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

Content Analysis of News Video by Feature Analyses of Foreground and Background Regions

井手一郎　浜田玲子　坂井修一　田中英彦
東京大学大学院 工学系研究科[1]
Ichiro Ide　Reiko Hamada　Shuichi Sakai　Hidehiko Tanaka
Graduate School of Engineering, The University of Tokyo

概要: 日々放送される映像量の増大につれ，それらを再利用や検索に供するための自動索引付けの需要が高まっている．筆者らは特に索引付けの価値が高いと思われるニュース映像を対象に，従来の類似研究で軽視されがちであった，画像内容と索引との対応を考慮した索引付け手法を提案している．本稿では，そのような対応の考慮に必要となる，画像特徴量からの画像内容推定手法を紹介し，推定対象をニュース番組に頻出する場面に限定した簡単な実験により，その有効性を検証する．

Abstract: Due to the increase of the amount of video data, automatic indexing is in high demand for their recycling and retrieval. We are proposing an news video indexing method that considers the correspondences between video contents and indices, which has not necessarily been considered in conventional methods. In this paper, viedo content analysis from graphical features required for such indexing is proposed, and the efficacy is verified by a preliminary experiment applied to limited scenes that appear frequently in news video.

1 はじめに

1.1 背景と目的

日々放送される映像量の増大につれ，それらの映像を再利用や検索に供するために，自動索引付けの必要性が高まっている．特に，内容の実用的価値や速報性の点から，ニュース映像への自動索引付けの需要が高いと思われる．

このような需要に応えるべく，Informedia プロジェクト [Wactler99] の News-on-Demand システム [Wactler98] のように，ニュース映像への自動索引付けに関する研究が盛んに行なわれている．しかし，それらの手法の多くは，映像に付随する言語情報から比較的単純な手法で索引となり得る語句を抽出するのみであり，画像内容と索引との対応が考慮されていない．

そこで我々は，索引を付与する際に，画像内容と索引との対応を考慮した索引付け手法を提案している．本稿では，そのために必要となる，画像特徴量からの画像内容推定手法を紹介し，推定対象をニュース映像に頻出する場面に限定した簡単な実験により，その有効性を検証する．

[1] 〒 113-8656 東京都文京区本郷 7-3-1
TEL:(03)5841-7413, FAX:(03)5800-6922
{ide,reiko,sakai,tanaka}@mtl.t.u-tokyo.ac.jp

1.2 画像内容を考慮した索引付け機構

図 1に，画像内容を考慮した索引付け機構の全体像を示す．この機構では，画像と索引の候補となる映像に付随するテキストの両メディアに対し,(1) いつ (when),(2) どこで (where),(3) 誰が (who),(4) 何を (what) の 4 種類，いわゆる 4W と呼ばれる属性レベルでの対応を考慮することで，画像内容と索引との対応を保証する．これらの属性のみでは一般的な映像に適用するには不十分であるが，ニュース映像に対しては，これらによる検索に限定することは妥当であると考える．

このような対応を考慮する際には，画像特徴量と画像内容との関係に関する知識の利用が不可欠となる．これらの属性のうち,(1) に関しては比較的画像との関連が薄い属性であるため，議論から省く.(3) に関しては，Satoh らによる Name-It システム [Satoh99] により,(4) に関しては Nakamura ら [Nakamura97] や筆者ら [井手 99] がショット分類による大まかな内容推定と対応付けを実現している．

そこで，本稿では,(2) に関する対応を考慮するために必要となる知識，すなわち画像特徴量と画像内容 (この場合は場面) の関係の獲得と，獲得した知識に基づく実映像の場面推定実験の結果を示す．知識ベースの内容は，画像特徴量と画像内容との関係に関する知識であれば，明示的に記述された規則でも良いが，本

