教師なし表情分類法に基づく ライフログ映像からの印象的なシーンの検出

野宮　浩揮† 　森國　淳司† 　宝珍　輝尚†

† 京都工芸繊維大学大学院工芸科学研究科
〒 606–8585 京都府京都市左京区松ヶ崎御所海道町
E-mail: †{nomiya,hochin}@kit.ac.jp

あらまし　ライフログ映像から印象的なシーンを検索することを目的として，映像中の人物に何らかの表情が現れているシーンを検出する手法を提案する．一般的な表情認識では，幸福，悲しみ，驚きなどの典型的な表情を認識対象としているが，一般に，ライフログ映像にはそれらでは認識が困難な表情が多く含まれている．そこで，表情表出時の顔面上の変化をとらえることができると考えられる，顔特徴点から得られる特徴量を用いて，教師なし学習法により，予め認識対象の表情を定めることなく，表情表出シーンを検出する手法を構成する．また，ライフログを想定した映像データセットを用いて，表情表出シーン検出性能の評価を行う．

キーワード　ライフログ，表情認識，教師なし学習，映像検索

Impressive Scene Detection from Lifelog Videos based on Unsupervised Facial Expression Classification

Hiroki NOMIYA†, Atsushi MORIKUNI†, and Teruhisa HOCHIN†

† Graduate School of Science and Technology, Kyoto Institute of Technology
Goshokaido-cho, Matsugasaki, Sakyo-ku, Kyoto 606–8585 Japan
E-mail: †{nomiya,hochin}@kit.ac.jp

Key words　lifelog, facial expression recognition, unsupervised learning, video retrieval

1. はじめに

個人の行動や体験をマルチメディアデータとして記録するライフログ [1] [2] が近年注目を集めている．ライフログは，テキスト，画像，映像，音声など様々なデータ形式で記録することができるが，特に近年では，映像記録機器の小型化・高性能化により，時間や場所を問わず手軽に撮影を行うことができるため，映像の形でライフログデータを作成する機会が非常に多くなっているといえる．その一方で，大量に作成された映像データの中から有用な映像データを検索することが難しいため，多くのライフログ映像が退蔵されることが問題となっている．

そこで本研究では，ライフログ映像から，有用と考えられる印象的なシーンの効率的な検索を行うことを目的とする．ライフログ映像は人物を主体とすることが多く，印象的なシーンでは，その人物に何らかの表情が表出している可能性が高いと考えられるため，ライフログ映像中から表情表出シーンを検出する手法について検討を行う．

著者らはこれまでに，ライフログ映像からの表情表出シーンの検出手法を提案してきている [3] [4]．これらの手法では，顔特徴点から得られる様々な特徴量を組み合わせて，効率的に表情表出シーンを検出することができる．

しかし，これらの先行研究では，検出対象とする表情を予め定めておき，それらの表情を識別するための学習を行う必要がある．先行研究や表情認識に関する多くの既存研究 [5] [6] [7] では，基本 6 表情（怒り，嫌悪，恐怖，幸福，悲しみ，驚き）などの典型的な表情を識別対象としているが，ライフログ映像中に現れる表情には，このような典型的な表情とは異なる表情（例えば，苦笑いや，驚きと幸福が入り交じったような表情など）が多く含まれており，予め検出対象とする表情を適切に定めることは難しい．さらに，表情の識別モデルを構築するために，学習データを作成する必要があるが，十分な量の学習データを作成するためには，多くの人的コストがかかることも問題である．

そこで本論文では，クラスタリングに基づく教師なし学習を用いて表情分類モデルを構築することにより，検出対象とする表情を予め定める必要がなく，学習データも必要としない表情

表出シーンを検出手法を構成する．さらに，本手法では，顔特徴点の位置関係から得られる少数の特徴量を用いて表情の分類を行うため，多数の特徴量の中から特徴選択を行う必要のある先行研究の手法と比べて，より効率的な表情表出シーン検出を可能とする．

以降，2 章では提案手法で用いる特徴量を示す．3 章で教師なし学習に基づく表情分類モデルの生成法について述べ，4 章で表情表出シーンの検出法について述べる．5 章でライフログを想定した映像を用いた表情表出シーン検出実験を行い，最後に 6 章でまとめを行う．

2. 特 徴 量

表情表出時の顔面上の変化をとらえるために，表情が表出する際に顕著な動きが見られると考えられる，顔面上の特徴点（以後「顔特徴点」と呼ぶ）を用いて 10 種類の特徴量を定める．

2.1 顔特徴点

顔特徴点として，図 1 に示す，左右の眉に沿ってほぼ等間隔に配置した点 10 点（p_1〜p_{10}），左右の目の周囲の点 18 点（p_{11}〜p_{28}），口の内周上と外周上の点 14 点（p_{29}〜p_{42}）の計 42 個の点を用いる．なお，これらの顔特徴点は，顔特徴点抽出ソフトウェア Luxand FaceSDK 4.0 [8] を用いて抽出する．

