれ以外の背景領域とに領域を分割し，各々について内容解析を行う．

このうち,「人物 (Who)」に関しては，顔認識に関する研究が盛んに行われているうえ，画像中の人物の顔領域を抽出し，主音声を書き下したテキスト (クローズドキャプション) 中の人物名と対応付ける研究も既に行われている [11]．また,「行為 (What)」に関しても，前述のように大雑把な推定手法は存在する [2, 7]．そこで，ここでは「場面 (Where)」に関する画像内容を考慮した索引付けを実現する際に必要となる画像処理部の機能として，人物領域を分離した背景領域の画像特徴量に基づく場面推定手法の紹介と，実際のニュース映像を用いた実験による有効性の評価結果を示す．

2.2　人物領域と背景領域の分割

前節で述べたような画像内容の解析を行うためには，人物領域と背景領域を分割する必要がある．

そこで，ニュース画像中の人物は,(1) 良好な照明条件下で,(2) 正面から撮影されることが多い，という仮定に基づき，現在の技術でも比較的実現が容易な顔領域抽出を利用し，図 4 に示すような，顔領域を基準としたテンプレートにより人物領域を決定する簡便

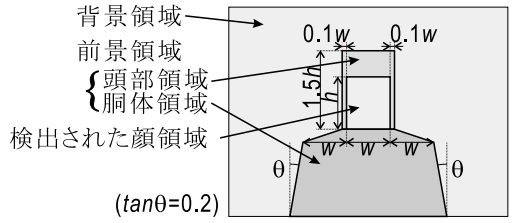

図 4　顔領域を基準とした人物領域抽出テンプレート

な手法を適用してみた．顔領域抽出には，CMU において開発されたニューラルネットワークによる訓練を用いた Face Detector [10] を利用した．

しかし，現実的には様々な状況下で撮影される画像において，上記の仮定は必ずしも満たされず，顔領域の正しい抽出や，正面からの撮影を前提としたテンプレートの適用ができず，以下の実験では人手により人物領域を切り出したものを用いている．ただし，スタジオのキャスタに限り，実験に用いた全事例 (231 件) について顔領域抽出もテンプレートの適用も過不足なく適切に行えたため，自動的に分割した結果を用いた．

3　背景の画像特徴による場面推定

3.1　事例に基づく画像内容推定

以上のように前景の人物領域を分離して得られる背景領域について，その画像内容，つまりどのような場面であるか，を推定する．そのためには，背景領域の特徴量と内容 (場面) との関係に関する知識が必要である．一般の画像に対しては，このような知識の獲得や記述は著しく困難である．しかし，ニュース画像においては，周期的・集中的に特定の事象に関連する話題を取り上げるため，一定期間あるいは一定量以上の事例が集まれば，頻出場面に関しては，特徴量と場面との関係に関する知識ベースを構築することは現実的であると予想される．

本研究では，このような予想のもとに，背景領域の特徴量と場面の関係に関する事例に基づく場面推定を行った．このように単純な事例に基づく推定は，本稿における実験のように比較的小規模の事例群を想定したものである．より大規模な事例群を対象とする際には，弁別能力や計算量の観点から，代表ベクトルの設定などによる他のより効率的な推定手法の導入が必要となろう．

3.2　関連研究：画像特徴に基づく内容推定

画像の特徴量と内容の関連付けを行う研究としては，初期のものとして，形容詞を中心とした印象語と画像特徴量の対応を心理実験から統計的に求める栗田らによる絵画データベースに関する研究 [6] や，木本による感性語を用いた画像検索の試み [5] などがある．しかし，これらの感性工学における試みは，ニュース画像のように具体的事象 (主に名詞) を対象とする場合とは問題点やその解決法が異なる．