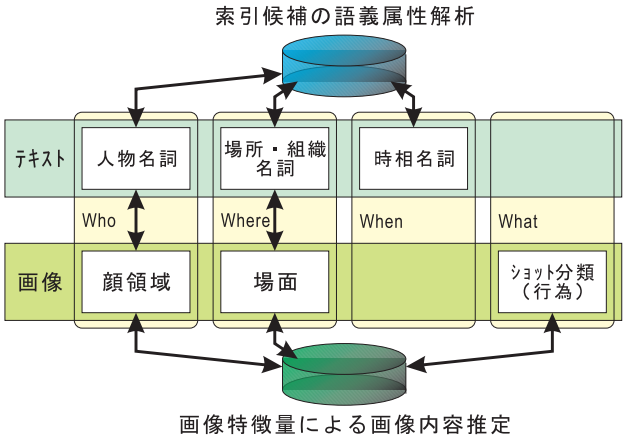

図 1: 画像内容を考慮した索引付け機構

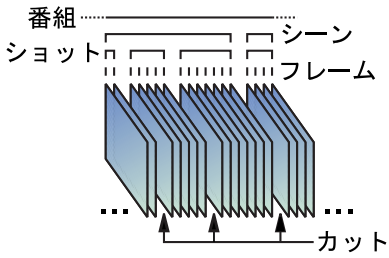

図 2: 映像の階層的構成と用語の定義

稿では多数の訓練事例から得られる知識を用いることにする．

一方，テキストに関しては，索引語の候補は字幕から得る．そのため，字幕のなかでも具体的な画像内容を端的に示すことが多い名詞句であるものに限定した属性解析を行う．この際，日本語において名詞句の末尾の名詞が名詞句全体の語義を決定する傾向が強いという性質を利用する [Ide99]．本稿では画像内容推定を中心に論じるので，このような名詞句の語義属性解析手法については詳述しないが，370 分間のニュース番組に出現する 2,546 の字幕に関する属性解析実験では，人物に関しては適合率 72.47%，再現率 82.35%，場所・組織に関しては適合率 54.77%，再現率 88.47%，時相に関しては適合率 41.93%，再現率 93.50%の精度で解析に成功している．

1.3 用語の定義

図 2に，映像の階層的構成と用語の定義を示す．映像（番組）はフレームと呼ばれる静止画像の連続から構成される．画像的に連続なフレームの集合をショットと呼び，**1.2** 節で紹介した索引付け機構では，ショット単位での索引付けを考えている．画像的あるいは内容的に類似したショットの集合をシーンと呼び，ニュース映像では後者は一つの話題に相当する．また，ショットとショットの境界の不連続点をカットと呼ぶ．

2 画像特徴量解析による内容推定

前章で述べたような索引付けを実現するためには，画像特徴量から画像内容を推定する必要がある．そこで，関連研究をいくつか紹介したうえで，そのような機能を実現するために本稿で提案する，人物領域と背景領域を分割した画像内容推定について述べる．

2.1 関連研究：特徴量と内容の関連付け

画像特徴量と画像内容を描写する索引と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との対応関係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 [栗田 92] がある．しかし，ニュース映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．

また森らによる手法 [森 98] は，まず語の一般的な共起関係により単語クラスタ空間を形成し，次に単語クラスタ空間中の単語間距離に基づき百科事典中の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がクラスタ化させる．しかし，この手法では単語ではなく文との関係を見ているため，獲得された関係を一般的に用いるのは難しい．

この他にも様々な画像分類に関する研究があるが，汎用性の高い分類規則，つまり画像特徴量と内容との関連を自動的あるいは統計的に獲得する手法は少ない．

2.2 人物領域・背景領域分割による内容推定

ニュース映像の特徴として，人間の社会的行為を扱った内容が多い点が挙げられる．そのため，画像中に人物，特に上半身が大きく映っていることが非常に多い．一方，頻出する話題については，それらの人物が登場する背景の場面（場所）が共通であることが多い．

このような特徴を考慮すると，映っている人物の領域を除外した背景領域の画像特徴量を参考にすることにより，場面を推定することができると考えられる．また，頻出する話題に関しては，背景領域が類似した画像特徴を示すはずであり，推定のための知識ベースの構築も比較的容易であると考えられる．