2.2 特 徴 量

2.1 節に示した顔特徴点を用いて，次に示す 10 種類の特徴量 ($f_1 \sim f_{10}$) を定める．

- f_1 ... 左右の眉の傾き

右眉上の点 ($p_1 \sim p_5$) ならびに左眉上の点 ($p_6 \sim p_{10}$) を用いて，最小二乗法により右眉の傾き a_r と左眉の傾き a_l を求め，左右の眉をほぼ左右対称であるとみなして，式 (1) に示すように特徴量 f_1 を定める．

$$f_1 = (a_l - a_r)/2 \tag{1}$$

- f_2 ... 眉と目の間の距離

眉と目の上側にある顔特徴点を結んでできる線分の長さの平均値を用いて，式 (2) に示すように特徴量 f_2 を定める．

$$f_2 = \frac{\sum_{i=1}^{10} ||\vec{p}_i - \vec{p}_{i+10}||}{10 \cdot l_N} \tag{2}$$

ここで，l_N は画像中の顔の大きさの違いなどによる影響を低減するための正規化に用いる項であり，左目の中心と右目の中心の顔特徴点を結ぶ線分の長さである．すなわち，$l_N = ||\vec{p}_{27} - \vec{p}_{28}||$ と表される．

- f_3 ... 眉間の面積

眉間の周辺にある 4 つの顔特徴点 p_5, p_6, p_{16}, p_{15} をつないで構成される四角形の面積に基づいて，式 (3) に示すように特徴量 f_3 を定める．

$$f_3 = S(p_5, p_6, p_{16}, p_{15})/l_N^2 \tag{3}$$

ここで，$S(p_{P_1}, \ldots, p_{P_m})$ は，p_{P_1} から p_{P_m} までの m 個の顔特徴点をつないで構成される m 角形の面積を表す．

- f_4 ... 目の面積

左右の目の周囲の顔特徴点をつないで構成される八角形の面積に基づいて，式 (4) に示すように特徴量 f_4 を定める．

$$\begin{aligned} f_4 \quad = \quad & \frac{1}{2 \cdot l_N^2} \{S(p_{11}, p_{12}, p_{13}, p_{14}, p_{15}, p_{23}, p_{22}, p_{21}) \\ & + S(p_{16}, p_{17}, p_{18}, p_{19}, p_{20}, p_{26}, p_{25}, p_{24})\} \end{aligned} \tag{4}$$

- f_5 ... 目の縦横比

目の上端と下端を結ぶ線分の長さと，目の左端と右端を結ぶ線分の長さの比に基づいて，式 (5) に示すように特徴量 f_5 を定める．

$$f_5 = \frac{1}{2}\left(\tan^{-1}\frac{||\vec{p}_{22} - \vec{p}_{13}||}{||\vec{p}_{15} - \vec{p}_{11}||} + \tan^{-1}\frac{||\vec{p}_{25} - \vec{p}_{18}||}{||\vec{p}_{20} - \vec{p}_{16}||}\right) \tag{5}$$

- f_6 ... 口の外側の面積

口の外側の顔特徴点をつないで構成される八角形の面積に基づいて，式 (6) に示すように特徴量 f_6 を定める．

$$f_6 = S(p_{29}, p_{30}, p_{31}, p_{32}, p_{33}, p_{34}, p_{35}, p_{36})/l_N^2 \tag{6}$$

- f_7 ... 口の内側の面積

口の内側の顔特徴点をつないで構成される八角形の面積に基づいて，式 (7) に示すように特徴量 f_7 を定める．

$$f_7 = S(p_{29}, p_{37}, p_{38}, p_{39}, p_{33}, p_{40}, p_{41}, p_{42})/l_N^2 \tag{7}$$

- f_8 ... 口の外側の縦横比

口の内側の上端と下端を結ぶ線分の長さと，口の左端と右端を結ぶ線分の長さの比に基づいて，式 (8) に示すように特徴量 f_8 を定める．

$$f_8 = \tan^{-1}\frac{||\vec{p}_{35} - \vec{p}_{31}||}{||\vec{p}_{33} - \vec{p}_{29}||} \tag{8}$$

- f_9 ... 口の外側の縦横比

口の外側の上端と下端を結ぶ線分の長さと，口の左端と右端を結ぶ線分の長さの比に基づいて，式 (9) に示すように特徴量 f_9 を定める．

$$f_9 = \tan^{-1}\frac{||\vec{p}_{41} - \vec{p}_{38}||}{||\vec{p}_{33} - \vec{p}_{29}||} \tag{9}$$

- f_{10} ... 口角の上昇度

口角がどの程度上がっているかを表す特徴量として，f_{10} を式 (10) のように定める．

$$f_{10} = \frac{\{y(p_{38}) + y(p_{41})\} - \{y(p_{29}) + y(p_{33})\}}{|y(p_{38}) - y(p_{41})|} \tag{10}$$

