K-046

映像特徴に基づく撮影者が意図した人物被写体の推定
Inferring Camcorder User's Intended Subject of Persons Based on Visual Features

上柿　普史† 中島　悠太† 馬場口　登†
Hiroshi Uegaki Yuta Nakashima Noboru Babaguchi

1. まえがき

デジタルビデオカメラなどの普及に伴い，個人が大量の映像・画像を所有するようになった．しかし，大量の映像・画像から目的の映像を探し出し，視聴する際には，多大な労力を強いられることが多い．これに対して，映像・画像の効率の良い検索・要約を目的とした映像・画像コンテンツの解析手法が多く提案されている．

映像・画像コンテンツの解析手法の一つに ROI（Region of Interest）推定がある．ROI 推定は，映像・画像の特徴に基づいて関心領域を推定するものであり，効率の良い検索・要約を目的とした研究で広く利用されている [1, 2, 5, 3]．Oge らは多数の画像に対して，画像中の ROI を抽出し，抽出された ROI の特徴に基づいて画像を分類している [5]．Wang らは，映像中の ROI を抽出し，人間の視覚的な特性に基づく重要度によって大きさを変更して適切に配置することで 1 枚の画像を生成・提示する Video Collage を提案している [2]．これらの技術は，目立つ色や急激な変化に注意が向くという人間の視覚的な特性に基づいて Saliency Map を構築し，この Saliency Map を利用して ROI を推定・抽出している [5, 4]．

個人が撮影した映像・画像においては，撮影者はどのオブジェクトを撮影するかを決めている場合が多い．これを撮影者が意図したオブジェクトと呼ぶ．従って，撮影者が意図したオブジェクトに注目した映像の検索や要約は非常に有用であると考えられる．しかし，撮影者が意図して撮影したオブジェクトは人間の視覚的特性に基づく ROI とは必ずしも一致するものではない．

そこで，本稿ではオブジェクトとして人物被写体に注目し，撮影者が意図した人物被写体に基づく映像の検索・要約の実現を目指し，撮影者が意図した人物被写体を推定する手法を提案する．提案手法では，まず映像中の顔領域を抽出し，そこから得られる特徴量について撮影者の意図した人物被写体の映像特徴モデルを構築し，このモデルに基づいて撮影者の意図した人物被写体を推定する．さらに，実験により提案手法の有効性や課題を明らかにする．

2. 撮影者が意図した人物被写体について

本研究では，撮影者の撮影したかった人物被写体を撮影者の意図した人物被写体と呼び，それ以外の人物被写体を撮影者の意図していない人物被写体と呼ぶ．個人が撮影した映像においては，撮影者が意図した人物被写体のみではなく，撮影者が意図していない人物被写体も映像に映り込む．例えば，通行人などは撮影者が意図していない人物被写体である．

†大阪大学大学院工学研究科, Graduate School of Engineering, Osaka University

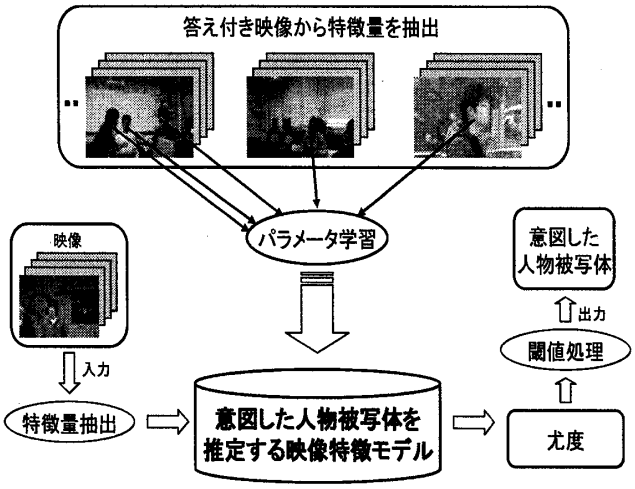

図 1: 提案手法の概要

本研究では，フレーム内の各人物被写体に対する撮影者の意図の有無は，

- 人物被写体のフレーム内での位置
- 人物被写体がフレーム内に占める面積
- 人物被写体が追いかけて撮影されているか

に反映されると考え，これらに関連する特徴量を用いて映像特徴モデルを構築し，撮影者が意図した人物被写体を推定する．映像特徴モデルのパラメータは答えつきの映像を用いて学習する．答えつきの映像とは，映像中の意図した人物被写体が撮影者によって指定された映像である．