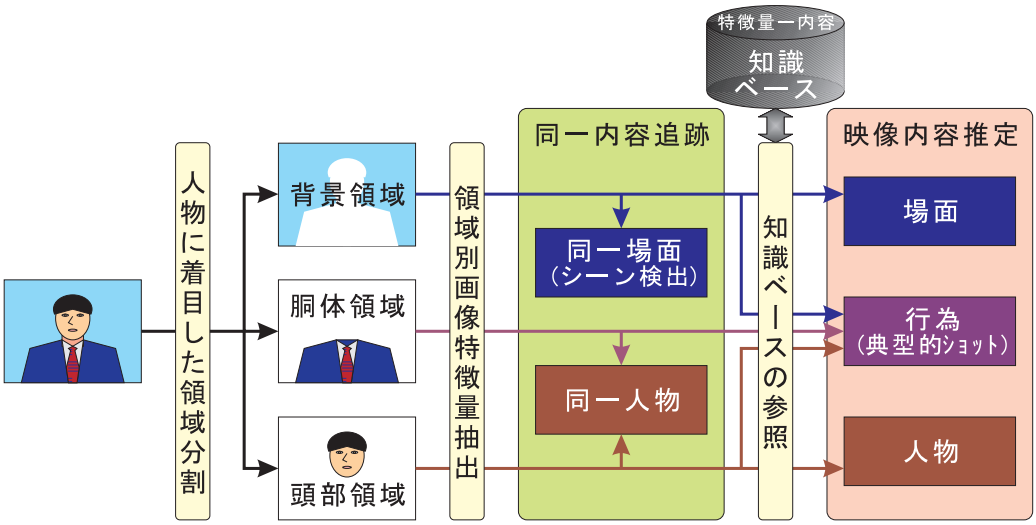

図 3: 人物領域と背景領域の分割による画像内容推定

そこで本研究では，画像中に一定以上の大きさの人物が存在する場合は，その人物の頭部と胴体の領域を分離し，背景領域の画像特徴量と知識ベース中の訓練事例との類似度に基づき，場面の推定を行う．図 3に，背景領域と人物領域（頭部領域と胴体領域からなる）の分割による画像内容推定機構を示す．なお，本稿ではこの機構のうち，場面の推定を中心に論じる．

以下に各処理とそれに先立つ前処理について簡単に記す．

2.2.1 前処理

内容推定に先立ち，以下の前処理を行う．

画像のディジタル化

本実験では，以下の条件で画像のディジタル化を行った．

- **空間解像度：** 横 320×縦 240 ピクセル
- **色解像度：** RGB 各色 8bit，合計 24bit (16,777,216 色)
- **時間解像度：** 15 フレーム毎秒
- **フレーム圧縮：** JPEG
- **動画像圧縮：** なし

カット検出

カット検出には様々な手法が提案されているが，本研究では離散余弦変換（DCT）特徴による手法 [有木 97] を採用している．この手法により，370 分間のニュース番組中の 1,541 箇所のカットに対して，適合率 59.78%，再現率 92.05%の精度で検出に成功している．しかし本稿では，場面推定手法単独での評価を行うため，目視により検出した正しいカットを利用した．

2.2.2 人物に着目した領域分割

前処理により分割されたショット中の各フレームに対して，人物領域と背景領域の分割を行う．ショット中の全フレームに対して内容推定を行うのが望ましいが，計算量の点から，ショットを代表する 1 フレームを推定対象とする．ここでは，ショットの冒頭の 1 フレームを代表フレームとして採用した．

人物領域の抽出は，基本的に顔領域を手がかりとして行う．本来は，正確な輪郭抽出を行うべきだが，ニュース映像中に映っている大きな人物は良好な照明条件下の正面顔という理想的な状態であることを期待し，図 4に示すような実測に基づく簡単なテンプレートマッチングを行って頭部・胴体領域を決定する．

顔領域抽出は，肌色領域抽出を含めて多くの既存研究が存在するため，それらを利用する．顔領域抽出に，ニューラルネットワークによる学習を用いたツール：*face detector* [Rowley98] を用い，図 4のテンプレートに基づき頭部・胴体領域の決定を試みたところ，良好な照明条件下の正面顔という条件を満たすスタジオ内のキャスタの画像 173 件に対して，100%の精度で過不足なく顔領域の抽出に成功し，頭部・胴体領域の決定もほぼ正確であることが目視で確認できた．しかし，他のより一般的な画像中の人物に関しては，必ずしも照明条件が良好でないことや正面顔でなかった．このことから，本稿では，場面推定手法単独での評価を行

