カメラの動きと映像特徴からの撮影者が意図した領域の推定

上柿　普史†　　中島　悠太†　　馬場口　登†

† 大阪大学 大学院工学研究科　〒 565–0871 大阪府吹田市山田丘 2–1
E-mail: †{uegaki,nakashima,babaguchi}@nanase.comm.eng.osaka-u.ac.jp

あらまし　映像の要約や編集などのアプリケーションでは，重要な領域を推定し，その領域に基づいて映像を処理することから，重要な領域の推定手法が必要とされている．一方，YouTube などにはモバイルカメラで撮影された映像が多く投稿されており，これらの映像には「自分の子供を撮影したい」などの撮影意図が存在する．このとき「子供の領域」のような撮影者が意図的に撮影した領域（意図領域）は撮影意図に不可欠である．本稿では，意図領域を重要な領域として推定する手法について述べる．提案手法では，撮影者はフレーム中で意図領域を適切に配置するようにカメラを動かすことから，撮影者の行動が反映されるカメラの動きと映像特徴のモデルを構築し，このモデルに基づいて意図領域を推定する．さらに，実験により提案手法の有効性を示す．

キーワード　意図領域，意図マップ，慣性計測デバイス，顕著性マップ

1. はじめに

近年，映像の要約 [1], [2]，編集 [3]，リサイズ [4], [5]，圧縮 [6]〜[8]，プライバシー保護 [9] など，映像処理のアプリケーションが多く提案されている．これらのアプリケーションでは，映像中の重要な領域を定義し，その定義に従って推定された領域に基づいて映像を処理する．例えば映像の要約では，重要な領域を抽出し，それらを適切な大きさ，位置で配置することにより一枚の画像を生成する [1]．また，映像のリサイズでは，映像を携帯電話のような小さな画面で快適に視聴するために，重要な領域のみを切り出した映像を生成する [5]．このように，映像中の重要な領域の定義と推定手法は，これらのアプリケーションの根幹をなす要素技術である．

これらのアプリケーションの多くでは，視聴者の視覚が注意を向ける領域が重要であると定義し，視覚的に顕著な領域を視覚的注意モデルと呼ばれる技術を利用して推定する．これは，映像・画像の明るさや色，動きなどの低レベル特徴を利用して，視覚的に顕著な領域に注意を向けるという生物の視覚システムの持つ性質を模倣することから [10]〜[13]，視聴者の観点からの重要な領域であると言える．

一方，YouTube に代表される動画投稿サイトには，モバイルカメラで撮影された映像が数多く投稿されている．このような映像において，撮影者は，例えば「自分の子供を撮影したい」「風景を記録として残したい」などの撮影意図を持つ．このとき,「子供」や「風景」などの被写体は撮影意図の達成に必要不可欠である．本稿では，これらの被写体に対応する映像中の領域を撮影者が意図した領域（意図領域）と呼ぶ．意図領域は撮影者の観点からの重要な領域であり，前述の映像の要約や編集などのアプリケーションにおいて撮影意図を損なわないためには，意図領域に基づく処理が必要となる．

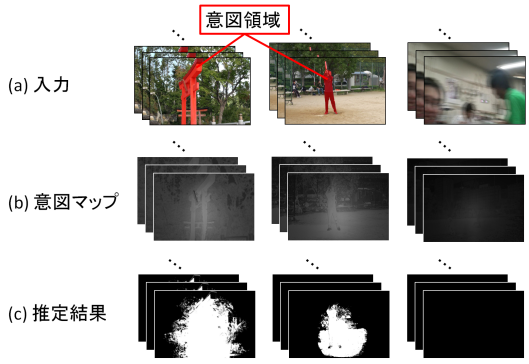

図 1　推定結果の例

そこで上柿ら [14], [15] や Nakashima ら [16] は，特に人物被写体に着目した意図領域の推定手法として，撮影者が意図した人物被写体の推定を提案した．しかし，これらの手法は人物検出に基づくものであるため，一般の被写体が意図領域になり得る映像には適用できない．

そこで本稿では，カメラの動き，および人物検出に依らない映像特徴を用いることで，映像中の一般の意図領域を推定する手法について述べる．提案手法では，撮影者は意図領域を映像中で適切な位置・大きさとするようカメラを動かすことに着目する．すなわち，意図領域を撮影する際のカメラの動きには一定の傾向があり，また，意図領域とその他の領域の位置・大きさの分布に差があると考え，訓練集合を用いてカメラの動きと映像特徴のモデルをそれぞれ構築する．入力された映像のフレームに対して，図 1(b) に示すような各画素の意図度合いを表す意図マップを生成し，閾値処理により意図領域を推定する．また，映像からのカメラの動きの推定 [21] は動物体の影響などにより困難であるため，スマートフォンのように慣性計測デバイスを搭載した機器で映像を撮影

する機会が増えたことを考慮して，カメラに取り付けられた慣性計測デバイスから得られる加速度，角速度をカメラの動きとして利用する．

提案手法の新規性は以下のようにまとめられる．(i) 撮影意図に不可欠な領域，すなわち，意図領域という概念を導入する．意図領域は撮影者の観点からの重要な領域であり，人物被写体に限定しない一般の意図領域は本稿で初めて提案する新しい概念である．(ii) 提案手法では慣性計測デバイスから得られる加速度，角速度をカメラの動きとして利用する．これは慣性計測デバイスの新しい応用法を提案するものである．

2. 関 連 研 究

視覚的注意モデルに関する研究は多くなされている [10]〜[13]．Itti らは，霊長類の視覚システムが明るさ，色，方向などの変化に強く反応する性質を模倣する視覚的注意モデルを構築した [12]．また，Liu らは，映像中の明るさや色の分布などの特徴量を Conditional Random Field を利用して統合することで，視覚的に顕著な領域を抽出する手法を提案した．このような視覚的注意モデルは多くのアプリケーションで用いられている [1], [3]〜[8]．Wang らは，視覚的注意モデルを用いて映像中のフレームや画像領域を選択し，適切な大きさ，位置で一枚の画像に配置することで映像を要約する手法を提案した [1]．Liu ら [5] や Fan ら [4] は，視覚的注意モデルを用いて決定した映像の一部を切り出すことで携帯電話のような小さな画面に適した映像の提示を可能にした．しかし，視覚的注意モデルは撮影意図を考慮しないため，これらのアプリケーションにおける領域の抽出や圧縮の際に撮影意図が損なわれる可能性がある．提案手法は，カメラの動きなどの利用により直接的に意図領域を推定するため，このような問題を軽減できる．