3. 撮影者が意図した人物被写体の推定

図 1 に提案手法の概要を示す．映像特徴モデルは，映像が入力されると，映像中のそれぞれの人物領域に対して，その人物領域が撮影者が意図した人物被写体であるかに関する尤度をフレームごとに算出する．得られた尤度に対して閾値処理をすることにより，撮影者が意図した人物被写体を推定する．

3.1 特徴量抽出

提案手法では，撮影者が意図した人物被写体は顔を含めて撮影されるものとし，顔領域を人物領域として検出する．顔領域の検出には，Haar-like 特徴と AdaBoost を利用した手法 [6, 7] などが提案されているが，この手法は顔向きなどの変化に脆弱であり，フレーム内に含まれ

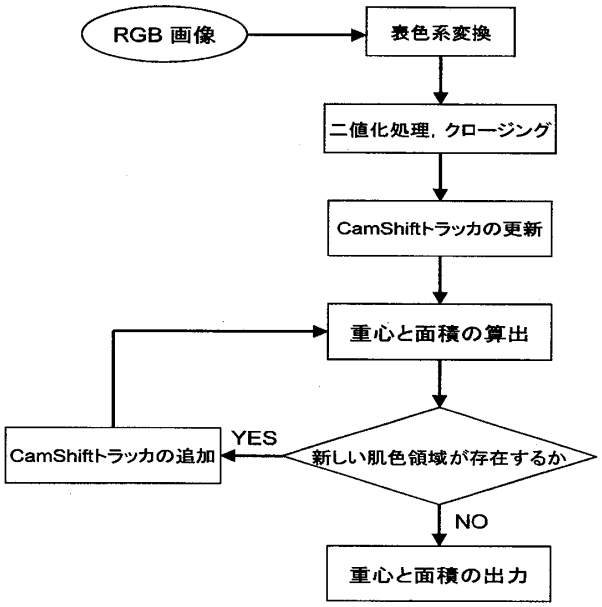

図 2: 肌色領域の検出と追跡の流れ

る全ての顔の検出は困難である．そこで提案手法では，フレームに含まれる肌色領域を顔領域として検出する．これにより，顔の向きなどの変化に頑健な顔の検出が可能となる．

提案手法では，入力映像中の肌色領域 $r_a(a = 1, 2, \cdots)$ に対して，第 n 番目のフレーム $f_n(n = 1, 2, \cdots)$ における以下の特徴量を抽出する．

- 面積 $S_{n,a}$
- 重心 $(g^x_{n,a}, g^y_{n,a})$
- カメラの動きに対する r_a の移動量 $F_{n,a}$

$F_{n,a}$ は，連続するフレーム間における重心 $(g^x_{n,a}, g^y_{n,a})$ の変化と，カメラの動きの大きさを利用して算出するため，顔領域の追跡が必要である．追跡には CamShift [8, 9] を用いる．CamShift は初期領域として与えられた領域の色特徴に基づきオブジェクトを追跡する手法であり，色特徴を逐次更新することによって，追跡対象のオブジェクトの見た目の時間的な変化に対して頑健な追跡が可能である．また，CamShift トラッカからは，追跡対象のオブジェクトが含まれる矩形の面積，重心座標が得られることから，これらを $S_{n,a}$，$(g^x_{n,a}, g^y_{n,a})$ として利用する．

入力映像中の肌色領域 r_a を追跡する CamShift トラッカの集合を **Tracker** $= \{t_a | a = 1, 2, \cdots\}$ とすると，f_n における各肌色領域の面積と重心座標は，以下のアルゴリズムを適用することで求められる（図 2）．

1. f_n を色相，彩度，明度の成分を持つ HSV 表色系に変換する．この変換により，色相値のみに対する閾値処理によって肌色領域を抽出できる．
2. 色相値に対する閾値処理により肌色領域を表す二値画像を生成する．フレーム f_n の i 番目の画素の色相値を $h_{n,i}$ とすると，

$$b_{n,i} = \begin{cases} 0 & (h_{n,i} > T_H \text{ のとき}) \\ 1 & (h_{n,i} \leq T_H \text{ のとき}) \end{cases} \quad (1)$$