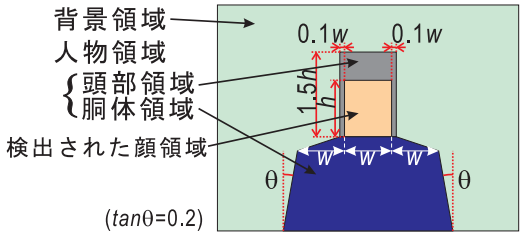

図 4: 顔領域を基準とした頭部・胴体領域決定

うため，スタジオ内のキャスタ以外は，目視により正しい領域を抽出したものを実験に用いた．

2.2.3 領域別画像特徴量抽出

次に，各領域毎に画像特徴量を抽出する．画像特徴量としては，計算量の点から処理が容易なものを中心に用いる．領域によって内容推定に適した特徴量は異なり，本稿で取り上げる場面推定に対しては，具体的事物の認識は行わず，色彩などの抽象度の高い特徴量を用いることによりロバストな推定を行うことを目指す．

具体的には，以下の 2 種類の色彩に関する画像特徴量を各々別に用いる．

色ヒストグラム（出現頻度分布）

色ヒストグラム $H(c_i)$ とは，画像中における色 c_i のピクセルの出現確率であり，次式のように定義される．

$$H(c_i) \equiv \frac{\text{色 } c_i\text{のピクセル数}}{\text{全ピクセル数}} \quad (i = 1, 2, ..., 64)$$

色コリログラム（共起頻度分布）

色コリログラム $C(c_{j1}, c_{j2}, d)$ とは，画像中における色 c_{j1} と色 c_{j2} のピクセルが距離 d 離れて出現する確率であり，次式のように定義される．

$$C(c_{j1}, c_{j2}, d) \equiv \frac{\text{距離 } d \text{ 離れた色 } c_{j1} \text{と } c_{j2}\text{のピクセルの組の数}}{H(c_{j1}) \times 8d} \quad (j1, j2 = 1, 2, ..., 16; d = 1, 2, 3, 4)$$

色ヒストグラムが画像全体のマクロな色彩の特徴を表すのに対し，色コリログラムはミクロな色彩の特徴を表す．具体例として，図 5に示すように，着色面積が等しい大円と小円の集合について，色ヒストグラムは区別できないのに対し，色コリログラムは区別できる．

以上の定義式中の色の階調（$i, j1, j2$）及び距離（d）の最大値は，次章の実験で用いる値を示した．色の階調は RGB 色空間を線型に分割したものを用い，距離としては簡単のために 8 近傍（チェス盤）距離を採用した．

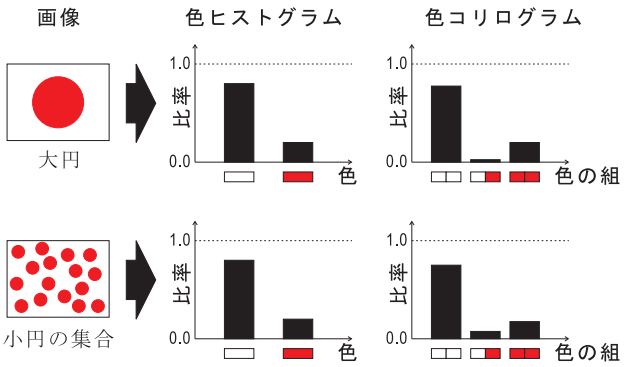

図 5: 色ヒストグラムと色コリログラムの差異の例

2.2.4 知識ベースの利用による内容推定

以上のようにして，各ショットの領域毎の画像特徴量が得られ，次にこれをもとにして，画像内容の推定を行う．内容推定は，推定を行おうとするデータと知識ベース中の知識との類似度に基づいて行う．具体的には，特徴量ベクトル間の類似度，すなわち，場面推定を行いたい画像の画像特徴量ベクトル $\vec{F}_e$ と，知識ベース中の訓練事例の画像特徴量 $\vec{F}_t$ との類似度を求める．類似度としては，次式で定義されるベクトル同士のなす角度 θ の余弦（$0 \leq \cos\theta \leq 1$）を用いる．

$$\vec{F}_e \text{と} \vec{F}_t \text{の類似度} \equiv \cos\theta = \frac{\vec{F}_e \cdot \vec{F}_t}{|\vec{F}_e||\vec{F}_t|}$$