映像中の意図領域を推定する手法として，撮影者が意図した人物被写体の推定手法が提案されている [14]〜[17]．例えば上柿らは，肌色検出により人物の顔領域を検出し，顔領域の面積，重心，カメラの動きと顔領域の動きの比を用いて，撮影者が意図した人物被写体を推定する手法を提案した [14], [15]．また，Nakashima らは，映像中の撮影者が意図しない人物被写体に対するプライバシー保護システムを提案した [9]．撮影者が意図した人物被写体の推定手法は，肌色検出や人物検出に基づくものであるため，人物領域のみを対象とする．しかし，映像には様々な意図領域が存在することから，さまざまなアプリケーションに適用するためには，人物領域のみでは不十分である．提案手法は一般の意図領域の推定が可能であり，これらのアプリケーションにおいて，視覚的注意モデルと置き換えて利用できる．

3. 意図領域の推定

撮影者はフレーム中で意図領域が適切な位置・大きさ

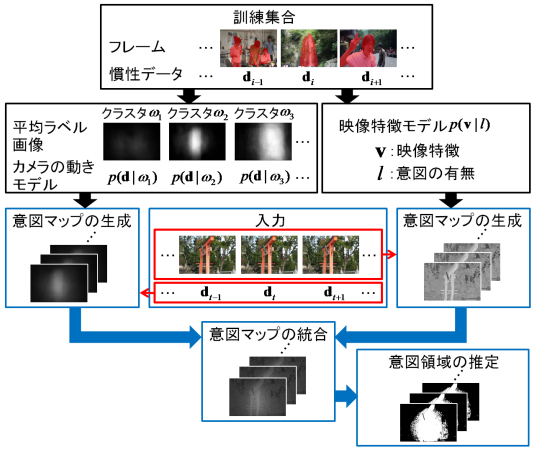

図 2　提案手法の概要

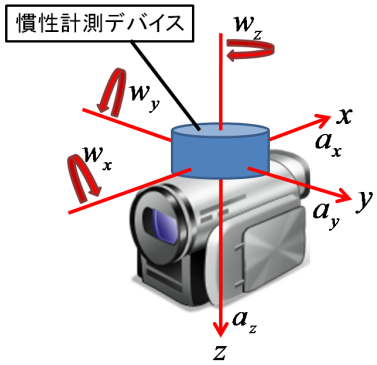

図 3　カメラ座標系と加速度，角速度の向き

となるようカメラを動かす．例として，動く人物を撮影する際にはその人物を追うようにカメラを動かすことでフレームの中央付近に配置し，また，風景を撮影する際には風景全体を映像に収めるためにゆっくりとカメラを左右に動かすなどが挙げられる．撮影者のこのような行動は，カメラの動き，および意図領域を構成する物体の領域の位置・大きさに関する映像特徴に反映され，撮影者ごとのカメラの動きなどの差は小さいものと考えられる．従って，カメラの動きの分布は意図領域の位置・大きさなどによって異なり，フレームを構成する物体領域の映像特徴の分布は，その物体を意図的に撮影したか否かによって異なる．

この考察に基づき，提案手法では図 2 に示すように，意図領域を示すラベル画像が付与されたフレームとそのフレームに対応する慣性データからなる訓練集合を用いて慣性データが従う分布を表すカメラの動きモデル，および物体の領域の位置・大きさに関する映像特徴が従う分布を表す映像特徴モデルを構築する．慣性データは，カメラに取り付けられた慣性計測デバイスから得られる加速度，角速度から構成されるものであり，提案手法では慣性データをカメラの動きとして用いる．入力としてフレームと慣性データが与えられると，構築されたそれぞれのモデルを用いて，各画素における撮影者の意図度合いを表す二つの意図マップを生成し，それらの統合により得られる意図マップに閾値処理を適用することで意

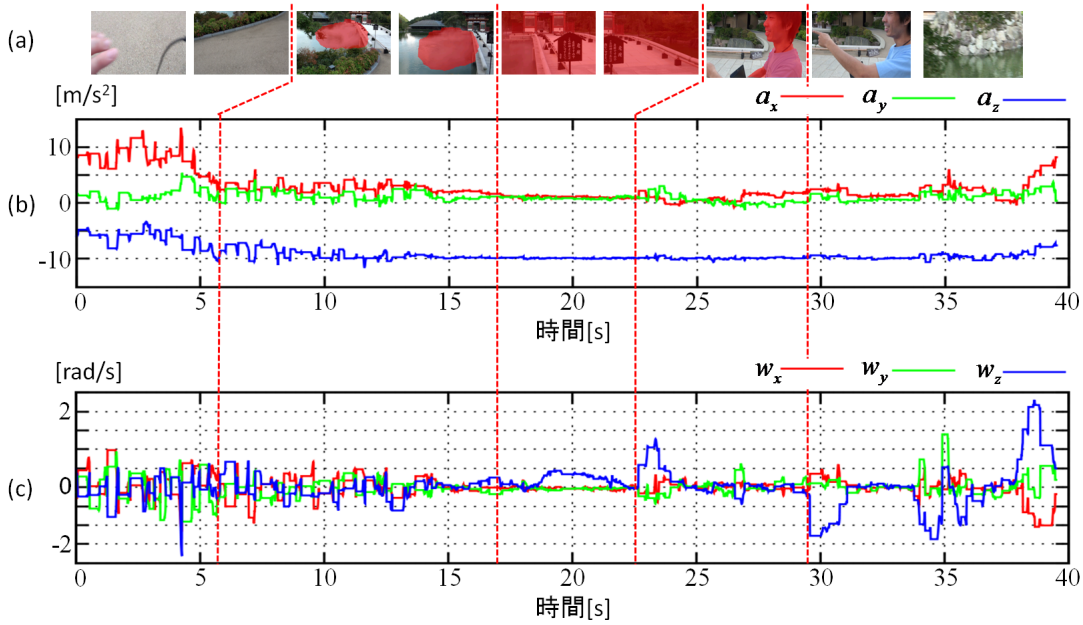

図 4　(a) 撮影された映像の例．赤色の領域は意図領域を表す．(b) 加速度 (a_x, a_y, a_z) の例．(c) 角速度 (w_x, w_y, w_z) の例．