ここで，$y(p)$ は特徴点 p の y 座標を表す．p_{29} と p_{33} の y 座標の平均値が p_{38} と p_{41} の y 座標の平均値より大きければ，f_{10} は正の値をとる．したがって，f_{10} の値が大きいほど口角が上がっているといえる．なお，式 (10) の分母は正規化のための項である．

3. 表情分類法

ライフログ映像から表情表出シーンを検出するために，まず映像をフレーム画像（静止画像）に分割し，各フレーム画像に対して表情の識別を行う．既存の表情認識手法では，基本 6 表情など，予め識別対象とする表情を定めて，それらの表情が表

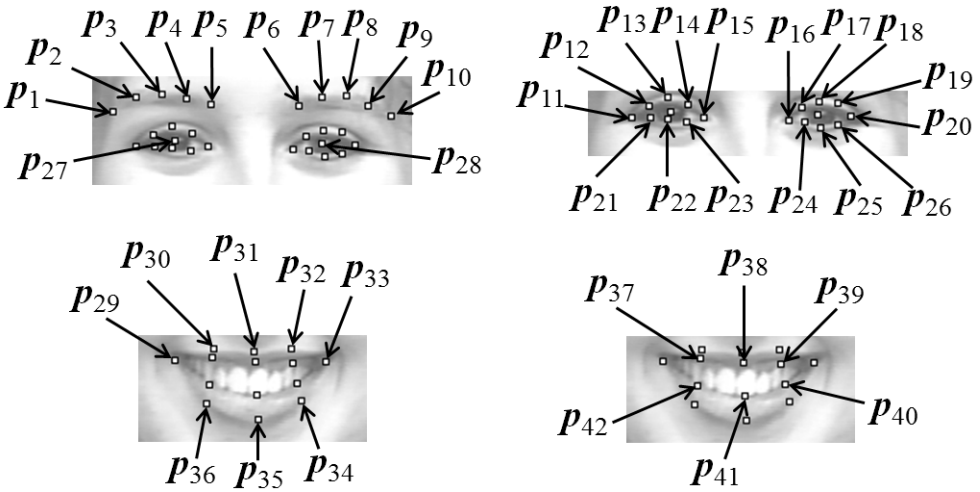

図 1　顔 特 徴 点

出している学習データを用いて表情分類モデルを構築し，表情の識別を行う．しかし，ライフログ映像中に現れる表情は様々であり，基本 6 表情のような典型的な表情とは異なる表情も多く含まれるため，予め検出対象とする表情を適切に定めることは難しい．さらに，十分な量の学習データを作成するためには，多くの人的コストがかかる．

そこで，本論文では，クラスタリングに基づく教師なし学習を用いて，識別対象の表情を定める必要がなく，学習データも必要としない表情分類モデルを構成する．まず，ライフログ映像から抽出した各フレーム画像に対して，2.2 節に示した 10 種類の特徴量を求め，これらを並べて 10 次元の特徴ベクトルを構成する．

ここで，ライフログ映像から抽出したフレーム画像の集合を $I = \{I_1, \ldots, I_n\}$ と表す．ここで，n はフレーム画像の枚数である．各フレーム画像 I_i に対する特徴ベクトル F_i を式 (11) のように定める．

$$F_i = [f_1(I_i), \ldots, f_{10}(I_i)] \tag{11}$$

ここで，$f_i(I_j)$ は，j 番目のフレーム画像から得られた特徴量 f_i の値を表す．

表情表出時の顔面上の特徴をより捉えやすくするため，これらの特徴ベクトル $F_1, \ldots, F_n$ に対して主成分分析を行い，式 (12) により F_i から新たな特徴ベクトル X_i を生成する．なお，これ以降 X_i を単に「特徴ベクトル」と呼ぶこととする．

$$X_i = [x_1(I_i), \ldots, x_{10}(I_i)] \tag{12}$$

ここで，

$$x_j(I_i) = \sum_{k=1}^{10} \ell_{jk} f_k(I_i) \tag{13}$$

であり，ℓ_{jk} は第 j 主成分の k 番目の主成分負荷量の値を表す．

特徴ベクトル X_i を用いて，図 2 に示す k-means 法に基づくクラスタリングを行い，表情を分類する．クラスタリングにより生成された各クラスタが，それぞれ異なる表情を表す．し

入力：特徴ベクトルの集合 $X = \{X_1, \ldots, X_n\}$
　　生成するクラスタの数 K

（1） X の中からランダムに K 個の特徴ベクトル $X_{c_1}, \ldots, X_{c_K}$ を選択し，初期クラスタ $C_1, \ldots, C_K$ を生成する．なお，$\forall i\ X_{c_i} \in C_i$ とする．