従来は，詳細なモデルに基づく具体的事物の認識や，対象分野を極めて限定した特徴量と内容の関係を用いた内容推定が行われてきたに過ぎず，より一般的な内容推定を行う手法は少ない．そのようなものの一つとして森らは，百科事典中の画像と説明文から特徴量と内容の関係を獲得する手法 [9] を提案している．この手法では，まず一般的な語の共起関係により単語クラスタ空間を形成しておく．次に単語クラスタ空間中の単語間距離に基づき画像に付随する説明文間の類似度を計算し，それを反映させて構築した画像特徴量空間中で，説明文が類似した画像をクラスタ化する．これにより，問い合わせ画像に関連した説明文や語句の検索が可能になるが，百科辞典が扱う対象が一般的過ぎるため，良好な検索性能が発揮されない．

一方，孟らは適切にクラスタ化した訓練事例画像との画像特徴量の比較による，スポーツ映像を応用例としたジャンル推定手法を提案している [8]．この手法では，事前に訓練事例を分類して最適なクラスタ形成を行っておき，評価事例と各クラスタとの類似度に基づく分類を行う．手段や目的は類似しているが，画像全体に対する内容推定であり，筆者らの提案手法のように画像の構成内容に立ち入った推定は行っていない．

4　場面推定実験

以下の実験条件及び実験手順に従って，実際のニュース映像に対する場面推定実験を行った．

4.1　実験条件

4.1.1　用いた画像

実験には，2 つの異なる時期に集中的に録画した 15 分の全国版ニュース映像 20 本中の一部，817 ショットの先頭フレームの画像を用いた．場面の違いによる性能の差異をみるために，いくつかの特定の頻出場面を中心に選び，評価の際の正解判定のため (のみ) に，事

表 1　設定場面毎の事例数

場　面	人物あり	人物なし	合　計
(1) 閣議控室	22	10	32
(2) 国会議場	31	21	52
(3) 報道会見場	11	6	17
(4) 法廷	6	23	29
(5) スタジオ	231	0	231
その他	124	332	456
合　計	425	392	817

表 2　画像キャプチャの際の条件

項　目	条　件
空間解像度	横 320× 縦 240 ピクセル四方
色解像度	24bit (R, G, B 各色 8bit)
時間解像度	15 フレーム毎秒
フレーム圧縮	JPEG
ビデオ圧縮	なし (非圧縮)

前に目視により正解を定めておいた．具体的には，

1. 閣議控室 (cabinet)
2. 国会議場 (parliament)
3. 報道官会見場 (press conference)
4. 法廷 (court)

の 4 種類の国内報道に関係する頻出場面と，

5. スタジオ (studio)

を設定した．表 1 に各場面毎の事例数を，人物像の有無に分けて示す．なお，画像中に人物が 3 人まで前向きに大きく写っている場合に「人物あり」とみなして領域分割の対象とし，それ以外は「人物なし」とした．

また，画像は家庭用アナログビデオテープレコーダを用いて地上波放送を録画したものを，ビデオキャプチャボードを用いて最低限の JPEG 圧縮によりフレーム単位でディジタル化した．キャプチャの際の詳細な条件を表 2 に示す．

4.1.2　画像特徴量

様々な画像特徴量が存在するなかで，本実験では，以下の 2 通りの色に関する統計的特徴量を用いた：

1. 色出現頻度分布 (ヒストグラム)
   画像中のピクセルの色の出現確率の分布．本実験では，RGB 空間を線形に 64 分割した色ブロック毎の確率を求めたため，64 次元ベクトルとして表現される．

2. 色共起頻度分布 (コリログラム) [1]
   画像中の一定距離離れたピクセル間の色の組合せの出現確率の分布．本実験では，RGB 空間を線形に 16 分割した色ブロックの組合せ毎に距離 1〜4 の確率を求めたため,($16 \times 16 \times 4 =$) 1,024 次元ベクトルとして表現される．