図領域を推定する．

次節から，慣性データ，カメラの動きモデル，映像特徴モデル，および意図マップの生成について詳述する．

3.1　慣性データ

提案手法では，カメラに取り付けられた慣性計測デバイスから，図 3 に示すカメラ座標系における加速度 (a_x, a_y, a_z) と角速度 (w_x, w_y, w_z) を取得する．それぞれの向きは図 3 に示す通りである．

図 4 に映像と映像に対応する加速度 (a_x, a_y, a_z) と角速度 (w_x, w_y, w_z) の例を示す．なお，a_z は重力加速度のため -10 m/s^2 付近の値となる．この映像では，図 4(a) に示すように 0 秒から 6 秒までは意図領域は存在しない．このとき，加速度，角速度は大きく変化する．6 秒から 29 秒までは意図領域が存在し，加速度，角速度の変化が小さい．特に，17 秒から 23 秒のように風景を撮影しているときは，意図領域はフレーム全体となり，このとき，加速度は大きな変化がなく，角速度はカメラの左右の回転を示す w_z が緩やかに変化する．29 秒からは再び意図領域がなくなり，加速度，角速度の変化が大きくなる．この例から，慣性計測デバイスから得られる加速度，角速度は意図領域と相関があると考えられる．

提案手法では，カメラの動きの時間的な変化を考慮するために，図 5 のように各フレームに対して N_{D} 個の加速度，角速度のサンプルを割り当てる．映像のフレームレートと加速度，角速度のサンプリングレートが異なることから，映像の t 番目のフレームを f_t，τ 番目の加速度，角速度のサンプルを

$$\mathbf{u}_\tau = (a_{x,\tau}, a_{y,\tau}, a_{z,\tau}, w_{x,\tau}, w_{y,\tau}, w_{z,\tau})^\top \qquad (1)$$

とする．ただし，$\top$ はベクトルの転置を表す．加速度，角速度は手ぶれなどの影響を受けるので，次式により $\mathbf{u}_\tau$ の移動平均をとることにより，影響を軽減する．

$$\mathbf{v}_\tau = \frac{1}{T} \sum_{\tau'=\tau-T/2}^{\tau+T/2-1} \mathbf{u}_{\tau'} \qquad (2)$$

ただし，T は平均されるサンプル数を表す定数である．フレーム f_t に割り当てられる慣性データ $\mathbf{d}_t$ は次式により与えられる．

$$\mathbf{d}_t = (\mathbf{v}_{\tau_t}^\top, \mathbf{v}_{\tau_t+1}^\top, \cdots, \mathbf{v}_{\tau_t+N_{\mathrm{D}}-1}^\top)^\top \qquad (3)$$

ただし，慣性計測デバイスから得られる加速度，角速度のサンプリングレートを S サンプル毎秒，撮影された映像のフレームレートを F フレーム毎秒とすると，τ_t は $(\tau_t + N_{\mathrm{D}}/2)/S$ が t/F と最も近くなるように選ぶ．

3.2　カメラの動きモデル

本研究では，意図領域の位置，大きさ，形状と慣性データには相関があり，類似する位置，大きさ，形状の意図領域を撮影する際の慣性データの分布は類似すると仮定する．この仮定から，提案手法では意図領域の位置，大きさ，形状ごとにカメラの動きをモデル化する．具体的には，訓練集合中のラベル画像を意図領域の位置，大きさ，形状に関する類似度によりクラスタリングし，各クラスタごとに慣性データの従う分布を推定することでカメラの動きモデルを構築する．

訓練集合中の i 番目のフレームに対するラベル画像を f_i により表し，ラベル画像の n 番目の画素における意図の有無を $f_{i,n} \in \{0, 1\}$ で表す．ただし，$f_{i,n} = 0$ のとき意図なし，$f_{i,n} = 1$ のとき意図ありを表す．提案手法では，ラベル画像 f_i を初期値依存がなく，安定した結果が得られる Affinity Propagation [18] を用いてクラスタリ

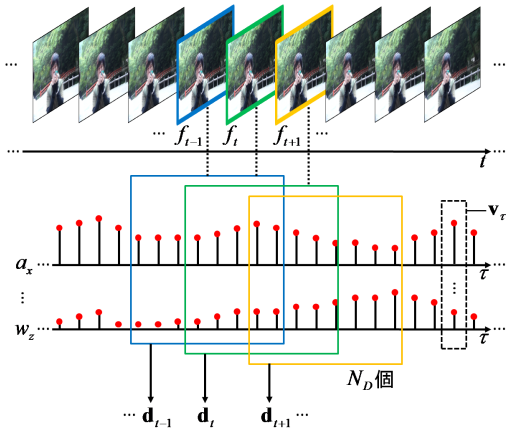

図 5　慣性データの生成

ングする．Affinity Propagation では，ラベル画像 f_i と f_j 間の類似度 $s_{i,j}$，およびクラスタ数を決定する定数であるプリファレンス $s_{i,i}$ を利用する．$s_{i,j}$ $(i \neq j)$ は次式により定義される．

$$s_{i,j} = \frac{1}{|\Omega_{\mathrm{A}}|} \sum_{n \in \Omega_{\mathrm{A}}} (f_{i,n} \oplus f_{j,n}) \tag{4}$$