（2） 各特徴ベクトル $X_1, \ldots, X_n$ を，それぞれ $X_{c_1}, \ldots, X_{c_K}$ の中で最も距離の短いものが属しているクラスタに割り当てる．なお，特徴ベクトル間の距離は，主成分分析で得られた固有値による重み付けを行ったユークリッド距離により定める．すなわち，X_i は，式 (14) をみたすクラスタ C_{j^*} に割り当てられる．

$$j^* = \arg\min_j \sqrt{\sum_{k=1}^{10} \lambda_k \left\{ x_k(I_i) - x_k(I_{c_j}) \right\}^2} \tag{14}$$

ここで，λ_k は第 k 主成分の固有値である．

（3） $C_1, \ldots, C_K$ の重心点を求め，それぞれ $Z_1, \ldots, Z_K$ とする．また，$Z_i = [z_1(i), \ldots, z_{10}(i)]$ と表す．

（4） 各特徴ベクトル $X_1, \ldots, X_n$ に対して，各クラスタの重心点の中から最も近いものを求め，そのクラスタに割り当てる．すなわち，X_i は，式 (15) をみたすクラスタ $C_{j^{**}}$ に割り当てられる．

$$j^{**} = \arg\min_j \sqrt{\sum_{k=1}^{10} \lambda_k \left\{ x_k(I_i) - z_k(j) \right\}^2} \tag{15}$$

（5） クラスタリングを行った後と行う前とで，各クラスタに属する事例が変わっていなければ，各クラスタを出力としてクラスタリングを終了する．変わっていれば，(3) に戻り再度クラスタリングを行う．

図 2　クラスタリングアルゴリズム

たがって，K 個のクラスタを生成することは，ライフログ映像から得られたフレーム画像を K 種類の表情に分類することに相当する．

4. 表情表出シーン検出

前述の表情分類法を用いることにより，ライフログ映像中の各フレーム画像を K 個のクラスタ，すなわち K 種類の表情に

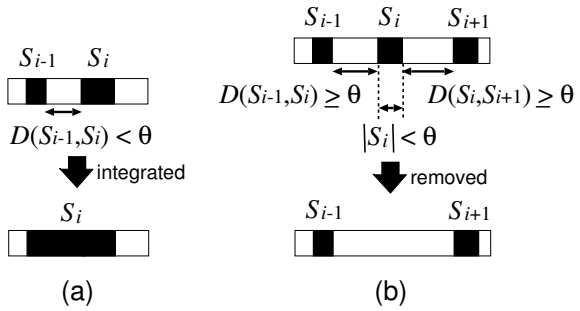

図 3　フレーム集合の統合と削除

分類できる．この分類結果を用いて，映像中から各表情が表出しているシーンを検出する．

1 つのクラスタが 1 つの表情を表しているため，表情表出シーン検出はクラスタごとに行う．はじめに，クラスタに含まれる個々のフレームを 1 つのシーンとみなす．そして，近接しているシーン同士を統合し，孤立しているシーンを削除していくことにより，表情表出シーンを形成していく．

シーンが近接しているかどうかを判断するために，2 つのシーン S_i と S_j の間の距離 $D(S_i, S_j)$ を式 (16) のように定める．

$$D(S_i, S_j) = s_j^{min} - s_i^{max} - 1 \tag{16}$$

ここで，s_j^{min} は S_j に含まれるフレームの中での最小のフレーム番号であり，s_i^{max} は S_i に含まれるフレームの中での最大のフレーム番号である．なお，映像の先頭から n 番目のフレームのフレーム番号を n とする．また，シーン S_i はシーン S_j よりも前にある（$s_i^{max} < s_j^{min}$）とする．

あるシーン S_i について，1 つ前にあるシーン S_{i-1} との距離を求め，それが閾値 θ よりも小さい場合は，図 3(a) に示すように，これらのシーンを統合して単一のシーンとする．同様に，S_i の 1 つ後ろにあるシーン S_{i+1} との距離を求め，それが θ よりも小さい場合は，これらのシーンを統合して単一のシーンとする．

S_i と S_{i-1} の距離ならびに S_i と S_{i+1} の距離が θ 以上であり，さらに S_i の長さ（$|S_i|$ と表す）が θ を下回る場合は，S_i は孤立した短いシーンであり，重要性は低いと考えられるので，図 3(b) に示すように S_i を削除する．したがって，検出されるシーンは，少なくとも θ の長さを持つことになる．

このようにして，シーンの統合と削除を繰り返していき，統合・削除することのできるシーンがなくなった時点でのシーンを最終的な表情表出シーンとして出力する．複数の表情について表情表出シーン検出を行う場合は，各表情ごとに，この表情表出シーン検出手順を実行する．

最後に，表情表出シーン検出アルゴリズムを図 4 に示す．

入力： 同一のクラスタ C に属しているフレーム番号の集合 $\{q_1 \ldots, q_{n_C}\}$ および閾値 θ.
なお，n_C は C に属しているフレーム画像の数であり，$i < j$ である任意の i, j について $q_i < q_j$ が成り立つとする．