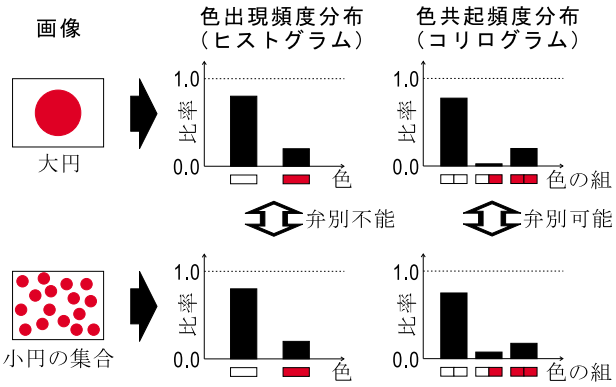

図 5　ヒストグラムとコリログラムの差異

両者には図 5 に示すような差異があり，ヒストグラムは画像全体のマクロな色彩的特徴を，コリログラムはミクロな特徴を反映する．これらの特徴量を採用した理由は，人間の目視による場面の推定の際に，色彩的な特徴が大きな手がかりになると思われることと，両特徴量の性質の差異による推定性能への影響を見るためである．

以下の実験では，各々の特徴量を独立に用いて評価しているが，より多くの場面の導入や，場面毎の弁別能力の向上のためには，他の特徴量を含めた，統合的な利用が必要である．

4.1.3　類似度評価尺度

以上の 2 つの特徴量ベクトルの各々同士を比較する類似度評価尺度として，次式で求められるような，ベクトル $\vec{F_1}, \vec{F_2}$ 間の角度 θ の余弦 (値域 : 0〜1) を用いた：

$$\vec{F_1}\text{と}\vec{F_2}\text{の類似度} \equiv \cos\theta = \frac{\vec{F_1} \cdot \vec{F_2}}{|\vec{F_1}||\vec{F_2}|} \qquad (1)$$

4.1.4　推定性能評価尺度

場面推定性能の評価にあたり，以下に定義するような類似度上位 n 位過半率を用いる．実際の場面推定の際には，最も類似している場面か，上位 n 位中の過半数を占める場面を推定結果とするため，この指標により場面推定性能が示される．

- 類似度上位 n 位過半率（$n = 1, 3, 5, 7, 9$）:
  各設定場面に属す全評価事例のうち，類似度上位 n 位中に正答が過半数を占めたものの割合．結果的に $n = 1$ の値は，最も類似している場面が正答であったものの占めた割合を示す．

4.2　実験手順

以上の条件のもとに，以下の手順により実験を行った:

1. 背景領域分離
   「人物あり」の事例について，目視により一定の向き，大きさの人物領域を切り出す．スタジオのキャスタのみ，Face Detector と図 4 のテンプレートによる自動的な分離結果を用いる．
2. 特徴量抽出
   切り出した人物領域を分離し，背景領域のヒストグラムとコリログラムを求める．人物領域が存在しなければ，画像全体を背景領域とみなす．
3. 類似度評価
   評価事例と訓練事例（全事例から評価事例を除いた余集合）の類似度の評価を行う．
4. 場面推定
   3. を全事例に対して相互に適用し，設定場面別に類似度上位 n 位過半率を集計する．

なお，図 6 に示すように，2. から 4. の処理において，人物領域分割の効果を見るために,「人物あり」の事例に対しても，人物領域の分離を行わずに，画像全体での比較も併せて行った．

このようにして，以下の 3 通りの類似度評価条件，2 通りの画像特徴量について実験を行った：

- 類似度評価
  - Without segmentation
    人物領域分割を行わない場合（従来手法）．
  - With segmentation
    人物領域分割を行い，全訓練事例との類似度を評価した場合（提案手法）．