ただし，Ω_{A} はラベル画像中の全ての画素の集合，$|\Omega_{\mathrm{A}}|$ は Ω_{A} に属する画素の数を表す．演算子 $\oplus$ は以下のように定義される排他的論理和を表す．

$$f_{i,n} \oplus f_{j,n} = \begin{cases} 0, & \text{if } f_{i,n} = f_{j,n} \\ 1, & \text{otherwise} \end{cases} \tag{5}$$

Affinity Propagation によりラベル画像は N_{CLS} 個のクラスタに分類されるものとし，各クラスタのラベルを ω_k $(k = 1, 2, \cdots, N_{\mathrm{CLS}})$ とする．

次に，各クラスタについて慣性データの分布をカーネル密度推定により求める．ラベル画像 f_i に対応する慣性データを $\mathbf{d}_i$ とすると，クラスタ ω_k が与えられたときの慣性データ $\mathbf{d}$ の分布は次式により与えられる．

$$p(\mathbf{d}|\omega_k) = \frac{1}{|C_k|} \sum K(\mathbf{d} - \mathbf{d}_i) \tag{6}$$

ただし，上式において総和は $f_i \in C_k$ を満たす i について計算される．また，$K(\mathbf{d})$ はガウシアンカーネルであり，

$$K(\mathbf{d}) = \frac{1}{\sqrt{(2\pi)^{N_{\mathrm{CLS}}-1}|\Phi|}} \exp\left(-\frac{\mathbf{d}^T \Phi^{-1} \mathbf{d}}{2}\right) \tag{7}$$

で与えられる．ただし，Φ は $6N_{\mathrm{D}} \times 6N_{\mathrm{D}}$ の分散共分散行列であり，σ^2 を定数として $\Phi = \mathrm{diag}(\sigma^2, \sigma^2, \cdots, \sigma^2)$ により与えられる．

3.3　映像特徴モデル

撮影者は意図領域がフレーム中で適切な位置・大きさとなるように映像を撮影するため，意図領域はその他の領域と比べて，領域の位置・大きさに特徴があると考えられる．提案手法では，色に基づいて分割した領域が物体の領域に対応するものとし，分割された領域に属する画素の位置の平均を物体の領域の位置を表す特徴量として，分散を物体の大きさ・形状を表す特徴量として抽出する．また，これらの特徴量が従う分布を訓練集合から推定し，映像特徴モデルとする．

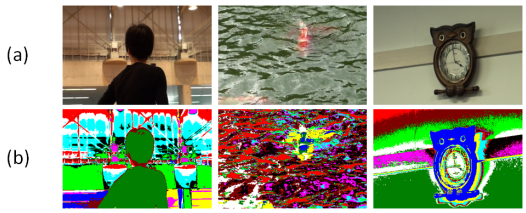

図 6　画素の色に基づく領域分割の例．(a) フレーム (b) 画素の色に基づいて領域分割した画像．分割した領域ごとに異なる色で表示している．

まず，フレーム中の物体の領域を抽出するために，画素の色に基づいて k 平均法によりフレーム中の画素を M 個の領域にクラスタリングする．フレーム中の n 番目の画素における色を

$$\mathbf{c}_n = (R_n, G_n, B_n) \tag{8}$$

と表す．ただし，R_n，G_n，B_n はそれぞれ n 番目の画素における赤色成分，緑色成分，青色成分である．提案手法では，k 平均法で用いられる $\mathbf{c}_n$，$\mathbf{c}_{n'}$ 間の距離として，次式のユークリッド距離 $\Delta(\mathbf{c}_n, \mathbf{c}_{n'})$ を用いる．

$$\Delta(\mathbf{c}_n, \mathbf{c}_{n'}) = |\mathbf{c}_n - \mathbf{c}_{n'}| \tag{9}$$

図 6(a) に例示するフレームを画素の色に基づいて領域分割した結果を図 6(b) に示す．これより，フレームを画素の色に基づく領域分割結果が，物体の領域とおおまかに対応することがわかる．分割された m 番目の領域の画素の集合を Ω_m と表す．ただし，$m = 1, 2, \cdots, M$ である．

次に，Ω_m に属する画素の位置の平均 $\mathbf{g}_m$，および分散 Σ_m を以下のように算出する．

$$\mathbf{g}_m = \frac{1}{|\Omega_m|} \sum_{n \in \Omega_m} \mathbf{X}_n \tag{10}$$

$$\Sigma_m = \frac{1}{|\Omega_m|} \sum_{n \in \Omega_m} (\mathbf{X}_n - \mathbf{g}_m)(\mathbf{X}_n - \mathbf{g}_m)^\top \tag{11}$$

ただし，$\mathbf{X}_n$ は n 番目の画素の座標を表す列ベクトル，$|\Omega_m|$ は Ω_m に属する画素の数を表す．$\mathbf{g}_m$，Σ_m の要素をそれぞれ

$$\mathbf{g}_m = (g_m^X, g_m^Y), \qquad \Sigma_m = \begin{pmatrix} \Sigma_m^{XX} & \Sigma_m^{XY} \\ \Sigma_m^{YX} & \Sigma_m^{YY} \end{pmatrix} \tag{12}$$

とすると，$\Sigma_m^{XY} = \Sigma_m^{YX}$ を考慮して，映像特徴 $\mathbf{v}_m$ は次式により与えられる．

$$\mathbf{v}_m = (g_m^X, g_m^Y, \Sigma_m^{XX}, \Sigma_m^{YY}, \Sigma_m^{XY})^\top \tag{13}$$

ここで，分割された領域に対する意図の有無を表す

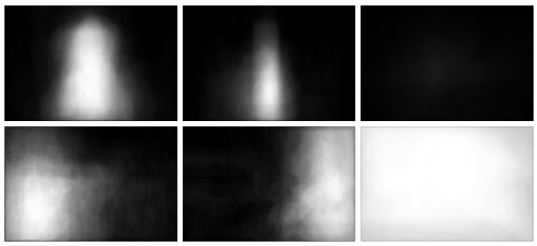

図 7　平均ラベル画像の例

ラベル $l \in \{0,1\}$ が与えられたときの映像特徴 $\mathbf{v}$ の分布は次式のガウス混合モデル (GMM: Gaussian Mixture Model) に従うと仮定する．

$$p(\mathbf{v}|l) = \sum_{k=1}^{K} w_l^k \mathcal{N}(\mathbf{v}|\mu_l^k, \Psi_l^k) \tag{14}$$