（1） $S_i = \{q_i\}$ $(i = 1, \ldots, n_C)$，$S = \{S_1, \ldots, S_{n_C}\}$ とする．
（2） $\nu \leftarrow n_C$.
（3） S の全ての要素を，$|S_i|$ の値について昇順に整列する．
ここで，p 番目にこの値の小さい要素を S_{u_p} と表す．
（4） $p \leftarrow 1$.
（5） If $u_p > 1 \wedge D(S_{u_p-1}, S_{u_p}) < \theta$ Then
S_{u_p-1} を S_{u_p} に統合する．
$S_{u_p} \leftarrow S_{u_p} \cup S_{u_p-1}$.
$S \leftarrow \{S_1, \ldots, S_{u_p-2}, S_{u_p}, \ldots, S_\nu\}$.
Go to (9).
End If
（6） If $u_p < \nu \wedge D(S_{u_p}, S_{u_p+1}) < \theta$ Then
S_{u_p+1} を S_{u_p} に統合する．
$S_{u_p} \leftarrow S_{u_p} \cup S_{u_p+1}$.
$S \leftarrow \{S_1, \ldots, S_{u_p}, S_{u_p+2}, \ldots, S_\nu\}$.
Go to (9).
End If
（7） If $|S_{u_p}| < \theta$ Then
If $\nu = 1$ Then
表情表出シーンなしとしてシーン検出を終了する．
Else
S_{u_p} を削除する．
$S \leftarrow \{S_1, \ldots, S_{u_p-1}, S_{u_p+1}, \ldots, S_\nu\}$.
Go to (9).
End If
End If
（8） If $p = \nu$ Then
For $i = 1$ To ν
s_i^{min} から s_i^{max} までの間を表情表出シーンとして出力する．
End For
シーン検出を終了する．
Else
$p \leftarrow p + 1$.
Go to (5).
End If
（9） S の要素の添字を以下のように振り直す．
$S = \{S_1, \ldots, S_{\nu-1}\}$
$\nu \leftarrow \nu - 1$
Go to (3).

図 4　表情表出シーン検出アルゴリズム

5. 実　　験

5.1 ライフログ映像からの表情表出シーン検出

5.1.1 ライフログ映像データセット

ライフログ映像を想定して，いずれも 20 歳代の男子大学生である 5 名の被験者（以後，被験者 A, B, C, D, E と呼ぶ）のそれぞれに対して，被験者を含む 2 名でトランプ（神経衰弱）を行ってもらい，その様子を Web カメラを用いて被験者の顔が映るように撮影した．なお，被験者にはできるだけ普段通りに振る舞うよう指示している．

各被験者に対して，それぞれ 10 分間の映像を 1 本作成し，全部で 5 本の映像をデータセットとして使用した．これらの映像は，いずれも解像度が 640×480 画素，フレームレートが 25 フレーム／秒である．各映像について，先頭から 10 フレームお

表 1　検出された表情表出シーン

被験者	シーン数	平均シーン長（秒）	標準偏差（秒）
A	4	14.4	2.29
B	4	26.8	17.3
C	2	26.8	16.4
D	2	33.6	24.9
E	7	17.4	8.66

きに 1 フレームを選択して得られた 1500 フレームの画像を実験に用いている．

5.1.2　表情表出シーン検出精度

本実験で使用したライフログ映像内で表出している表情はほとんどが笑顔であり，笑顔が表出しているシーン以外では，被験者はほぼ無表情である．すなわち，ライフログ映像中で表出している表情は 2 種類（無表情と笑顔）であるとみなすことができるため，クラスタ数 K を 2 として，3 章に示した手法により表情分類を行った．ここでは，表情が表出しているシーンを検出することが目的であるため，生成された 2 つのクラスタのうち，何らかの表情が表出している可能性の高いクラスタを 1 つ選択した．選択基準として，2.2 節に示した 10 種類の特徴量を平均 0，分散 1 に正規化した際の特徴量の平均値を用いた．無表情に近い表情では，これらの特徴量の値が小さくなりやすいと考えられるため，2 つのクラスタのうち，特徴量の平均値が高い方のクラスタを選択し，4 章で述べた手法により表情表出シーン検出を行った．なお，検出されるシーンには，そのシーンで起こっている出来事がわかる程度の長さが必要であると考えられる．ここでは，そのために少なくとも 1 つのシーンにつき 10 秒程度は必要であると考え，閾値 θ を 25 と定めることにより，検出されるシーンの長さの最小値が 10 秒になるようにしている．

まず，各被験者の映像から検出された表情表出シーンの数，平均シーン長ならびにシーン長の標準偏差を表 1 に示す．

いずれの被験者からも，合計して 1〜2 分程度の表情表出シーンが得られた．検出されたシーンの数やシーン長は，被験者によって大きく異なっているが，これは映像中で表情が表出しているシーンの長さや数が被験者によって異なっているためであると考えられる．

次に，検出されたシーンがどの程度正確に表情表出シーンを捉えているかを評価するために，表情表出シーンの検出精度を求めた．ここでは，式 (17) により表情表出シーンの検出精度 A を求めている．

$$A = \frac{|T_e \cap \hat{T}_e| + |T_n \cap \hat{T}_n|}{|T_e| + |T_n|} \tag{17}$$

