オーディオビジュアル複合情報処理 32−11
（2001. 3. 2）

動画像インデキシングを目的とした人物顔追跡に関する検討

奥野 洋平† 朱 青†† 小舘 亮之†† 富永 英義†††

† 早稲田大学 理工学部
〒169−8555 東京都新宿区大久保 3−4−1 55 号館 -N0602

†† 早稲田大学 国際情報通信研究センター
〒169−0051 東京都新宿区西早稲田 1−3−10 29-7 号館

okn@tom.comm.waseda.ac.jp

デジタルビデオカメラの普及およびPCの発達により，一般人でも映像の撮影，保存を行なえる環境が整いつつある．しかし編集および閲覧するための設備はいまだ不十分であり．より容易に高速に所望のシーンにアクセスするため，映像の特徴を示す情報（インデックス）を自動的に抽出し，ラベル付けする技術（インデキシング）が必要とされている．一般人の撮影する映像には撮影対象が人物である場合が多く，インデックスとして登場人物や人数，人物の動きなどが求められる．本稿では，これらのインデックスを抽出するために，人物顔を映像から検出するシステムを検討している．中でも顔追跡の方法にCamShiftアルゴリズムを用いており，インデキシングに適した応用方法について述べる．

キーワード

動画像インデキシング，人物，顔追跡，CamShift，HSV表色系

Face Tracking Method for Video Indexing

Yohei OKUNO† Qing ZHU†† Akihisa KODATE†† Hideyoshi TOMINAGA†††

† Dept. of Science and Engineering, Waseda University
3−4−1 Ohkubo, Bldg.55-N0602, Shinjuku-ku, 169-8555, Tokyo JAPAN

†† Global Information and Telecommunication Institute, Waseda University
1−3−10 Nishi-Waseda, Bldg.29-7, Shinjuku-ku,169−0051, Tokyo JAPAN

okn@tom.comm.waseda.ac.jp

By the spread of a digital video camera and the development of PC, we can film and store it easily now. However, the facility to edit and watch is not enough. In order to access a desired scene at high speed more easily, a technology extractinng information to show video contents is needed. There are many cases that the target is a person to the picture which we film as a hobby, and a character and the number of people, movement of a person are needed as an index. I examine a system detecting human faces from a picture in order to extract these indexes in this paper. I use CamShift algorithm in a method of a face traking and describe a suitable method to index.

Keywords

Video Indexing, Person, Face Tracking, CamShift, HSV Color System

1. はじめに

映像メディアの利用が進むにつれて，膨大な映像素材から所望の映像を効率良く検索するために，検索の手がかりとなるように映像にその特徴を表すインデックス情報を付与，抽出していく必要がある．

本稿では特に一般家庭用ビデオカメラによって撮影される映像（以下，素人映像）を対象としている．一般人(素人)の利用が主なこのような映像は，子供の行事・成長記録，旅行，アウトドア・レジャー，結婚式などの記録，学校サークルなどの記録，日常生活のスナップ，スポーツの記録，業務・取材用などの目的で撮影される．そしてこれらの多くが撮影対象を人物としており，人物インデックスの有用性は特に高いと言える．登場人物や人数，人物の動きなどのインデックス情報(図1)が得ることができれば，映像を素早く検索することが可能になり，映像閲覧，編集を効率化が図れる．

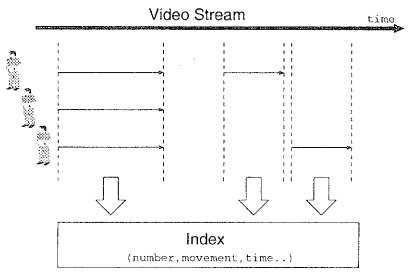


図 1: 人物インデックス

顔画像認識における課題は，対象とする顔の可変性が高いという点である．従来の研究では，照明条件，顔の向きなどといったいくつかのパラメータを限定することにより検討が行なわれてきた．本稿ではそれらをふまえ，素人映像という自由度の高い映像における自動インデキシングを目指している．ここでは素人映像をサンプルとし人物インデックスの特長を整理し，現在の技術で補える部分および不足した技術について調査，検討を行うものである．