ただし，$K, w_l^k, \mu_l^k, \Psi_l^k$ はそれぞれ混合数，重み，平均，分散共分散行列を表す．また，$\mathcal{N}(\cdot|\mu, \Psi)$ は平均 μ，分散共分散行列 Ψ の多次元正規分布を表す．

GMM のパラメータは，訓練集合中のフレームを同様に分割して得られる領域 Ω_m から算出した映像特徴 $\mathbf{v}_m$ と，その領域における意図の有無を示すラベル l_m から EM アルゴリズムを用いて推定する．このとき，l_m は訓練集合中のラベル画像 f から以下のように決定する．Ω_{GT} を f によって与えられる意図領域の画素の集合，$|\Omega_{\mathrm{GT}}|$ を Ω_{GT} に属する画素数とすると，Ω_m と Ω_{GT} の積集合の画素数 $|\Omega_m \cap \Omega_{\mathrm{GT}}|$ と $|\Omega_m|$ の比率 Q_m は

$$Q_m = \frac{|\Omega_m \cap \Omega_{\mathrm{GT}}|}{|\Omega_m|} \tag{15}$$

と表される．Q_m を用いて，以下のように l_m を決定する．

$$l_m = \begin{cases} 0, & \text{if } Q_m < D \\ 1, & \text{otherwise} \end{cases} \tag{16}$$

ただし，D は閾値であり，$l_m = 0$ のとき意図なし，$l_m = 1$ のとき意図ありを表す．EM アルゴリズムによる学習では，上記の操作を訓練集合中のすべてのフレームに対して適用して得られる映像特徴と意図の有無を表すラベルを用いる．

3.4　意図マップの生成

入力として与えられたフレーム f_t と慣性データ $\mathbf{d}_t$ から意図マップを生成する．

(i) カメラの動きモデルを用いた意図マップの生成：カメラの動きモデルを用いた意図マップは，ラベル画像のクラスタごとに平均ラベル画像を生成し，これらを $\mathbf{d}_t$ が与えられたときのクラスタ ω_k の事後確率の重みで足し合わせることにより生成される．

クラスタ ω_k の平均ラベル画像 I_k は，ω_k に分類されたラベル画像の集合を C_k として，次式により算出される．

$$I_{k,n} = \frac{1}{|C_k|} \sum_{f \in C_k} f_n \tag{17}$$

ただし，$|C_k|$ は C_k に含まれるラベル画像の数，f_n はラベル画像 f の n 番目の画素を表す．図 7 に平均ラベル画像の例を示す．

ここで，式 (6) とベイズの定理より，$\mathbf{d}_t$ が与えられたときのクラスタ ω_k の事後確率は

$$p(\omega_k|\mathbf{d}_t) = \frac{p(\omega_k)p(\mathbf{d}_t|\omega_k)}{p(\mathbf{d}_t)} \tag{18}$$

と表される．ただし，ω_k の事前確率 $p(\omega_k)$ は

$$p(\omega_k) = \frac{|C_k|}{\sum_{k=1}^{N_{\mathrm{CLS}}} |C_k|} \tag{19}$$

とする．また，$p(\mathbf{d}_t)$ は

$$p(\mathbf{d}_t) = \sum_{k=1}^{N_{\mathrm{CLS}}} p(\omega_k)p(\mathbf{d}_t|\omega_k) \tag{20}$$

により与えられる．f_t に対応するカメラの動きモデルを用いた意図マップ CIM_t は，平均ラベル画像の事後確率による重み付き和として次式により与えられる．

$$CIM_{t,n} = \sum_{k=1}^{N_{\mathrm{CLS}}} I_{k,n}\, p(\omega_k|\mathbf{d}_t) \tag{21}$$

ただし，$CIM_{t,n}$ は CIM_t における n 番目の画素値を表す．

(ii) 映像特徴モデルを用いた意図マップの生成：f_t における映像特徴を $\mathbf{v}_{t,m}$，意図の有無を示すラベルを $l_{t,m}$ とする．式 (14) とベイズの定理より，$\mathbf{v}_{t,m}$ が与えられたときの $l_{t,m}$ の事後確率は

$$p(l_{t,m}|\mathbf{v}_{t,m}) = \frac{p(l_{t,m})p(\mathbf{v}_{t,m}|l_{t,m})}{p(\mathbf{v}_{t,m})} \tag{22}$$

で与えられる．ただし，$l_{t,m} \in \{0,1\}$ に対して，$p(l_{t,m}) = 0.5$ とする．また，$p(\mathbf{v}_{t,m})$ は

$$p(\mathbf{v}_{t,m}) = \sum_{l_{t,m} \in \{0,1\}} p(l_{t,m})p(\mathbf{v}_{t,m}|l_{t,m}) \tag{23}$$

と表される．

この事後確率から，f_t に対応する映像特徴モデルを用いた意図マップ VIM_t を次式により生成する．

$$VIM_{t,n} = p(l_{t,m} = 1|\mathbf{v}_{t,m}) \quad \text{for all } n \in \Omega_{t,m} \tag{24}$$

ただし，$\Omega_{t,m}$ は f_t における分割された領域の画素の集合，$VIM_{t,n}$ は VIM_t における n 番目の画素値を表す．

(iii) 意図マップの統合：カメラの動きモデルを用いた意図マップ CIM_t と映像特徴モデルを用いた意図マップ VIM_t を次式により統合することで，統合した意図マップ UIM_t を生成する．

$$UIM_{t,n} = \alpha CIM_{t,n} + (1-\alpha)VIM_{t,n} \tag{25}$$

ただし，$UIM_{t,n}$ は UIM_t における n 番目の画素値，$\alpha \in [0,1]$ は $CIM_{t,n}$ と $VIM_{t,n}$ のそれぞれの寄与率を決定する定数を表す．