ここで，T_e, T_n は，それぞれ実際に表情が表出しているフレームの集合と，それ以外（無表情）のフレームの集合である．これらは，著者のうち 2 名の評価者により決定しており，評価者のうち少なくとも一方が表情が表出していると判断したフレームの集合を，実際に表情が表出しているフレームの集合と定めている．また，$\hat{T}_e, \hat{T}_n$ は，それぞれ提案手法により検出された表情表出シーンに含まれるフレームの集合と，表情表出シーンに含まれないフレームの集合である．各被験者に対する表情表出シーン検出精度を図 5 に示す．

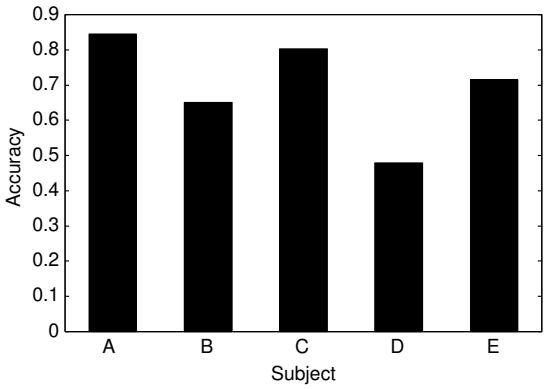

図 5　表情表出シーン検出精度

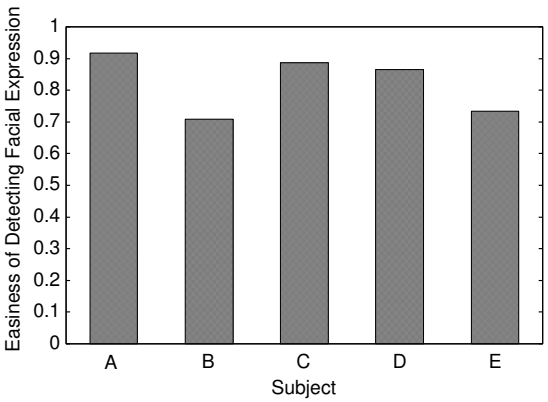

図 6　表情識別の容易性

表情表出シーン検出精度には，被験者によって違いが見られるが，これは個人差による影響が大きいと考えられる．被験者によって，はっきりと表情を表出することが多い場合と，(例えば微笑程度の笑みを浮かべるなど）表情が表出しているかどうかの判断が難しいことが多い場合があり，表情表出シーン検出精度は，後者に比べて前者の方が高くなっていると考えられる．

このことを明らかにするために，表情が表出しているかどうかを判断することの容易さを表す指標である，表情識別の容易性を定め，各被験者の表情識別の容易性を求めた．表情識別の容易性 ε は，2 名の評価者による目視での表情分類結果に基づいて，式 (18) のように定める．

$$\varepsilon = \frac{|T_e^1 \cap T_e^2| + |T_n^1 \cap T_n^2|}{|T_e^1| + |T_n^1|} \tag{18}$$

ここで，T_e^1, T_e^2 は，それぞれ 1 人目の評価者，2 人目の評価者が表情が表出していると判断したフレームの集合であり，T_n^1, T_n^2 は，それぞれ 1 人目の評価者，2 人目の評価者が表情が表出していないと判断したフレームの集合である．なお，$|T_e^1| + |T_n^1|$ はフレームの総数に相当し，本実験では 1500 である（$|T_e^1| + |T_n^1| = |T_e^2| + |T_n^2|$ である）．各被験者に対する表情識別の容易性を図 6 に示す．

図 5 と図 6 より，被験者 A, B, C, E については，表情識別の容易性と表情表出シーン検出精度がほぼ比例していることから，人の目で見ても表情が表出しているかどうかの判断が比較

図 7　表情が表出していると評価者が判断したシーンのフレーム画像例 (a), (b) と表情が表出していないと評価者が判断したシーンのフレーム画像例 (c), (d)

表 2　表情表出シーン検出に要した時間（秒）

処理	平均所要時間	標準偏差
特徴量算出	0.108	0.012
表情分類	0.040	0.000
シーン検出	0.001	0.001
合計	0.149	0.013

的難しい被験者では，表情表出シーン検出精度もやや低くなるといえる．

しかし，被験者 D は，表情識別の容易性は高いが，表情表出シーン検出精度は低くなっている．この原因として，この被験者は，笑顔が表出している時の顔面上の変化が非常に大きいことが考えられる．

図 7(a), (b) は，被験者 D の映像に含まれている，2 名の評価者がいずれも表情が表出していると判断したフレーム画像の一部である．目が細められる，口が大きく開かれるといった，顔面上の変化がはっきりと現れていることがわかる．一方，図 7(c), (d) は，2 名の評価者がいずれも表情が表出していないと判断したフレーム画像の一部である．(c) は無表情に近いといえるが，(d) は (c) と比べると目が細くなっており，口も開かれていることから，単独の画像として見ると表情が表出しているようにも見える．しかし，(a), (b) のように，表情がより強く現れているフレームが映像中に多く含まれているため，評価者は相対的に判断して (d) には表情が現れていないと判断していると考えられる．