2. 従来技術

映像中から特定の人物を検出する手法として，以下の手法が挙げられる．

1. 映像の時間的変化を利用する方法
2. 形状，色，テクスチャ等からの認識，理解を利用する方法

1. の方法には，動画像の連続するフレーム間での差分を利用する方法や，あらかじめ用意しておいた背景との差分をとる方法がある．監視の分野で主に用いられている方法である．人物を移動物体とほぼ限定できるためこのような方法を用いることができる．人物が動いていない場合や背景が動いている場合は検出が困難となる．

2. の方法は，主に顔の形状や色の特徴を個々の方法で定義することで検出するものである．色を用いた手法で代表的なものは，HSV 表色系や独自の表色系を用いて，顔の肌色領域とそれ以外とを分離する方法である[(1)(2)]．これらの方法では顔領域の色(肌色)が一様に安定して得られるという環境が撮影条件として必要となる．主に，顔認識を目的として検出を行なうため，撮影条件や顔領域の解像度などに条件が加わる．

一方，コンピュータービジョンの分野でも顔検出が試みられている．Gary.R[(3)]は，ゲーム用のアプリケーションとしてディスプレイの前にビデオカメラを設置し，人物の顔の動きをゲーム上に反映させるという目的で顔の検出を試みている．始めに顔領域を矩形で指示し，指示された領域中の色ヒストグラムからフィルタとなる色テーブルを作成し，その後のフレームにおいて追跡を行うというものである．また，この技術は Intel[(4)] からフリーツール（CamShift[(3)]）として発表されている．

2.1 動画像インデキシングを想定した顔検出

Henry A.Roeley[(5)]，Smith[(6)] らは，本稿と同様の目的で顔検出を検討している．Roeley らはニューラルネットワークを用いて，正面顔の複数のサンプルを学習させる方法で検出を試みている．この方法では顔が正面を向いていなくてはならないという制限はあるものの，同一静止画像中の異なる大きさの複数の顔の検出に成果が出ている．130 個の静止画像から 78.9 %～ 90.0 %という検出結果も出ている．Smith らは，この手法を動画像インデキシングに応用しており．計算量を削減するため，同様の処理をフレーム毎ではなく 15 フレーム毎に行なっている．

この方法における問題点は正面を向いている顔が存在するフレームのみしか検出できないという点である．正面を向いていないフレームにおいての検出が必要となる．

3. 提案

3.1 人物インデックス

まずは動画像インデキシングという目的をふまえ，人物インデックスについての特長を整理した．

分析 1 従来では，目的に応じて，動き情報や色情報といった固定パラメータを設け映像中における人物を特徴付けることで，検出が試みられてき

た．固定パラメータとして以下のようなものが考えられる．

- 顔のサイズ
- 顔の位置 (動き)
- 顔の向き
- 顔の個数
- 顔の表情
- 照明条件 (照明，日光，陰影)
- 背景

本稿の目的はインデキシングであるので，個人認証や監視の分野のようには，これらのパラメータを固定することはできない．より多くの映像に対してインデキシングできるような方法が必要となる．

分析 2 得られたインデックスをどのように利用するかによって，必要とされるインデックスの質も変わってくる．

- 低：人物の存在
- 中：人数，顔の位置，サイズ
- 高：個人認証

低・中レベルにおいては顔を囲む矩形領域程度の情報が得られれば十分だが，高いレベルでの利用になると，追跡結果として高解像，詳細な顔形状が必要とされる．

分析 3 動画像とは静止画像を時系列上に並べたものと考えることができるが，時間的な連続性をもつという点で，単なる静止画像アーカイブとは異なる．つまり同一ショット (連続して撮影された映像区間，図 2) の中においてであれば，静止画単位の検出ではなくて，連続したフレーム間の共有する画像特徴から，顔を検出できる可能性があるということである．