により，肌色領域における画素値が 1 となる二値画像を生成する．ただし，T_H は色相値に対する閾値を，$b_{n,i}$ は二値画像の i 番目の画素値である．

次に，生成された二値画像に対して，クロージングを行う．クロージングとは二値画像に対する膨張処理の後に収縮処理を行う操作のことであり，この操作によって二値画像中の雑音を除去する．

3. 二値画像中の連結成分のうち，面積が閾値 T_S 以上のものを肌色領域 r_k とし，その面積を $S_{n,k}$，重心座標を $(g^x_{n,k}, g^y_{n,k})$ で表す．
4. **Tracker** 中の全 CamShift トラッカを更新し，$t_{a'} \in$ **Tracker** より得られる矩形の座標から追跡中の肌色領域の面積 $S_{n,a'}$ と重心座標を $(g^x_{n,a'}, g^y_{n,a'})$ を求める．
5. 二値画像から得られた各肌色領域中に CamShift トラッカが追跡していない肌色領域が存在するかを T_G を閾値として判定する．

$$|g^x_{n,a'} - g^x_{n,k}| \geq T_G \text{ または } |g^x_{n,a'} - g^y_{n,k}| \geq T_G \quad (2)$$

上式が成り立つとき，連結成分 r_k が追跡されていない肌色領域であると判定し，f_n において $(g^x_{n,k}, g^y_{n,k})$ を中心とする 1 辺 10 画素の正方形を初期領域として CamShift トラッカを生成して **Tracker** に追加する．

6. すべての $t_a \in$ **Tracker** について，t_a から得られる矩形の座標の面積 $S_{n,a}$ と重心座標を $(g^x_{n,a}, g^y_{n,a})$ を求め，出力する．

移動量 $F_{n,a}$ は，f_n における r_a の重心と，追跡によって対応付けられた f_{n-1} における r_a の重心の変化，およびカメラの動きの大きさを利用して次式で定義される．

$$F_{n,a} = \frac{\sqrt{(g^x_{n,a} - g^x_{n-1,a})^2 + (g^y_{n,a} - g^y_{n-1,a})^2}}{\sqrt{(m^x_n)^2 + (m^y_n)^2 + 1}} \quad (3)$$

ただし，m^x_n，m^y_n はそれぞれフレーム f_n におけるカメラの x 軸方向と y 軸方向の動きの大きさを表す．m^x_n，m^y_n の算出には Frederic らの手法 [10] を用いる．$F_{n,a}$ は，カメラが肌色領域を追いかけて撮影しているとき，すなわち肌色領域 r_a の重心の位置の変化がカメラの動きに対して小さいときに，小さな値をとる．

3.2 映像特徴モデル

撮影者が肌色領域 r_a に対応する人物被写体を意図して撮影した場合には，抽出された特徴量 $\mathbf{v}_{n,a} = \{g^x_{n,a}, g^y_{n,a}, S_{n,a}, F_{n,a}\}$ がそれぞれ独立であり，正規分布に従うものと仮定する．これより，撮影者がフレーム f_n

において，肌色領域 r_a に対応する人物被写体を意図して撮影した場合の条件付き確率密度分布は次式のように表される．

$$p(\mathbf{v}_{n,a}|I_{n,a}) = \frac{1}{\{(2\pi)^4|\Sigma|\}^{\frac{1}{2}}} \exp\left\{-\frac{(\mathbf{v}_{n,a}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{v}_{n,a}-\boldsymbol{\mu})}{2}\right\} \quad (4)$$

ただし，$\boldsymbol{\mu}$，Σ はそれぞれ $\mathbf{v}_{n,a}$ の平均と共分散を表し，$\boldsymbol{\mu} = (\mu_x, \mu_y, \mu_S, \mu_F)$，$\Sigma = \mathrm{diag}(\sigma_x^2, \sigma_y^2, \sigma_S^2, \sigma_F^2)$ である．