知識として，訓練事例の特徴を統計的処理により抽象化した代表ベクトルを用いる手法も考えられるが，同一内容の訓練事例が特徴量空間中で必ずしも稠密な分布を示すとは限らない．そこで，本稿では，訓練事例を抽象化せずに全て保存しておき，それら一つ一つとの類似度から判定する，記憶に基づく推論方式を用いる．

2.2.5 同一画像内容の追跡

図 6に示すように，複数のショットに跨って，同一人物や同一場面が存在することがあり，これらを追跡することにより，ショットに跨って索引の付与範囲を決定する必要がある．そのために，分割された各領域の画像特徴量の類似度に基づき，ショットを跨った同一人物・同一場面の追跡を行う．

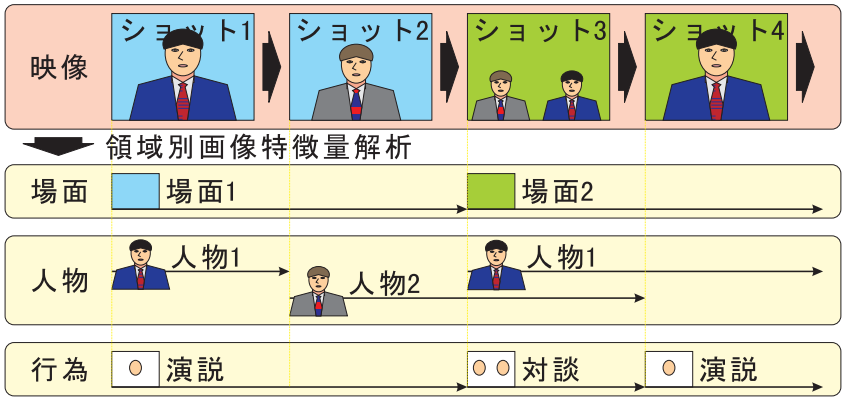

図 6: 同一画像内容の追跡

3 場面推定実験

以上のような画像特徴量から画像内容を推定する機構の実現可能性を調べるために，背景領域の画像特徴量から場面を推定する実験を行った．

実験には，15 分間の全国版ニュース映像 20 本，合計 300 分間の映像を用いた．この映像中には 1,542 ショットが存在したが，訓練事例として同一場面毎に一定量のデータが必要なため，この中で頻出する国内政治関連の場面とスタジオに限定した場面推定を行った．具体的には，(1) 閣議前室，(2) 記者会見室，(3) スタジオ，(4) その他の場面が登場する合計 659 ショットを用い，(4) を除く各場面から評価用事例を 1, 2 ショット選択し，残りの 654 ショットを訓練事例として用いた．表 1に，各場面分類毎の事例数を人物領域の有無に分けて示す．

表 1: 場面分類毎の事例数：カッコ内は評価事例数

場面分類		人物有	人物無	合計
(1)	閣議前室	18 (1)	10 (1)	28 (2)
(2)	記者会見室	11 (1)	6 (1)	17 (2)
(3)	スタジオ	240 (1)	0 (0)	240 (1)
(4)	その他	126 (0)	248 (0)	374 (0)
	合計	395 (3)	264 (2)	659 (5)

なお，画像特徴量として，(1) 64 次元の色ヒストグラム，(2) 1,024 次元の色コリログラムの 2 通りを用いた場合に分けた実験結果を示す．

図 7に示すように，人物領域がある画像については領域分割による背景領域の特徴量を，ない画像については全領域の特徴量を用いて類似度評価を行った．なお，領域分割の効果を検証するために，前者についても，分割を行った場合と行わなかった場合を比較した．また，分割を行った場合でも，人物の存在の有無を含めた場面の推定という観点から，分割を行ったものに限定した類似度評価も併せて行った．