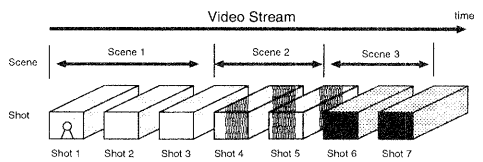


図 2: ショット単位

3.2 システム構成

上記の分析をまとめると，動画像としての時間的連続性をうまく利用して，分析 1 に示す多くのパラメータに対しどれだけロバストに検出が行なえるかということになる．

具体的に説明すると，ます時間的連続性が応用できるように映像をショット単位に分割する（図 2）．そしてショットの中から静止画単位で顔検出を行なう．この検出は従来技術で補えるものとする．例えば，正面顔のある程度の大きさをもった顔ならば従来技術で発見することができ，本稿ではこれを条件とする．そしてショット中のいずれかのフレームにおける顔が発見できたら，そのフレームを基準として追跡を行なう．追跡には，静止画単位では発見することのできない横を向いた顔や，動き，大きさが変化する顔領域に対しても追跡できる性能が必要とされる．本稿では追跡のための特徴量として色が適していると判断し，図 3に示すような構成のもと，色情報を用いた顔追跡について検討を行なった．

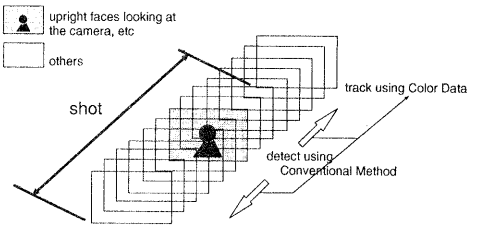


図 3: システム構成

3.3 色情報を用いた顔追跡

色を特徴量とした場合，探索対象物の形状に依らないので，顔の向きや表情変化にロバストであると言える．また単にある色を持つ領域を探索するだけなので，探索対象の静，動は関係なく追跡することができる．その反面，肌色と近い色をもった無関係の物体が背景中に存在する場合，検出が困難となる (もしくは，検出ができない) という欠点がある．

色情報を用いた顔追跡として 2 章に示した Intel(?) の CamShift(3) というツールがある．本稿の目的に適した従来技術として検討を行なった．

3.3.1 Camshift アルゴリズム

1. 基準フレーム (初期フレーム) において顔領域を矩形で指示 (図 4(1))
2. 以下の方法で追跡用画像を作成する
   (a) 原画像を HSV 表色系に変換する
   (注)HSV が値をとる範囲は 0 ～ 255 に変換しておく
   (b) (a) で指示した領域中において H(色相) のヒストグラムを求める (図 4(2))
   (c) 最も頻度の高い H の輝度値を 255 とし，他の H について頻度の割合に応じて輝度値

を算出する (追跡用画像変換テーブル)

L_h：H の値が h である画素に対応する追跡用画像の輝度値
F_h：H の値が h である画素の指示領域中における頻度
h：0 ～ 255．H の値

$$L_h = F_h / Max(F) * 255 \qquad (1)$$

(d) さらに S と V に輝度画像出力のための最低のしきい値 (Th_S, Th_V) を固定値として与える

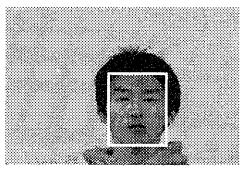
$$L_h = \begin{cases} L_h & S > Th_S \ and \ V > Th_V \\ 0 & otherwise \end{cases} \qquad (2)$$

(e) (1)(2) 式に従って追跡用輝度画像 (図 4(3)) を生成

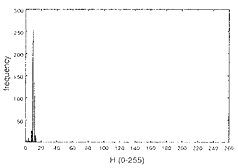
3. すべてのフレームにおいて追跡用輝度画像を生成する

4. 輝度画像から meanshift 法 (詳細は省略) を用いて探索する

5. フレーム毎の矩形領域情報が得られる

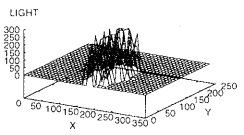


(1) 原画像



(2) 指示領域 histogram

(3) 追跡用輝度画像



(4) (3) の輝度値

図 4: CamShift アルゴリズム

ズームやパンにより顔領域が変化しても，図 4(4) に示すような輝度の山を探索する手法 (meanshift 法) で追跡が可能である．また顔の向きが変わっても顔の皮膚領域が残っていれば追跡が続行できる．すべての映像に対して追跡が可能とは言えないが，顔の向きのおよび静止・移動に対し，比較的ロバストに追跡が行える．