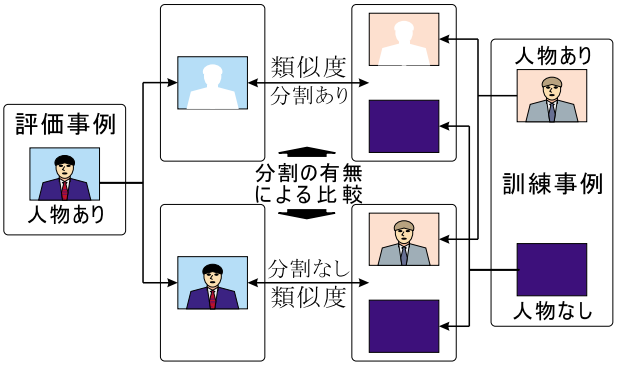

図 6　人物領域分割の有無による類似度評価

  - With only segmentation
    人物領域分割を行い，訓練事例中の人物領域が存在する事例のみとの類似度を評価した場合（提案手法，人物像の存在そのものも考慮）．

- 画像特徴量
  - 色ヒストグラム
  - 色コリログラム

4.3　実験結果と考察

図 7 は，評価のために設定した表 1 中の場面（1）から（5）の推定結果を総合して評価したグラフである．この結果から，以下のことが言える：

- n が小さいときに，コリログラムはヒストグラムよりもわずかに良い結果を示し，n が大きくなるにつれ，逆の傾向が見られる．これは，コリログラムは非常に類似した画像との類似度を高く示すものの，わずかな変動に敏感であり，ヒストグラムにはその逆の性質があることを示している．これらの性質は，**4.1.2** で述べた両特徴量のミクロ・マクロな性質から説明することができる．

- 提案手法である人物領域分離により，2%から 3%の性能向上がみられる．さらに，訓練事例中の人物像が存在しないものを除いた類似度評価の結果，7%から 8%の性能向上が見られた．後者の結果から，人物像の存在そのものも画像特徴量として考慮することが重要であることが示唆される．

次に，画像特徴量としてヒストグラムあるいはコリログラムを用いた場合，n を 1 あるいは 3 に設定して評価した場合における,（1）閣議控室（cabinet）,（2）国会議

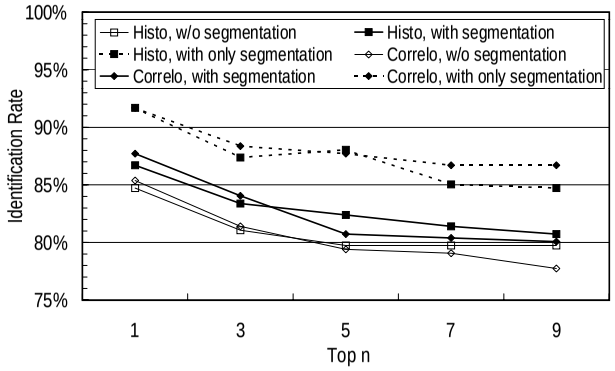

図 7　場面推定結果（総合評価；n 位過半率）

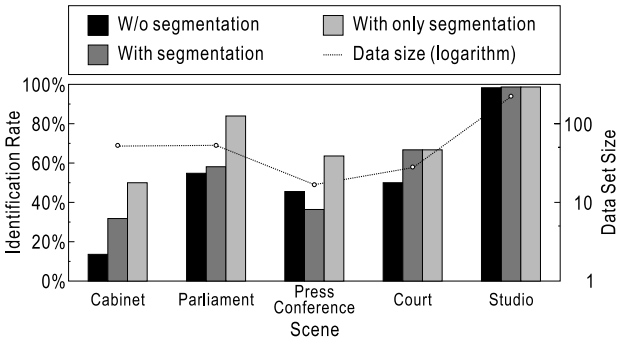

図 8　場面推定結果（1 位過半率，ヒストグラム）

場（parliament），(3) 報道会見場（press conference），(4) 法廷（court），(5) スタジオ（studio），の各設定場面別の推定性能を図 8 から図 11 に示す．また，グラフ中には，各設定場面毎の事例数を線グラフで対数表示してある．