提案手法では，図 7(d) のようなフレーム画像はほぼ表情が表出していると判断されている（図 7(a), (b) と同じクラスタに属している）ため，(d) のようなフレームを含むシーンが表情表出シーンとして検出され，結果として表情表出シーン検出精度が低下したと考えられる．

5.1.3　表情表出シーン検出速度

提案手法の表情表出シーン検出速度を評価するため，表情表出シーン検出に要する時間を計測した結果を表 2 に示す．計測には，CPU が Xeon W3580 (3.33GHz)，メモリが 8GB の計算機を使用した．なお，顔特徴点抽出には既存のソフトウェアを用いているため，顔特徴点の抽出に要した時間は含めていない．表 2 には，2 章に示した特徴量を算出する際に要した時間，3 章に示した表情分類に要した時間，4 章に示した表情表出シーン検出に要した時間，ならびにこれらの合計の時間を示している．表中の「平均所要時間」と「標準偏差」は，それぞれ，5 名の被験者に対する所要時間の平均値および標準偏差である．

表 3　CK データセットの分類結果

	クラスタ 1	クラスタ 2	クラスタ 3	クラスタ 4
怒り	43	15	6	13
幸福	16	76	2	9
悲しみ	34	27	13	7
驚き	23	5	2	70

表 4　JAFFE データセットの分類結果

	クラスタ 1	クラスタ 2	クラスタ 3	クラスタ 4
怒り	14	4	11	1
幸福	9	11	1	10
悲しみ	12	2	17	0
驚き	6	0	6	18

いずれの被験者についても，10 分間の映像からの表情表出シーン検出に要した時間はおよそ 0.15 秒程度であり，効率的に表情表出シーン検出ができているといえる．したがって，提案手法は大規模なライフログ映像データベースに対しても十分に適用可能であると考えられる．

5.2　複数の表情の識別

5.2.1　表情識別実験に使用したデータセット

前述の実験に使用したライフログ映像では，表出している表情の大部分が笑顔であるため，ここでは提案手法の笑顔以外の表情の識別性能を評価する．そのためのデータセットとして，Cohn-Kanade AU-Coded Facial Expression Database [9]（以下 CK データセットと記述する）ならびに The Japanese Female Facial Expression Database [10]（以下 JAFFE データセットと記述する）を用いた．

CK データセットは，18 歳から 30 歳の男女の被験者 97 人の表情データが，無表情の状態から表情が表出されるまでの連続した画像として収められている．この中から，怒り，幸福，悲しみ，驚きの表情が表出している画像をそれぞれ 77, 103, 81, 100 枚使用した．

JAFFE データセットは，日本人女性の被験者 10 人の表情データが，表情表出時および無表情時の画像として収められている．この中から，怒り，幸福，悲しみ，驚きの表情が表出している画像をそれぞれ 30, 31, 31, 30 枚使用した．

いずれのデータセットに対しても，静止画像を用いているため，表情表出シーン検出は行わず，表情分類のみを行った．また，表情分類の際に生成するクラスタの数 K は 4 とした．

5.2.2　表情分類結果

CK データセットの分類結果を表 3 に，JAFFE データセットの分類結果を表 4 に示す．各表の列は，4 つのクラスタそれぞれに含まれていた，怒り，幸福，悲しみ，驚きの事例の数を示している．例えば，CK データセットでは，クラスタ 1 には全部で 116 の事例が含まれており，そのうち 43 の事例に怒り，16 の事例に幸福，34 の事例に悲しみ，23 の事例に驚きの表情が表出していたことを表している．

CK データセットでは，怒り，幸福，驚きの表情に関しては，多くの事例が特定のクラスタ（それぞれクラスタ 1, 2, 4）に分類されている．したがって，提案手法は，分類対象の表情に関

(a) (b) (c) (d)

図 8　JAFFE データセットでクラスタ 1 に分類された事例 (a), (b) とクラスタ 3 に分類された事例 (c), (d)

する事前知識を与えられなくても，これらの表情をある程度分類することのできる分類モデルを形成しているといえる．

一方，悲しみの表情は様々なクラスタに分類されており，十分に識別できていないといえる．悲しみの表情が表出する際には，眉や口角が下がるなどの動きが顔面上に見られるが，他の表情と比べると顔面上の動きが小さいため，十分に特徴を捉えることができなかったのではないかと考えられる．