このアルゴリズムを応用する上で考慮すべきなのは，自動化である．インデキシングを行う上でのひとつの課題は，「どれだけ人の手を介さずにそのインデックスを抽出することができるか」である．CamShift では，より顔領域の対比をとるため，S と V にしきい値を設定し，しきい値に満たない画素については輝度を 0 としている．そしてこのしきい値の決定は，人間に委ねられており，映像を再生しながらその追跡用輝度画像を判断し，しきい値を調整しなくてはならない．そこでしきい値の決定を自動で行う方法について検討した．

3.3.2 しきい値決定法

図 5(a)(b) は図 4(1) の画像から追跡用画像を出力したものである．(a)(b) はそれぞれ式 (2)のしきい値を (a) Th_S:0 Th_V:0　(b) Th_S:56 Th_V:96 とした画像である．(a) では，探索領域以外にも高い輝度値をもっている領域が多く存在するため追跡は誤ってしまう．しきい値を設定し，(b) のように探索領域以外の輝度を除外する必要がある．

(a)Th_S:0,　Th_V:0

(b)Th_S:56　Th_V:96

図 5: しきい値の影響

CamShift では追跡実行中（再生中）に映像を見ながらしきい値を少しずつ変更し，5(b) のような状態（画像）を生成していた．人間の場合，「探索領域中 (顔) の輝度が落ちすぎず，それ以外の領域の輝度が落ちるように」という方針のもとしきい値を調整する．同様の考え方でしきい値を算出する方法を提案する．

まず，図 6に示すような比較領域を定義する．探索領域（図中の白色部分）とその $\sqrt{2}$ 倍に拡大した矩形領域とに挟まれた領域を外側領域（図中の斜線部分）とする．ただし，拡大率は，明確な理由は示すことはできないが，比較する上での利便性のため外側領域と探索領域の面積が同じになる様に $\sqrt{2}$ 倍とした．

探索領域および外側領域にそれぞれについて，領域内におけるすべての輝度の合計値 W を算出する．ただし，領域の縦横の長さを w，h，領域内の xy 座標を i，j，点における追跡用輝度値を L_{ij} とする．

$$W[Th_S][Th_V] = \sum_{i=0}^{w} \sum_{j=0}^{h} L_{ij}[Th_S][Th_V] \qquad (3)$$

さらに式 (4)に示す比率 R を求める．R はしきい値の変化によって探索領域と外側領域とに輝度の対比がどのように変化するかを計るための式である．分かりやすいように $W[0][0]$ で割ることで，初期状

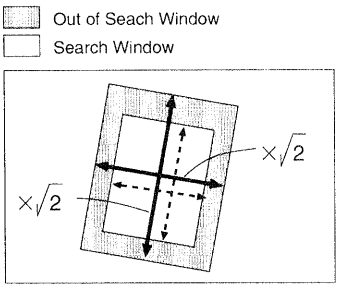


図 6: 外側領域

態（$Th_S = 0, Th_V = 0$）における R を 1 としている．初期状態に比較してより探索領域に輝度が集中すると，1 より大きな値とる．またしきい値を上げていくと，探索領域，外側領域共に W は減少する．しかしその減少の仕方がそれぞれで異なり，比率 R にその変化が表われてくる．

$$R[Th_S][Th_V] = \frac{W_{in}[Th_S][Th_V]/W_{in}[0][0]}{W_{out}[Th_S][Th_V]/W_{out}[0][0]} \quad (4)$$