提案手法は，$UIM_{t,n}$ が次式を満たすとき，f_t における n 番目の画素が意図領域に含まれると判定する．

$$UIM_{t,n} > TH \tag{26}$$

ただし，TH は閾値である．

4. 実　　験

本章では，提案手法の有効性を示すために行った実験について述べる．

4.1 実験データ

映像はデジタルビデオカメラ SONY Handycam HDR-CX550V を用いて撮影した．映像のフレームレート F は 30 フレーム毎秒，フレームサイズは 720 × 480 画素である．慣性計測デバイスには，MicroStrain Inertia-Link を用いた．加速度，角速度のサンプリングレート S は 100 サンプル毎秒である．なお，映像と慣性計測デバイスから得られる加速度，角速度のサンプル列との同期は手作業で行った．

上述の機器を用いて 11 名の撮影者が，46 本の映像を撮影した．映像 1 本の長さは 1 分程度であり，映像には人物や風景などを撮影した様々なシーンが含まれる．ラベル画像を作成するために，11 名の撮影者が自分の撮影した映像に対して，意図領域を指定した．指定作業は，撮影者が意図領域を塗りつぶすことで行った．ただし，すべてのフレームに対して指定作業を行うのは時間がかかることと，隣接するフレームは類似していることを考慮して，10 フレーム毎にラベル画像を作成し，6227 枚のラベル画像を得た．

4.2 評 価 方 法

映像 46 本に対して，映像 1 本から得られるラベル画像が付与されたフレームとそのフレームに対応する慣性データをテスト集合，残り 45 本を訓練集合とする交差検定法により提案手法の推定精度を評価する実験を行った．パラメータ T，$s_{i,i}$ は経験的に $T = 15$，$s_{i,i} = -7$ とした．このとき，N_{CLS} は交差検定法の各試行において 12 または 13 となった．また，パラメータ N_{D} は慣性データ $\mathbf{d}_t$ がおよそ 1 秒間の加速度，角速度のサンプル列を含むよう $N_{\mathrm{D}} = 101$ と定めた．さらに，ラベル l_m の決定における閾値は経験的に $D = 0.9$ とした．これは，意図領域は多くの場合フレームの中央付近に集中するため，$p(\mathbf{v}|l=1)$ は $p(\mathbf{v}|l=0)$ よりも狭い分布となると考えられるが，D の値が小さいと意図領域でない領域を意図領域と判定する誤判定が増加して，$p(\mathbf{v}|l=1)$ と $p(\mathbf{v}|l=0)$ の差が小さくなり，推定精度が低下すると考えたためである．

図 8　(a) Ω_{GT} の例．赤色の領域は意図領域を表す．(b) Ω_{INF} の例．白色の領域は式 (26) を満たす領域を表す．

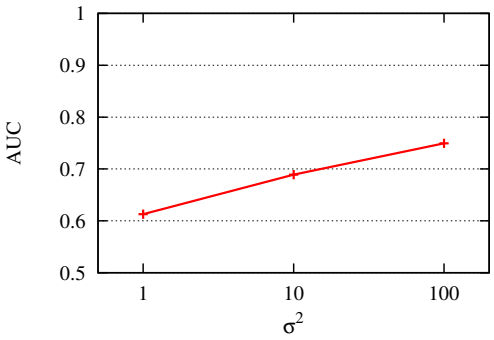

図 9　カメラの動きモデルを用いた意図マップの AUC

また，推定精度の評価には，True Positive Rate (TPR) と False Positive Rate (FPR) から Receiver Operating Characteristic (ROC) 曲線を生成し，ROC 曲線が囲む右下の領域の面積 (AUC: Area Under the Curve) を算出した．ただし，TRP と FPR は次式により与えられる．

$$\mathrm{TPR} = \frac{|\Omega_{\mathrm{GT}} \cap \Omega_{\mathrm{INF}}|}{|\Omega_{\mathrm{GT}}|}, \quad \mathrm{FPR} = \frac{|\bar{\Omega}_{\mathrm{GT}} \cap \Omega_{\mathrm{INF}}|}{|\bar{\Omega}_{\mathrm{GT}}|} \tag{27}$$

なお，AUC が 1 に近いほど推定精度が高いことを意味する．Ω_{INF} は式 (26) を満たす画素の集合であり，$\bar{\Omega}_{\mathrm{GT}}$ は Ω_{GT} の補集合を表す．図 8(a) と (b) にそれぞれ Ω_{GT} と Ω_{INF} の例を示す．

また，意図領域は多くの場合，フレームの中央付近に存在すると考えられるため，比較対象として，フレームの中心を原点とした半径 R の円領域を意図領域として推定した場合の ROC 曲線を R を変化させることで生成した．さらに，視覚的に顕著な領域と意図領域の相関を検証するために，Itti らが提案した顕著性マップ [12] を用いて推定した場合の ROC 曲線を生成した．

4.3 評価・考察

カメラの動きモデルにおけるパラメータ σ^2 が推定精度に及ぼす影響を検証するため，σ^2 を変化させたときのカメラの動きモデルを用いた意図マップの推定精度を AUC により評価した．図 9 に AUC の結果を示す．図 9 より，σ^2 が大きい場合に推定精度が高いことがわかる．これは慣性データ $\mathbf{d}_t$ が $N_{\mathrm{D}} = 101$ のとき 606 次元と高い次元数となるため，σ^2 が小さい場合に過学習状態になるためと考えられる．