JAFFE データセットでは，驚きの表情は特定のクラスタ（クラスタ 4）に分類されることが多いという傾向が見られるため，よく識別できているといえる．しかし，怒りと悲しみの事例が同じクラスタに含まれることが多くなっている．これは，JAFFE データセットでは，怒りと悲しみに類似した表情が多いことが原因と考えられる．図 8(a), (b) は，いずれもクラスタ 1 に分類された事例であるが，(a) は怒り，(b) は悲しみの表情である．また，図 8(c), (d) は，いずれもクラスタ 3 に分類された事例であるが，(c) は怒り，(d) は悲しみの表情である．いずれの事例も顔面上の変化が小さく，また (a) と (b) は互いに類似しており，(c) と (d) も互いに類似していることから，これらの事例が混同されたと考えられる．

幸福の事例には，口を閉じて笑っている事例や口を大きく開いて笑っている事例があるため，前者の一部が怒りや悲しみの事例が多く含まれるクラスタ（クラスタ 1）に分類され，後者が驚きの事例が多く含まれるクラスタ（クラスタ 4）に分類されたと考えられる．

表 3 と表 4 を混同行列とみなすと，CK データセットの表情認識精度は約 56 %であり，JAFFE データセットの表情認識精度は約 49 %である．JAFFE データセットには，前述の通り識別の難しい表情が含まれているため，やや低い値となっているが，およそ半数程度の表情は正しく分類できる可能性があるといえる．

6. ま と め

ライフログ映像から印象的なシーンを検索することを目的として，映像中から表情表出シーンを検出する手法を提案した．提案手法は教師なし学習法に基づいているため，予め検出対象とする表情を定めておく必要がなく，学習用のデータも必要としない．さらに，表情表出シーン検出の効率性も高いため，大規模なライフログ映像からの表情表出シーン検出を少ない人的コストで実現できる可能性を持つといえる．

ライフログ映像を想定した映像データセットに対して表情表出シーン検出実験を行ったところ，被験者によるばらつきはあるものの，ある程度正確に表情（主に笑顔）が表出しているシーンを検出することができた．

怒り，幸福，悲しみ，驚きの表情が含まれるデータセットに対しては，およそ半数程度の事例を適切に識別できる可能性を示した．しかし，これらは静止画像に対する表情分類にとどまっているため，今後は様々な表情が含まれるライフログ映像を用いて提案手法の性能評価を行うことを検討している．

また，提案手法が適用できるのは，遮蔽物が映っていない正面の顔のみであるため，遮蔽物や顔面の向きの変化に対する頑健性の向上も今後の課題である．さらに，個人差による表情表出シーン検出精度に対する影響の低減や，様々な表情をより正確に識別するために，表情分類法の改善を行っていく予定である．

謝辞　本研究の一部は，科学研究費補助金（課題番号：22700098）による．ここに記して謝意を表す．

文　献

[1] J. Gemmell, G. Bell, R. Luederand, S. Drucker and C. Wong, “MyLifeBits: Fulfilling the Memex Vision,” Proc. of the 10th ACM International Conference on Multimedia, pp.235–238, 2002.

[2] K. Aizawa, T. Hori, S. Kawasaki and T. Ishikawa, “Capture and Efficient Retrieval of Life Log,” Pervasive 2004 Workshop on Memory and Sharing Experiences, pp.15–20, 2004.

[3] 野宮浩揮，森國淳司，宝珍輝尚，“表情表出シーン検出を用いたライフログ映像からの印象的なシーンの検索”，第 4 回データ工学と情報マネジメントに関するフォーラム (DEIM2012)，B9-3, 2012.

[4] 野宮浩揮，宝珍輝尚，“アンサンブル学習を用いた効率的な映像からの表情表出シーン検出”，電子情報通信学会論文誌, Vol.J95-D, No.2, pp.193–205, 2012.

[5] G. Littlewort, M. S. Bartlett, I. Fasel, J. Susskind and J. Movellan, “Dynamics of Facial Expression Extracted Automatically from Video,” Image and Vision Computing, Vol.24, No.6, pp.615–625, 2004.

[6] D. Datcu and L. Rothkrantz, “Facial Expression Recognition in Still Pictures and Videos Using Active Appearance Models: A Comparison Approach,” Proc. of the 2007 International Conference on Computer Systems and Technologies, pp.1–6, 2007.

[7] G. Fanelli, A.Yao, P.-L. Noel, J. Gall, L. Van Gool, “Hough Forest-based Facial Expression Recognition from Video Sequences,” Proc. of International Workshop on Sign, Gesture and Activity, 2010.

[8] Luxand, Inc., “Luxand FaceSDK 4.0,” http://www.luxand.com/facesdk/

[9] T. Kanade, J. F. Cohn and Y. Tian, “Comprehensive Database for Facial Expression Analysis,” Proc. of the 4th IEEE International Conference on Automatic Face and Gesture Recognition, pp. 46–53, 2000.

[10] M.J. Lyons, S. Akamatsu, M. Kamachi and J. Gyoba, “Coding Facial Expressions with Gabor Wavelets,” Proceedings of the 3rd IEEE International Conference on Automatic Face and Gesture Recognition, pp. 200–205, 1998.