図 4(1) の画像に対して比率 R を求めた結果が図 7である．（ただし，Th_S, Th_V は 5 間隔）しきい値 $Th_S = 85, Th_V = 80$ の時に R は最大となり，その値は 110.6 である．対比はとれているものの，探索領域内の輝度が初期状態に比べて 38 % しか残っていない．図 9(a) にしきい値を $Th_S = 85, Th_V = 80$ とした場合の追跡用輝度画像を示す．顔の輪郭すら分からないぐらいに輝度が無くなってしまっており，追跡は誤ってしまった．そこで，輝度が落ちすぎないように以下の三つの式を定めた．ただし $Area_{out}$ は外側領域の面積とする．

$$R = Max(R[Th_S][Th_V]) \quad (5)$$

$$W_{in}[Th_S][Th_V]/W_{in}[0][0] \geq \alpha \quad (6)$$

$$\begin{cases} W_{out}[Th_S][Th_V]/W_{out}[0][0] \leq \beta \\ \quad or \\ W_{out}[Th_S][Th_V]/Area_{out} < 2 \end{cases} \quad (7)$$

式 (6)は探索領域中の輝度を保持するため，式 (7)は探索外領域中の輝度を除去するための式である．(6)(7)共に左辺値の最大は 1 となる．α と β は定数で初期値として，それぞれ 1，0.3（予備実験による）を与えた．この式 (6)(7)を満たすような Th_S，Th_V がひとつもない場合，$\alpha = \alpha - 0.05$，$\beta = \beta + 0.01$ として再度条件を満たす Th_S，Th_V が見つかるまで繰り返す．組が一つ以上見つかったら，それらのなかで R が最大となる Th_S，Th_V を最終的なしきい値とする．

図 7中で式 (6)(7)の条件を両方とも満たす R については値はそのまま，それ以外のどちらか一方でも条件を満たさない R については $R = 0$ として，再び比率 R をグラフにしたのが図 8である．また，$\alpha = 0.8$，$\beta = 0.34$ であった．図中の R の最大値は約 2.58 でしきい値は $Th_S = 44, Th_V = 65$ であった．このしきい値で追跡用輝度画像を出力したものが図 9(b) である．探索領域内の輝度は落ちすぎず．かつ外部との対比がとれている．

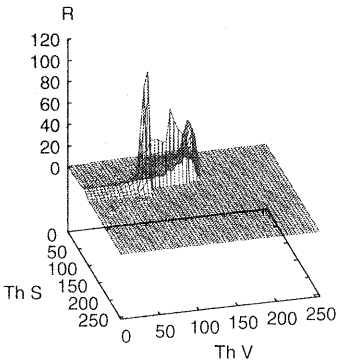


図 7: 比率 R

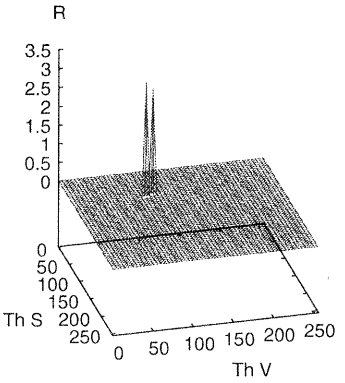
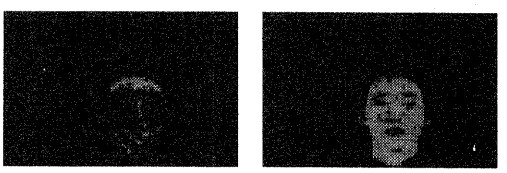


図 8: 比率 R($\alpha = 0.8 \quad \beta = 0.34$)



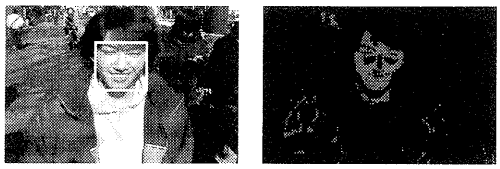
(a)Th_S:85, Th_V:80　(b)Th_S:44 Th_V:65

図 9: しきい値決定法

3.3.3 結果

しきい値自動決定法を用いて四つの画像について追跡用輝度画像（図10～図12）を出力した．

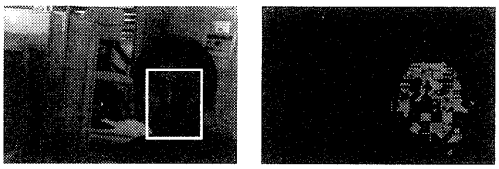
図10に示す画像は屋外で撮影し，日光の照射が非常に強い場合の映像である．特にV（明度）のしきい値が高い値となっている．顔の部分だけ特に，日光が強く反射していることがうかがえる．また対比もとれている．