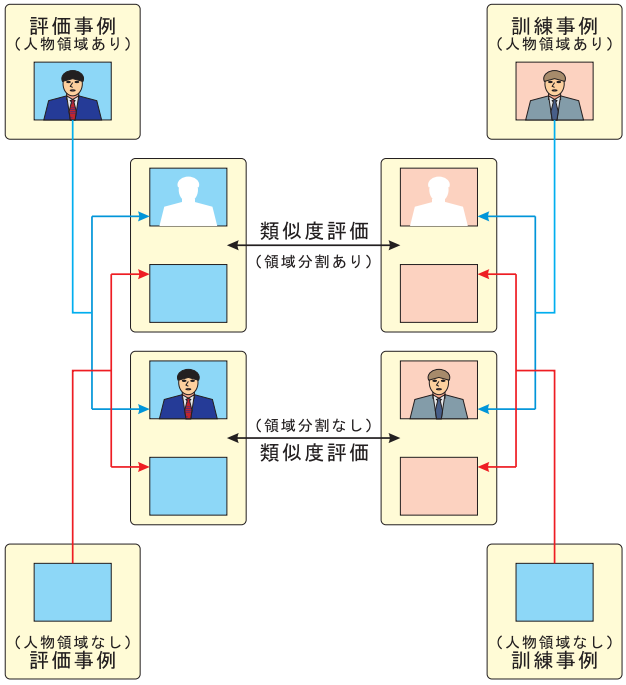

図 7: 場面推定実験の類似度評価条件

3.1 実験結果と考察

以上の実験条件のもとに，評価事例の訓練事例に対する類似度から，場面推定を行った．その結果を表 2, 3に示す．

表 2: 場面推定実験の結果（人物領域がある評価事例）

評価事例の場面分類			訓練事例数	類似度上位 k 位中の正解数							
				色ヒストグラム				色コリログラム			
				k= 1	k= 3	k= 5	k=10	k= 1	k= 3	k= 5	k=10
（1）	閣議前室	（分割なし）	26	0/ 1	0/ 3	0/ 5	0/10	0/ 1	0/ 3	0/ 5	1/10
		（分割あり）	26	0/ 1	0/ 3	1/ 5	1/10	0/ 1	1/ 3	1/ 5	2/10
		（分割あり限定）	17	0/ 1	1/ 3	2/ 5	4/10	**1**/ 1	1/ 3	**3**/ 5	**5**/10
（2）	記者会見室	（分割なし）	15	0/ 1	0/ 3	1/ 5	3/10	0/ 1	0/ 3	0/ 5	1/10
		（分割あり）	15	0/ 1	**2**/ 3	**3**/ 5	4/10	0/ 1	1/ 3	**3**/ 5	4/10
		（分割あり限定）	10	**1**/ 1	**3**/ 3	**3**/ 5	**5**/10	0/ 1	1/ 3	**3**/ 5	4/10
（3）	スタジオ	（分割なし）	239	**1**/ 1	**3**/ 3	**5**/ 5	**9**/10	**1**/ 1	**3**/ 3	**5**/ 5	**10**/10
		（分割あり）	239	**1**/ 1	**3**/ 3	**4**/ 5	**7**/10	**1**/ 1	**3**/ 3	**5**/ 5	**8**/10
		（分割あり限定）	239	**1**/ 1	**3**/ 3	**5**/ 5	**9**/10	**1**/ 1	**3**/ 3	**5**/ 5	**9**/10

表 3: 場面推定実験の結果（人物領域がない評価事例）

評価事例の場面分類			訓練事例数	類似度上位 k 位中の正解数							
				色ヒストグラム				色コリログラム			
				k= 1	k= 3	k= 5	k=10	k= 1	k= 3	k= 5	k=10
（1）	閣議前室	（分割なし）	26	**1**/ 1	1/ 3	1/ 5	2/10	**1**/ 1	1/ 3	1/ 5	1/10
		（分割あり）	26	**1**/ 1	1/ 3	1/ 5	2/10	**1**/ 1	1/ 3	1/ 5	3/10
（2）	記者会見室	（分割なし）	15	0/ 1	0/ 3	0/ 5	0/10	0/ 1	0/ 3	1/ 5	1/10
		（分割あり）	15	0/ 1	0/ 3	0/ 5	0/10	0/ 1	0/ 3	0/ 5	0/10