これらの結果から以下のことが言える：

- 総合評価で述べたように，提案手法の有効性が基本的には示された．しかし，場面によって有効性に差異が見られる．

- 色ヒストグラムと色コリログラムの優劣にも場面により差異が見られる．

  これらから，人物領域分離と用いる画像特徴量の有効性は，対象とする場面により異なることが分かる．ここで，場面毎の特徴には，撮影しているカメラに関する制約（設置可能な位置・動かせる角度など）も含まれ，頻出場面においても，このような制約や場面固有の様々な画像的特徴により，画像的な典型度は無視できない範囲で一定の変動を示すものと思われる．

- 場面推定性能はおおまかに事例数の対数に比例している．上位 n 件の類似ショットの過半数を占める場面に基づいて推定を行ったため，この傾向は特に $n = 3$ の場合（図 10 及び図 11）に顕著である．訓練事例の規模が大きく多様であるほど，より多くの類似事例が存在する確率が高くなるはずであり，小規模で多様性の低い事例群を用いて，類似度上位 n 位の過半数を正しい場面で占めることは困難である．

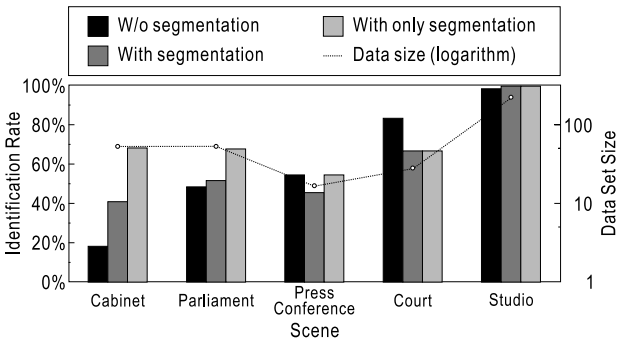

図 9　場面推定結果（1 位過半率，コリログラム）

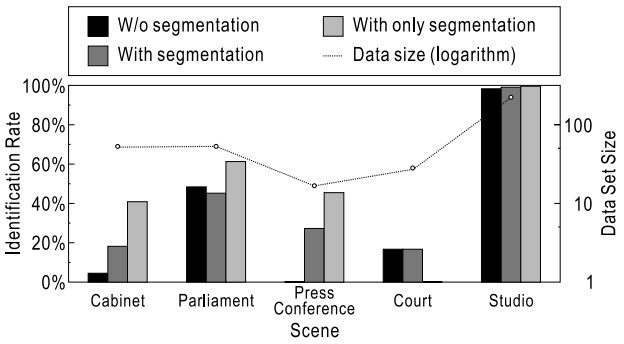

図 10　場面推定結果（3 位過半率，ヒストグラム）

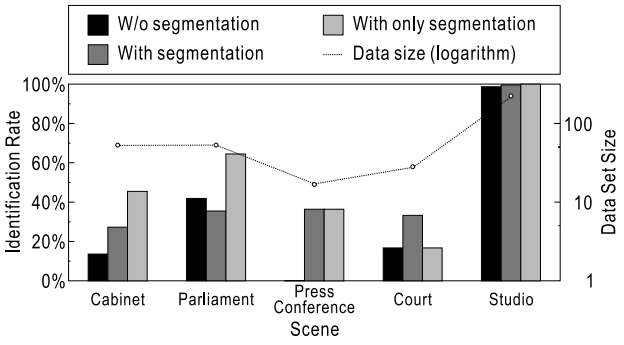

図 11　場面推定結果（3 位過半率，コリログラム）

5　おわりに

本稿では，ニュース映像の自動索引付け機構の一部として，背景領域から人物領域を分離して場面を推定する手法を紹介した．提案手法は，ニュース映像において，特定の場面は頻出するという特徴を利用し，事前に索引付けしておいた事例との類似度評価により推定を行うものである．