次に，映像特徴モデルにおけるパラメータ M，K が推定精度に及ぼす影響を検証するために，M，K を変化させたときの映像特徴モデルを用いた意図マップの推定

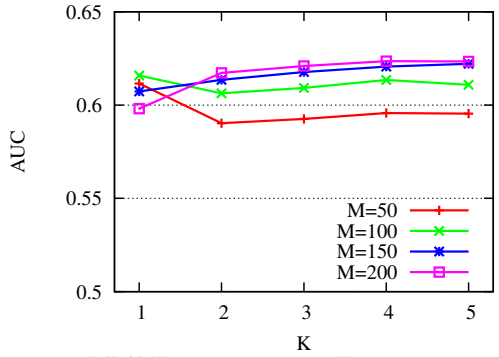

図 10　映像特徴モデルを用いた意図マップの AUC

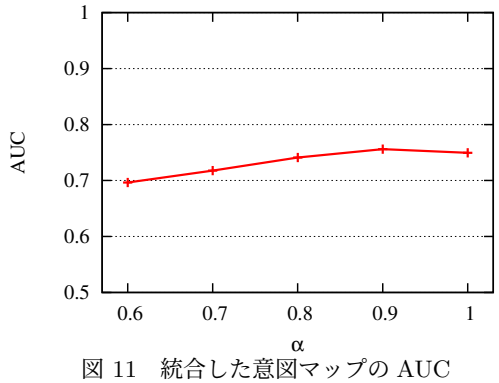

図 11　統合した意図マップの AUC

精度を AUC により評価した．図 10 に AUC の結果を示す．図 10 より，$M = 50, 100$ については K の値が小さいときに推定精度が高く，$M = 150, 200$ については K の値が大きいときに推定精度が高いことが分かる．これは M が小さい場合は一つの領域に属する色が多くなるため，各領域における映像特徴間の差が小さくなり，単峰性の分布に近づく一方，M が大きい場合は一つの領域に属する色が少なくなるため，各領域における映像特徴間の差が大きくなり，多峰性の分布の方がより実際の分布を再現できたためと考えられる．

次に，α を変化させたときの統合した意図マップの推定精度を評価した．図 11 に AUC の結果を示す．図 11 より $\alpha = 0.9$ の場合が最も推定精度が高い．

図 12 に提案手法，円領域，顕著性マップそれぞれの ROC 曲線を示す．なお，提案手法のパラメータは前述の結果に基づき $\sigma^2 = 100$，$M = 200$，$K = 4$，$\alpha = 0.9$ とした．これよりカメラの動きモデルを用いた意図マップと映像特徴モデルを用いた意図マップを統合した場合の推定精度が最も高いことがわかる．また，円領域による推定が高い推定精度を示していることから，意図領域はフレームの中央付近に多く存在すると考えられる．図 13(a) に意図領域を赤色で示したフレーム例を，(b), (c), (d), (e), および (f) に (a) のフレームに対応するカメラの動きモデルのみを用いた意図マップ，映像特徴モデルのみを用いた意図マップ，統合した意図マップ，提案手法の推定結果，顕著性マップの推定結果の例を示す．

カメラの動きモデルを用いた意図マップは，図 13(b) のように多くの場合フレームの中央付近の画素値が大

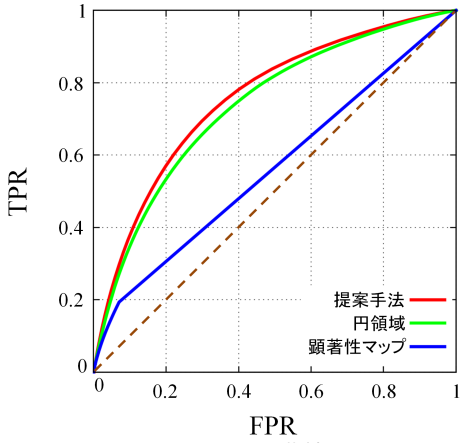

図 12　ROC 曲線

きくなる．一方で，カメラの動きが大きい場合には，図 13(b) の 3 段目の画像例のように，フレーム全体の画素値が小さくなる．撮影者は，意図領域の移り変わりや撮影対象を探しているとき，カメラを大きく動かす場合が多く，このとき意図領域は存在しない．これらのことから，カメラの動きのみを用いた意図マップは意図領域の有無を推定可能であり，フレームの中心を原点とした円領域を用いた推定よりも推定精度が優れているといえる．また，映像特徴モデルのみを用いた意図マップの場合は最も精度が低いが，物体の領域の形状を考慮した推定が可能であるため，意図マップを統合した際に性能が向上したと考えられる．顕著性マップを用いて推定される視覚的に顕著な領域は，図 13(f) の 2 段目の画像例のように，意図領域と相関が高い場合も存在するが，図 12 に示す結果から，多くの場合，意図領域と視覚的に顕著な領域の相関は低いと考えられる．

5. おわりに

本稿では，カメラの動きと映像特徴に基づく意図領域の推定手法を提案した．実験により，カメラの動きモデルを用いた意図マップと映像特徴モデルを用いた意図マップを統合した場合，AUC が 0.76 の推定精度が得られた．提案手法は特に，カメラの動きが大きくフレーム中に意図領域が存在しない場合において，円領域による推定より優れた結果を出力する．今後の課題として，慣性データの次元削減や映像特徴の抽出におけるフレームの領域分割法の改善，および新たな特徴量の追加などが挙げられる．本研究の一部は、科学研究費補助金による．

文　献

[1] T. Wang, T. Mei, X.-S. Hua, X.-L. Liu, and H.-Q. Zhou, "Video collage: a novel presentation of video sequence," *In Proc. Int'l Conf. Multimedia and Expo*, pp. 1479–1482, July 2007.

[2] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," *IEEE Trans. Multimedia*, vol. 7, no. 5, pp. 907–919, October 2005.