$Th_S = 50$　　$Th_V = 144$

図 10: 屋外で撮影・日光強し

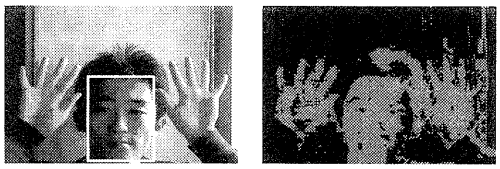
図11は図4(1)と同じ部屋で照明を落し，暗くして撮影した映像である．この場合，しきい値が比較的低い値となっている．つまり顔領域におけるS（彩度）およびV（明度）の特徴が少なくなったということである．照明の影響で，特徴量が減り，検出は困難になることがうかがえる．



$Th_S = 0$　　$Th_V = 20$

図 11: 屋内の暗い部屋で撮影

図12は顔の周りに両手を置いて撮影した映像である．手のように顔と似ている色特徴をもった物体が近くにある場合は，色による分離は困難であり，別の方法の検討が必要である．



$Th_S = 72$　　$Th_V = 70$

図 12: 屋内で撮影，周りに手

屋内で照明が安定している環境で撮影した映像に対しては比較的，顔の色特徴が強く，しきい値がおおよその値であってもある程度の対比をとることができ，手動でも設定が可能である．しかし屋外で日光の照射が強い場所や逆に照明の少ない場所では，その影響で色特徴は不安定となり，しきい値の設定にも正確な指定が必要となる．

またしきい値の自動決定は映像への影響とは別に時間の短縮にも効果がある．しきい値の設定を手動で行なうには，やはりある程度の時間が必要となる．少なくともショットの開始部分ではしきい値が定まっていないので追跡を開始できない．場合によってはショットが終了するまで適したしきい値を決めることができないこともある．しきい値の決定を計算機で行なうことで，これらを解決し基準フレームから追跡を開始することができる．

4. まとめ

動画像インデックスとして特に人物に着目し，そのインデックス抽出方法について検討した．映像をショット単位に分割し，静止画レベルで顔検出を行なう．ここまでで検出できないフレームに対しては色情報を用いて追跡する方法がよりロバストで適していると判断し，従来技術であるCamShiftアルゴリズムを応用した．インデキシングに応用する上での課題は自動化であり，CamShiftではしきい値の設定を人間が行なう必要があった．そこで本稿ではしきい値決定法を提案し，その自動化を試みた．その結果として，自動化の実現だけでなく，人間では微妙な判断がつきにくい映像や時間の短いショットに対して有効性を示すことができた．

今後の課題として，時間の経過により探索領域の色特徴が変化する場合の対処が考えられる．

参考文献

(1) 松橋 聡；顔領域抽出に有効な修正HSV表色系の提案，テレビジョン学会誌 Vol.49 No.6 pp.787-795(1995)

(2) 石橋 聡；背景参照画像と両眼視を用いた任意背景中の人物像抽出，テレビジョン学会誌 Vol.45 No.10 pp.1270-1276(1991)

(3) Gary R ; Computer Vision Face Traking For Use in a Perceputual User Interface ,Intel Technology Journal Q2 '98

(4) Intel ; Open Source Computer Vision Libraly, http://www.intel.com/research/mrl/research/cvlib/download.htm

(5) Henry A.Roeley ; Neural Network-Based Face Detection, IEEE, 1063-6919/96 203-208 1996

(6) Michael A. Smith ; Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques, IEEE, 1063-6919/97 775-781 1997