評価実験として，この手法を実際のニュース映像中の 817 のショットに適用したところ，総合評価では実用的な水準の推定性能を示し，領域分割のみでなく，人物の存在自体も画像特徴量として併せて考慮することが重要であることが示唆された．実験において，人物領域の切り出しは主に人手で行ったものの，特定の良好な条件下では顔領域を基準とした単純なテンプレートの適用による自動化が可能であり，将来のこの分野における要素技術の進展により，より一般的な自動化も可能になるものと思われる．

一方で，事前に設定した場面毎の評価において，以下のように今後検討の余地が残される点が洗い出された：

- 設定場面毎に特徴をより良く表す画像特徴量が異なり，複数の特徴量を重み付きで組み合わせて用いる必要がある．重みは訓練事例群から事前に得ておくことが考えられる．

- 場面推定性能が事例数の対数におおまか比例する傾向にあることから，様々な状況に対応し得るに十分な事例を確保するためには，事例数を指数規模で増やす必要がある．
  同時に，事例数を増やすことにより場面間の弁別能力の低下が懸念されるため，前項に挙げたような重み付けを用いて様々な画像特徴量を導入する必要がある．

謝辞

Face detector[10] の使用を快諾して下さった，Henry Rowley 博士に感謝する．

参考文献

[1] J. Huang, S. R. Kumar, and R. Zabih, “Image Indexing Using Color Correlograms,” Proc. IEEE Conf. on Computer Vision and Pattern Recognition ’97, pp.762–768, June 1997.

[2] 井手一郎，山本晃司，浜田玲子，田中英彦，“ショット分類に基づく映像への自動的索引付け手法，” 信学論（D-II），Vol.J82-D-II，No.10，pp.1543–1551, Oct. 1999.

[3] I. Ide, R. Hamada, S. Sakai, and H. Tanaka, “Semantic Analysis of Television News Captions Referring to Suffixes,” Proc. Fourth Intl. Workshop on Information Retrieval with Asian Languages (IRAL’99), pp.37–42, Nov. 1999.

[4] 井手一郎，浜田玲子，坂井修一，田中英彦，“ニュース映像における人物・背景領域を分割した特徴量解析による内容推定，” 第 5 回信学知能情報メディアシンポジウム論文集，pp.45–51, Dec. 1999.

[5] 木本晴夫，“感性語による画像検索とその精度評価，” 情処学論，Vol.40，No.3，pp.886–898, March 1999.

[6] 栗田多喜夫，加藤俊一，福田郁美，板倉あゆみ，“印象語による絵画データベースの検索，” 情処学論，Vol.33，No.11，pp.1373–1383, Nov. 1992.

[7] Y. Nakamura, and T. Kanade, “Semantic Analysis for Video Contents Extraction –Spotting by Association in News Video–,” Proc. Fourth Intl. Multimedia Conf. (ACM Multimedia ’97), pp.393–402, Nov. 1997.

[8] 孟　洋，佐藤真一，坂内正夫，“事例画像を用いたシーン分類による映像索引付け手法，” 第 5 回信学知能情報メディアシンポジウム論文集，pp.53–60, Dec. 1999.

[9] 森　靖英，高橋裕信，岡　隆一：“画像と単語の空間配置データベースに基づく画像理解の試み，” 第 4 回信学知能情報メディアシンポジウム論文集，pp.127–132, Dec. 1998.

[10] H. D. Rowley, S. Baluja, and T. Kanade, “Neural Network-Based Face Detection,” IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.20, No.1, pp.23–38, Jan. 1998.

[11] 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, March 1999.

[12] H. D. Wactler, A. G. Hauptmann, M. G. Christel, R. A. Houghton, and A. M. Olligschlaeger, "Complementary video and audio analysis for broadcast news archives," Commun. ACM, Vol.43, No.2, pp.42–47, Feb. 2000.