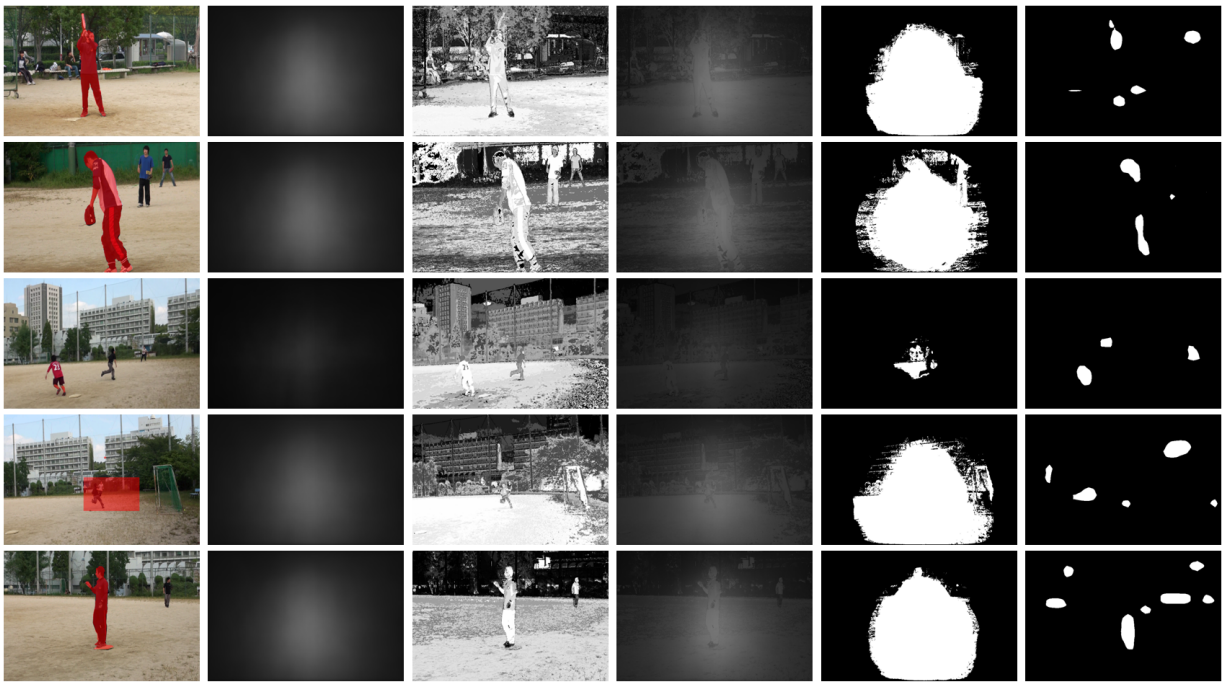

図 13　出力例．(a) 入力フレーム．赤色の領域が意図領域を表す．(b) カメラの動きモデルのみを用いた意図マップ．(c) 映像特徴モデルのみを用いた意図マップ．(d) 統合した意図マップ．(e) 提案手法の推定結果 (f) 顕著性マップの推定結果．

[3] X.-S. Hua, L. Lu, and H.-J. Zhang, "Optimization-based automated home video editing system," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 14, no. 5, pp. 572–583, May 2004.

[4] X. Fan, X. Xie, H.-Q. Zhou, and W.-Y. Ma, "Looking into video frames on small displays," *In Proc. the 11th ACM Int'l Conf. Multimedia*, pp. 247–250, November 2003.

[5] F. Liu and M. Gleicher, "Video retargeting: automating pan and scan," *In Proc. the 14th ACM Int'l Conf. Multimedia*, pp. 907–919, October 2005.

[6] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Processing*, vol. 13, no. 10, pp. 1304–1318, October 2004.

[7] W. Lai, X.-D. Gu, R.-H. Wang, W.-Y. Ma, and H.-J. Zhang, "A content-based bit allocation model for video streaming," *In Proc. Int'l Conf. Multimedia and Expo*, pp. 1315–1318, June 2004.

[8] M.-H. Hsiao, Y.-W. Chen, H.-T. Chen, K.-H. Chou, and S.-Y. Lee, "Content-aware video adaptation under low-bitrate constraint," *EURASIP Journal on Advances in Signal Processing*, vol. 2007, no. 2, June 2007.

[9] Y. Nakashima, N. Babaguchi, and J. Fan, "Automatically protecting privacy in consumer generated videos using intended human object detector" *In Proc. ACM Int'l Conf. Multimedia*, pp. 1135–1138, October 2010.

[10] T. Liu, N. Zheng, W. Ding, and Z. Yuan, "Video attention: learning to detect a salient object," *In Proc. Int'l Conf. Pattern Recognition*, pp. 1–4, December 2008.

[11] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," *In Proc. Int'l Conf. Computer Vision*, pp. 2106–2113, September 2009.

[12] L. Itti, C. Koch, and E. Niebur, "A model of saliency based visual attention for rapid scene analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, November 1998.

[13] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision Research*, vol. 49, no. 10, pp. 1295–1306, June 2009.

[14] H. Uegaki, Y. Nakashima, and N. Babaguchi, "Discriminating intended human objects in consumer videos," *In Proc. Int'l Conf. Pattern Recognition*, pp. 4380–4383, August 2010.

[15] 上柿普史，中島悠太，馬場口登，"映像特徴に基づく撮影者が意図した人物被写体の推定"，第 8 回情報科学技術フォーラム (FIT2009), K-046, pp. 639–642, September 2009.

[16] Y. Nakashima, N. Babaguchi, and J. Fan, "Detecting intended human objects in human-captured videos," *In Proc. Int'l Conf. Computer Vision and Pattern Recognition Workshop*, pp. 33–40, June 2010.

[17] 中島悠太，上柿普史，馬場口登，"映像中の撮影者が意図した人物被写体の検出"，電子情報通信学会 2010 年総合大会, D-12-41, pp. 152, March 2010.

[18] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, January 2007.

[19] T. Mei, X.-S. Hua, H.-Q. Zhou, and S. Li, "Modeling and mining of users' capture intention for home video," *IEEE Tran. Multimedia*, vol. 9, no. 1, pp. 66–77, January 2007.

[20] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *In Proc. Int'l Conf. Computer Vision and Pattern Recognition*, pp. 886–893, June 2005.

[21] F. Dufaux and J. Konrad, "Efficient, robust, and fast global motion estimation for video coding," *IEEE Trans. Image Processing*, vol. 9, no. 3, pp. 497–501, March 2000.