また，$I_{n,a}$ はフレーム f_n において肌色領域 r_a が撮影者の意図した人物被写体であるという事象を表す．これらのパラメータは，パラメータ推定用の答え付き映像から撮影者が意図した顔領域のみの特徴量を抽出し，得られた特徴量から最尤推定することにより決定する．

3.3 撮影者の意図した人物被写体の推定方法

撮影者が意図した人物被写体の推定は，

$$L_{n,a} = p(\mathbf{v}_{n,a}|I_{n,a}) \quad (5)$$

で定義される尤度に対して次式により行われる．

$$\begin{cases} A_{n,a} = 1 & (L_{n,a} \geq T_I \text{ のとき}) \\ A_{n,a} = 0 & (L_{n,a} < T_I \text{ のとき}) \end{cases} \quad (6)$$

ただし，T_I は閾値であり，$A_{n,a}$ はフレーム f_n における領域 r_a が撮影者の意図した人物被写体であると推定されたとき 1，意図した人物被写体でないと推定されたとき 0 の値をとる．

4. 実験

4.1 実験概要

実験に使用した映像は，3 人の被験者がビデオカメラ (SONY HDR-SR1) を使用し，長さ 1 分程度で撮影したもので，解像度は 720 × 480，30fps である．これらの映像はすべて人物を主に撮影した映像である．実験では 5 本の答えつき映像（映像 1〜映像 5）に対して交差検証を行った．すなわち，5 本の答え付き映像のうち，4 本をパラメータ推定用映像，1 本を評価用映像として推定を行うことを，すべての映像が評価用映像となるように繰り返した．閾値は経験的に $T_H = 30$，$T_A = 200$，$T_S = 2500$，$T_I = 3 \times 10^{-12}$ とした．

正解は，撮影後にそれぞれの映像の各フレームに対して，意図して撮影した人物を撮影者が指定したものである．ここで，提案手法は追跡された肌色領域に対して意図の有無を推定するため，正解として指定された人物と追跡された肌色領域を対応付ける必要がある．ここでは，CamShift トラッカが追跡する肌色領域の重心が，正解として指定された人物の顔領域に含まれている場合に，正しく追跡しているものとし，手作業で対応付けを行った．提案手法は，次式によって定義される適合率，再現率，追跡成功率によって評価する．

表 1: 適合率と再現率と追跡成功率

	映像 1	映像 2	映像 3	映像 4	映像 5	平均
適合率	0.62	0.56	0.91	0.91	0.99	0.80
再現率	0.87	0.94	0.80	0.96	0.90	0.89
追跡成功率	0.94	1.00	0.91	1.00	1.00	0.97

$$\text{適合率} = \frac{\text{正解推定数}}{\text{推定数}} \quad (7)$$

$$\text{再現率} = \frac{\text{正解推定数}}{\text{正解対象数}} \quad (8)$$

$$\text{追跡成功率} = \frac{\text{追跡成功数}}{\text{正解対象数}} \quad (9)$$

ただし，推定数を意図ありと推定した数，正解対象数を，ある人物が正解と指定されたフレーム数の，すべての人物についての合計，追跡成功数を，正解と指定された人物を正しく追跡していたフレーム数のすべての人物についての合計，正解推定数を，正解と指定された人物を正しく追跡しており，かつ意図ありと推定したフレーム数のすべての人物での合計とする．

4.2 実験結果

表 1 に適合率，再現率，追跡成功率を示す．追跡成功率についてはすべて高い値を示していることから，撮影者が意図した人物被写体の顔領域は正しく追跡できており，適合率，再現率は追跡の失敗による変動をほぼ含まないと考えられる．

映像 4，5 では，適合率，再現率ともに高い値となった．これらの映像は，映像中に登場する人物被写体の数が少なく，人物被写体以外の肌色領域が少ない映像である．図 3 に肌色追跡の結果も含めた映像 5 の画像例を示す．