類似度からの場面推定には様々な基準が考えられるが，ここでは，k 近傍法で類似度が上位 k 位のデータ中の過半数を占めるものを解答とし，この基準に従って正解と判定される結果を表 2, 3中に太字で強調して示した．

この結果から，以下に列挙するような事項が言え，提案手法の有効性が示された．

- 色ヒストグラムに対する色コリログラムの優位性
- 人物領域がある場合の人物領域分割による背景領域分離の有効性
- 背景領域の画像特徴量のみならず人物の存在の有無を含めた「場面」の定義の優位性

4 おわりに

本稿では，筆者らが提案する，画像内容を考慮した索引付け手法を紹介し，その実現のために必要となる，画像特徴量と画像内容との関係に関する知識ベースに基づいた画像内容推定手法を提案するとともに，簡単な実験を通じてその有効性を確認した．特に理想的な証明条件下で正面顔で撮影されたスタジオ内のキャスタについては，領域分割以降の処理を完全に自動化したうえで正しい推定が行えた．

実験の結果，人物領域がある場合に，人物領域分割なしでは画像特徴量の類似度を用いて推定し得なかった場面が，人物領域分割を行うことにより推定できる場合があることが示された（特に類似度上位 5 位の過半数で判定した場合）．

また，単純に背景領域の画像特徴量のみから判断するのではなく，前景における人物の存在の有無という特徴量を含めた「場面」の定義が有効であることがうかがえた．

このような特徴量の組合せや適性を考慮した類似度判定手法を導入する必要があり，今後は，場面分類毎の各特徴量の重要性を各特徴量の分布などから判定し，

それらを考慮して各特徴量の類似度に重み付けをした類似度判定法の導入を検討している．

更に他の画像特徴量の導入も考えており，色彩に関するものでも，現行の RGB 色空間を線型に分割したものではなく，HSI 空間に変換し，色相（H）での色ヒストグラムや色コリログラム，明度（I）の利用などを考えている．

なお，本稿では同一画像内容の追跡実験は行っていないが，同様の類似度評価により，同一内容の遷移を追跡できるものと考えている．

謝辞

顔領域検出ツール *face detector* を快く提供して下さった，元米国 Carnegie Mellon 大学（CMU）の Henry D. Rowley 博士に感謝する．

参考文献

[有木 97] 有木康雄：“DCT 特徴のクラスタリングに基づくニュース映像のカット検出と記事切出し”，信学論（D-II），vol.J80-D-II，no.9，pp.2421-2427（Sep 1997）．

[Huang97] Huang, J., Kumar, S. R., Mitra, M.: “Combining supervised learning with color correlograms for content-based image retrieval”, Proc. fifth ACM intl. multimedia conf., pp.325-334 (Nov 1997).

[井手 99] 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）．

[Ide99] Ide, I., Hamada, R., Sakai, S., Tanaka, H.: “Semantic Analysis of Television News Captions Referring to Suffixes”, Proc. fourth intl. workshop on information retrieval with Asian languages, pp.37-42 (Nov 1999).

[栗田 92] 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）．

[森 98] 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 回知能情報メディアシンポ論文集，pp.127-132（Dec 1998）．

[Nakamura97] Nakamura, Y., Kanade, T.: “Semantic analysis for video contents extraction – Spotting by association in news video–”, Proc. fifth ACM intl. multimedia conf., pp.393-402 (Nov 1997).

[Rowley98] Rowley, H. D., Baluja, S., Kanade, T.: “Neural network-based face detection”, IEEE Trans. on pattern analysis and machine intelligence, vol.20, no.1, pp.23-38 (Jan 1998).

[Satoh99] Satoh, S., Nakamura, Y., Kanade, T.: “Name-It: Naming and detecting faces in news video”, IEEE Multimedia, vol.6, no.1, pp.22-35 (Mar 1999).

[Wactler99] Wactler, H. D., Christel, M. G., Gong, Y., Hauptmann, A. G.: “Lessons learned from buliding a terabyte digital video library”, IEEE Computer, vol.32, no.2, pp.66-73 (Feb 1999).

[Wactler98] Wactler, H. D., Hauptmann, A. G., Witbrock, M. J.: “Informedia News-on-Demand: Using speech recognition to create a digital video library”, CMU tech. rep., CMU-CS-98-109 (Mar 1998).