映像 1，2，3 は，顔以外の肌色領域が多数存在する映像である．映像 2 では，顔以外の肌色領域の重心が，学習によって得られた正規分布の平均と近い位置に多数存在したため，顔以外の肌色領域に対して $L_{n,a}$ が高い値となり，適合率が下がった．しかし，正解と指定された人物被写体は，高い精度で推定できており，再現率は高い値となった．映像 1 では，顔以外の肌色領域に対して $L_{n,a}$ が高い値となったことに加え，正解と指定された人物被写体の肌色領域の重心が，学習によって得られた正規分布の平均と離れた位置に多数存在したため，適合率，再現率がともに下がった．図 4 に肌色追跡の結果も含めた映像 1 の画像例を示す．この結果から，特徴量間の独立を仮定した正規分布では，さまざまな撮影技法で撮影された映像における，撮影者の意図した被写体から得られる特徴量を十分にモデル化できないことが分かる．従って，適合率，再現率のさらなる向上のためには，映像特徴モデルとしてガウス混合分布などの利用を検討する必要がある．

また，ズームして撮影された人物被写体は，撮影者が意図した人物被写体であると考えられるが，提案手法では，図 3 の 260 フレームのように，人物被写体をズーム

図 3: 映像 5 の画像例

図 4: 映像 1 の画像例

して撮影した場合に，正解と指定された人物被写体に対して，意図なしと推定する誤りが多い．この原因は，前述のモデルの問題に加え，実験で用いたパラメータ学習用映像中に人物被写体をズームして撮影した映像が少かったため，肌色領域の面積 $S_{n,a}$ の分散が広がらず，$S_{n,a}$ の大きな値に対して $L_{n,a}$ が小さな値となったためであると考えられる．

5. まとめ

本稿では，撮影者の意図した人物被写体による映像の効率的な検索・要約を目指し，映像中の顔領域から抽出される特徴量についての映像特徴モデルに基づいて撮影者の意図した人物被写体を推定する手法を提案した．実験の結果，適合率は平均 0.80，再現率は平均 0.89 で推定が可能であった．顔以外の肌色領域が少ない映像に対しては適合率，再現率はともに高い値となり，良好な結果が得られた一方，顔以外の肌色領域が多数存在する映像に対しては，多くの場合，適合率，再現率が下がった．これらの問題に対して，1) 抽出された特徴量のモデルの検討，2) 人物被写体の人数に応じた映像特徴モデルの構築，3) 人物領域検出・追跡手法の改善などが考えられる．

なお，本研究の一部は，科研費の補助による．

参考文献

[1] Tang Wang, Tao Mei, Xian-Sheng Hua, Xue-Liang Liu, and He-Qin Zhou, "Video collage: a novel presentation of video sequence," *In Proc. IEEE International Conference on Multimedia and Expo*, pp.1479–1482, July, 2007.

[2] Yusuo Hu, Xing Xie, Zonghai Chen, and Wei-Ying Ma, "Attention model based progressive image transmission," *In Proc. IEEE International Conference on Multimedia and Expo*, Vol.2, pp.1079–1082, June, 2004.

[3] Yu-Fei Ma, Lie Lu, Hong-Jiang Zhang, and Mingjing Li, "A user attention model for video summarization," *In Proc. The Tenth ACM International Conference on Multimedia* , pp.533–542, 2002.

[4] Laurent Itti, Christof Koch, and Ernst Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.20, No.11, pp.1254–1259, November, 1998.

[5] Oge Marques, Liam M. Mayron, Gustavo B. Borba, and Humberto R. Gamba, "Using visual attention to extract regions of interest in the context of image retrieval," *In Proc. The ACM Southeast Regional Conference*, pp.638–643, 2006.

[6] Rainer Lienhart and Jochen Maydt, "Rapid object detection using a boosted cascade of simple features," *In Proc. The 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol.1, pp.511, 2001.

[7] Rainer Lienhart and Jochen Maydt, "An extended set of haar-like features for rapid object detection," *In Proc. IEEE International Conference Image Processing*, Vol.1, pp. 900–903, 2002.

[8] Dorin Comaniciu, Visvanathan Ramesh, and Peter Meer, "Real-time tracking of non-rigid objects using mean shift," *In Proc. IEEE Conference on Computer Vision and Pattern Recognition*, Vol.2, pp.142–149, 2000.

[9] Gary R. Bradski, "Computer vision face tracking for use in a perceptual user interface," *Intel Technology Journal*, 1998.

[10] Frederic Dufaux and Janusz Konrad, "Efficient, robust, and fast global motion estimation for video coding," *IEEE Transactions on Image Processing*, Vol.9, No.3, pp.497–501, March, 2000.